FROM THE EDITOR

This "Special Issue" is devoted mainly to papers on **classification and data analysis** prepared by the authors participating in the International Federation of Classification Societies Conference (IFCS-2002) held in Cracov, Poland, in July 16th-19th, 2002. This part of the journal was organized and edited by our Associate Editor, Professor Krzysztof Jajuga from the Wroclaw Academy of Economics, Poland. He has also prepared the ***Foreword,*** which is presented after my comments upon this issue. I would like to remind you that this is already a second issue of our journal devoted to classification and data analysis of the IFCS. The first issue was published in 1995 (vol.2, Number 2, June 1995).

There are also two articles in section **Other Articles** in this issue. Section **Reports** ends this issue with two reports.

There are following eight articles devoted to some problems of classification and data analysis:

1. H. H. Bock, *Clustering Methods: From Classical Models to New Approaches.*
2. W. J. Krzanowski, *Orthogonal Components for Grouped Data: Review and Applications.*
3. Cz.. Domański, *Some Remarks on the Tasks of Statistics on the Verge of the Twenty First Century.*
4. A. Zeliaś, *Some Notes on the Selection of Normalisation of Diagnostic Variables.*
5. .D. Banks and R. T. Olszewski, *Combinatorial Search in Multivariate Statistics*.
6. M. Greenacre and J. G. Clavel, *Simultaneous Visualization of Two Transition Tables.*
7. E. Gatnar, *What is Data Mining?*
8. D. Larsen, *Impact of Latent Class Clustering of NSF Doctoral Survey Data on Adjusted Rand Index Values.*

The second part of this issue under the title **Other articles** contains two articles:

1) A. Młodak, *Some Approach to the Problem of Spatial Differentiation of Multi–feature Objects using Methods of Game Theory*.
2) P. Singh and N. Mathur, *An Alternative to an Improved Randomized Response Strategy.*

At the end of this issue there are two reports:

- The first report, prepared by K. Jajuga, is from the ***International Federation of Classification Societies Conference*** – IFCS-2002, Cracow, July 16th-19th, 2002.
- *The second report, prepared by A. Zelias, is from the* **Ninetieth Anniversary of the Foundation of the Polish Statistical Association - the Scientific Conference,** *Cracow, Poland, 14th-15th July 2002.*

Jan Kordos

The Editor

STATISTICS IN TRANSITION, October 2002
Vol. 5, No. 5, pp. 723

CLASSIFICATION AND DATA ANALYSIS

*
* *

FOREWORD

The present special issue of "*Statistics in Transit*ion" is devoted to the problems of classification and data analysis. In particular this issue is connected to the international conference "*Classification, Data Analysis and Related Methods*" which took place in Cracow, Poland, in July 16-19, 2002. This conference was organized by the International Federation of Classification Societies (IFCS).

International Federation of Classification Societies is the scientific organization with the aim of development of the theoretical and practical issues related to the classification and data analysis methods. IFCS was founded in Cambridge in 1985. Currently there are 12 member societies, including Section of Classification and Data Analysis of Polish Statistical Association (SKAD).

This volume contains several papers in the area of classification and data analysis. Some of these papers were presented during the conference. Among them are two papers presented as Keynote Lectures during the conference. These articles were written by Hans-Hermann Bock and Wojtek J. Krzanowski. Some other contributions were written by the authors (including Polish authors), who participated in the conference and in their scientific work made significant contributions in the area of statistical data analysis.

In addition, this volume contains the report from the mentioned conference, prepared by the Chairman of the Scientific Program Committee, Krzysztof Jajuga and the Chairman of the Local Organizing Committee, Andrzej Sokołowski.

I would like to thank Professor Jan Kordos for his great contribution in the preparation of this volume. My great thanks go also to Daniel Papla, who provided valuable help in the editorial work on this volume.

Krzysztof Jajuga

STATISTICS IN TRANSITION, October 2002
Vol. 5, No. 5, pp. 725–758

CLUSTERING METHODS: FROM CLASSICAL MODELS TO NEW APPROACHES

Hans-Hermann Bock[1]

ABSTRACT

This paper provides a survey on recent advances in clustering and points to new applications. We describe probabilistic clustering models for noisy and outlying data and survey some theorems on principal points and self-consistent center systems. We introduce 'convexity-based' clustering criteria, provide a corresponding optimization algorithm which uses maximum-support-plane partitions and show situations where these criteria are useful, e.g., in two-way clustering for contingency tables. Testing for a clustering structure is considered in the framework of mixture models and multimodality where new asymptotic results are obtained by Gaussian approximations. Finally, we comment on some recent and efficient clustering algorithms and mention new data types where clustering methods are needed.

***Key words*:** Clustering, probability models, k-tangent algorithm, clustering tests, convexity-based clustering, micro arrays.

1. Introduction

Clustering methods provide a powerful tool for analyzing data sets with applications in quite different domains such as, e.g., marketing, web commerce, biology, pattern recognition and image processing, document retrieval, and linguistics. In this article, 'clustering' means subdividing a set O = {1, ..., *n*} of *n* objects into a system of 'homogeneous', hopefully well separated classes C_1, ..., $C_m \subset$ O which will be called *clusters*, thereby using a matrix $X = (x_{kj})_{n\times p}$ of observed data where each row corresponds to an object and each column to one of *p* observed variables. We will consider here only real-valued variables such that the properties of an object *k* are characterized by the (column) vector $x_k = (x_{k1}, ..., x_{kp})'$, a data point in the *p*-dimensional space P^p. It is expected that objects from

[1] Hans-Hermann Bock, Institut für Statistik, RWTH Aachen, D-52056 Aachen, Germany, *bock@stochastik.rwth-aachen.de*.

the same cluster C_i comprise mainly 'similar' or neighbouring data points, whereas data points from different clusters will be 'dissimilar' and not too close to each other, as a rule.

During the last 40 years, a plethora of *clustering methods* have been proposed, differing, e.g., in the type of underlying data, the classification type (disjoint or overlapping classes, hierarchical systems, ...), the precise definition of the term 'cluster' (center-oriented, connected components, density clusters, ...), the given optimality criterion, etc. On the other hand, a lot of *theoretical investigations* has been conducted with the idea to reveal the structural and computational properties of the proposed methods in typical 'standard' situations, to derive a 'good' clustering criterion for an underlying 'clustering model', to evaluate the resulting clusters and to check the relevance of the constructed classification by 'homogeneity' or 'clustering tests'. Another direction of research was devoted to the generalization of classical criteria, to the relationship between clustering approaches, computational learning strategies and neural networks, to the adaptation of methods to new data types or applications (e.g., symbolic data, micro arrays, web data, ...), and to the graphical visualization of clustering results (projection pursuit, Kohonen maps, ...).

This paper is devoted to a review of some recent advances which have been obtained in four sub-domains of clustering methodology and which characterize some typical lines of research today. Section 2 will report on two clustering models which comprise noise and outliers, respectively, and lead both to a new 'robust' clustering method. In section 3 we discuss a 'continuous' clustering problem (for distributions instead of data) and describe some recent results on principal and self-consistent point systems. A relatively new 'convexity-based' clustering criterion is considered in Section 4 which generalizes the classical variance (SSQ) clustering criterion and the corresponding *k*-means method, replacing 'minimum-distance partitions' by 'maximum-support-plane partitions'. This method leads to various interesting applications relating, e.g., to data compression, optimum discretization of P^p, to the χ^2 non-centrality parameter and Csiszár's ϕ-divergence, and to two-mode clustering for contingency tables. In Section 5 we address the problem of testing for 'homogeneity': First we consider a maximum likelihood ratio test under the mixture model where non-identifiability poses a theoretical problem, and review some recent progress in overcoming this difficulty. Moreover, we describe a modification of Silverman's test for multimodality which resides on a new bootstrap strategy and leads to an exact asymptotic significance level. In section 6 we point to some clustering heuristics which were recently developed in the data mining and computer science community. Finally, we comment on the emergence of new data types in fields such as web mining, microbiology, and survey statistics, thereby tracing some new lines of development.

2. Probabilistic clustering models involving noise and outliers

There exist many probabilistic models which describe a clustering structure of data (Bock 1996a, 1996d). In this section we consider three basic models and describe their implications when designing a reasonable clustering method, especially in the case of noisy data and outliers. For ease of presentation we will consider only real-valued data such that the objects are characterized by n data points $x_1, ..., x_n$ in P^p. Adopting a probabilistic approach, they are considered as realizations of n independent random vectors $X_1, ..., X_n$ in P^p.

Our clustering models assume that there exists a given number m of *clusters* or *classes* labelled by $i = 1, ..., m$ and that each cluster i is characterized by a class-specific probability density $f_i(x)$ on P^p which has typically the form $f_i(x) = f(x; \vartheta_i)$ with a density family $\{f(\cdot; \vartheta) | \vartheta \text{ in } \in \Theta\}$ where the parameter ϑ belongs to a suitable parameter space Θ.

The fixed-partition model:

A first model assumes that there exists a *fixed* but unknown m-partition $\mathrm{X} = (C_1, ..., C_m)$ of the set $\mathrm{O} = \{1, ..., n\}$ and m unknown class-specific parameter values $\vartheta_1, ..., \vartheta_m \in \Theta$ such that

$$X_k \sim f_i(x) = f(\cdot; \vartheta_i) \text{ for all } k \in C_i \text{ and } i = 1, ..., m. \qquad (1)$$

Under this model, the maximum likelihood (m.l.) method for 'estimating' the unknown partition X and the parameter vector $\theta = (\vartheta_1, ..., \vartheta_m)$ yields the *m.l. clustering criterion*:

$$g_n(\mathrm{X}; \theta) := \frac{1}{n} \cdot \sum_{i=1}^{m} \sum_{k \in C_i} \left[-\log f\left(x_k; \vartheta_i\right) \right] \rightarrow \min_{\mathrm{X}, \theta} \qquad (2)$$

Various well-known clustering criteria can be derived from this model by suitable specification of the densities $f(\cdot; \vartheta)$; see, e.g., Bock (1974, 1996a, 1996b, 1996c).

