Journal of Retailing and Consumer Services 43 (2018) 39–45

Contents lists available at ScienceDirect

Journal of Retailing and Consumer Services

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

Identifying superfluous survey items

Kylie Brosnan[a,*], Bettina Grün[b], Sara Dolnicar[a]

[a] University of Queensland, The Faculty of Business, Economics and Law, 39, Blair Dr, St Lucia, QLD 4067, Australia
[b] Johannes Kepler University Linz, Department of Applied Statistics, Altenbergerstraße 69, 4040 Linz Austria

ARTICLE INFO

Keywords:
Survey length
Survey item reduction
Measurement

ABSTRACT

Surveys provide critical insights into consumer satisfaction and experience. Excessive survey length, however, can reduce data quality. We propose using constrained principle components analysis to shorten the survey length in a data-driven way by identifying optimal items with maximum information. The method allows assessing item elimination potential, and explicitly identifies which items provide maximum information for a specified number of items. We use artificial data to explain the method, provide two illustrations with empirical survey data, and make code freely available in an online tool

1. Introduction

Retailers and service providers rely on consumer surveys for decision-making (Reichheld and Covey, 2006). Most consumer surveys are administered online with the average questionnaire taking 15 min to complete (GRIT, 2015). More than a third of surveys take longer than 15 min. Short surveys have a number of advantages over long surveys: they are less expensive (Lavrakas, 2008), return higher completion rates (Crawford et al., 2001; Galesic and Bosnjak, 2009; Fan and Yan, 2010; Hoerger, 2010; Stanton et al., 2002), and have less random or systematic error associated with fatigue or boredom (Galesic and Bosnjak, 2009; Herzog and Bachman, 1981).

Consumer surveys typically contain a range of questions with groups of questions forming different item batteries. Unlike most surveys designed, for example, for psychometric, behavioural medicine or social science research, these questions are often not the result of a scale development procedure where multiple questions are typically asked to measure one construct. Rather, each item asks a specific question the answer to which is directly relevant for managerial decision making, but groups of questions are connected and form an item battery. The total number of items is frequently high, higher than recommended in terms of minimising respondent fatigue to maximise data quality.

We propose a statistical method for the assessment of item elimination potential in such survey contexts, and the determination of the optimal set of items per item battery for a given number of items. The method builds on principal components analysis, and extends it to allow for straightforward identification of redundant items with minimal information loss or the optimal items for maximum information gain.

Using a statistical method to reduce survey length provides a sound data-driven solution to optimising batteries of items not developed using a scale development procedure. This method is of immediate practical relevance to solve the managerial problem of obtaining optimal information for minimal cost, especially in the context of survey studies which involve data collection across multiple points in time, both longitudinal and repeat cross-sectional. This paper contributes to the knowledge and use of statistical techniques for reducing survey burden to improve data quality.

2. Identifying elimination potential and the optimal set of items

The variability of responses to a survey item reflects the information contained in the item. Survey items can have low information content because (a) they contribute less to understanding the variability between respondents than other items; or (b) they can contain information redundant to that of other survey items within the item battery.

A standard method for determining reduction potential for a set of a survey items is principal components analysis (McHorney, Ware and Raczek, 1993; Sweeney and Soutar, 2001; Fodor, 2002; Vyas and Kumaranayake, 2006; Boyes, Girgis, and Lecathelinais, 2009). Principal components analysis (Jolliffe, 2002; Abdi and Williams, 2010) solves the following maximisation problem to determine the first k principal components for each $k = 1, ..., p$: For a centred data matrix $X \in \mathbb{R}^{n \times p}$ consisting of n observations and p items determine a k-dimensional linear combination of the data matrix given by XA_k, with $A_k \in \mathbb{R}^{p \times k}$ to maximise the following criterion:

* Corresponding author.
E-mail address: k.brosnan@business.uq.edu.au (K. Brosnan).

https://doi.org/10.1016/j.jretconser.2018.02.007
Received 16 November 2017; Received in revised form 26 January 2018; Accepted 15 February 2018

$$\sqrt{\frac{tr(A_k^T X^T X A_k)}{tr(X^T X)}},$$

where $tr(\bullet)$ denotes the trace of a matrix and X^T is the transpose of the matrix X. The matrix A_k is in general uniquely identified (up to the sign) under the restriction that A_k is an orthogonal matrix, i.e., $A_k^T A_k = I_k$, and $A_k^T X^T X A_k$ is a diagonal matrix with decreasing entries. This restriction also implies that the first k columns of A_{k+1} are identical to A_k. The columns of the matrix XA_k are also referred to as principal components and the criterion used ensures that the k principal components maximise the variance retained in the k-dimensional subspace. To reduce the number of variables used in a subsequent analysis, the k principal components retaining a sufficient amount of variance are selected. However, these k principal components generally contain information from all items.

Principal components analysis indicates the potential for eliminating survey items. If a small number of principal components contains most of the variability in the data, there is substantial potential for eliminating redundant survey items. If each principal component explains a substantial amount of variability, none of the items are superfluous. Using principal components analysis leads to two possible conclusions: superfluous survey items exist in the item battery or they do not. But traditional principal components analysis does not identify which items are superfluous and which provide optimal information for a given number of survey items. The method we propose does.

