Nonmetric Unfolding of Marketing Data: Degeneracy and Stability

Michel van de Velden, Alain De Beuckelaer, Patrick J.F. Groenen, and Frank M.T.A. Busing

ERIM REPORT SERIES *RESEARCH IN MANAGEMENT*	
ERIM Report Series reference number	ERS-2011-006-MKT
Publication	March 2011
Number of pages	35
Persistent paper URL	http://hdl.handle.net/1765/22725
Email address corresponding author	vandevelden@ese.eur.nl
Address	Erasmus Research Institute of Management (ERIM) RSM Erasmus University / Erasmus School of Economics Erasmus Universiteit Rotterdam P.O.Box 1738 3000 DR Rotterdam, The Netherlands Phone: + 31 10 408 1182 Fax: + 31 10 408 9640 Email: info@erim.eur.nl Internet: www.erim.eur.nl

Bibliographic data and classifications of all the ERIM reports are also available on the ERIM website: www.erim.eur.nl

ERASMUS RESEARCH INSTITUTE OF MANAGEMENT

REPORT SERIES

RESEARCH IN MANAGEMENT

ABSTRACT AND KEYWORDS	
Abstract	Nonmetric unfolding is a powerful (nonparametric) analytical tool generating a preference-based joint display of subjects (e.g., customers) and objects (e.g., brands or products). Systematic patterns in customers' preferences can be directly inferred from this display, and may provide valuable input for making important marketing decisions such as deciding what new product to launch. Unfortunately, nonmetric unfolding frequently produces degenerate unfolding solutions (i.e., unfolding solutions showing close-to-perfect model fit irrespective of the data analyzed). As a degenerated display shows ill-positioned customers and brands/products, the chance of making an incorrect marketing decision (e.g., launching the wrong product) is very high. To solve this problem adequately, we combine bootstrapping with penalized nonmetric unfolding (Prefscal) to obtain an accurate, nondegenerate and stable unfolding solution.
Free Keywords	perceptual mapping, customer preference modeling, nonmetric multidimensional unfolding, bootstrap analysis
Availability	
Classifications	

Nonmetric unfolding of marketing data: degeneracy and stability

Michel van de Velden#
Erasmus University Rotterdam
Erasmus School of Economics, Econometric Institute
The Netherlands
vandevelden@ese.eur.nl

Alain De Beuckelaer
Ghent University, Department of Personnel Management, Work and Organizational Psychology and
Department of Sociology, Belgium;
Renmin University China, School of Sociology and Population Studies;
& Radboud University Nijmegen
Institute for Management Research
The Netherlands
A.DeBeuckelaer@fm.ru.nl

Patrick J.F. Groenen
Erasmus University Rotterdam
Erasmus School of Economics, Econometric Institute
The Netherlands
groenen@ese.eur.nl

Frank M.T.A. Busing
Universiteit Leiden
Faculty of Social Sciences, Psychological Institute,
The Netherlands
BUSING@fsw.leidenuniv.nl

#Corresponding author; Address: P.O. Box 1738, 3000 DR Rotterdam, The Netherlands

Acknowledgements. We would like to thank Bas Hillebrand for his valuable suggestions on a previous version of this paper.

Nonmetric unfolding of marketing data: degeneracy and stability

Abstract

Nonmetric unfolding is a powerful (nonparametric) analytical tool generating a preference-based joint display of subjects (e.g., customers) and objects (e.g., brands or products). Systematic patterns in customers' preferences can be directly inferred from this display, and may provide valuable input for making important marketing decisions such as deciding what new product to launch. Unfortunately, nonmetric unfolding frequently produces degenerate unfolding solutions (i.e., unfolding solutions showing close-to-perfect model fit irrespective of the data analyzed). As a degenerated display shows ill-positioned customers and brands/products, the chance of making an incorrect marketing decision (e.g., launching the wrong product) is very high. To solve this problem adequately, we combine bootstrapping with penalized nonmetric unfolding (Prefscal) to obtain an accurate, nondegenerate and stable unfolding solution.

Keywords: perceptual mapping, customer preference modeling, nonmetric multidimensional unfolding, bootstrap analysis

1 Introduction

Perceptual mapping has gained much popularity in marketing research (e.g., Bijmolt and Wedel, 1999, Cornelius, Wagner and Natter, 2010; Faure and Natter, 2010; Green, Carmone, and Smith, 1989; Ho, Chung and Lau, 2010). Perceptual maps are useful for marketers as they provide a means to get a better understanding of product differentiation, product positioning, and customer preferences for brands and/or products (Chaturvedi and Carroll, 1998). Perceptual mapping comprises a wide variety of statistical techniques, such as methods based on principal component analysis, factor analysis, cluster analysis, multiple (polynomial) regression, partial least squares, correspondence analysis, and multidimensional scaling. In this paper, the focus is on a special form of multidimensional scaling: multidimensional unfolding. Originating in the field of psychology (Coombs, 1950), the method has been studied extensively by scholars in marketing (e.g., DeSarbo and Rao, 1986; DeSarbo, Young and Rangaswamy, 1997; Ho et al., 2010) as it allows a joint display of customers and brands in one single map.

In multidimensional unfolding, subjects (e.g., customers) and objects (e.g., brands) are jointly plotted in a low dimensional display in such a way that distances between subjects and objects optimally represent subjects' preferences. A recent study by Cornelius et al. (2010) showed that managers often prefer graphical formats over numbers and tables when evaluating alternative product positionings. Joint displays of customers and brands facilitate a better understanding of customers' preferences. Using preference relationships as input marketers may take important marketing decisions such as deciding which product to launch first.

An important distinction concerns the use of either *metric* or *nonmetric* multidimensional unfolding. The appropriateness of these two unfolding models depends upon the measurement properties of the data. If the choice alternatives are rated on an interval or ratio scale the so-called metric unfolding model is appropriate. However, if the choice alternatives are rated on an ordinal scale, for instance, a Likert-scale, nonmetric unfolding is more adequate. In the nonmetric model, transformations of the original data are allowed provided that their original order is preserved. The nonmetric unfolding model is also appropriate when customers have rank-ordered the choice alternatives or when modeling network data in which distances are calculated by counting. Such count data occur, for instance, when examining the minimum number of steps between nodes in

the network (e.g., Ho et al., 2010). By allowing the data to be transformed, better fitting low dimensional representations are obtained. This paper is focused on nonmetric unfolding.

Results from unfolding should only be used when the resulting maps accurately represent customer preferences. In this respect, two issues should be considered within the context of *nonmetric* unfolding:

1. *The issue of degeneration.* Nonmetric unfolding may lead to so-called degenerate solutions. Degenerate unfolding solutions are unfolding solutions where the extent of misfit (that is the amount of Stress in the unfolding solution) becomes arbitrarily small irrespective of the data (Borg and Groenen 2005, Section 14.4). Perceptual maps describing a degenerate unfolding solution contain many ill-positioned customers, brands or products, and are thus meaningless from a marketing point of view. As a consequence, marketing decisions (e.g., the decision as to what new product to launch) based on the interpretation of such an untrustworthy map are likely to be completely wrong as they are not supported by the preference relationships manifested in the data.
2. *The issue of stability assessment.* As nonmetric unfolding is essentially a nonparametric method, no specific distributional assumptions are made concerning the model parameters (e.g., the location of points in the map or the errors). As a consequence, one cannot rely on statistical inference to make an adequate assessment of the statistical significance and stability of the unfolding solution. At present, no statistical measures are available to assess the quality/stability of the individual points included in a nonmetric unfolding solution. Although the method produces point estimates of the positioning of subjects and objects, there are no estimates available about the uncertainty of these positions. Obviously, basing marketing decisions on poorly represented subjects and/or objects, is undesirable and should be avoided.

The degeneracy issue is a long-standing issue in nonmetric unfolding. Several authors have suggested different solutions (see, for instance, Busing, Groenen, & Heiser, 2005; DeSarbo and Rao, 1984; Heiser, 1989; and Kim, Rangaswamy, and DeSarbo, 1999). We refer to Busing et al. (2005) for an extensive overview of the degeneracy problem and proposed solutions. In a recent paper, Ho et al. (2010) also considered the degeneracy issue in the context of large-scale marketing data. Unfortunately, their proposal for dealing with degeneracy did not involve the

nonmetric unfolding case. In fact, their suggestion amounted to applying metric unfolding without transformations—which inherently lacks degeneracy problems—to nonmetric data. In this paper, we jointly deal with the two issues mentioned above, that is, the degeneracy problem in nonmetric unfolding and the absence of stability estimates of an unfolding solution. As the exact location of points in a degenerate solution is arbitrary, degenerate solutions are likely to be instable. The reverse is also likely to hold. Hence, by being able to assess the stability of an unfolding solution, we are able to differentiate between stable, nondegenerate solutions and instable, degenerate solutions.

Concerning stability it should be noted that several researchers (e.g. MacKay and Dröge, 1990, MacKay and Zinnes, 1986) have proposed model-based unfolding methods that allow hypotheses testing However, for the nonparametric unfolding methods, it appears that stability has not been studied.

To find stable, nondegenerate, unfolding solutions, measures that quantify stability are required. We first construct such stability measures for nonmetric unfolding solutions. Such stability measures may then help selecting an appropriate (nondegenerate) solution. Moreover, the stability measure will also enable an assessment of the stability of each point corresponding to either subjects (e.g. customers) or objects (e.g. brands) in the map. The methodology to visualize stability that we introduce in this paper can also be used when plotting unfolding results of large data sets, in particular, data sets with many objects and/or subjects.

The remainder of this paper is organized as follows: In the next section, we present a brief technical account of nonmetric unfolding and the degeneracy problem. Next, we consider the stability issue of nonmetric unfolding solutions, introduce stability measures, and illustrate how stability measures may be used to determine optimal nondegenerate unfolding solutions. We also show ways of directly visualizing stability in an unfolding map. To illustrate our approach, we apply the proposed methodology to several marketing data sets. In Section 7, we briefly discuss how our approach can be applied to marketing data sets of a much larger size. We conclude our paper with a summary of the main results.

2 Nonmetric Multidimensional Unfolding

The goal of multidimensional unfolding is to obtain a low-dimensional (spatial) map with subjects and objects, in such a way that distances between subjects and objects in the map best represent the preferences stored in the data. Hence, if a subject has a strong preference for an object, the corresponding distance should be small. Analogously, if a subject has a weak preference for an object, the distance should be relatively large. To construct such a map we seek coordinates for both subjects and objects.