The random-partition model:

This model assumes that m classes $C_1, ..., C_m \in \mathrm{O}$ of objects are generated by sampling n objects randomly and independently from a global population Π of objects which is subdivided into m 'sub-populations' $\Pi_1, ..., \Pi_m$. Let $p_i \in [0,1]$ denote the probability that a sampled object belongs to the population Π_i (with $p_i \geq 0$ and $\sum_i p_i = 1$) and define the random binary 'class indicators' I_k by $I_k := 1$ (or 0) iff $k \in C_i$ (or $k \notin C_i$) for $k = 1, ..., n$. Then the m classes $C_i := \{k \in \mathrm{O} | I_k = 1\}$ build a *random partition* $\mathrm{X} = (C_1, ..., C_m)$ of the set of objects O. Our data are now n random pairs $(I_1, X_1), ..., (I_n, X_n)$ where the I_k cannot be observed and where, conditional on $k \in C_i$ (or $I_k = i$), X_k has the density $f(x; \vartheta_i)$ as in (1).

For this clustering model the m.l. method leads to the log-likelihood criterion

$$\tilde{g}_n(\mathrm{X},\theta,\pi) = g_n(\mathrm{X},\theta) - \sum_{i=1}^{n} \frac{|C_i|}{n} \cdot \log p_i \rightarrow \min_{\mathrm{X},\theta,\pi} \tag{3}$$

Partial minimization with respect to the unknown probability vector $\pi = (p_1, ..., p_k)$ yields $\hat{p}_i = |C_i|/n$ and leads, after substituting $\hat{p}_i$ for p_i in (3), to the modified clustering criterion

$$\tilde{g}_n(\mathrm{X},\theta) = g_n(\mathrm{X},\theta) - \sum_{i=1}^{n} \frac{|C_i|}{n} \cdot \log \frac{|C_i|}{n} \rightarrow \min_{\mathrm{X},\theta} \tag{4}$$

The mixture model:

In the previous model, the observable data X_1, ..., X_n are obviously independent, all with the same marginal density

$$f(x;\ \theta,\ \pi) := \sum_{i=1}^{m} p_i\, f(x;\vartheta_i)\ \ x \in \mathrm{P}^p. \tag{5}$$

Thus, if we restrain our consideration to the marginal distributions and to the parameters only, we can obtain estimates for θ and π by maximizing the log-likelihood function:

$$L_n(\theta,\pi) := \sum_{k=1}^{n} \left[\sum_{i=1}^{m} p_i\, f(x_k;\vartheta_i \right] \rightarrow \max_{\theta,\pi} \tag{6}$$

There exists a large literature on mixture models (see, e.g., Everitt 1981, Titterington and Makov 1985, Titterington 1990, Bock 1996a, McLachlan and Peel 2000). Even if the mixture model provides no clustering approach in the strong sense, we can at least obtain a *fuzzy classification* of the objects by calculating, from the estimates $\hat{\vartheta}_i$ and $\hat{p}_i$, the posterior probabilities for k belonging to the population Π_i, i.e., the 'membership degrees' $\pi_{ik} := \hat{p}_i\, f(x_k;\hat{\vartheta}_i) / f(x_k;\hat{\theta},\hat{\pi})$, for all classes i and each object k.

The classical normal distribution clustering model

It is well-known that the fixed-partition clustering model (1) yields, in the case of m spherical normal distributions $f_i \sim \mathrm{N}_p(z_i,\ \sigma^2 I_p)$ with variance σ^2 and the unit matrix $I_p = \mathrm{diag}(1, ..., 1)$, the classical *variance (SSQ) clustering criterion*

$$g_n(\mathrm{X},\ \mathrm{Z}) := \frac{1}{n} \sum_{i=1}^{m} \sum_{k \in C_i} \left\| x_k - z_i \right\|^2 \rightarrow \min_{\mathrm{X,Z}} \tag{7}$$

where $Z = (z_1, ..., z_m) \in \mathrm{P}^{mp}$ comprises the class-specific expectations. Here partial minimization with respect to Z leads to the system $Z^* = (z_1^*,..., z_m^*)$ of class centroids $z_i^* = \bar{x}_{C_i}$ (i = 1, ..., m), and therefore to the equivalent criterion:

$$g_n(\mathrm{X}) := g_n(\mathrm{X}, Z^*) = \frac{1}{n}\sum_{i=1}^{m}\sum_{k\in C_i}\left\|x_k - \bar{x}_{C_i}\right\|^2$$

$$= \frac{1}{n}\sum_{k=1}^{n}\left\|x_k\right\|^2 - \sum_{i=1}^{m}\frac{|C_i|}{n}\cdot\left\|\bar{x}_{C_i}\right\|^2 \rightarrow \min_{\mathrm{X}} \qquad (8)$$

In a similar way we may derive many other clustering criteria by assuming special distribution types and parameter configurations in (1), e.g.,

- allowing for different variances and/or covariance matrices in the classes,
- by replacing the class-specific 'center points' z_i by class-specific 'central subspaces' ($\Rightarrow$ *principal component clustering*),
- by adding restrictions to the class centers z_1, ..., z_m (e.g.: all are comprised in the same subspace of P^p ($\Rightarrow$ *projection pursuit clustering*), etc.

For details the reader may refer to Bock (1974, 1996a, 1996b, 1996c, 1996c), Banfield and Raftery (1993). Various fuzzy analogues are described, e.g., in Bock (1979a, 1979b), Rousseeuw, Kaufmann and Trauwaert (1996) and Höppner, Klawonn and Kruse (1997).

However, when applying these methods in practice, we face the problem that not all data points follow strictly the assumed model. Often the set of 'relevant' data points is contaminated by *noisy data*, and there may be aberrant observations, i.e. *outliers*. There are various heuristic proposals in the literature in order to cope with such situations (e.g., Jolion and Rosenfeld 1989, Wishart 2003). In the following we mention two recently developed, *model-based* approaches which provide a theoretically well-based clustering criterion.

(1) Modeling noisy data in clustering

Banfield and Raftery (1993) defined 'noisy' data as data points which have no clustering structure insofar as they are uniformly distributed in some (sufficiently large) domain $Q \in \mathrm{P}^p$. Their clustering model assumes that there is, in the set O of all available objects, an unknown set $C_0 \subset \mathrm{O}$ of 'noisy' objects and that the remaining set $\mathrm{P} := \mathrm{O} - C_0$ of 'regular' objects is partitioned into m homogeneous classes C_1, ..., C_m. The probabilistic assumptions are as follows:

- All data vectors X_1, ..., X_n are independently distributed in P^p.
- Noisy data points form a *Poisson process* in P^p with intensity $\lambda > 0$. As a consequence, the number $|C_0|$ of noisy points in Q is random with a Poisson distribution $P(|C_0| = s) = e^{-\lambda V}\,(\lambda V)^{s/s!}$ for s = 0, 1, 2, ..., where V is the p-

dimensional volume of the domain G, and the noisy data are uniformly distributed in Q.

- For regular objects, i.e., for $k \in \mathrm{P}$, we have:

$$X_k \sim f_i(x) = f(\cdot;\ \vartheta_i) \text{ for all } k \in C_i,\ i = 1, ..., m.$$

Invoking now the m.l. approach, we find that the partition $\mathrm{X} = (C_0, C_1, ..., C_m)$ of the set of objects O, the class-specific parameters ϑ_i, and the intensity λ are to be determined from the *clustering criterion with noisy data*:

$$g_n(\mathrm{X},\ \theta,\ \lambda) := \sum_{i=1}^{m} \sum_{k \in C_i} \left[-\log f(x_k; \vartheta_i) \right] + \lambda V - \log\left[\lambda V^{|C_0|} / |C_0|! \right] \rightarrow \min_{\mathrm{X},\theta,\lambda} \qquad (9)$$

Banfield and Raftery describe examples which identify the noisy observations and group the regular data in a satisfactory way.

(2) Modeling outliers and robust clustering

In practical work, another annoying situation is provided by the occurrence of outliers in the data set. This necessitates some type of 'robust' clustering. A corresponding model and method have been proposed in the articles by Gallegos (2001, 2002, 2003) from which we describe a situation where 'regular' data are normally distributed with the same covariance matrix Σ in all classes.

First let us recall a classical, fixed-partition normal clustering model:

$$X_k \sim \mathrm{N}_p(z_i, \Sigma) \text{ for all } k \in C_i \text{ and } i = 1, ..., m \qquad (10)$$

with an unknown k-partition $\mathrm{X} = (C_1, ..., C_m)$, m class centers $z_1, ..., z_m \in \mathrm{P}^p$ and the unknown covariance matrix Σ. Denoting by

$$W(\mathrm{X}) := \sum_{i=1}^{m} \sum_{k \in C_i} (x_k - \bar{x}_{C_i})(x_k - \bar{x}_{C_i})' \qquad (11)$$

the within-cluster scatter matrix of the data, it can be shown (Bock 1974, Späth 1985) that for this special model the m.l. clustering criterion (?) is equivalent to the *determinantal criterion*

$$g(\mathrm{X}) := \det(\mathrm{W}(\mathrm{X})) \Rightarrow \min_{\mathrm{X}}. \qquad (12)$$

Gallegos designs a corresponding outlier clustering model by modifying a classical outlier model (see, e.g., Mathar 1981). She assumes that there is a known number s of outliers in O which are compiled in the (unknown) outlier set $C_0 \subset \mathrm{O}$. Let $\mathrm{P} := \mathrm{O} - C_0$ denote the set of $r = n - s$ 'regular' data and $\Pi_m(\mathrm{P})$ the family of all m-partitions $\mathrm{X} = (C_1, ..., C_m)$ of P with $|C_i| \geq p + 1$ for all i. Then her *fixed-partition outlier clustering model* is given by:

- All data vectors $X_1, ..., X_n$ are independently distributed in P^p.
- For regular objects, i.e., for $k \in \mathrm{P}$, we have:

$$X_k \sim \mathrm{N}_p(z_i,\Sigma) \text{ for all } k \in C_i \text{ and } i = 1, ..., m.$$

- For outlier objects, i.e., for $k \in C_0$, we have:

$$X_k \sim \mathrm{N}_p(\zeta_k,\Sigma)$$

with $m + s$ unknown centers $z_1, ..., z_m$ and ζ_k (for $k \in C_0$). In this case, the m.l. method leads to the following *determinantal clustering criterion with outliers*:

$$\min_{\mathrm{P} \subset \mathrm{O}, |\mathrm{P}|=r} \min_{\mathrm{X} \in \Pi_m(\mathrm{P})} \det(W(\mathrm{X})) \tag{13}$$

It is obvious that for each outlier data point x_k (i.e., with $k \in C_0$) the m.l. estimate for the corresponding center ζ_k is given by $\hat{\zeta}_k = x_k$. Moreover, for a fixed outlier set C_0 (i.e., a fixed set P), we may minimize $\det(W(\mathrm{X}))$ by classical algorithms. However, it is not at all obvious how to overcome the combinatorial problems related to the determination of an optimal outlier set C_0 and thereby to obtain a generalized *k*-means clustering method. A suitable combinatorial algorithm has been developed by Gallegos (2002). It uses, when modifying C_0, the ordering of the objects according to their Mahalanobis distances to the current class centers. This algorithm seems to work for sample sizes such as $n = 5500$ with $s = 500$ outliers, in 2 to 5 dimensions, and resulting in a small percentage of wrongly classified outliers.

