British Journal of Nutrition (2013), **109**, 772–774

Letter to the Editor

Categorisation of input variables for deriving dietary patterns

(First published online 8 January 2013)

We would like to congratulate Smith *et al.*[1] for their recent paper on the relevant issue of input variable quantification when deriving dietary patterns by principal component analysis (PCA). Indeed, previous studies have compared different methods of deriving empirical patterns such as cluster- *v.* factor-based analysis[2], while some authors[3] have compared exploratory factor methods, such as PCA, with confirmatory methods, such as factor analysis. But as Smith *et al.*[1] make clear, few studies have specifically dealt with the issue of how the choice of input variables influences the derived dietary patterns.

In their paper, Smith *et al.*[1] compared dietary patterns derived from PCA of four types of input variables, of which three were interval variables (weight in g/d, mean weight for total energy intake and percentage contribution to total energy intake), the fourth one being binary yes/no variables, coding whether the food groups had been consumed or not. From our understanding, their main conclusion was that there were no obvious differences between the patterns derived using any of the three interval variables, while the patterns derived using the binary variable appear somewhat different, reflecting more general food preferences. This result seems meaningful, but we would like to make a few comments that we think could be useful regarding this matter of input variable quantification for deriving dietary patterns.

Do assumptions that underlie correlation coefficients always hold for food consumption data?

We could not agree more with Smith *et al.*[1] when they state that 'dichotomising food intakes does not capture the complexity of eating behaviour'. It is likely with the creditable intention to take into account this complexity as much as possible, that most authors deriving empirical dietary patterns with factor-based methods do analyse interval-scale variables such as the first three quoted by Smith *et al.*[1]. PCA and related methods, such as confirmatory factor analysis[3], are based on measures of association between interval-type food groups variables, such as covariances or Pearson correlation coefficients[4]. These measures were initially derived to assess associations in a bivariate Gaussian context (for which a zero value of the coefficient is equivalent to probabilistic independence). But, even in a descriptive context, one must keep in mind that these measures of association are appropriate to assess only linear or at least approximately linear and strongly monotonous relationships (which is quite restrictive an assumption when one wishes to account for the 'complexity of eating behaviour'); also, they are quite sensitive to outlying values and/or high leverage observations (quite common in observational studies, all the more when dealing with dietary intake)[5,6]. Smith *et al.*[1] do make somewhat indirect reference to this when they state that skewed distributions can be an issue. There are numerous methods or techniques that have been put forward to deal with these various issues, such as variable transformations, rank correlation coefficients and non-linear PCA amongst others. But a rather straightforward, and often used, way to deal with the discrepancies between the assumptions that underlie assessment of associations by Pearson correlation coefficients and the actual data is to categorise the variables prior to analysis and assess inter-relationships using suitable measures of association and related multivariate methods[7].

Does categorisation need to be only binary (consumers *v.* non-consumers)?

Regarding this issue of categorisation, Smith *et al.*[1] did discuss binary coding of food group variables. However, categorisation need not be only binary, based on whether the subject consumed the food group or not; many possibilities do exist to categorise a food group interval variable after examination of its distribution. If that appears to be rather impractical because the number of food groups is too high, it can also be done in a more automatic manner by deriving quantiles (tertiles, quintiles or else depending on sample size), possibly after having derived a special category for the non-consumers. This would appear as a midway between analysing interval variables 'to account for the complexity of eating behaviour', but with the risk of biased measures of associations due to departure from assumptions, which underlie interpretation of the Pearson correlation coefficients, and the quite coarse dichotomisation, which only distinguishes between consumers and non-consumers. Regarding measures of association between categorical variables, there are a number of them, but many of which are related to the 'χ^2 distance' computed from the contingency table resulting from the cross-tabulation of the two variables[5]. When applied to categorised interval variables, these non-parametric measures of association reduce the influence of potential outliers and do not require any distributional assumption, nor specific hypothesis on the shape of the relationship. From an inferential point of view, it is indeed well known that when all the assumptions are met, such measures of association on categorised variables have less power to detect associations than measures such as Pearson correlation coefficients to detect linear associations on initial interval variables. However, we have discussed earlier that many of these assumptions either do not always hold good or are practically difficult to check when analysing high-dimensional food consumption data (e.g. sixty-six groups in the example of Smith *et al.*[1] result in more than 2000 pairs of variables); hence, the tradeoff of less power for more robustness may be sometimes worthwhile, even if some authors do advocate avoiding categorisation as far as possible[8].

Which statistical analysis methods Could be used to derive factor-type patterns from categorised variables?

For interval variables, PCA, which Smith *et al.*[1] focus on, derives factor patterns that are weighted linear combinations of food group input variables, which maximise a variance criterion based on eigenvalues/eigenvectors decomposition of the correlation or covariance matrix. Analogous methods do exist to derive factors, taking into account inter-relationships between categorical or categorised interval variables on the basis of the robust and non-parametric measure of association based on the χ^2 distance; for instance, a method such as multiple correspondence analysis (MCA)[6,9] derives factors as a series of orthogonal-weighted linear combinations, of decreasing order of importance, based on eigen decomposition of a generalised contingency table. It can deal with any combination of intrinsically categorical variables and/or categorised interval variables; applied to categorised interval variables, it is a quite straightforward way to assess the structure of interdependencies between variables without any hypothesis, neither pertaining to their distribution, nor to the shape of their interdependencies, nor to the leverage of outliers. From the end-user point of view, input variables of MCA are binary variables coding to which categories of the different food groups variables the subject belongs (e.g. if one would categorise food groups in quintiles, there would then be five input binary variables for each food group). MCA can be viewed as a generalisation of PCA for categorical variables; for each subject, her/his score on a given factor is thus a weighted linear combination of the binary variables coding her/his categories of input variables. It is widely available in most statistical packages such as Stata (Stata Corporation LP), SAS (SAS Institute, Inc.) and R (R Foundation for Statistical Computing), and it has been used in a variety of research fields[10–13]. As for the specific issue of dietary patterns, Smith *et al.*[1] did reference use of MCA, but only on dichotomised data[14], while some authors[15] have used MCA to derive factor-type patterns from quintiles of forty-three food groups (initially in g/1000 kcal).