Let us first introduce some notation. Throughout this paper the indices *i* and *j* correspond to subjects and objects respectively, and the total number of subjects and objects in the sample are denoted by *n* and *p* respectively. Now, let $\mathbf{x}_i$ and $\mathbf{y}_j$ denote the $k\times 1$ coordinate vectors for subject *i* and object *j*, respectively, where *k* is the (user-supplied) dimensionality of the solution. The Euclidean distance between $\mathbf{x}_i$ and $\mathbf{y}_j$ as represented in the map is defined as: $d_{ij} = \sqrt{\left(\mathbf{x}_i - \mathbf{y}_j\right)'\left(\mathbf{x}_i - \mathbf{y}_j\right)}$.

The preferences can be measured in several ways. For example, subjects may indicate their preferences either by means of ratings, rankings, or through paired comparisons. In this paper, we consider preference data that indicate an ordering of preferences. Moreover, we shall assume, without loss of generality, that the preference of subject *i* for object *j* is coded in such a way that it is represented by the dissimilarity δ_{ij}. Hence, a low value of δ_{ij} indicates a high preference and a high value corresponds to a low preference. As only rank order information is used, we may replace the observed preferences by any monotonically nondecreasing transformation $\hat{d}_{ij} = f_i\left(\delta_{ij}\right)$ yielding so-called pseudo-distances $\hat{d}_{ij}$. Thus, $f_i\left(\delta_{ij}\right)$ transforms the original dissimilarities ordinally to $\hat{d}_{ij}$'s, with a separate transformation function f_i for each individual *i*. This case is referred to as the *row-conditional* case in nonmetric unfolding.

The objective of the unfolding analysis is to find coordinate matrices **X** and **Y**, with as rows the transposed subject and object coordinate vectors respectively, in such a way that the distances d_{ij} match the pseudo distances $\hat{d}_{ij}$ in some optimal manner. This objective is formalized by the so-called *normalized* Stress function

$$\sigma_n^{\,2} = \sigma_n^{\,2}\left(\hat{\mathbf{D}}, \mathbf{X}, \mathbf{Y}\right) = \frac{\sum_{ij}\left(\hat{d}_{ij} - d_{ij}\right)^2}{\sum_{ij}\hat{d}_{ij}^{\,2}}, \qquad (1)$$

where $\hat{\mathbf{D}}$ is the matrix with elements $\hat{d}_{ij}$. Objective (1) is minimized (to indicate a closer 'match') over the set of functions (transformations) and configurations **X** and **Y**.

2.1 *The Degeneracy Problem*

Busing et al. (2005) showed that a degenerate unfolding always exist when transformations include a constant term. To see this, consider the transformation $\hat{d}_{ij} = f_i\left(\delta_{ij}\right) = c + g_i\left(\delta_{ij}\right)$ where c is a constant. Then $\sum_{ij}\left(\hat{d}_{ij} - d_{ij}\right)^2 = \sum_{ij}\left(c + g_i\left(\delta_{ij}\right) - d_{ij}\right)^2$. Hence, upon choosing $g_i\left(\delta_{ij}\right) = 0$, an optimal unfolding solution would result from choosing coordinates in such a way that $d_{ij} = c$ for all points so that value of the objective function becomes zero, indicating perfect model fit. In a two-dimensional setting, such a perfect solution can be obtained by choosing all the points **X** to lie on a circle with all the points **Y** at its center, or vice versa.

One solution to the degeneracy problem is use transformations that do not include a constant term. Several recommendations have been made in this respect, see for instance, Heiser (1981, 1989), Kim et al. (1999) and Borg and Lingoes (1987). Other recommendations have been made by DeSarbo and Rao (1984) and Kruskal and Carroll (1969). For a more detailed treatment of the degeneracy issue we refer to Busing et al. (2005). In the same paper, a penalized approach is suggested that offers an adequate solution for the degeneracy problem. This approach is further elaborated on in this paper.

2.2 *Penalized Nonmetric Unfolding*

The main idea of penalized nonmetric unfolding is to steer the unfolding solution away from a degenerate solution. To do so, a force is added to $\sigma_n^{\,2}$ that assigns a penalty to unfolding solutions that are degenerate. The penalty is incorporated as an increase in a penalized Stress criterion. To illustrate how this is actually carried out, recall the normalized Stress criterion defined in (1). A

degenerate solution is characterized by $\hat{d}_{ij}$'s that are all the same. To be effective, a good penalty term should have high values when the $\hat{d}_{ij}$'s are close to constant, and small values when the average of the $\hat{d}_{ij}$'s differs greatly from their variation. An objective measure that compares variation to average is Pearson's coefficient of variation which is defined as the standard deviation divided by the mean: $\nu(\mathbf{a}) = s(\mathbf{a}) / \bar{a}$, where $\bar{a}$ and $s(\mathbf{a})$ denote, respectively, the sample mean and standard deviation for vector **a**. In the degenerate case, the $\hat{d}_{ij}$'s are constant and the variation coefficient becomes zero. The penalized Stress criterion can now be formulated as

$$\sigma_p^{\,2}\left(\mathbf{P}, \hat{\mathbf{D}}, \mathbf{X}, \mathbf{Y}\right) = \sigma_n^{\,2\lambda}\left(\hat{\mathbf{D}}, \mathbf{X}, \mathbf{Y}\right)\left(1 + \omega \frac{\nu^2(\boldsymbol{\delta})}{\nu^2(\hat{\mathbf{d}})}\right), \tag{2}$$

where $\boldsymbol{\delta}$=vec($\boldsymbol{\Delta}$) and $\hat{\mathbf{d}}$=vec($\hat{\mathbf{D}}$) are vectors with the observed dissimilarities (preferences) and pseudo-distances, respectively, and the penalty parameters λ and ω are user-supplied constants which determine the strength of the penalty. This formula differs slightly from Formula (8) presented in Busing et al. (2005) that does not contain the constant $\nu^2(\boldsymbol{\delta})$ and uses *Raw* Stress rather than *Normalized* Stress. Criterion (2), however, is implemented in SPSS and is used throughout this paper. Note that the term $\nu^2(\boldsymbol{\delta})$ is merely a scaling constant that is useful but not essential (see also Busing, 2010). Formula (2) shows that a low variation coefficients for $\hat{\mathbf{d}}$, lead to high penalized stress values, making unfolding solutions unattractive. If the variation coefficient for $\hat{\mathbf{d}}$ is equivalent to the variation coefficient for $\boldsymbol{\delta}$, the minimization of penalized Stress becomes similar to the minimization of σ_n^2. Therefore, if there exists a nondegenerate perfect solution of (1), it will also be a solution of the penalized stress criterion (2).

In Busing et al. (2005), a simulation study revealed that low values for λ (that is, strong penalties) lead to near-linear transformations of the observed dissimilarities. The effect of ω on the unfolding solutions appeared to be relatively weaker. However, if the chosen value of ω is too low, degenerate solutions may still occur. Busing et al. (2005, p. 82), suggest to fix λ=0.5 and to consider different values for ω starting from 0.5 (which in (2) should be adjusted depending upon

the observed variation coefficient $\nu^2(\boldsymbol{\delta})$[1]). Unfortunately, it is not trivial to determine which penalty value is more appropriate or desirable. In their recent paper, Ho et al. (2010) , for example, have shown that solutions obtained using the penalized approach with $\lambda = \omega = 0.5$ (the current SPSS default values) may still yield degenerate solutions.

3 Stability of Nonmetric Unfolding Solutions

The stability of unfolding configurations is of great practical importance. Unfolding solutions that are greatly influenced by small changes in the data are undesirable. In addition, if a solution is relatively stable but is located far away from the true configuration, thus a solution with *strong bias*, its interpretation will be distorted. To assess the stability and bias of a nonmetric unfolding solution we propose to use a nonparametric bootstrap procedure (Efron, 1982; Efron and Tibshirani, 1993).

In a bootstrap analysis, the statistical method is applied repeatedly to resampled data. That is, from the original sample, *B* new samples of the same size, the so-called 'bootstrap samples', are randomly drawn with replacement. Drawing subjects with replacement, implies that subjects may be observed more than once (or not at all) in a bootstrap sample. The objects, however, are observed in every bootstrap sample. To avoid this imbalance, we use a *balanced* bootstrap. In the balanced bootstrap, individual subjects may be drawn once, repeatedly, or not at all in a given bootstrap sample. However, after drawing the *B* bootstrap samples, each subject is drawn exactly *B* times.

Each bootstrap sample is analyzed by means of nonmetric unfolding, yielding a configuration of subjects and objects. However, each bootstrap configuration is based on a different set of subjects. Moreover, as the configurations only represent relative distances, a direct comparison of the location of point coordinates representing subjects' and objects' location across different bootstrap solutions is not meaningful; each unfolding configuration is *nonunique* as it can be freely rotated, translated and scaled, altering the location of subjects and objects without changing the distances. To account for this nonuniqueness, we apply Procrustean similarity

[1] The values used and suggested in Busing et al. (2005) should be multiplied by the inverse of the squared variation coefficient to obtain results which are comparable to the ones presented in this paper .

transformations (Schönemann and Carroll, 1970; Borg and Groenen, 2005) in such a way that the bootstrap coordinates for objects are as close as possible to their coordinates in the unfolding solution of the original data. For convenience, this solution is simply referred to as "the unfolding solution" in the remainder of this paper.

When using a two-dimensional space, one can integrate the unfolding solution as well as all (rotated) bootstrap configurations in one single two-dimensional plot. In this way, a configuration is obtained where each individual subject and object is represented by a cloud of points. The sizes of subject and object clouds provide a measure for stability; the smaller the size, the higher the stability. Plotting all bootstrap points, however, leads to cluttered plots that make it virtually impossible to identify individual subject or object clouds. To avoid such cluttered plots we shall use density plots and confidence ellipses for plotting the subject and object clouds respectively.

3.1 *Confidence Ellipses*

For the object points, it is important to clearly indicate which cloud belongs to which object. Therefore, plotting all bootstrap points is not a viable option and it is more insightful to display (1-α)% confidence ellipses around the bootstrap means. These ellipses are constructed in such a way that for each object, the ellipse contains exactly (1-α)% of the corresponding bootstrap points. Using confidence ellipses, the relative positions of the objects points are clearly depicted, and -at the same time- the sizes and shapes of the ellipses nicely visualize stability and dependencies among the points. Based on earlier work by Meulman and Heiser (1983), Linting, Meulman, Groenen, and Van Der Kooij (2007) have described a nonparametric procedure for calculating confidence ellipses that exhibits greater flexibility than producing confidence ellipses based on the bivariate normal distribution.