Remark: The criterion (13) and the resulting estimate for the covariance matrix Σ are closely related to Rousseeuw's *minimum covariance determinant estimator* (Rousseeuw 1983, Pesch 1999, Rousseeuw and van Driessen 1999). In fact, for $m = 1$, i.e., one single class, the previous clustering model is identical to Rousseeuw's classical outlier model.

3. Principal and self-consistent center systems

Whereas the clustering criteria from section 2 are designed to find an optimum *m*-partition X^* of the finite set of data points $\{x_1, ..., x_n\}$, we may ask if we can formulate an analogous partitioning or *segmentation problem* for the whole space P^p, thereby assuming a given probability distribution P on P^p and generalizing the previous criteria and clustering methods in a suitable way. For ease of presentation, we will consider typical cases where P has a Lebesgue density $f(x)$ for $x \in \mathrm{P}^p$ and speak of a *continuous* clustering (segmentation, quantization) problem.

In this article we will concentrate on the continuous version of the *discrete* variance criterion (7): Find a partition B = (C_1, ..., C_m) of P^p and a system Z =(z_1, ..., z_m) of class centers $z_i \in P^p$ such that

$$g(\mathrm{B}, \mathrm{Z}) := \sum_{i=1}^{m} \int_{B_i} \|x - z_i\|^2 \, dP(x) \rightarrow \min_{\mathrm{B,Z}} \tag{14}$$

In fact, this problem has been considered by Bock (1974, chap. 15.b) who shows that – similarly to the discrete case – partial minimization leads to the equivalent problems

$$g(\mathrm{B}) := \sum_{i=1}^{m} \int_{B_i} \|x - E[X \mid X \in B_i]\|^2 \, dP(x) \rightarrow \min_{\mathrm{B}} \tag{15}$$

where the optimum centers $z_i^* := E[X \mid X \in B_i]$ have been substituted and

$$\gamma(\mathrm{Z}) := \int_{R^p} \min_{j=1,\ldots,m} \left\{ \|x - z_j\|^2 \right\} dP(x) = g(\mathrm{B}(\mathrm{Z})) \rightarrow \min_{\mathrm{Z}} \tag{16}$$

where B(Z) := (B_1^*,..., B_m^*) denotes the Voronoi (minimum-distance) partition of P^p which is generated by the fixed center system Z with classes $B_i^* := \{x \in P^p \mid \|x - z_i\|^2 = \min_{j=1,\ldots,m}\{\|x - z_j\|^2\}\}$ for $i = 1, \ldots, m$.

In fact, this latter problem has been popularized by Flury (1990, 1993) and any solution $Z^* = (z_1^*, \ldots, z_m^*)$ of (14) or (16) has been termed there a *principal point system (PPS)*. The PPS approach can be considered as a modification of *principal components* in multivariate analysis, as a trivial variant of *principal component clustering* (Bock 1974, chap. 17, Bock 1987), and a special case of *principal planes* or *curves* (Hastie and Stützle 1989). In the clustering context, the continuous problems (15) to (16) find their applications in survey statistics, multivariate stratification, data compression and digitalization, telecommunication, and image analysis.

From a theoretical point-of-view we would expect that the optimal center system $Z^{(n)}$ obtained from the discrete variance criterion (7) and (8) converges, for an increasing number *n* of samples, to a PPS Z^* obtained from (14) or (16) if the data vectors X_1, X_2, ... are all distributed with the same distribution *P*. This conjecture has, in fact, been proved under various assumptions (see, e.g., Bryant and Williamson 1978, Hartigan 1978, Pollard 1982, Bock 1985) which all require that there exists only one single asymptotic PPS Z^*. So it would be interesting to know more about the uniqueness of a PPS.

In recent years, this question has been intensively investigated. It is closely related to the concept of a *self-consistent point system* Z, i.e., a center system Z =

$(z_1, ..., z_m) \in P^{mp}$ with the property that the corresponding Voronoi partition $B^* := B(Z)$ generates these centers in the sense that

$$z_i = E[X \mid X \in B_i^* \text{ for } i = 1, ..., m \tag{17}$$

(Flury 1990). A pair (B, Z) with this property has also been called a *stationary pair* and it is well-known that stationarity is a necessary condition for any PPS (see, e.g., Bock 1974). For an arbitrary density *f*, Flury (1990) and Mizuta (1998) provide necessary and sufficient conditions for an optimum with $m = 2$ classes, for $m = 3$ classes see Shimizu, Mizuta and Sato (1997, 1998). Stationary pairs and configurations for normal distributions have been determined and displayed by many authors: For the univariate normal case $N_1(0, 1)$ see Cox (1957), and Bock (1974, p. 179), for the two-dimensional spherical normal distribution $N_2(0, I_2)$ see Flury (1990), Bock (1998), Shimizu et al. (1998) with $m = 3, ..., 11$ classes from which we may guess the corresponding PPS (which is not unique due to the radial symmetry of N_2). For ellipsoidal normals see Kipper and Pärna (1992), Tarpey, Li and Flury (1995), Shimizu et al. (1998), and Tarpey (1998). Zoppè (1997) considers also the case of an exponential distribution. For mixtures see Zoppè (1997) and Yamamoto and Shinozaki (2000).

The unicity problem has often been investigated in parallel with the symmetry problem, i.e., the question if a *symmetric* density $f(x)$ (with center $x = 0$) admits also non-symmetric PPSs. The current state-of-the-art in this domain is characterized by the following theorems:

Theorem 3.1: The one-dimensional case p = 1

(a) The PPS is unique if $\log f(x)$ is concave (Trushkin 1982, Karlin 1982, Tarpey 1994); examples are provided by N(0,1), Pearson II and Kotz-type distributions etc. (Tarpey 1998).

(b) Symmetry of a PPS (w.r.t. the origin $x = 0$) does **not** follow from the symmetry of *f*. Counterexamples are provided, e.g., by Karlin (1982), Mizuta (1995), Zoppè (1997) and (g) below.

(c) For $m = 2$ classes, symmetry and uniqueness hold if
- *f* is symmetric,
- the distribution corresponding to the density *f* is 'new better than used',
- the function $x/(2F(x) - 1)$ is nondecreasing for $x > 0$ (Yamamoto and Shinozaki 2000).

Here $F(x) = P(X \le x)$ is the distribution function corresponding to *f*.

(d) The last condition in (c) holds if *f* is symmetric and unimodal.

(e) Mixture densities of the type $f(x) = (g(x - m) + g(x + m))/2$ with a symmetric and log-concave density *g* fulfil the conditions in (c), thus have a unique PPS with $m = 2$ points.

(f) For $s \ge 3$ degrees of freedom, Student's t_s distribution fulfils (c) (but is not log-concave).

(g) A scale mixture N(0,1) + N(0, σ^2))/2 has a unique pair of principal points if and only if $\sigma < 4.03537$.

See also Li and Flury (1995) and Shimizu et al. (1999) for a sufficient condition if $k \geq 2$.

For the multivariate case, there are still no satisfactory results. Some structural properties are provided by

Theorem 3.2: The multivariate case

(a) For p-dimensional spherical normal distributions with $p \geq 2$ the PPS is *not* unique (since rotations do not change the PPS property).

(b) Consider a density of the form $f(x) = c \cdot g(x'\Sigma^{-1}x)$ with a positive definite matrix Σ, i.e. an elliptical distribution centered at 0. Then the subspace $S := [z_1, ..., z_m]$ in P^p which is spanned by the points of a PPS $Z = (z_1, ..., z_m)$ is spanned by q eigenvectors (principal factors) $v_1, ..., v_q$ of the matrix Σ where $q := \dim(S)$ (Tarpey, Li and Flury 1995).

In this domain there exists a large number of unresolved problems. In particular, we have no results about the structure of a PPS or a stationary configuration for a density f describing a clustering structure (e.g., a mixture of m normal distributions).

4. Convexity-based clustering criteria

In this section we consider some type of clustering criteria which has been developed during the last years and involves a convex function ϕ. It will be seen in the sequel that by choosing this function in a suitable way we can formulate and resolve several interesting problems in statistics and data analysis.

Some motivation is provided by the classical variance criterion (7) where formula (8) shows that it is equivalent to the maximization of the following criterion:

$$G_n(\mathrm{X}) := \sum_{i=1}^{m} \frac{|C_i|}{n} \cdot \left\| \bar{x}_{C_i} \right\|^2 \rightarrow \max_{\mathrm{X}}$$

Replacing the square $\|x\|^2$ here by an arbitrary *convex* function $\phi(x)$ we obtain a quite general *convexity-based clustering criterion*:

$$G_n(\mathrm{X}) := \sum_{i=1}^{m} \frac{|C_i|}{n} \cdot \phi\left(\bar{x}_{C_i}\right) \rightarrow \max_{\mathrm{X}} \qquad (18)$$

There exists also an 'continuous' analogue of this 'discrete' problem which is essentially a *segmentation*, *discretization* or *quantization problem*: Consider a a random vector X in P^p with distribution P. We want to find an m-partition $B = (B_1, ..., B_m)$ of the whole space P^p which is optimal in the sense:

$$G(\mathrm{B}) := \sum_{i=1}^{m} P(X \in B_i) \cdot \phi(E[X \mid X \in B_i]) \to \max_{\mathrm{B}} \quad (19)$$

It is obvious that the discrete problem (18) is returned from (19) when substituting $P = P_n$, the empirical distribution of the data $x_1, ..., x_n$. So we will mainly concentrate on the continuous problem.

Convexity-based clustering criteria were introduced by Bock (1983, 1991, 2002a) who developed an iterative *k*-means-like clustering method (e.g., for optimum stratification). Strasser and Pötzelberger (2001) generalized the criterion, investigated the mathematical properties of the resulting partitions and provided an asymptotic theory for $n \to \infty$, similar to the classical SSQ case.