Conclusion

Regarding the issue of input variable quantification to derive factor-type empirical dietary patterns, the ultimate goal would be use or development of suitable non-linear, distribution-free dimension reduction methods to fully take into account the complexity of distributions and inter-relationships of interval food consumption variables. However, to broaden the choice between analysing interval food group variables, based on linear correlation coefficients that have their own limitations, and the quite coarse analysis of dichotomised variables, we do think that analysing (e.g. by MCA) suitably categorised variables is a relatively simple alternative to be considered. It is also to be noted that this issue of input variable quantification or categorisation may also arise when trying to extract not only factor-type dietary patterns, but also patterns derived by cluster analysis or analogous methods[2]; indeed, a number of the earlier-discussed limitations and/or sensitivity to depart from the required assumptions (such as sensitivity to outliers, shape of the distributions) also partly apply. Input variable quantification may also be an issue when trying to derive predictive patterns[16].

We hope the earlier comments will be of interest to the researchers in the field, so that they are as informed as possible about the choices regarding input variable quantification when deriving factor-type empirical dietary patterns. Beyond the points discussed earlier, we wholeheartedly agree with Smith *et al.*[1] that more research is needed to better assess the impact of this quantification on diet–disease associations and we thank them for sharing with us their insight regarding this matter.

Pierre Traissac and Yves Martin-Prével
Institut de Recherche pour le Développement (IRD),
NUTRIPASS Research Unit,
IRD-UM2-UM1, Montpellier,
France
email pierre.traissac@ird.fr

doi:10.1017/S000711451200551X

References

1. Smith AD, Emmett PM, Newby PK, *et al.* (2012) Dietary patterns obtained through principal components analysis: the effect of input variable quantification. *Br J Nutr* (epublication ahead of print version 6 september 2012).
2. Newby PK & Tucker KL (2004) Empirically derived eating patterns using factor or cluster analysis: a review. *Nutr Rev* **62**, 177–203.
3. Varraso R, Garcia-Aymerich J, Monier F, *et al.* (2012) Assessment of dietary patterns in nutritional epidemiology: principal component analysis compared with confirmatory factor analysis. *Am J Clin Nutr* **96**, 1079–1092.
4. Jolliffe IT (2002) *Principal Components Analysis*, 2nd ed. New York: Springer.
5. Armitage P & Colton T (1998) *Encyclopedia of Biostatistics*. Chichester: Wiley.
6. Dodge Y (2008) *The Concise Encyclopedia of Statistics*. New York: Springer.

7. Turner EL, Dobson JE & Pocock SJ (2010) Categorisation of continuous risk factors in epidemiological publications: a survey of current practice. *Epidemiol Perspect Innov* **7**, 9.
8. Bennette C & Vickers A (2012) Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents. *BMC Med Res Methodol* **12**, 21.
9. Greenacre MJ (1984) *Theory and Applications of Correspondence Analysis*. London: Academic Press.
10. Sourial N, Wolfson C, Bergman H, *et al.* (2010) A correspondence analysis revealed frailty deficits aggregate and are multidimensional. *J Clin Epidemiol* **63**, 647–654.
11. Tessier S, Traissac P, Bricas N, *et al.* (2010) Food shopping transition: socio-economic characteristics and motivations associated with use of supermarkets in a North African urban environment. *Public Health Nutr* **13**, 1410–1418.
12. Fillol F, Dubuisson C, Lafay L, *et al.* (2011) Accounting for the multidimensional nature of the relationship between adult obesity and socio-economic status: the French second National Individual Survey on Food Consumption (INCA 2) dietary survey (2006–07). *Br J Nutr* **106**, 1602–1608.
13. Traissac P & Martin-Prevel Y (2012) Alternatives to principal components analysis to derive asset-based indices to measure socio-economic position in low- and middle-income countries: the case for multiple correspondence analysis. *Int J Epidemiol* **41**, 1207–1208, author reply 1209–1210.
14. Guinot C, Latreille J, Malvy D, *et al.* (2001) Use of multiple correspondence analysis and cluster analysis to study dietary behaviour: food consumption questionnaire in the SU.VI.-MAX. cohort. *Eur J Epidemiol* **17**, 505–516.
15. Aounallah-Skhiri H, Traissac P, El Ati J, *et al.* (2011) Nutrition transition among adolescents of a south-Mediterranean country: dietary patterns, association with socio-economic factors, overweight and blood pressure. A cross-sectional study in Tunisia. *Nutr J* **10**, 38.
16. Hoffmann K, Schulze MB, Schienkiewitz A, *et al.* (2004) Application of a new statistical method to derive dietary patterns in nutritional epidemiology. *Am J Epidemiol* **159**, 935–944.