3.2 *Density Plots*

For the subject points, it is typically less important to distinguish between individual subjects. Moreover, as most applications involve many subjects, plotting confidence ellipses leads to a cluttered plot in which it is difficult to disentangle the ellipses. However, it is informative to spot areas with small and large concentrations of subjects. Simply plotting all the points may already

show this to some extent, but as identical coordinates are depicted only once, density effects are ignored. A smooth depiction of the density can be obtained by using some form of two-dimensional density estimation. Here, we estimate the density using a bivariate kernel density estimation procedure proposed by Botev (2009). The densities are indicated by color intensity.

The suggested plotting procedures will be illustrated in Section 6.

4 Stability Measures

The graphical procedure proposed in Section 3, allows for a visual inspection of the stability of individual points. However, the bootstrap results may also be used to measure the overall stability of an unfolding solution. We propose the following measures: total variation and mean squared error.

4.1 *Total Variation*

Total variation can be calculated by considering, for each point, the squared Euclidean distance between the bootstrap points and their 'point of gravity'. The point of gravity is determined by the mean location of that particular point across all bootstrap unfolding solutions. Assuming that all points (subjects and objects) are equally important, a simple measure of the total variation of the unfolding solution would be the average squared deviation from all bootstrap points to their means. However, as the number of subjects typically exceeds the number of objects by a considerable margin, the subject bootstrap variation is likely to represent the largest component contributing to the total bootstrap variation. Therefore, we first examine the two sources of variation separately, and then propose an overall measure based on equal weights for both sources of variation.

To calculate the total subject bootstrap variation, we first define the total bootstrap variation for subject i as

$$TV_i = \frac{1}{B}\sum_{b=1}^{B}\left(\mathbf{x}_{ib} - \bar{\mathbf{x}}_i\right)'\left(\mathbf{x}_{ib} - \bar{\mathbf{x}}_i\right),$$

where, $\mathbf{x}_{ib}$ denotes the coordinate vector for subject i in the bth bootstrap configuration, B denotes

the number of bootstrap samples, and $\bar{\mathbf{x}}_i$ is the mean bootstrap coordinate vector for subject i. The mean total subject variation becomes:

$$MTV_{subjects} = \frac{1}{n}\sum_{i=1}^{n} TV_i.$$

The total object variation for object j can be calculated in a similar fashion. That is,

$$TV_j = \frac{1}{B}\sum_{b=1}^{B}\left(\mathbf{y}_{jb} - \bar{\mathbf{y}}_j\right)'\left(\mathbf{y}_{jb} - \bar{\mathbf{y}}_j\right),$$

where $\mathbf{y}_{jb}$ denotes the coordinate vector for object j in the bth bootstrap configuration, $\bar{\mathbf{y}}_j$ is the mean bootstrap coordinate vector for object j. The mean total object variation may be defined as:

$$MTV_{objects} = \frac{1}{p}\sum_{j=1}^{p} TV_j.$$

4.2 *Mean Squared Error*

In addition to variance, bias is of key importance to assess the validity of an unfolding solution. To gain an estimate of the bias, one may consider the deviation of the mean bootstrap configuration to the unfolding solution. Hence, the squared bias for the ith subject point is

$$Bias_i^{\,2} = (\bar{\mathbf{x}}_i - \hat{\mathbf{x}}_i)'(\bar{\mathbf{x}}_i - \hat{\mathbf{x}}_i),$$

where $\hat{\mathbf{x}}_i$ is the coordinate vector for subject i in the unfolding solution.

To assess stability of an unfolding configuration, one should take into account both variance and bias. This may be achieved by using the mean squared error. The mean squared error (*MSE*) for the subject points can be calculated as:

$$MSE_i = \frac{1}{B}\sum_{b=1}^{B}(\mathbf{x}_{ib} - \hat{\mathbf{x}}_i)'(\mathbf{x}_{ib} - \hat{\mathbf{x}}_i), \tag{3}$$

and the mean squared error for the object points, as

$$MSE_j = \frac{1}{B}\sum_{b=1}^{B}(\mathbf{y}_{jb} - \hat{\mathbf{y}}_j)'(\mathbf{y}_{jb} - \hat{\mathbf{y}}_j),$$

where, $\hat{\mathbf{y}}_j$ is the coordinate vector for object *j* in the original unfolding solution.

Note that (3) can be rewritten as

$$MSE_i = (\bar{\mathbf{x}}_i - \hat{\mathbf{x}}_i)'(\bar{\mathbf{x}}_i - \hat{\mathbf{x}}_i) + \frac{1}{B}\sum_{b=1}^{B}(\mathbf{x}_{ib} - \bar{\mathbf{x}}_i)'(\mathbf{x}_{ib} - \bar{\mathbf{x}}_i) = Bias_i^{\,2} + TV_i, \tag{4}$$

showing that the mean squared error can be decomposed in a bias and variance part. Hence, for unbiased estimators, the mean squared error equals the variance, whereas for biased estimators the mean squared error is equal to the sum of the squared bias and the variance. The total mean squared error for the subjects then becomes

$$TMSE_{subjects} = \sum_{i=1}^{n} MSE_i = \sum_{i=1}^{n}\left(Bias_i^{\,2} + TV_i\right),$$

and the total mean squared error for the objects is

$$TMSE_{objects} = \sum_{j=1}^{p} MSE_j = \sum_{j=1}^{p}\left(Bias_j^{\,2} + TV_j\right).$$

The mean squared error measure proposed in (4) considers the deviations of the bootstrap samples from the unfolding solution. However, as the scale of unfolding solutions is arbitrary (because only relative positions are important in unfolding analysis), the actual size of the mean squared error is not very informative and cannot be used to compare different solutions. To overcome this

indeterminacy, we propose a relative mean squared error measure.

Define the total sum of squares for the subject and object coordinates as

$$TSS_{subjects} = \sum_{i=1}^{n} \hat{\mathbf{x}}_i' \hat{\mathbf{x}}_i$$

and

$$TSS_{objects} = \sum_{j=1}^{p} \hat{\mathbf{y}}_j' \hat{\mathbf{y}}_j ,$$

respectively. For each set of points we can define the relative mean squared error *(RMSE)* as the total mean squared error divided by the total sum of squares

$$RMSE_{subjects} = \frac{TMSE_{subjects}}{TSS_{subjects}} = \frac{\sum_{i=1}^{n}\sum_{b=1}^{B} (\mathbf{x}_{ib} - \hat{\mathbf{x}}_i)'(\mathbf{x}_{ib} - \hat{\mathbf{x}}_i)}{B\sum_{i=1}^{n} \hat{\mathbf{x}}_i' \hat{\mathbf{x}}_i}.$$

and

$$RMSE_{objects} = \frac{TMSE_{objects}}{TSS_{objects}} = \frac{\sum_{j=1}^{p}\sum_{b=1}^{B} (\mathbf{y}_{jb} - \hat{\mathbf{y}}_j)'(\mathbf{y}_{jb} - \hat{\mathbf{y}}_j)}{B\sum_{j=1}^{p} \hat{\mathbf{y}}_j' \hat{\mathbf{y}}_j}$$

These measures consider the bootstrap variation around the unfolding solution relative to the bootstrap variation around the origin. These measures may become larger than 1, in which case a solution that places all points at the origin has a smaller mean squared error than the unfolding solution. It is useful to calculate the measures for the two sets of points separately as overall stability may be dominated by stability, or lack thereof, in one of the two sets. For example, in the case of a degenerate solution with all subject points placed in the origin and the object points

placed on a circle around them, we may find stability for the subjects but large instability for the objects as their locations on the circle are arbitrary. Therefore, a joint measure needs to be constructed. We consider the average of the two relative mean squared errors, that is,

$$RMSE = \frac{1}{2}\left(RMSE_{subjects} + RMSE_{objects}\right). \qquad (5)$$

5 Stability and Degeneracy

In nonmetric unfolding, variance and bias are likely to depend upon the choice of $\boldsymbol{\lambda}$ and ω. For example, by choosing a weak penalty, a degenerate unfolding solution may be avoided in the original sample but not in some bootstrap samples. Hence, the final unfolding solution may become unstable and biased. On the other hand, if the optimal transformations differ significantly from linear transformations, a strong penalty (enforcing such linear transformations) may also lead to higher variance and/or bias. Thus, the variance and bias of nonmetric unfolding configurations are a function of the penalty parameters. We use this relationship to find appropriate values for the penalty parameters. More specifically, our aim is to find $\boldsymbol{\lambda}$ and ω so that the *RMSE* in (5) is as small as possible.

5.1 *Local Search Algorithm*

Unfortunately, it is not possible to determine analytically how the stability measures discussed earlier on are related to the parameters $\boldsymbol{\lambda}$ and ω. So, one way to determine $\boldsymbol{\lambda}$ and ω that minimize the *RMSE* is by employing a *grid search*. For example, for λ and ω we consider combinations of the values [0.1, 0.2, …, 1.0] and [0.10, 0.20, 0.50, 1, 2, 5, 10, 20, 50, 100] for λ and ω respectively. For each combination in this grid, bootstrap analyses are performed. Depending on the size of the sample data, the number of bootstrap replications for each combination of λ and ω, and the size of the grid, the grid search may become too time consuming.

Alternatively, if stability decreases more or less monotonically when the penalty becomes either too strong or too weak, a *greedy search* algorithm over the space of λ and ω to find a minimum is computationally more feasible. The greedy search does not necessarily yield the

global minimum in terms of the lowest *RMSE*. However, it is likely to yield a solution of comparable and near optimal stability. Our proposal is to use the following greedy search algorithm: (1) start with some initial values for the penalty parameters, say, $\lambda = \lambda_0$ and $\omega=\omega_0$, and (2) move to neighboring positions on the grid until the improvement in the relative mean-squared error measure is larger than some (small) predetermined threshold value. The choice of the initial values (λ_0, ω_0) may be critical for the effectiveness of the greedy search algorithm. Since the measures introduced in this paper are all new and no prior study has examined the stability of nonmetric unfolding solutions, it is still an open question as to what values for the relative mean squared error are reasonable, and what initial values for the algorithm are recommendable. However, in their simulation study, Busing et al. (2005) found that degeneracy is typically avoided when $\lambda \leq 0.5$ and $\omega \geq 0.5$. Given the (previously mentioned) changes in the objective function with respect to the method described in Busing et al. (2005) and relying on the assumption that nondegenerate solutions are more stable than degenerate solutions, we propose to use as start values:

$$\lambda_0 = 0.5 \text{ and } \omega_0 = 1/(2\,v^2(\delta)). \qquad (6)$$

6 Marketing Applications

To illustrate different aspects of the proposed methodology, we present three marketing applications. First, we make use of the so-called breakfast data (Green and Rao, 1972). This is a well-known and often used (see, for instance, Busing et al., 2005; Borg and Groenen, 2005) data set consisting of preference rankings for 42 individuals (the subjects) on 15 breakfast items (the objects). Next, we re-analyse the citation data presented in Ho et al. (2010) to show how stability and degeneracy are related. Moreover, our analysis offers additional insights into the results obtained by Ho et al. (2010). Finally, we show how our approach to nonmetric unfolding can be employed with data sets collected in industry. To this end, we analyzed a data set consisting of preference rankings for soup-ideas.