It is obvious that an exact optimization is impossible for the criterion (19). However, we can design a heuristic maximization algorithm in analogy to the classical *k*-means algorithm for the variance criterion if we can find a *two*-parameter criterion $K(\mathrm{B}, Z)$ such that optimization of $K(\mathrm{B}, Z)$ with respect to *both* parameters B and Z is equivalent to the minimization of $G(\mathrm{B})$ with respect to B alone (just as with (7) and (8)). Then we can apply an iterative alternating optimization strategy to $K(\mathrm{B}, Z)$ as in the classical case.

The construction of K is not at all self-evident. First we recall that for each $z \in \mathrm{P}^p$ the function $u = \phi(x)$ has a *support hyperplane* $u = t(x; z, a) := a'(x - z) + \phi(z)$ (with a coefficient vector $a \in \mathrm{P}^p$) with the properties

$$\phi(x) \geq t(x; z, a) \text{ for all } x \in \mathrm{P}^p \text{ and } \phi(z) = t(z; z, a) \text{ for } x = z. \quad (20)$$

For ease of presentation we will assume ϕ to be differentiable such that $a = a(z) := grad_x\, \phi(x)_{|x = z}$ and the hyperplane is the *tangent hyperplane* of the surface $u = \phi(x)$ at the support point $x = z$.

Returning to the clustering problem (19), we will introduce for each class $B_i \subset \mathrm{P}^p$ an (arbitrary) support point $z \in \mathrm{P}^p$ and compile these points in the *support system* $Z = (z_1, ..., z_m)$. We will also consider the tangent (support) hyperplane H_i of ϕ for the support point z_i which is described by $u = t(x; z_i) := a_i'(x - z_i) + \phi(z_i)$ (with $a_i := a(z_i)$) and is always below $u = \phi(x)$. Finally, we consider the piecewise linear function $u = p(x; \mathrm{B}, Z)$ which is defined in the domain B_i by the hyperplane *Hi*, i.e., $p(x; \mathrm{B}, Z) := t(x; z_i)$ for $x \in B_i$. Obviously $p(x; \mathrm{B}, Z) \leq \phi(x)$ for all x. – Now the following theorem specifies the required two-parameter criterion:

Theorem 4.1*: The continuous* ***one****-parameter maximization problem (19) is equivalent to the* ***two****-parameter 'minimum volume problem' defined by*:

$$K(\mathrm{B}, Z) := E[\phi(X) - p(X; \mathrm{B}, Z)] = \sum_{i=1}^{m} \int_{B_i} [\phi(x) - t(x; z_i)]\, dP(x) \to \min_{\mathrm{B},Z} \quad (21)$$

Similarly, the discrete clustering problem (18) is equivalent to:

$$K(\mathrm{X}, \mathrm{Z}) := \frac{1}{n}\sum_{i=1}^{m}\sum_{k\in C_i}\left[\phi(x_k) - t(x_k; z_i)\right] \rightarrow \min_{\mathrm{X,Z}} \qquad (22)$$

Note that the criterion $K(\mathrm{B}, Z)$ can be interpreted as a weighted volume between the surfaces $u = \phi(x)$ and $u = p(x)$ in P^{p+1}.

Now we can (approximately) maximize the clustering criterion (19) by partially minimizing the equivalent criterion (21), first with respect to the support system Z and then with respect to the partition B, and iterating these two steps as in the *k*-means algorithm (similarly for the discrete criterion (22)). It appears that partial minimization is possible in an explicit way such that we can formulate the following iterative optimization algorithm (with steps $t = 0, 1, 2, ...$) which is called the *maximum-support-plane (MSP) algorithm*. Since we have assumed a differentiable convex function ϕ, it is in fact a *k-tangent algorithm*.

The k-tangent or maximum-support-plane (MSP) algorithm

- $t = 0$: Select an arbitrary initial *m*-partition $\mathrm{B}^{(0)} = (B_1^{(0)},..., B_m^{(0)})$ of P^p.
- $t \rightarrow t + 1$:

(i) Determine the system $Z^{(t)} = (z_1^{(t)},..., z_{m1}^{(t)})$ of support points that minimizes the criterion $K(\mathrm{B}^{(t)}, Z)$ with respect to all Z. This system is given by the class-specific expectations (centroids) $z_i^{(t)} = E\left[X \mid X \in B_i^{(t)}\right]$ of the classes $B_i^{(t)}$, i.e., by $Z^{(t)} := (E[X \mid X \in B_1^{(t)}],..., E[X \mid X \in B_m^{(t)}])$.

(ii) Determine the partition $\mathrm{B}^{(t+1)} = (B_1^{(t+1)},..., B_m^{(t+1)})$ of P^p that minimizes the criterion $K(\mathrm{B}, Z^{(t)})$ with respect to all *m*-partitions B. This optimum partition $\mathrm{B}^{(t+1)}$ is given by the *maximum-support-plane partition* (here: a *maximum-tangent-plane partition*) that is generated by the current support system $Z^{(t)}$, i.e., with classes

$$B_i^{(t+1)} := \{x \in \mathrm{P}^p \mid t(x_k;\ z_i^{(t)}) = \max_{j=1,...,m} t(x_k;\ z_j^{(t)})\} \qquad (23)$$

for $i = 1, ..., m$ where $t(x;\ z_i^{(t)})$ corresponds to the tangent hyperplane H_i of ϕ in the support point $z_i^{(t)}$, the centroid of the class $B_i^{(t)}$.

- *Stopping criterion*:

Both steps are iterated for $t = 1, 2, ...$ until the partitions $\mathrm{B}^{(t)}$ do not change any more or until the support points $E[X \mid X \in B_i^{(t)}]$ attain an (approximately) stationary state.

Some applications and special cases

Various problems in statistics and data analysis lead to optimization criteria of the type (18) or (19). We mention four special cases:

(1) Maximizing the non-centrality parameter of the χ^2 goodness-of-fit test

If Y is a random vector in P^p with distribution P, the well-known χ^2 goodness-of-fit test checks the hypothesis H_0: $P = P_0$ for a given distribution P_0. It uses the test statistic $\chi^2 := n \sum_{i=1}^{m} \left(N_i / n - P_0(D_i)\right)^2 / P_0(D_i)$ where $\Delta = (D_1, ..., D_m)$ is a given partition (segmentation) of P^p and N_i the number of observations points $x_1, ..., x_n$ in D_i. The asymptotic performance of this test when testing against an alternative $P = P_1$ (e.g., its Pitman or Bahadur efficiency) can be characterized by the non-centrality parameter

$$\tilde{G}(\Delta) = \sum_{i=1}^{m} \frac{\left(P_1(D_i) - P_0(D_i)\right)^2}{P_0(D_i)} = \sum_{i=1}^{m} P_0(D_i) \left(\frac{P_1(D_i)}{P_0(D_i)} - 1 \right)^2 = \sum_{i=1}^{m} P_0(D_i) \cdot \phi(\lambda(D_i)) \quad (24)$$

where $\lambda(D_i) := P_1(D_i)/P_0(D_i)$ is a discretized likelihood ratio and ϕ the convex function $\phi(\lambda) := (\lambda - 1)^2$.

Looking for an optimum stratification Δ means maximizing the non-centrality parameter (24). It appears that this problem is a special instance of the 'convexity-based' clustering problem (19). In fact, if we introduce the likelihood ratio $x = \lambda(y) = f_1(y)/f_0(y)$ of the densities of P_1 and P_0 and consider the random variable $X := \lambda(Y)$ (of dimension $p = 1$) we see easily that maximizing $\tilde{G}(\Delta)$ is equivalent to finding an optimum partition $\mathrm{B} = (B_1, ..., B_m)$ of the λ-space P_+^1 in the sense that

$$G(\mathrm{B}) := \sum_{i=1}^{m} P_0(X \in B_i) \cdot \phi\left(E_0[X \mid X \in B_i]\right) \to \max_B \quad (25)$$

Here the partitions Δ of P^q and B of P_+^1 are related by $D_i = \{y \in \mathrm{P}^q \mid \lambda(y) \in B_i\}$ and we have used the fact that the discretized likelihood ratio can be written as a conditional expectation of the form

$$E_0[X \mid X \in B_i] = E_0[\lambda(Y) \mid Y \in D_i] = P_1(D_i)/P_0(D_i) = \lambda(D_i). \quad (26)$$

Thus, the previous k-tangent algorithm can be used in order to find an optimum stratification Δ for the χ^2 test (Bock 1983).

(2) Maximizing the ϕ-divergence between discretized distributions

More generally, we may look for a partition Δ of P^q which maximizes the performance of other tests between H_0: $P = P_0$ and H_1: $P = P_1$. Corresponding measures have often the form of Csiszár's ϕ-divergence (Csiszár 1967):

$$G_\phi(\Delta) := \sum_{i=1}^{m} P_0(D_i)\phi(\lambda(D_i)) = \sum_{i=1}^{m} P^X(B_i)\phi(E_0[X \mid X \in B_i]) = G(\mathrm{B}) \to \max_{\mathrm{B}} \quad (27)$$

with a convex ϕ and B as before. Special cases include the *Kullback-Leibler discriminating information*:

$$G_{KL}(\Delta) := \sum_{i=1}^{m} P_0(D_i) \cdot \log \frac{P_0(D_i)}{P_1(D_i)} = \sum_{i=1}^{m} P_0(D_i) \cdot \phi(\lambda(D_i)) \to \max_{\mathrm{B}} \quad (28)$$

with $\phi(\lambda) := -\log \lambda$ and the *variation distance*

$$G_{VD}(\Delta) := \sum_{i=1}^{m} |P_0(D_i) - P_1(D_i)| = \sum_{i=1}^{m} P_0(D_i) \cdot \phi(\lambda(D_i)) \to \max_{\mathrm{B}} \quad (29)$$

with $\phi(\lambda) := |\lambda - 1|$. Again, maximization can be performed by the MSP algorithm (for details and examples see Bock 1991, 1994, 2002a).