The underlying statistical approach is a constrained principal components analysis. The added constraint ensures that the number of principal components is equal to the number of items included (for an overview see Cadima et al., 2004). In this case, if k principal components are selected for subsequent analysis, these also only contain information from k items. The extension of traditional principal components analysis to the case where only k items are allowed for a k components solution is achieved by imposing the restriction that A_k only uses k items from X. Formally this restriction can be imposed by requiring:

$$\#\{i = 1,..., p : a_{k,i}^T a_{k,i} > 0\} \leq k,$$

where $a_{k,i} \in \mathbb{R}^k$ corresponds to the ith row of the matrix A_k, i.e., there are $p - k$ rows consisting only of zeros in A_k.

Traditional principal components analysis can be performed in a computational efficient way by determining the singular value decomposition of the matrix X and using the first k right-singular vectors to define the matrix A_k. This implies that all principal components analysis solutions from 1 to p components are simultaneously obtained given the singular value decomposition. Imposing the additional constraint of using only k items for the k component solution, forces principal components analysis to solve a different optimisation problem for each value of k. The reason is that the single item capturing most of the variability is not necessarily contained in the set of the best two items. Therefore, to obtain the best set of items for each number of items, the best subset from the total number of possible subsets given by $\binom{p}{k}$ needs to be determined. An exhaustive search of all possible subsets quickly becomes prohibitively computational expensive. For example, finding the best 10 out of 30 items requires checking more than 30 million possible sets.

Alternative improved computational strategies to solve the constrained principal components analysis problem exist. An exact efficient procedure is the branch-and-bound algorithm proposed by Duarte Silva (2002) based on the leaps-and-bounds algorithm by Furnival and Wilson (1974) for variable selection in linear regression. This algorithm first creates a branch where it evaluates promising sets of items for each number of items. The performance criteria for these promising sets form the bounds. Then a second branch is created where sets of poorly performing items are investigated. All subsets of these sets can be discarded, and do not need to be explicitly evaluated if the criterion for these sets is worse than the current best bound. These sets can be excluded from further consideration because of the monotonicity condition: the variance retained can only be reduced if less items are included.

Performing constrained principal components analysis is more computationally demanding than traditional principal components analysis. But the additional computational effort pays off: imposing this constraint allows the identification of the best set of survey items for each step of the principal components analysis; for each number of survey items the set of items capturing most variability is identified.

Based on traditional and constrained principal components analysis we propose the following method for identifying item elimination potential and selecting optimal items:

(1) Perform *traditional and constrained principal components analysis* for all number of items (components) and calculate the explained variance for each number of items (components).
(2) Assess the item elimination potential by calculating the *area under the curve (AUC) values* of the curve given by plotting the number of items (components) on the x-axis against the cumulative explained variance for each number of items (components) on the y-axis.
The area under the curve takes values between 0.5 and 1. The smallest possible value of 0.5 results if each additional item increases the explained variance by the same amount and indicates no elimination potential. The highest value results when all variables are perfectly correlated and one single survey item contains all the information contained in the entire set of items.
The AUC values are always higher or the same for traditional principal components analysis than for constrained principal components analysis for the same data set. The difference indicates how much explained variance is sacrificed by imposing the constraint on the number of items to include.
(3) Visualise the curves in an *elimination plot* and determine a suitable number of items to retain. Fig. 1 provides examples of elimination plots for three artificial data sets. Three exemplary elimination plots are given for the scenarios where there is no item elimination potential (panel on the left), some item elimination potential (middle panel) and substantial item elimination potential (right panel). As can be seen, in the elimination plot, the curve for traditional principal components analysis is always above the curve for constrained principal components analysis.
Optimally, the curve exhibits a distinctive kink from a steeply increasing linear function to only a slightly increasing linear or horizontal function. The number of items where the kink occurs is the optimal number of items. A steep increase before this point indicates that a large amount of variability would be sacrificed if fewer items were selected. A slight increase afterwards implies that additional items contain little variance. Fig. 1 shows on the right side (in the panel labelled “Substantial item elimination potential”) an exemplary elimination plot for such a scenario.
The extreme cases are a straight line from the left lower corner to the right upper corner and a triangle curve from the left lower corner to the left upper corner to the right upper corner. The first case corresponds to AUC = 0.5 and thus to no item reduction potential and the second case to AUC about 1 and thus indicates that a single item is sufficient. Fig. 1 shows on the left side (in the panel labelled “No item elimination potential”) an exemplary elimination plot for the case where AUC is close to 0.5. The middle plot of Fig. 1 visualises the case where some elimination potential is present, but there is no distinct kink discernible. In this case the elimination plot indicates how many items are required to retain a certain amount of variability, e.g., 80%. In the middle panel of Fig. 1, eight items are required to retain about 80% of the variability. This number of items is determined by assessing the point where a horizontal line inserted at the value of 0.8 on the y-axis intersects the curves in the plot.