6.1 *Breakfast Data*

In our analysis of the breakfast data, we try to identify those values for the penalty parameters that yield the most stable joint configuration of objects and subjects. We used the full grid search with $\lambda \in$ [0.1,0.2,…,1.0], and $\omega \in$ [0.10, 0.20, 0.50, 1, 2, 5, 10, 20, 50, 100]. For each pair (λ, ω) we performed penalized nonmetric unfolding and a bootstrap analysis with B=1,000 replications. The resulting *RMSE*, as defined in equation (5), are found in Table 1.

Examination of the values in Table 1 reveals that the most stable unfolding solution corresponds to penalty parameters $\lambda = 0.7$ and $\omega = 10$. The solution corresponding to the current default values, $\lambda = 0.5$ and $\omega = 1$, is considerably less stable. Note that by decreasing the penalty (that is, choosing higher values for λ and lower values for ω), the *RMSE* increases significantly. This result indicates that the corresponding unfolding solutions are exceedingly unstable, suggesting the possible occurrence of degeneracies. Similarly, the effect of imposing a stronger penalty by increasing the ω parameter is limited, but it generally leads to more stable unfolding solutions. The values in the lower left corner of Table 1 show that once the penalty becomes too weak, the *RMSE* increases substantially, in some cases even exceeding 1.

Table 1 shows that differences in stability among the most stable solutions are small. The median *RMSE* value over the grid is 0.0936. To see to what extent differences and similarities in stability have influence on the final configurations, a quantitative comparison of different configurations is needed. For this purpose, the alienation coefficient as described by Borg and Leutner (1985) is computed. The alienation coefficient, which lies between zero and one, can be interpreted as a measure of dissimilarity between two unfolding configurations. It directly compares the Euclidean distances within the unfolding configurations. A low value for the alienation coefficient indicates that the two configurations are similar (with zero indicating a perfect match). In Table 2, alienation coefficients between the optimal unfolding configuration and the unfolding configurations corresponding to all other parameter combinations are presented. We see that the optimal solution is more similar to other stable solutions than to the default solution. In general, it appears that as stability decreases, solutions become less similar**.**

Table 1: Relative mean squared errors for different penalty settings based on 1,000 bootstrap samples of the breakfast data.

	ω									
λ	0.1	0.2	0.5	1	2	5	10	20	50	100
0.1	0.1037	0.1023	0.0929	0.0928	0.0934	0.0937	0.0939	0.0939	0.0940	0.0941
0.2	0.1188	0.0925	0.0819	0.0848	0.0871	0.0885	0.0896	0.0898	0.0900	0.0901
0.3	0.2417	0.1396	0.0866	0.0757	0.0777	0.0818	0.0835	0.0839	0.0854	0.0855
0.4	0.6864	0.2756	0.1240	0.0798	0.0735	0.0741	0.0774	0.0771	0.0776	0.0778
0.5	1.0384	0.5592	0.2104	0.1175	**0.0679**	0.0734	0.0736	0.0734	0.0740	0.0769
0.6	0.8343	0.6708	0.4018	0.1601	0.0808	**0.0684**	0.0690	0.0697	0.0730	0.0755
0.7	0.8350	0.7106	0.4617	0.2842	0.1314	0.0739	**0.0676**	0.0802	**0.0680**	**0.0680**
0.8	1.0004	0.9613	0.7594	0.5596	0.2161	0.0968	0.0793	0.0730	0.0714	0.0716
0.9	1.0623	0.9361	0.9922	0.8198	0.5909	0.2368	0.1519	0.0938	0.0868	0.0850
1	0.9315	0.8541	0.5822	0.8640	0.9760	0.4917	0.4842	0.2999	0.2148	0.1995

Notes. Five smallest values are printed in boldface. Values on the lower left side of the separation line are generally larger than 0.20 and/or at least twice as large as the values on the other side of the line. Shaded cells indicate combinations considered when the local search algorithm is used.

Figure 1 provides the most stable configuration for objects and subjects, with 90% confidence ellipses. We see that the stability of different breakfast items differs considerably. Certain breakfast items, in particular "toast pop-up" (TP) and "cinnamon toast" (CT) have larger ellipses around their bootstrap means than other breakfast items indicating that the locations of these breakfast items vary more over the different bootstrap samples. From the density clouds, it is clear that "danish pastry" (DP) and "cinnamon bun" (CB) are the most popular breakfast items. As far as statistical information is concerned, the confidence areas offer a means to assess the stability of each individual object or subject positioned in the unfolding solution. For instance, non-overlapping areas of groups of objects may indicate significant (or substantial) differences in brand or product perceptions and, conversely, overlapping areas indicate insignificant differences in brand or product perceptions. For example, in Figure 1 we see that the breakfast items hard rolls and butter (HRB) and toast and margarine (TMn) show great overlap indicating similar perception of these breakfast items.

Table 2: Alienation coefficients between solutions of the breakfast data and the optimal configuration with penalty parameters λ =0.7 and ω=10.

	ω									
λ	0.1	0.2	0.5	1	2	5	10	20	50	100
0.1	0.1933	0.1898	0.1635	0.1614	0.1616	0.1620	0.1622	0.1623	0.1624	0.1624
0.2	0.1827	0.1467	0.1398	0.1457	0.1523	0.1554	0.1564	0.1569	0.1571	0.1572
0.3	0.2828	0.1773	0.1242	0.1227	0.1309	0.1415	0.1462	0.1479	0.1487	0.1490
0.4	0.3960	0.2495	0.1471	0.1063	0.1108	0.1197	0.1253	0.1281	0.1293	0.1302
0.5	0.4509	0.4208	0.1955	0.1267	**0.0756**	0.1051	0.1091	0.1109	0.1110	0.1122
0.6	0.7074	0.5956	0.3258	0.1882	0.0580	**0.0599**	0.0721	0.0736	0.0960	0.1025
0.7	0.7621	0.6902	0.4638	0.2483	0.1426	0.0475	**0**	0.0196	**0.0452**	**0.0502**
0.8	0.8164	0.7604	0.5871	0.3902	0.2095	0.0798	0.0593	0.0469	0.0244	0.0211
0.9	0.8574	0.8191	0.7012	0.5817	0.2654	0.1681	0.1431	0.0773	0.0718	0.0670
1	0.8923	0.8573	0.7969	0.7587	0.6976	0.2231	0.1846	0.1552	0.1480	0.1459

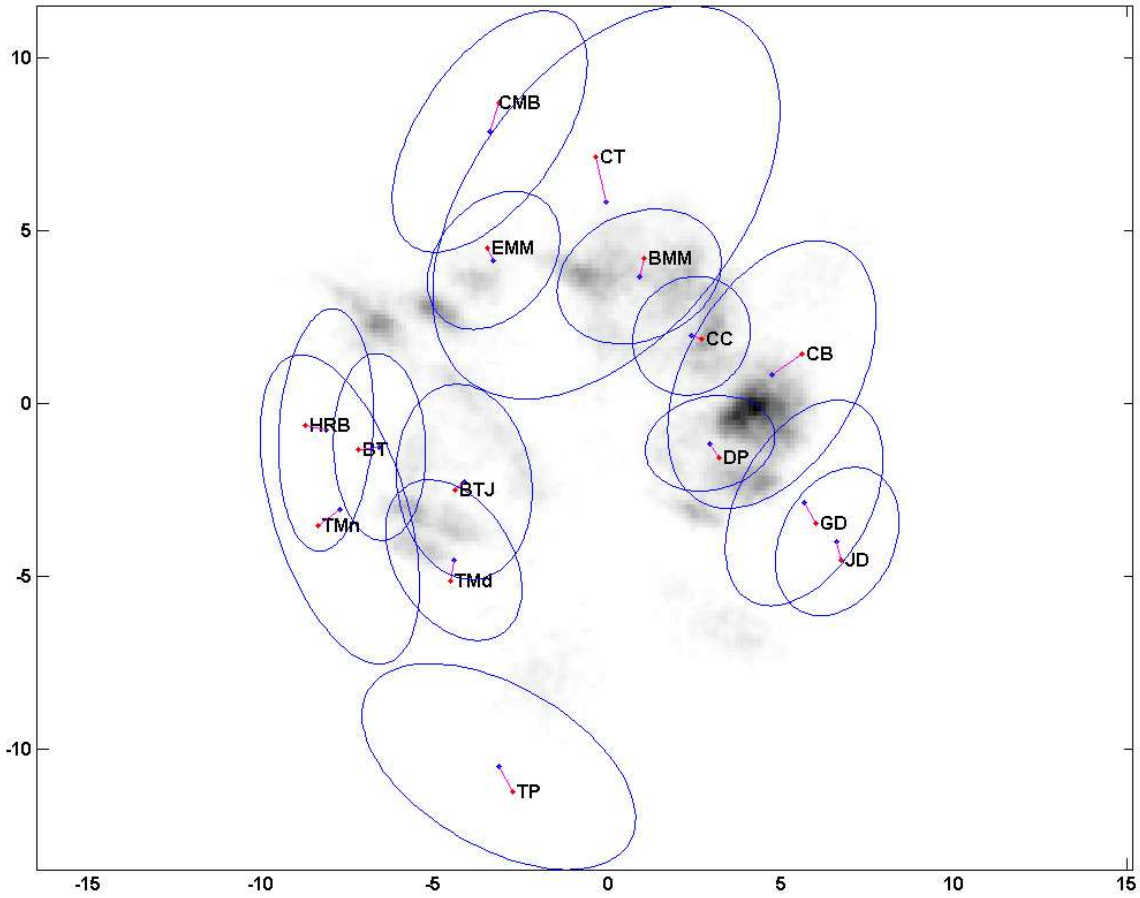
Notes. Low values indicate similarity. Bold faced values correspond to the 5 most stable solutions. Shaded cells indicate combinations considered in the local search algorithm.