(3) Kohonen's iterative projection method

Strasser and Pötzelberger (2001) have shown that for an arbitrary convex function ϕ the maximization problem (19) is equivalent to the problem

$$F_m(a_1, ..., a_m) := \int_{R^p} \max_{j=1,...,m} \left\{ a_j' x - \phi^c(a_j) \right\} dP(x) \to \max_{a_1,...,a_m \in R^p} \quad (30)$$

where $\phi^c(a) := \sup_{x \in R^p} \left\{ a_j' x - \phi(x) \right\}$ is the conjugate convex function of ϕ (with a direction vector $a \in \mathrm{P}^p$). Since for the Euclidean norm $\phi(x) := \|x\|$ we have $\phi^c(a) = 0$ ($=\infty$) for $\|a\| \le 1$ (for $\|a\| > 1$) the problem (19):

$$G(\mathrm{B}) = \sum_{i=1}^{m} P(X \in B_i) \cdot \|E[X \mid X \in B_i]\| \to \max_{\mathrm{B}} \quad (31)$$

is equivalent to the problem (30), i.e.

$$F_m(a_1, ..., a_m) := \int_{R^p} \max_{j=1,...,m} \left\{ a_j' x \right\} dP(x) = \sum_{i=1}^{m} \int_{B_i^*} a_i' x dP(x) = \to \max_{a_1,...,a_m \text{ with } \|a_j\| \le 1 \text{ for all } j} \quad (32)$$

where $B_i^* := \{x \in \mathrm{P}^p \mid a_i' x = \max_{j=1,...,m} \{ a_j' x \} \}$ for $i = 1, ..., m$.

This latter problem corresponds to a problem investigated by Kohonen (1984): Consider an m-partition $\mathrm{B} = (B_1, ..., B_m)$ of P^p and for each class B_i a direction $a_i \in \mathrm{P}^p$. Project all $x \in B_i$ on the direction a_i. Look for a B and m directions $a_1, ..., a_m$ such that the 'average projection length' is maximized and, insofar, the subspaces spanned by the a_i discriminate optimally the m classes of B. Whereas Kohonen has presented a stochastic approximation method for solving

(32) we can use here the previous MSP algorithm here for resolving the equivalent problem (31) (with a modification allowing for the non-differentiability of $\phi(x) = ||x||$).

(4) Simultaneous clustering of the rows and columns of a contingency table

Consider two random variables U and V with categories in Y = {1, ..., a} and ς = {1, ..., b}, respectively and denote by $f(u, v) := P(U = u, V = v)$ the joint probability function of (U, V). This distribution is typically compiled in a contingency table N $=(f(u, v))_{a\times b}$ (and could, e.g., be an observed empirical distribution). If the numbers a and b of categories are large, it may be desirable to reduce the large table N to a smaller one of size $m\times l$, say, just by aggregating the categories of Y into m classes A_1, ..., A_m, i.e., to consider an m-partition A = (A_1, ..., A_m) of Y and, similarly, an l-partition B = (B_1, ..., B_l) of ς. We want to perform this aggregation in a way such that a possible *dependence structure* of the original variables (U, V) is optimally unfolded in the reduced table $\mathrm{N(A, B)}_{m\times l}$ with entries $P(A_i\times B_j) := P(U \in A_i, V \in B_j)$. One way to quantify this idea is to maximize the Kullback-Leibler information

$$G_{KL}(\mathrm{A, B}) := \sum_{i=1}^{m}\sum_{j=1}^{l} P(A_i)P(B_j)\cdot\left[-\log\frac{P(A_i\times B_j)}{P(A_i)P(B_j)}\right] \rightarrow \max_{\mathrm{A,B}} \qquad (33)$$

i.e., the divergence between the distributions given by $P(A_i\times B_j)$ and $P(A_i)\cdot P(B_j)$ (independence case).

A basic strategy in order to (approximately) resolve this simultaneous optimization problem is by alternating partial maximization: We improve a given initial pair $(\mathrm{A}^{(0)}, \mathrm{B}^{(0)})$ of partitions first by maximizing G_{KL} with respect to B, then with respect to A and iterating these steps in turn. It has been shown by Bock (2001, 2003) that each partial maximization step corresponds to a convexity-based clustering problem and that therefore the MSP algorithm can be used for partial maximization. In the case of B this can be seen as follows:

For a fixed partition A of the rows of N, we denote

- by P_0 the marginal distribution of V with counting density $f^V(v) = P(V = v)$,
- by P_{1i} the conditional distribution of V under $U \in A_i$ with counting density $f_{1i}(v) := P(V = v \mid U \in A_i) = \sum_{u\in A_i} f(u,v)/P(a_i)$

  such that

- $\lambda_i(v) := f_{1i}(v)/f^V(v) = \left[\sum_{u\in A_i} f(u,v)\right]/\left[P(a_i)f^V(v)\right]$ is a class-A_i-specific likelihood ratio (i = 1, ..., m) and
- $\lambda(v) := (\lambda_1(v),...,\lambda_m(v)' \in \mathrm{P}_+^p$ the vector of all likelihood ratios.

With this notation we can write the criterion G_{KL}(A, B) in the form

$$G_{KL}(\mathrm{A},\mathrm{B}) = \sum_{j=1}^{l} P(B_j)\cdot \sum_{i=1}^{m} P(A_i)\cdot\left[-\log\frac{P(B_j \mid A_i)}{P(B_j)}\right] \tag{34}$$

$$= \sum_{j=1}^{l} P(B_j)\cdot \sum_{i=1}^{m} P(A_i)\cdot\left[-\log E_0\left[\lambda_i(V)\mid B_j\right]\right]$$

$$= \sum_{j=1}^{l} P(B_j)\cdot \sum_{i=1}^{m} P(A_i)\cdot \phi_{\mathrm{A}}\left(E_0\left[\lambda(V)\mid B_j\right]\right)$$

where the function $\phi_{\mathrm{A}}(\lambda) := \sum_{i=1}^{m} P(A_i)\cdot\left[-\log\lambda_i\right]$ is obviously convex for $\lambda \in \mathrm{P}^m$. Therefore (34) has just the form of the convexity-based basic criterion (19) and we can maximize (33) or (34) with respect to B by using the former MSP algorithm. – For other two-way simultaneous clustering criteria see Bock (2003) and Krolak-Schwerdt (2003) .

5. Testing for homogeneity versus clustering

Applying a clustering algorithm to some data presupposes, at least implicitly, that there exists a clustering structure in the data. In the alternative case of a 'homogeneous' population, any clustering method will typically produce more or less artificial clusters without substantial importance. This raises the problem of testing the 'homogeneity' of the data, i.e., the hypothesis H_1: $m = 1$, versus an alternative of 'clustering' in the framework of the models presented in section 2.

This problem and corresponding 'cluster tests' have been intensively investigated in the past (see, e.g., Hartigan 1978, 1985b, Bock 1985, 1996a, Bryant 1991). In the following we will describe two situations where we think that some important success has been attained in the recent past.

(1) Testing for the number of mixture components

The first one starts from the mixture model (5) with a given density family $\{f(\cdot;\vartheta) \mid \vartheta \in \Theta\}$ and considers the hypotheses H_m that the number of components is a preset integer $m \geq 1$. Various choices for the testing problem are possible, e.g., $H_1 \leftrightarrow H_2$, and $H_m \leftrightarrow H_{m+1}$. Here we consider the problem $H_1 \leftrightarrow H_m$ (with a given $m \geq 2$) and the corresponding classical maximum likelihood ratio test (MLRT) which rejects 'homogeneity H_1' in favour of 'clustering H_m' if the test statistic

$$\lambda_n := 2 \cdot \log \frac{\sup_{H_m} \prod_{k=1}^{n} f(x_k; \theta, \pi)}{\sup_{H_1} \prod_{k=1}^{n} f(x_k; \theta, \pi)} \tag{35}$$

is larger that a critical threshold $c = c(\alpha)$ for a given significance level α.

The determination of the percentage point $c(\alpha)$ is a serious problem. In fact, it has been observed, e.g. by Hartigan (1985a, 1985b), that the classical likelihood theory breaks down in this case such that the classical χ^2 approximation for λ_n is wrong. This is due to the fact that the homogeneity hypothesis H_1 is *not identifiable* in the usual parametrisation of H_m by $\theta = (\vartheta_1, ..., \vartheta_m)$ and $\pi = (p_1, ..., p_m)$ and, moreover, H_1 is located at the boundary of the parameter space. This can be seen from an example with $m = 2$ components:

H_2: $f(x) = (1 - \pi_2)\, f(x;\ \vartheta_1) + \pi_2\, f(x;\ \vartheta_2)$ with $\pi_2 \in [0, 1],\ \vartheta_1,\ \vartheta_2 \in \mathrm{P}^p$

H_1: $f(x) = f(x;\ \vartheta_1)$ with $\vartheta_1, \in \mathrm{P}^p$

Obviously H_1 results from H_2 for: $\begin{cases} \bullet\ \pi_2 = 0 & \vartheta_1, \vartheta_1 \in \mathrm{P}^{p^\pi} \\ \bullet\ \pi_2 = 1 & \vartheta_1, \vartheta_1 \in \mathrm{P}^{p^\pi} \\ \bullet\ \vartheta_1 = \vartheta_2 \in \mathrm{P}^p & 0 < \pi_2 < 1 \end{cases}$

There were many attempts to find suitable χ^2 approximations for normal mixtures, beginning with Wolfe's basic paper (Wolfe 1971) and later on by simulation or bootstrapping (McLachlan 1987, Thode et al. 1988, Mendell et al. 1991, 1993); see also Böhning et al. (1994). Theoretical results concerned, e.g., some type of consistency of the m.l. estimates $\hat{\theta}_n$, $\hat{\pi}_n$ saying that for any neighbourhood W of the *set* of parameter values (θ, π) describing H_1 (under H_m), we have $\lim_{n \to \infty} P_{H_1}((\hat{\theta}_n, \hat{\pi}_n) \in W)$ (Redner 1981). Bickel and Chernoff (1993) have shown that for any *unbounded* parameter space the MLR statistic approaches ∞ as fast as $\log(\log n)$ if $n \to \infty$. Ghosh and Sen (1985) were the first to derive an asymptotic distribution of λ_n under the homogeneity hypothesis H_1, assuming a two-component normal mixture model $(1 - p)\ \mathrm{N}_1(\vartheta_1, 1) + p\mathrm{N}_1(\vartheta_2, 1)$ with a fixed ϑ_1. Their results require a bounded interval $\vartheta_1, \vartheta_2 \in \Theta = [a, b]$ and a separation condition $|\vartheta_1 - \vartheta_2| > \varepsilon > 0$ in order to restore identifiability. This is not very realistic in practice.