Local search algorithm

The performance of the local search algorithm can be traced in Table 1 by considering the shaded cells. The variation coefficient for these data is 0.59, hence, using (6), we choose $\lambda = 0.5$ and $\omega = 2$ as initial values. In this case, the stability of only four (adjacent) combinations needs to be considered (see shaded cells in Table 1). The corresponding solution ($\lambda = 0.5$ and $\omega = 2$) is a local minimum. However, from Tables 1 and 2, we know that the *RMSE* of this solution as well as its corresponding (spatial) configuration are nearly equivalent to the solution and configuration as determined by the optimal values ($\lambda = 0.7$ and $\omega = 10$).

Figure 1: Nonmetric unfolding solution for breakfast items with penalty parameters λ=0.7 and ω=10

Notes. Ellipses represent 90% bootstrap confidence ellipses. Lines between points and centers of the ellipses depict biases. The gray to black clouds depict the density of the subjects' bootstrap points. Darker shades of gray indicate higher densities. The breakfast items (and labels) are: toast pop-up (TP), buttered toast (BT), English muffin and margering (EMM), jelly donut (JD), cinnamon toast (CT), blueberry muffin and margerine (BMM), hard rolls and butter (HRB), toast and marmelade (TMd), buttered toast and jelly (BTJ), toast and margarine (TMn), cinnamon bun (CB), Danish pastry (DP), glazed donut (GD), coffee cake (CC), and corn muffin and butter (CMB).



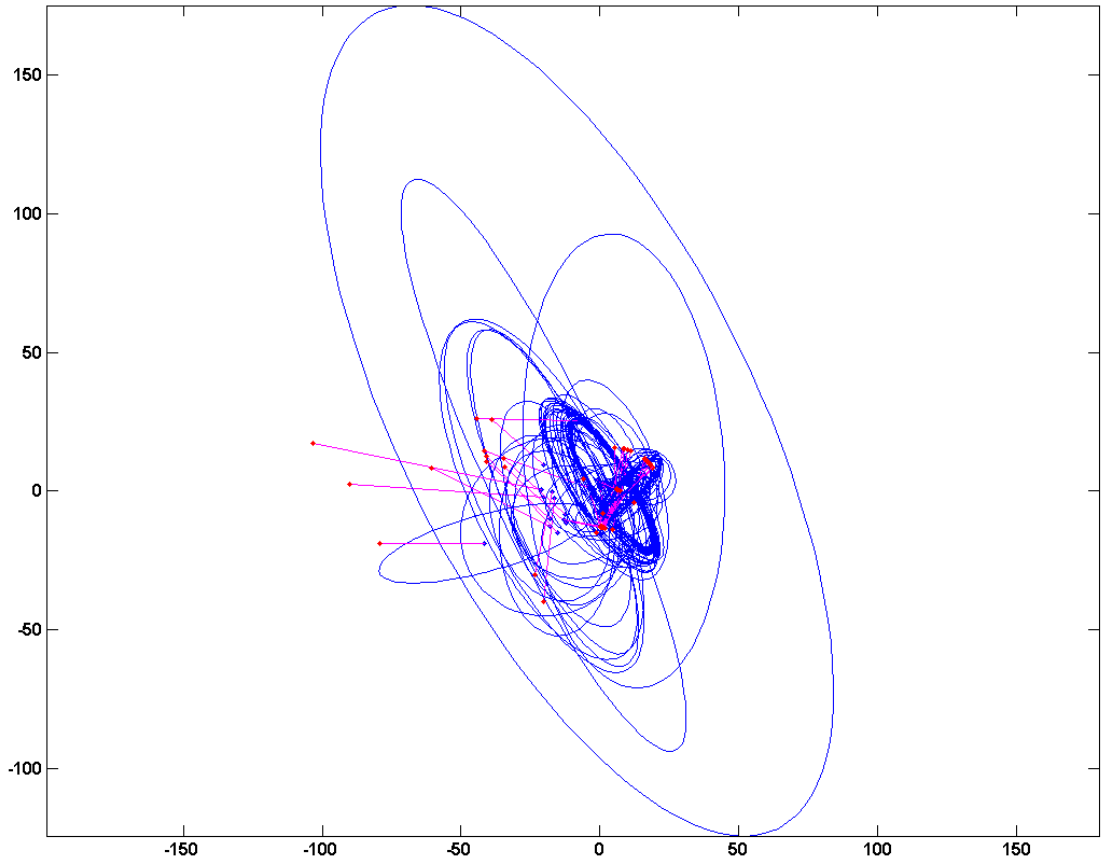
6.2 *Citation Network Data*

Next, we analyzed data concerning path distances of citations of scholars in marketing research. This data was taken from Ho et al. (2010). For 50 leading researchers, distance from one researcher to another is measured by counting the shortest path linking the citations. Thus, if author A refers to author B, but not to author C, and author B refers to author C, the distance between author A and C is 2. The 50 researchers are listed in Table 3. We started with the solution proposed (and presented) in Ho et al. (2010) in which both penalty parameter values were set to 0.5. This solution turned out to be a degenerate solution (see also Fig.2 as included in Ho et al. 2010). Interpreting the citing authors as observations and the cited authors as variables, we used a bootstrap analysis with B=1,000 replications. With a *RMSE* of 0.9094, this solution is not stable. To illustrate the poor stability of such a degenerate solution, Figure 2 gives the corresponding plot with 90% confidence ellipses around all points.

Table 3: Names and labels of researchers in the citation data

Upper case label	Last name	First name	Upper case label	Last name	First name
GA	Allenby,	Greg	GI	Iyer,	Ganesh
RA	Andrews,	Rick	CJ	Janiszewski,	Chris
DA	Ariely,	Dan	RK	Kivetz,	Ran
EA	Arnould,	Eric	DL	Lehmann,	Donald
WB	Bearden,	William	PM	Manchanda,	Puneet
EB	Bradlow,	Eric	VM	Mittal,	Vikas
BB	Bronnenberg,	Bart	NM	Morgan,	Neil
SB	Brown,	Stephen	VMO	Morwitz,	Vicki
JB	Burroughs,	James	RN	Netemeyer,	Richard
MC	Campbell,	Margaret	LP	Peracchio,	Laura
YC	Chen,	Yuxin	PR	Rossi,	Peter
AC	Chernev,	Alexander	JS	Sherry	John Jr.
PC	Chintagunta,	Pradeep	SS	Shugan,	Steven
JC	Cohen,	Joel	DS	Simester,	Duncan
WD	Desarbo,	Wayne	IS	Simonson,	Itamar
RD	Dhar,	Ravi	SST	Stremersch,	Stefan
JD	Dube,	Jean-Pierre	KS	Sudhir,	Karunakaran
PF	Fader,	Peter	BS	Sun,	Baohong
GF	Fitzsimons,	Gavan	GT	Tellis,	Gerard
VF	Folkes,	Valerie	GT	Thompson,	Craig
PHF	Franses,	Philip Hans	HV	VanHeerde	Harald
GD	Grewal,	Dhruv	MB	Villas-Boas,	Miguel
SG	Gupta,	Sachin	MW	Wedel,	Michel
CH	Homburg,	Christian	DW	Wittink,	Dick
JH	Huber,	Joel	JZ	Zhang,	John

Figure 2: Degenerate Nonmetric unfolding solutions with penalty parameters λ =0.5 and ω=0.5. Ellipses represent 90% bootstrap confidence regions. Lines between points and centers of the ellipses depict biases. Labels are omitted as interpretation is not possible.



Ho et al. (2010) suggested a solution to the degeneracy that amounts to using metric unfolding. Taking into account the nonmetric nature of the data, we shall seek a nonmetric unfolding solution that is both stable and nondegenerate. To achieve this, we use the greedy local search algorithm outlined in Section 5.1. The variation coefficient for this data is 0.32. Hence, using (6), we set λ=0.5 and ω=10 as initial values, where we rounded the ω value for convenience. The *RMSE* corresponding to these parameters is 0.1307. Next, we relied on the use of a local search algorithm as outlined in Section 5.1 with 0.0002 as threshold value for an improved solution (i.e., only a solution with an improved *RMSE* of at least 0.0002 is considered). This small threshold value is chosen here for illustration purposes only. The greedy search algorithm

identified, after having considered 20 combinations, the optimal values λ=0.3 and ω=200. In Table 4, the *RMSE* values obtained using the local search algorithm are provided as well as some additional values for specific combinations which are close to the search path.

To evaluate whether small changes in stability, cause substantial changes in the configurations, we considered the alienation coefficients with respect to the λ=0.3 and ω=200 configuration. The results are presented in Table 5. As with the breakfast data, we see that solutions with nearly equivalent *RMSE*'s yield configurations which are nearly identical. Hence, although the local search algorithm does not yield the global minimum, the obtained configuration is adequate.

By raising the threshold value, we could further decrease the number of combinations that need to be considered. For example, if we only consider improvements in *RMSE* of at least 0.001 rather than the previously used 0.0002 value, only 15 combinations would be evaluated to find as optimal values λ=0.3 and ω=50. We can see in Table 5, that this solution is very similar to the λ=0.3 and ω=200 configuration.

Table 4: Relative mean squared errors for penalty settings considered following the local search algorithm, based on 1,000 bootstrap samples of the citation data.

					ω			
λ	2	5	10	20	50	100	200	500
0.1			0.1154					
0.2		0.1156	0.1120	0.1104	0.1096	0.1094	0.1093	
0.3			0.1154		0.1066	0.1063	0.1061	0.1060
0.4			0.1180			0.1126	0.1124	
0.5		0.1465	0.1307	0.1234				
0.6			0.1678					

The resulting configuration for objects (destination nodes) and subjects (source nodes) are shown in Figure 3. In Figure 4, 90% confidence ellipses were added for destination node points and stability of source node points was indicated using the density plot. We used both of these plots to interpret the citation network data.

The group of destination node points (uppercase labels in Figure 4) on the left-hand-side

of the plot corresponds to authors that, overall, do not receive many citations from other authors in the set. The citations that these authors receive are mostly self-citations or citations from other authors positioned on the left-hand-side. If we take a closer look at the research interests of these authors we see that they have a strong interest in topics such as semiotics, symbolic consumption (James Burroughs: JB, John Sherry: JS), customer anthropology, lifestyle and culture (Craig Thompson: CT, Laura Peracchio: LP, Eric Arnould: EA). The research methods employed by these authors tend not to have such a strong focus on quantitative modeling in marketing.

Table 5: Alienation coefficients between solutions of the citation data and the optimal configuration with penalty parameters λ=0.3 and ω=200.

				ω				
λ	2	5	10	20	50	100	200	500
0.1	0.0835	0.0419	0.0319	0.0281	0.0263	0.0258	0.0256	0.0254
0.2	0.0409	0.0235	0.0182	0.0163	0.0156	0.0154	0.0153	0.0153
0.3	0.0216	0.0125	0.0096	0.0032	0.0011	0.0004	0	0.0002
0.4	0.0873	0.0764	0.0766	0.0769	0.0771	0.0772	0.0772	0.0772
0.5	0.1411	0.0984	0.0914	0.0894	0.0885	0.0882	0.0880	0.0879
0.6	0.7639	0.1289	0.1216	0.1170	0.1137	0.1128	0.1123	0.1120

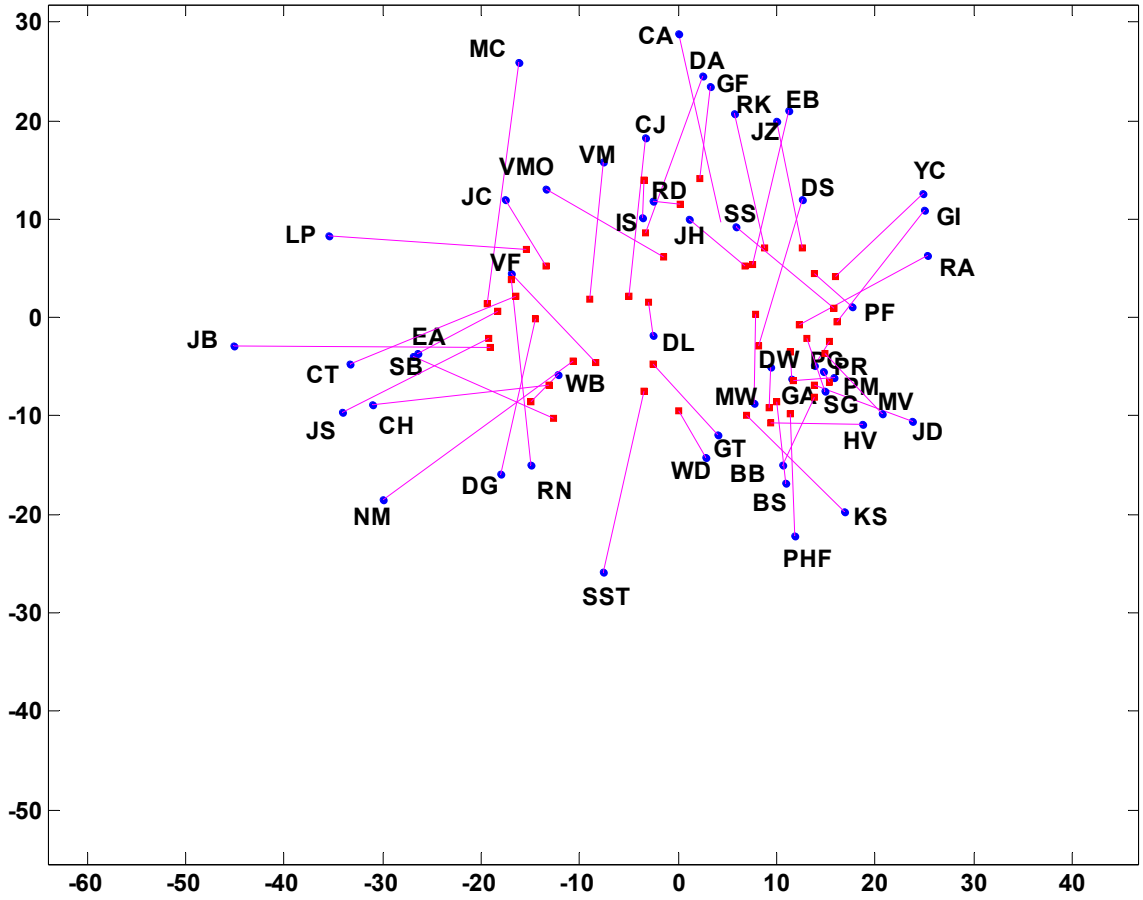
Notes. Low values indicate similarity. The shaded cells indicate the combinations considered following the local search algorithm.

If we look at the right-hand-side of the plot we see a large group of intermixed source and destination nodes. The mixture of both types indicates that these authors frequently cite both their own and each other's work. Furthermore, they generally do not cite the work of the authors on the left-hand-side of the plot. At the center of this group we find authors such as Dick Wittink (DW), Pradeep Chintagunta (PC), Peter Rossi (PR), Sachin Gupta (SG), Greg Allenby (GA) and Michel Wedel (MW). In general, the work of people in this group, and the surrounding scholars, typically involves quantitative modeling of customer preferences and/or purchase behavior. Several of these authors are close collaborators and co-authored several papers. The destination nodes slightly further removed from the center (e.g., Philip Hans Franses: PHF, Jean-Pierre Dube: JD, and Rick Andrews: RA) correspond to authors that are cited less frequently (by the 50 marketing

scholars included in this map) than those located near the center of this cluster.

A closer inspection of the different destination nodes also illustrates that distances between them are good indicators of the similarity/relatedness of the research of the corresponding authors. For example, the proximity of the three destination nodes of Harald Van Heerde: HV, Jean-Pierre Dube: JD, and Miguel Villas-Boas: MB, aligns quite nicely with their research interests and publications, which often involve a fair amount of econometrics applied to issues concerning price competition.

Figure 3: Nonmetric unfolding solutions for objects (destination nodes, labeled using author's initials) and subjects (source nodes, not labeled) of the citation data with penalty parameters λ =0.3 and ω=200. The lines connect an author's destination node to the source node of the same author. See Table 3 for the list of authors' initials.



The points corresponding to destination nodes located towards the centre top of the plot involve authors for which the source nodes appear to be less stable. Apparently, these authors do not typically cite each other's work as frequent as they cite the work of the other marketing scholars shown in the plot. If we look at the research interests of some of the authors in this group, we see that they tend to deal with consumer psychology, irrationality and behavioral economics (e.g., Dan Ariely: DA, Gaven Fitzsimmons: GF, and Ran Kivetz: RK).

Finally, the confidence ellipses make it possible to immediately detect authors that are difficult to position in the two dimensional plot. For example, the rather large ellipses around the destination node points corresponding to Stefan Stremersch (SST) and Wayne Desarbo (WD), suggests that one should be careful in interpreting distances from and to these points in Figures 3 and 4.

6.3 *Soup Idea Data*

We analyzed data from seventy-six untrained customers, all between 18 and 35 years old, who were invited in a testing laboratory owned by a commercial research agency. All respondents were given the name of 11 product ideas for ready-made soups as well as a short description including a list of (special) ingredients of each soup. They were asked to rank-order 11 ideas for new soups (without tasting). In addition, purchase intention for each soup idea was measured using a 5-point scale including the scale points 'certainly won't buy' (1), 'probably won't buy' (2), 'don't know' (3), 'probably will buy' (4) 'certainly will buy'. The cumulative percentage of '4'and '5' scores (that is, respondents who consider buying the product/soup idea) was referred to in this study as the top 2-box percentage of purchase intention.

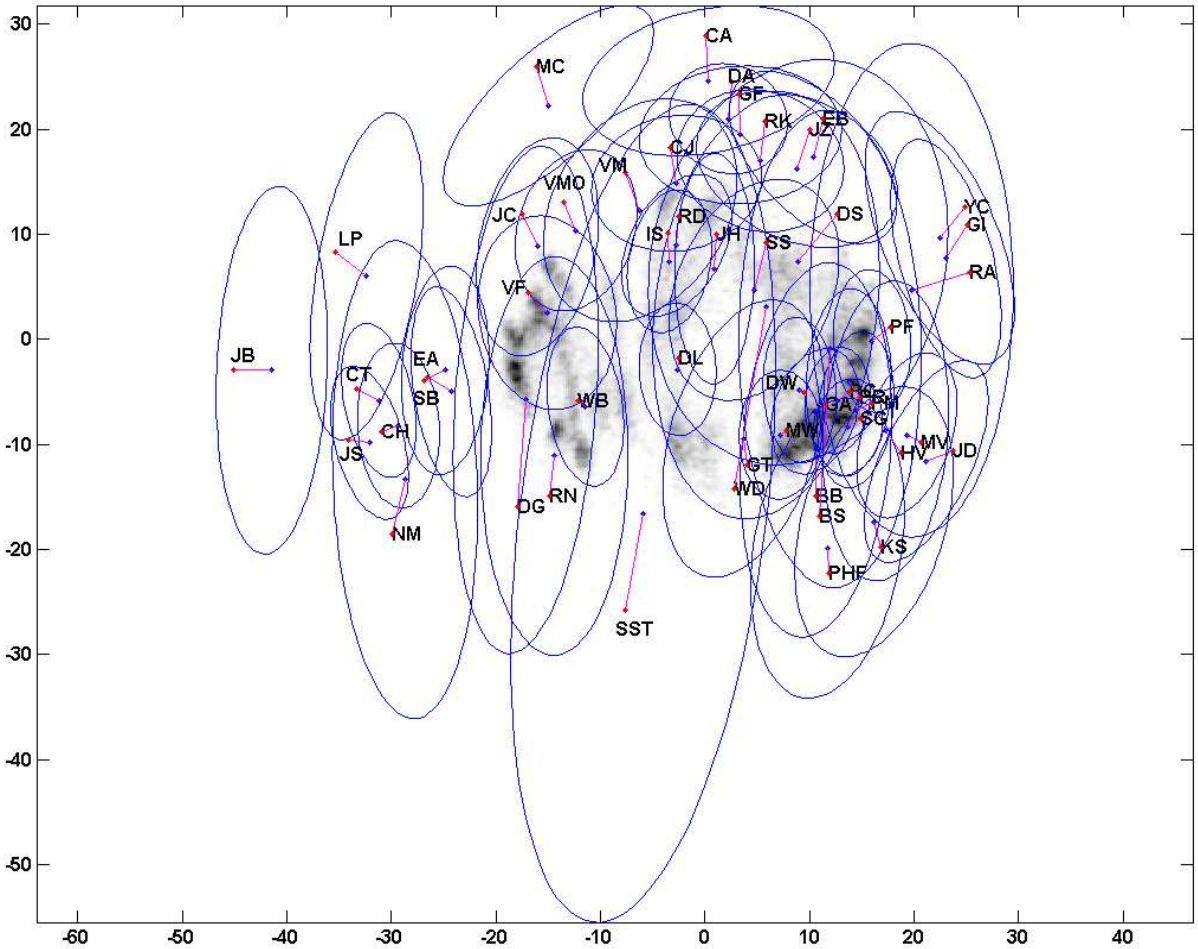
The following product ideas for soups were included (with labels between brackets):