It was only in recent years that a modified asymptotic likelihood theory has been proposed, e.g. by Dacunha-Castelle and Gassiet (1999), Berdai and Garel

(1994, 1996) and Garel (2001, 2002), and the separation condition was eliminated. In the case $H_1 \leftrightarrow H_2$ the general approach of Garel consists in deriving, for a given parameter (θ, π) from the domain Φ describing H_2, a suitable asymptotic expansion $T_n(\theta, \pi)$ for the likelihood ratio and to show that the process $(T_n(\theta, \pi))_{(\theta, \pi)\in\Phi}$ converges to a Gaussian process $(T(\theta, \pi))_{(\theta, \pi)\in\Phi}$, and that this approximation still holds when considering the maximum over $(\theta, \pi) \in \Phi$.

For illustration we describe a result from Garel (2001) who considers $H_1 \leftrightarrow H_2$ for a two-component normal mixture (5) with a known first expectation $\vartheta_1 = 0$, a constrained second expectation $\vartheta_2 \in \Theta = [-a, a]$ (a bounded interval) and the probability $p := p_2 \in [0, 1]$. A Taylor expansion approximates, for a fixed $u = \vartheta_2$, the likelihood ratio by the value

$$T_n(u) := (n)^{-1/2} \sum_{k=1}^{n} \frac{e^{uX_k - u^2/2} - 1}{\left(e^{u^2} - 1\right)^2}$$

and it is shown:

***Theorem 4.1**:*

For $n \to \infty$ we have

$$\lambda_n = \sup_{u\in\Theta-\{0\}} T_n^2(u) \cdot I_{n(u)\geq 0+o_p(1)} \,. \tag{36}$$

Theorem 4.2:

For $n \to \infty$, the stochastic process $(T_n(u))_{u\in\Theta}$ converges weakly to a centered Gaussian process $(T(u))_{u\in\Theta}$ with unit variance and covariance function

$$C(u, v) := \frac{e^{uv} - 1}{\sqrt{\left(e^{u^2} - 1\right)\left(e^{v^2} - 1\right)}} \,.$$

Therefore, percentage points for λ_n can be obtained, for a large n, from the distribution of the maximum of the Gaussian process $(T(u))_{u\in\Theta}$, restricted to the parameter values $u\in\Theta - \{0\}$ with $T(u) \geq 0$. In practice, this latter distribution can be determined from a sufficiently large number of simulations of the process T. – Additionally, Garel (2001) proposes some useful bounds for percentage points and develops similar approximations for the general case where more or all mixture parameters ϑ_1, ϑ_2, σ_1^2, σ_2^2 are unknown. The case of three normal mixture components is considered in Garel (2002).

(2) Testing for multimodality – Silverman's test revisited

Whereas the clustering models reported in section 2 are prototypes for a *parametric* approach, we may formulate clustering structures in a *nonparametric* framework as well and characterize clusters by the modes and troughs of the underlying (typically unknown) probability density $f(x)$ or, in an empirical version, of its estimate $\hat{f}_n(x)$ obtained from the data points $x_1, ..., x_n$. Then each 'cluster' corresponds to a *mode* (local maximum) of the density f and comprises all (data) points which can be assigned to this mode by a 'hill climbing process' ('clusters of relatively large density' in Bock (1974), 'mode clusters').

In particular, when testing the hypothesis H_1 of 'homogeneity' versus the alternative H_m of 'a clustering structure with m classes' (with a fixed $m \geq 2$) we are led to the following specifications:

H_1: f is unimodal (homogeneity)

H_m: f has m modes (clustering).

A suitable multimodality test has been proposed by Silverman (1981, 1983) for the one-dimensional case. He considers, for n data points sampled from f, the classical kernel density estimate

$$\hat{f}_n(x;h) := \frac{1}{nh} \cdot \sum_{k=1}^{n} K\left(\frac{x - x_k}{h}\right) \quad x \in \mathrm{P}^1 \tag{37}$$

with the standard normal kernel $K(x) = (2\pi)^{-1/2} \; e^{-x^2/2}$ and a bandwidth $h > 0$. He uses the fact that (for this kernel) the number $M_n(h)$ of modes of $\hat{f}_n(\cdot;h)$ is a non-increasing function of h such that it makes sense to consider the *critical threshold*

$$\hat{h}_{crit}^{(n)} = \inf\{h > 0 \mid \hat{f}_n(\cdot;h) \text{ has exactly } m \text{ modes}\}. \tag{38}$$

Silverman's multimodality test of level (error probability) α rejects the homogeneity hypothesis H_1 in favor of H_m if $\hat{h}_{crit}^{(n)} > c(\alpha)$ with a critical threshold $c(\alpha)$.

Silverman proposed to approximate the percentage points $c(\alpha)$ of $\hat{h}_{crit}^{(n)}$ under H_1 by a bootstrap method based on random sampling from the data set $\{x_1, ..., x_n\}$. However, it is known that this estimate is not asymptotically accurate insofar as even for $n \to \infty$ the test's exact level deviates from the nominal one α. In fact, the bootstrap part of his method does not consistently estimate the distribution of $\hat{h}_{crit}^{(n)}$ under H_1, even up to scale and location transformation. This problem has initated several theoretical investigations on the behaviour of $\hat{h}_{crit}^{(n)}$. For instance, Mammen, Marron and Fisher (1992) have shown that $\hat{h}_{crit}^{(n)}$ is of the order $n^{-1/5}$ insofar as

$$\lim_{c_1 \downarrow 0, c_2 \uparrow \infty} \liminf_{n \to \infty} P_{H_0}\left(\frac{c_1}{n^{1/5}} < \hat{h}_{crit}^{(n)} < \frac{c_1}{n^{1/5}} \right) = 1$$

whereas under H_2 $\hat{h}_{crit}^{(n)}$ may even not converge to 0. Fisher, Mammen and Marron (1994) have addressed numerical properties of Silverman's technique.

Quite recently, Hall and York (2001) have analyzed the level inaccuracy of the test for H_1 versus H_2 in P^1 (theoretically and numerically). In particular, they proposed some calibration of the test statistic and of the critical threshold in order to attain a correct level accuracy, either by tables, by formulas, or by determining a 'critical distribution' of the modified test statistic. In order to cope with the occurrence of spurious modes caused by outlying data, they consider even the case where (even if the underlying density has an infinite support) the modes are confined to a finite interval I of P^1.

From a technical point of view the procedure resides on a bootstrapped version of $\hat{h}_{crit}^{(n)}$: For given data x1, ..., xn, let $\hat{f}_{crit}(\cdot)$:= fn(·; $\hat{h}_{crit}^{(n)}$) denote the corresponding density estimate (37) for h = $\hat{h}_{crit}^{(n)}$. Consider a resample X_1^*, ..., X_n^* drawn from the distribution $\hat{h}_{crit}^{(n)}$, and put

$$\hat{f}_n^*(x;h) := \frac{1}{nh} \cdot \sum_{k=1}^{n} K\left(\frac{x - X_k^*}{h} \right) \quad x \in P^1. \tag{39}$$

Now let $\hat{h}_{crit}^*$ denote the version of $\hat{h}_{crit}$ obtained in this setting for X_1, ..., X_n (i.e., the infimum of all $h > 0$ such that $\hat{f}_n^*(\cdot;h)$ has one mode only).

The new test statistic is the bootstrap distribution P_n^* of the rescaled threshold T_n^* := $\hat{h}_{crit}^* \setminus \hat{h}_{crit}$, and the level α test of H_2 against H_1 is to reject homogeneity H_1 if $P_n^*(T_n^* \le \tau \mid x_1, ..., x_n) \ge 1 - \alpha$ for a suitable threshold $\tau = \tau(\alpha)$.

It is shown that the bootstrap distribution function $\hat{G}_n(t)$:= $P_n^*(T_n^* \le t \mid x_1, ..., x_n)$ (with $t \ge 0$) converges weakly to a nondecreasing stochastic process $\left(\hat{G}(t)\right)_{t \ge 0}$ whose distribution does not depend on unknown parameters and whose paths (distribution functions) are continuous with probability 1. So we can calculate a constant $\tau(\alpha)$ which fulfils $P(\hat{G}(\tau_\alpha) \ge 1 - \alpha) = \alpha$ (Hall and York provide a table for τ_α). With this specification of τ_α (method 1), the modified bootstrap test is asymptotically correct, in that for $n \to \infty$

$$P_{H_1}(P_n^*(T_n^* \le \tau_\alpha \mid X_1, ..., X_n) \ge 1 - \alpha) \to P(\hat{G}(\tau_\alpha) \ge 1 - \alpha) = \alpha$$

Alternatively, τ_α may be estimated from simulations of the distribution $\hat{G}$ (method 2). Other cases such as $H_m \leftrightarrow H_{m+1}$ can be treated in a similar way.

For some related or other multimodality tests see, e.g., Good and Gaskins (1980), Izenmann and Sommer (1988), Fisher et al. (1994), Bock (1996a) and Minotte (1997).

5. New algorithms and new applications

When browsing through the recent literature, e.g., in the computer science and Data Mining domain, we find a lot of proposals for new clustering strategies, often designed for special purposes or improving on speed and storage space. However, it appears often that 'new' algorithms are no more than copies or adaptations of traditional clustering algorithms. For example:

- *DENCLUE* (Hinneburg and Keim 1998) is essentially the classical hill-climbing algorithm for locating modal clusters as described, e.g., in Schnell (1964), Ihm (1965), Bock (1974).
- *DBSCAN* (Ester et al. 1996) is based on the graph-theoretical linking methods described already by Wishart (1969) and the (*k, d*)-clusters proposed by Ling (1972).
- *K-Harmonic Means Clustering* (Zhang, Hsu and Dayal 2000) is identical to Bezdek's fuzzy clustering method (Bezdek 1974) with fuzziness exponent $r = 2$ (see Bock 1979a, 1979b).

Nevertheless, the indicated papers provide often some progress insofar as they are related to new software which uses the full possibilities and capacities of modern computer equipment and data base management with the result that the methods can now be applied routinely in practice whereas at the time of their invention they needed too much time and/or too much storage space.