- Tomato soup, creamy (Tl)
- Tomato soup with special herbs (T2)
- Spicy tomato soup (T3)
- French type of Mustard soup. (F1)
- French type of Mushroom soup (F2)
- Vegetable soup, asparagus (V1)
- Vegetable soup, broccoli (V2)

- Vegetable soup, Celery (V3)
- Pea soup (P)
- Oriental soup, Thai vegetable (O1)
- Orietnal soup, chicken tikka soup (O2)

Figure 4: Nonmetric unfolding solutions for objects (destination nodes) and subjects (source nodes) of the citation data with penalty parameters λ =0.3 and ω=200

Notes. Ellipses represent 90% bootstrap confidence regions around the destination node points (labels in Table 3). Lines between points and centers of the ellipses depict biases. The gray to black clouds depict the density of the source node bootstrap points. Darker shades of gray indicate higher densities.



Tables 6 and 7 provide the *RMSE*'s for all combinations of penalty parameters, and the alienation coefficients with respect to the best solution, respectively. A similar picture as seen in the previous two applications emerges. Increasing the penalty generally leads to a more stable solution. However, especially if the λ penalty parameter becomes too small (that is, too strong) stability eventually decreases. The lower left corner of the table again shows that a weak penalty leads to very unstable solutions. The smallest *RMSE* for this data is obtained when $\lambda = 0.6$ and $\omega = 20$. The resulting configuration, with 90% confidence ellipses, is provided in Figure 5. As with the previous applications, we found that the differences with the second best solution ($\lambda = 06$ and $\omega = 10$) are, in terms of stability and relative positions, quite small. On the other hand, the default values of $\lambda = 0.5$ and $\omega = 1$ (in SPSS), yielded a solution that is clearly less stable than the optimal solution.

Table 6: Relative mean squared errors for different penalty settings based on 1,000 bootstrap samples of the soup data.

	ω									
λ	0.1	0.2	0.5	1	2	5	10	20	50	100
0.1	0.1874	0.2091	0.1808	0.1811	0.1839	0.1861	0.1867	0.1872	0.1872	0.1872
0.2	0.1675	0.1409	0.1366	0.1455	0.1553	0.1635	0.1661	0.1671	0.1677	0.1691
0.3	0.2529	0.1741	0.1228	0.1168	0.1259	0.1373	0.1423	0.1457	0.1468	0.1470
0.4	0.6181	0.2938	0.1461	0.1163	**0.1084**	0.1159	0.1200	0.1228	0.1254	0.1262
0.5	1.3997	0.6824	0.2204	0.1350	0.1116	**0.1083**	**0.1099**	0.1112	0.1124	0.1126
0.6	2.4036	1.0084	0.3532	0.1858	0.1301	0.1129	**0.1077**	**0.1066**	0.1104	0.1111
0.7	1.3064	1.4642	0.4916	0.3066	0.1704	0.1304	0.1216	0.1205	0.1177	0.1140
0.8	1.1362	1.0423	0.7224	0.4794	0.2432	0.1581	0.1426	0.1395	0.1355	0.1354
0.9	1.4213	1.0480	0.5238	0.7691	0.4744	0.2094	0.1744	0.1585	0.1554	0.1527
1	1.2755	1.3754	0.8270	0.5822	1.1569	0.3964	0.2331	0.1969	0.1908	0.1838

Notes. Five smallest values are printed in bold face. Values on the left and below the separation line are generally larger than 0.20 and/or at least twice as large as the values on the other side of the line. Shaded cells indicate combinations considered when the local search algorithm is used.

Table 7: Alienation coefficients between solutions of the soup idea data and the optimal configuration with penalty parameters λ=0.6 and ω=20.

					ω					
Λ	0.1	0.2	0.5	1	2	5	10	20	50	100
0.1	0.2133	0.2128	0.1407	0.1271	0.1251	0.1252	0.1254	0.1255	0.1256	0.1257
0.2	0.1796	0.0978	0.0786	0.0864	0.0927	0.0969	0.0981	0.0987	0.0990	0.1146
0.3	0.3080	0.1887	0.0541	0.0502	0.0758	0.0819	0.0838	0.0846	0.0852	0.0853
0.4	0.4317	0.2692	0.1291	0.0672	**0.0425**	0.0689	0.0757	0.0776	0.0785	0.0787
0.5	0.4975	0.4281	0.2083	0.1238	0.0674	**0.0535**	**0.0584**	0.0626	0.0624	0.0624
0.6	0.5643	0.5069	0.3016	0.1901	0.1181	0.0584	**0.0144**	**0.0000**	0.0446	0.0485
0.7	0.7372	0.5951	0.5514	0.2860	0.1738	0.1033	0.0806	0.0588	0.0479	0.0284
0.8	0.8283	0.7373	0.6696	0.3409	0.2485	0.1554	0.1303	0.0948	0.0876	0.0859
0.9	0.8730	0.8417	0.7977	0.6211	0.3335	0.2007	0.1665	0.1445	0.1362	0.1412
1	0.9265	0.9152	0.8778	0.8309	0.4416	0.2784	0.2347	0.2238	0.1715	0.1741

Notes. Low values indicate similarity. Bold faced values correspond to the 5 most stable solutions. Shaded cells indicate combinations considered when the local search algorithm is used.

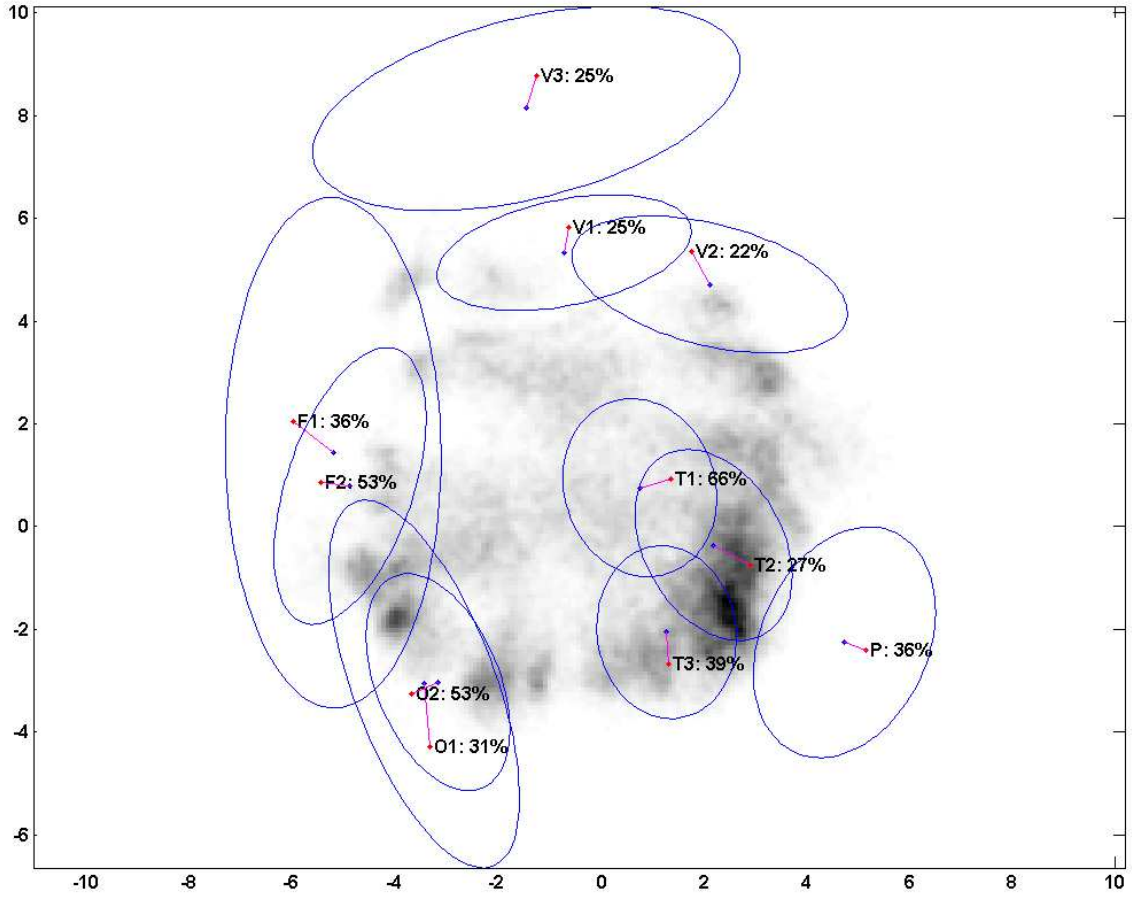
In Figure 5, we see that soups (soup ideas) are clustered quite naturally; vegetable soups are close to each other, tomato soups are near, the two oriental soups as well as the mushroom and mustard soups (French-type of soups) are close to each other with overlapping ellipses. It seems that they are difficult to distinguish from one another. Although the pea soup is in the proximity of the tomato soups, the small ellipses indicate that it is perceived differently. Inspecting the subjects, we see a large cluster of subjects close to the tomato soups. These respondents indicated a strong preference for either one of the tomato soups or the pea soup. A second cluster of subjects is situated between the oriental and French-type soups. For these consumers, the tomato soups are not their first choice but the tomato soups do receive larger rankings than the vegetable and pea soups, indicating that most respondents preferred the tomato soup ideas.

Based on the spatial map as presented in Figure 5 and the top 2-box percentages of purchase intention, the marketing manager was in a position to select those soup ideas that, on the basis of this concept test, seemed to have the highest market potential. The most promising soup ideas are: tomato soup T1 (top 2-box %: 66%, highest score obtained), oriental soup O2 and French-type of soup F2 (top 2-box in both cases: 53%, second highest score). This combination of 3 soup ideas would comprise a set that offers an interesting product offer to the great majority of

respondents.