A major issue in clustering is the occurrence of *high dimensional data* and the availability of *very large data sets*. The latter situation has always been a challenge for classificationists, and the former one is closely related to the inherent sparsity of the data points (e.g., when the sample size is of the same order as the dimension) where even the concept of proximity or clustering may break down. A plethora of corresponding algorithms has been developed in the last years in the fields of artificial intelligence, data mining, database management and informatics, but is not so well known in the communities of statisticians. Without going into details here, we mention two typical approaches:

(1) Projection methods

When dealing with high-dimensional data, a helpful strategy for locating clusters is to search for clusters which are characterized by low-dimensional

class-specific subspaces (as 'centers'), or for *global* projections where clustering can be detected. On the other hand, such projection methods are useful also when analyzing very large databases in order to reduce the large amount of information. Classical methods such as *principal component* or *subspace clustering* (Bock 1974) and *projection pursuit clustering* (Bock 1987) have been completed by heuristical search algorithms, for example:

– *ORCLUS* and *PROCLUS* are algorithms that proceed by an iterative search for subspaces and by suitable minimum-distance assignments, eventually combined with a hierarchical merging strategy which reduces iteratively the number of classes and the dimensions of the class-specific hyperplanes (Aggarwal, Park et al. 1999, Aggarwal and Yu 2000);

– *CLIQUE* identifies dense clusters in subspaces of maximum dimensionality (Agrawal et al. 1998). The algorithm begins with the dimension one and increases the dimension in several passes.

(2) Grid-based methods

These methods are density-oriented and look essentially for *mode clusters*. In order to tackle with a large amount of data and, simultaneously, a high dimension they partition the data space P^p into small cells by using a grid formed by appropriate $(p - 1)$-dimensional (not necessarily orthogonal or equidistant) cutting hyperplanes. Cutting planes are constructed by heuristic rules involving projections, marginal point densities and density troughs, eventually coarsening successively the grid and the data. For example:

– *DENCLUE* (Hinneburg and Keim 1998) uses a grid-based strategy, but stores only grid cells which contain data points. It connects all (non-empty) neighbouring cells of a highly-populated cell (density attractors).

– *BIRCH* (*Balanced Iterative Reducing and Clustering using Hierarchies*, see Zhang, Ramakrishnan and Linvy 1997) uses a hierarchical data structure called Cluster-Feature-tree. It stores similar data items in a node of the CF-tree.

– STING (*STatistical INformation Grid*, see Wang, Yang and Muntz 1997) divides the space into rectangular cells and stores the statistical properties (means, variance, ...) of the objects of a cell in the nodes of a quadtree-like structure.

– *WaveCluster* (Sheikhoeslami et al. 1998) is a wavelet-based approach which uses a rectangular grid. Its main idea is to apply a wavelet transformation to the (content of the) grid cells and to determine dense regions in the transformed domain be searching for connected components.

– *OptiGrid* (Hinneburg and Keim 1999) is a grid-partitioning technique designed for high-dimensional data. In several iterative loops, the algorithm selects some suitable contracting projection operators, determines in each of the projected data sets a density though and a corresponding local cutting plane, and then selects

some optimum cutting planes from which the grid and the clusters are constructed. Additional steps may lead to a refined clustering.

New data types in new fields of application

Finally, we want to point to some new domains where clustering problems occur, possibly formulated in terms of 'non-classical' data types or for quite specific applications.

A first remark concerns the field of *molecular biology* where entities (persons, bacteria, viruses, proteins) are characterized by DNA strains, molecular structures and gene constellations. During the last years, various clustering methods have been designed for such data, but the large amount of variables, the combinatorial structure of strain elements, and the underlying (chemical, spatial,...) constraints render this field very difficult. A particular field of research is the analysis of *gene expression data*} obtained from micro arrays; see, e.g., Hedge (2000), Dudoit et al. (2002), Allison (2002), Dudoit et al. (2002), Krause (2002), Lausen (2003), and Ziegler et al. (2003). Here clustering methods are indispensible in order to reveal useful groups of genes (e.g., related to specific diseases), groups of proteins, groups of animals, etc. Some recent approaches are described by Eisen (1998), Ben-Dor et al. (1999), Golub et al. 1999, Hartuv et al. (1999), Alizadeh et al. (2000), Hastie, Tibshirani and Eisen (2000), Heydebrinck et al. (2001), Krause (2002), Ramoni et al. (2002), Yeung et al. (2001), Markowetz and Heydebrinck (2003), and McLachlan et al. (2003).

Marketing and economics were always a classical domain for applying clustering methods, e.g., for detecting homogeneous groups of consumers, groups of (substitutable) products, enterprise types, etc., in the framework of *Data Mining* and *KDD*. Here clustering methods were mostly based on quantitative and qualitative variables. Nowadays, however, the emergence of internet applications, web technology and E-commerce has created new data types in the form of *navigation paths* which describe the way in which, e.g., a web user navigates backwards and forward in the website catalogue of an enterprise. Analysis and clustering of such data should reveal the needs and preferences of the user and the identification of user types in order to presenting and offering automatically suitable products which the user may like to buy (recommender services). Various methods have been proposed in this domain which is still in full progress; see, e.g., Gaul and Thieme-Schmidt (2000), Böhm et al. (2003), Punin et al. (2003).

Finally we point to the consideration of *symbolic data* where variables may be sets of categories, intervals, or frequency distributions. For example, a 'data vector' which describes the properties of a club (item k) in terms of the car types, the age and the gender of its members, could have the form x_k = ({*Audi*, *Peugeot*}, [35, 45], (*male*: 65%, *female*: 35%)) meaning that the club is composed by members who drive *Audi* or *Peugeot*, are aged between 35 and 45, and such that 65% of the club members are men. Such data occur, e.g., in survey statistics where individual data may be aggregated, e.g., on a local or regional level before

being integrated into a global, e.g., national or European database. There are many proposals for analyzing such 'symbolic' data and, e.g., to detect groups of similar clubs, cities, products, ... For details see, e.g., Bock and Diday (2000) and the recent electronic *Journal of Symbolic Data Analysis* (http://www.jsda.unina2.it).

REFERENCES

AGGARWAL, Ch.C., PARK, J.S., PROCPIUC, C., WOLF, J.L. and Ph.S. YU (1999): Fast algorithms for projected clustering. Proc. ACM SIGMOD International Conference on Management of Data, 1999, 61-72.

AGGARWAL, Ch.C., and Ph.S. YU (2000): Finding generalized projected clusters in hig dimensional spaces. Proc. ACM SIGMOD International Conference on Management of Data, 2000, 70-81.

AGRAWAL, R., GEHRKE, J., GUNOPULOS, D., and P. RAGHAVAN (1998): Proc. ACM SIGMOD International Conference on Management of Data, 1998, 94-105.

ALIZADEH, A.A., EISEN, M.B., et al. (2000): Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403, 503-511.

ALLISON, D.B., GADBURY, G.L., et al. (2002): A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics and Data Analysis* 39, 1-20.

BANFIELD, J.D., and A.E. RAFTERY (1993): Model-based Gaussian and non-Gaussian clustering. *Biometrics* 49, 803-821.

BEN-DOR, A., SHAMIR, R., and Z. YAKHINI (1999): Clustering gene expression patterns. *J. of Computational Biology* 6, 281-297.

BERDAI, A., and B. GAREL (1994): Performances d'un test d'homogénéité contre une hypothèse de mélange gaussien. *Revue de Statistique Appliquée* 42 (1), 63-79.

BERDAI, A., and B. GAREL (1996): Detecting a univariate normal mixture with two components. *Statistics and Decisions* 16, 35-51.

BICKEL, P., and H. CHERNOFF (1993): Asymptotic distribution of the likelihood ratio statistic in a prototypical non regular problem. In: J.K. Ghosh et al. (eds.): *Statistics and Probability*. Wiley Eastern Limited, New Delhi, 83-96.

BOCK, H.-H. (1974): *Automatische Klassifikation (Cluster-Analyse)*. Vandenhoeck & Ruprecht, Göttingen.

BOCK, H.-H. (1979a): Clusteranalyse mit unscharfen Partitionen. In: H.-H. Bock (ed.): *Klassifikation und Erkenntinis III: Numerische Klassifikation*. INDEKS-Verlag, Frankfurt, 137-163.

BOCK, H.-H. (1979b): Fuzzy clustering procedures. In: R. Tomassone: *Analyse de données et informatique*. INRIA, Le Chesnay, 205-218.

BOCK, H.-H. (1983): A clustering algorithm for choosing optimal classes for the chi-squared test. Bull. Intern. Statist. Inst., 44th Session, Madrid 1983, Vol II: Contributed papers, 758-762.

BOCK, H.-H. (1984): Statistical testing and evaluation methods in cluster analysis. In: J.K. Ghosh & J. Roy (Eds.): *Golden Jubilee Conference in Statistics: Applications and new directions*. Calcutta, December 1981. Indian Statistical Institute, Calcutta, 1984, 116-146.

BOCK, H.-H. (1985): On some significance tests in cluster analysis. *J. of Classification* 2, 77-108.

BOCK, H.-H. (1987): On the interface between cluster analysis, principal component analysis, and multidimensional scaling. In: H. Bozdogan and A.K. Gupta (eds.): *Multivariate statistical modeling and data analysis*. Reidel, Dordrecht, 17-34.

BOCK, H.-H. (1991): A clustering technique for maximizing ϕ-divergence, noncentrality and discriminating power. In: M. Schader (ed.): *Analyzing and modeling data and knowledge*. Springer-Verlag, Heidelberg, 1991, 19-36.

BOCK, H.-H. (1994): Information and entropy in cluster analysis. In: H. Bozdogan et al. (eds.): *The Frontiers of Statistical Modeling: An Informational Approach*. Proc. First US/Japan Conference on Statistical Modeling, Knoxville, Tennessee, May 1992. Kluwer Academic Press, Dordrecht, 1994, Vol. II, 115-147.

BOCK, H.-H. (1996a): Probability models and hypothesis testing in partitioning cluster analysis. In: Ph. Arabie, L. Hubert, and G. De Soete (eds.): *Clustering and classification*. World Science Publishers, River Edge/NJ, 1996, 377-453.

BOCK, H.-H. (1996b): Probabilistic models in partitional cluster analysis. In: A. Ferligoj and A. Kramberger (eds.): *Developments in data analysis*. FDV, Metodoloski zvezki, 12, Ljubljana, Slovenia, 1996, 3-25.

BOCK, H.-H. (1996c): *Probabilistic models and statistical methods in classification*. Tutorial organized by the Japanese Classification Society and the Japan Marketing Association, Tokyo, April 2-3, 1996. Japan Marketing Association, Tokyo, 1996, 50-68.

BOCK, H.-H. (1996d): *Probabilistic models in cluster analysis*. Computational Statistics and Data Analysis 23, 5-28.

BOCK, H.H. (1998): Probabilistic aspects in classification. In: C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H.-H. Bock, Y. Baba (eds.): *Data science, classification and related methods*. Springer-Verlag, Heidelberg, 1998, 3-21.