Figure 5: Nonmetric unfolding solution for subjects and objects of the soup idea data with penalty parameters λ=0.6 and ω =20

Notes. Ellipses represent 90% bootstrap confidence regions. Lines between points and centers of the ellipses depict biases. The gray to black clouds depict the density of the subjects' bootstrap points. Darker shades of gray indicate higher densities. The abbreviations used to indicate the soup ideas are explained above. Percentages indicate top 2-box scores for purchase intention (see text).



Local search algorithm

The combinations considered by the local search algorithm are again denoted by shaded cells in Table 6. As the variation coefficient of the proximities is 0.55, the initial values used were λ = 0.5

and $\omega = 2$. A local minimum *RMSE* is attained for $\lambda = 0.5$ and $\omega = 5$. As we can see in Tables 5 and 6, this solution is nearly equivalent, in terms of *RMSE* and relative positions of the points, to the solution obtained after searching the complete grid. However, the amount of computations required using the local search algorithm is much lower as instead of 100, only 10 combinations of values for λ and ω had to be evaluated.

7 Large Scale Marketing Data

In marketing, one frequently encounters large data sets. If there are many subjects (e.g., customers) and/or many objects (e.g., brands) two potential problems are encountered: 1) As the number of parameters to be estimated increases, the unfolding task may become computationally demanding, and 2) If there are (too) many points, plotting all of them may not yield an interpretable picture. Many points may be located close to each other leading to a clutter of points that cannot easily be interpreted.

Ho et al. (2010) have dealt with the first issue. In a simulation study, Ho et al. (2010) showed that, as far as computation times are concerned, the penalized nonmetric unfolding approach, works satisfactory even with large data sets. Calculating a two-dimensional solution for the largest data set considered (500 x 500) with a true dimensionality of 10 using 10 random starts took just over 4 minutes. Note that, without transformations, that is, applying metric unfolding as Ho et al. (2010), suggested, these times would decrease dramatically. Considering the size of the data and the number of parameters to be estimated, this seems acceptable. One may infer from this that a single bootstrap analysis with 1000 replications (without using random starts) would take approximately 6 hours and 40 minutes. This is not prohibitive yet but it may become so, especially if many λ and ω combinations for the penalty parameters need to be considered. In such cases, the greedy local search algorithm as proposed in this paper offers a viable alternative to a full grid search.

The second problem faced when analyzing large scale marketing data, concerns the display of many points. In their illustration of movie-by-evaluator data, Ho et al. (2010), produced a plot showing a massive clutter containing all movies and evaluators. As a result, the interpretation of the plot became problematical. The density plots proposed in Section 3.2 for depicting the

bootstrapped subject points, can be used to overcome this problem. Instead of a clutter of points, the depicted densities can be used to identify areas where groups of subjects have similar/dissimilar preferences.

8 Summary and Conclusions

Nonmetric multidimensional unfolding is a powerful and intuitive tool that can be applied in several marketing settings. In this paper, we considered two important issues in nonmetric unfolding—degeneracy and (in)stability of unfolding solutions—and we proposed a new methodology to resolve these issues. We evaluated the applicability and usefulness of our methodology, which relies on both the conduct of a balanced bootstrap analysis and the calculation of several stability measures, by means of three illustrative examples. Not only did we find our method to work well in practice, we were also able to empirically validate a greedy local search algorithm that can be used to find a stable, nondegenerate solution. Although this algorithm does not necessarily produce the most stable solution (i.e., the optimal solution), it is likely to yield a high quality solution, that is a solution that is close enough to the most stable solution without requiring excessive computational effort and time.

Currently, the bootstrap procedure presented in this paper is not yet available in standard statistical software. A procedure for carrying out the balanced bootstrap in SPSS, as well as Matlab routines for plotting the results (including ellipses and the density representations for the subjects) may be obtained from the first author.

Our approach to analyze nonmetric preference data using multidimensional unfolding is at present the only one that provides stable, nondegenerate solutions. Thereby, it offers market researchers an important analysis tool. Market researchers are now in a position to derive and interpret spatial maps showing: (1) information on the similarity or dissimilarity between various preference choice alternatives (e.g., brands or products) as well as (2) the extent to which these choice alternatives are in line with the needs as expressed by distinct clusters of customers. Especially choice alternatives that are located in areas showing a high density of individual customers are interesting from a marketing perspective as they are preferred by a subtantial number of customers. Such marketing conclusions can now be drawn without running the risk that

one's interpretation is not legitimate due the invalidity of the nonmetric unfolding solution.

References

Bijmolt, T.H.A. and Wedel, M. (1999). A Comparison of Multidimensional Scaling Methods for Perceptual Mapping. *Journal of Marketing Research*, 36 (May), 277-85.

Borg, I. and Groenen, P.J.F. (2005). *Modern Multidimensional Scaling: Theory and Applications.* 2nd ed. New York, NY: Springer.

Borg, I. and Leutner, D. (1985). Measuring the similarity between MDS configurations. *Multivariate Behavioral Research,* 20, 325–34.

Borg, I. and Lingoes, J.C. (1987). *Multidimensional Similarity Structure Analysis.* Berlin, Germany: Springer.

Botev, Z. (2009). *Kernel Density Estimation* (downloadable Matlab function with documentation): http://www.mathworks.de/matlabcentral/fileexchange/17204-kernel-density-estimation

Busing, F.M.T.A. (2010). Advances in Multidimensional Unfolding. Doctoral thesis, Leiden University, The Netherlands.

Busing, F.M.T.A., Groenen, P.J.F. and Heiser, W.J. (2005). Avoiding Degeneracy in Multidimensional Unfolding by Penalizing on the Coefficient of Variation. *Psychometrika*, 70 (1), 71-98.

Chaturvedi, A. and Carroll, J.D. (1998). A Perceptual Mapping Procedure for Analysis of Proximity Data to Determine Common and Unique Product-Market Structures. *European Journal of Operational Research*, 111 (2), 268-84.

Coombs, C.H. (1950). Psychological Scaling Without a Unit of Measurement. *Psychological Review*, 57 (3), 145-58.

Cornelius, B., Wagner, U., and Natter, M. (2010). Managerial Applicability of Graphical Formats to Support Product Positioning Decisions. *Journal für Betriebswirtschaft*, 60(3), 167-201.

D' Agostino, R. B. and Stephens, M.A. (1986) (Eds.). *Goodness-of-Fit Techniques*. New York: Marcel Dekker.

DeSarbo, W. S. and Rao, V. (1984). Genfold II: A Set of Models and Algorithms for the GENeral unFOLDing Analysis of Preference/Dominance Data. *Journal of Classification*, 1 (1), 147-

86.

DeSarbo, W. S., Young, M.R. and Rangaswamy, A. (1997). A Parametric Multidimensional Unfolding Procedure for Incomplete Nonmetric Preference / Choice Set Data in Marketing Research. *Journal of Marketing Research*, 34 (November), 499-516.

Efron, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans.* Philadelphia: Society for Industrial and Applied Mathematics.

Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap.* New York: Chapman & Hall.

Faure, C. and Natter, M. (2010). New Metrics for Evaluating Preference Maps. *International Journal of Research in Marketing,* 27(3), *261-270*

Green, P. E., Carmone, F.J. and Smith, S.M. (1989). *Multidimensional Scaling: Concepts and Applications.* Boston: Allyn and Bacon.

Green, P. E. and Rao, V. (1972), *Applied Multidimensional Scaling.* Hinsdale, IL: Dryden Press.

Heiser, W.J. (1981), *Unfolding Analysis of Proximity Data.* Unpublished doctoral dissertation, Department of Data Theory, Leiden University, Leiden, The Netherlands.

Heiser, W. J. (1989). Order Invariant Unfolding Analysis Under Smoothness Restrictions, in Geert De Soete, H. Feger, and K.C. Klauer (Eds.). *New Developments in Psychological Choice Modeling* (pp. 3-31), Amsterdam: North Holland.

Ho, Y., Chung, Y., and Lau, K. (2010). Unfolding of large-scale marketing data. *International Journal of Research in Marketing*, 27, 119-132.

Kim, C., Rangaswamy, A. and DeSarbo, W.S. (1999). A Quasi-Metric Approach to Multidimensional Unfolding for Reducing the Occurance of Degenerate Solutions. *Multivariate Behavioral Research*, 34 (2), 143-80.

Kruskal, J. B. and Carroll, J.D. (1969). Geometric Models and Badness of Fit Functions, in R.P. Krishnaiah, ed., *Multivariate Analysis – II* (pp. 639-671). New York: Academic Press.

Lilien, G. L. and Rangaswamy,A. (2003). *Marketing Engineering: Computer-Assisted Marketing Analysis and Planning.* Upper Saddle River, NJ: Prentice-Hall.

Linting, M., Meulman, J.J., Groenen, P.J.F. and Van Der Kooij, A. (2007). Stability of Nonlinear Principal Components Analysis: An Empirical Study Using the Balanced Bootstrap. *Psychological Methods*, 12 (3), 359–79.

MacKay, D. B. and Dröge, C. (1990). Extensions of probabilistic perceptual maps with

implications for competitive positioning and choice. *International Journal of Research in Marketing*, 7, 265-282.

MacKay, D. B. and Zinnes, J.L. (1986). A Probabilistic Model for the Multidimensional Scaling of Proximity and Preference Data. *Marketing Science*, 5 (4), 325-44.

Meulman, J.J. and Heiser, W.J. (1983). *The Display of Bootstrap Solutions in Multidimensional Scaling.* Murray Hill, NJ: Bell Laboratories.

Schönemann, P. H. and Carroll, R. M. (1970). Fitting One Matrix to Another Under Choice of a Central Dilation and a Rigid Motion, *Psychometrika*, 35, 245–255.

Publications in the Report Series Research* in Management

ERIM Research Program: "Marketing"

2011

Nonmetric Unfolding of Marketing Data: Degeneracy and Stability
Michel van de Velden, Alain De Beuckelaer, Patrick J.F. Groenen, and Frank M.T.A. Busing
ERS-2011-006-MKT
http://hdl.handle.net/1765/22725

---

* A complete overview of the ERIM Report Series Research in Management:
https://ep.eur.nl/handle/1765/1

ERIM Research Programs:
LIS Business Processes, Logistics and Information Systems
ORG Organizing for Performance
MKT Marketing
F&A Finance and Accounting
STR Strategy and Entrepreneurship