BOCK, H.-H. (2001): Méthodes de classification pour des critères basés sur la convexité. Actes du VIIIème Congrès de la Societé Francophone de Classification, Université des Antilles-Guyane, Pointe-à-Pitre, Guadeloupe, 1-5.

BOCK, H.-H. (2002a): Clustering methods with convexity-based clustering criteria with applications. *Statistical Methods and Applications* (submitted)

BOCK, H.-H. (2002b): Two-way clustering for probability distributions: maximally dependent clusters. (Preprint)

BOCK, H.-H. (2003): Two-way clustering for contingency tables: maximizing a dependence measure. In: W. Gaul and M. Schader (eds.): *Between data science and everyday web practice*. Springer Verlag, Heidelberg, 2003 (submitted).

BOCK, H.-H., and E. DIDAY (2000): *Analysis of symbolic data. Exploratory methods for extracting statistical information from complex data*. Springer Verlag, Heidelberg.

BÖHM, W., GEYER-SCHULZ, A., HAHSLER, M., and M. JAHN (2003): Repeat-buying theora and its application in recommender services. In: M. Schwaiger and O. Opitz (eds.): *Exploratory data analysis in empirical research*. Springer Verlag, Heidelberg, 229-239.

BÖHNING, D., DIETZ, E., SCHAUB, R., SCHLATTMANN, P., & LINDSAY, B.G. (1994): The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Mathematical Statistics* 46, 373-388.

BRYANT, P.G. (1991): Large-sample results for optimization-based clustering methods. *J. of Classification* 8, 31-44.

BRYANT, P.G., and J.A. WILLIAMSON (1978): Asymptotic behaviour of classification maximum likelihood estimates. *Biometrika* 65, 273-281.

CÉLEUX, G., & DIEBOLT, J. (1985): The SEM algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly*} 2, 73-82.

COX, D.R. (1957): A note on grouping. *J. Amer. Statist. Assoc*. 52, 543-547.

CSISZÁR, I. (1967): Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica* 2, 299-318.

DACUNHA-CASTELLE, D., and E. GASSIAT (1999): Testing the order of a model using locally conic parametrization: population mixtures and stationary ARMA processes. *Annals of Statistics* 27, 1178-1209.

DUDOIT, S., FRIDLYAND, J., and T.P. SPEED (2002): Comparison of discrimination methods for the classification of tumours using gene expression data. *J. American Statistical Association* 97, 77-87.

EISEN, M.B., SPELLMAN, P.T., BROWN, P.O., and D. BOTSTEIN (1998): Cluster analysis and display of genome-wide expression patterns. *Proc. National Academy of Science* 95, 14863-14868.

ESTER, M., KRIEGEL, H.-P., SANDER, J., and X. XU (1996): A density-based algorithm for discovering clusters in large spatial databases with noise. Proc. 2nd Intern. Conf. on Knowledge Discovery and Data Mining, AAAI Press, 1996.

EVERITT, B. S. (1981): A Monte Carlo investigation of the likelihood ratio test for the number of components in a mixture of normal distributions. *Multivariate Behavioural Research* 16, 171-180.

FISHER, N.I., MAMMEN, E., and J.S. MARRON (1994): Testing for multimodality. *Computational Statistics and Data Analysis* 18, 499-512.

FLURY, B.D. (1990): Principal points. *Biometrika* 77, 33-41.

FLURY, B.D. (1993): Estimation of principal points. *JRSS Series C: Applied Statistics* 42, 139-151.

GALLEGOS, M.T. (2001): Robust clustering under general normal assumptions. Technical report MIP-0103 September 2001. Fakultät für Mathematik und Informatik, Universität Passau.

GALLEGOS, M.T. (2002): Maximum likelihood clustering with outliers. In: K. JAJUGA, A. SOKOLOWSKI, and H.-H. BOCK (eds.): *Classification, clustering and data analysis – recent advances and applications.* Springer Verlag, Heidelberg, 247-255.

GALLEGOS, M.T. (2003): Clustering in the presence of outliers. In: M. Schwaiger and O. Opitz (eds.): *Exploratory data analysis in empirical research*. Springer Verlag, Heidelberg, 58-66.

GAREL, B. (2001): Likelihood ration test for univariate Gaussian mixture. *Journal of Statistical Planning and Inference* 96, 325-350.

GAREL, B., and F. GOUSSANOU (2002): Removing separation conditions in a 1 against 3-component Gaussian mixture problem. In: K. Jajuga, A. Sokolowski, H.-H. Bock (eds.): *Classification, clustering and data analysis*. Springer Verlag, Heidelberg, 61-73.

GAUL, W., and L. SCHMIDT-THIEME (2000): Frequent generalized subsequences - a problem from Web Mining. In: W. Gaul, O. Opitz and M. Schader (eds.): *Data analysis. Scientific modeling and practical applications*. Springer Verlag, Heidelberg, 2000, 429-445.

GHOSH, J. K., & SEN, P. K. (1985): On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results. In: L.M. LeCam, R.A. Ohlsen (Eds.): Proc. Berkeley Conference in honor of Jerzy Neyman and Jack Kiefer. Vol.II, Wadsworth, Monterey, 1985, 789-806.

GOFFINET, B., LOISEL, P., and B. LAURENT (1992): Testing in normal mixture models when the proportions are known. *Biometrika* 79, 842-846.

GOLUB, T.R., SLONIM, D.K., et al. (1999): Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286, 531-537.

GOOD, I.J., and R.A. GASKINS (1980): Density estimation and bump hunting by the penalized likelihood method exemplified by scattering and meteorite data. *J. American Statistical Association* 75, 42-73.

HALL, P., and M. YORK (2001): On the calibration of Silverman's test for multimodality. *Statistica Sinica* 11, 515-536.

HARTIGAN, J.A. (1978): Asymptotic distributions for clustering criteria. *Annals of Statististics* 6, 117-131.

HARTIGAN, J.A. (1985a): A failure of likelihood asymptotics for normal mixtures. In: L.M. Le Cam, and R.A. Ohlsen (eds.): Proc. Berkeley Conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II. Wadsworth, Monterey, 807-810.

HARTIGAN, J.A. (1985b): Statistical theory in clustering. *J. of Classification* 2, 63-76.

HARTUV, E., SCHMITT, A., et al. (1999): An algorithm for clustering cDNAs for gene expression analysis. Proc. Third International Conference on Computational Molecular Biology (RECOMB99), 188-197.

HASTIE, T., and W. STÜTZLE (1989): Principal curves. *J. American Statistical Association* 84, 502-516.

HASTIE, T., TIBSHIRANI, R., EISEN, M., et al. (2000): Gene shaving as a method for identifying sets of genes with similar expression patterns. *Genome Biology* 1, Research 0003.1-21.

HEDGE, P., et al. (2000): A concise guide to CDNA microarray analysis. *BioTechniques* 29, 548-562.

HEYDEBRECK, A. von, HUBER, W., POUSTKA, A., and M. VINGRON (2001): Identifying splits with clear separation: a new class discovery method for gene expression data. *Bioinformatics* 17, Suppl. 1, S107-S114.

HINNEBURG, A., and D.A. KEIM (1998): An efficient approach to clustering in large multimedia databases with noise. Proc. 4th International Conference on Knowledge Discovery and Data Mining, 1998, 58-65.

HINNEBURG, A., and D.A. KEIM (1999): Optimal grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering. Proc. 25th Conference on Very Large Data Bases (VLDB'99), Edinburgh, Scotland, 1999, 506-517.

HÖPPNER, F., KLAWONN, F., and R. KRUSE (1997): *Fuzzy-Clusteranalyse. Verfahren für die Bilderkennung, Klassifikation und Datenanalyse*. Vieweg, Braunschweig.

ZENMANN, A.J., and C. SOMMER (1988): Philatelic mixtures and multimodal densities. *J. American Statistical Association* 83, 941-953.

JOLION, J.-M., and A. ROSENFELD (1989): Cluster detection in background noise. *Pattern Recognition* 22, 603-607.

KARLIN, S. (1982): Some results on optimal partitioning of variance and monotonicity with truncation level. In: G. Kallianpur, P.R. Krishnaiah, J.K. Ghosh (eds.): *Statistics and probability: Essays in honor of C.R. Rao*. North-Holland, Amsterdam, 375-382.

KIPPER, S., and PÄRNA, K. (1992): Optimal *k*-centres for a two-dimensional normal distribution. *Acta et Commentationes Universitatis Tartuensis*, Tartu Ülikooli TOIMEISED 942, 21-27.

KOHONEN, T. (1984): *Self-organization and associative memory*. Springer Verlag, Berlin.

KRAUSE, A. (2002): *Large scale clustering of protein sequences*. Dissertation, Universität Bielefeld, Germany.

KROLAK-SCHWERDT, S. (2003): Two-mode clustering methods: compare and contrast. In: W. Gaul and M. Schader (eds.): *Between data science and everyday web practice*. Springer Verlag, Heidelberg, 2003 (submitted).

LAUSEN, B. (2002): Bioinformatics and classification: The analysis of genome expression patterns. In: K. Jajuga, A. Sokolowski, H.-H. Bock (eds.): *Classification, clustering and data analysis*. Springer Verlag, Heidelberg, 455-461.

LI, L., and B. FLURY (1995): Uniqueness of principal points for univariate distributions. *Statistics and Probability Letters* 25, 323-327.

MAMMEN, E., MARRON, J.S., and N.I. FISHER (1992): Some asymptotics for multimodality tests based on kernel density estimates. *Probability Theory and Related Fields* 91, 115-132.

MCLACHLAN, G.J. (1987): On bootstrapping the likelihood ratio statistic for the number of components in a normal mixture. *Applied Statistics* 36, 318-324.

MCLACHLAN, G., and D. PEEL (2000): *Finite mixture models*. Wiley, New York.

MCLACHLAN, G., NG, S.K., and D. PEEL (2000): On clustering by mixture models. In: M. Schwaiger and O. Opitz (eds.): *Exploratory data analysis in empirical research*. Springer Verlag, Heidelberg, 141-148.

MARKOWETZ, F., and A. von HEYDEBRECK (2003): Class discovery in gene expression data: characterizing splits by support vector machines. In: W. Gaul and M. Schader (eds.): *Between data science and everyday web practice*. Springer Verlag, Heidelberg, 2003 (submitted).

MATHAR, R. (1981): *Ausreißer bei ein- und mehrdimensionalen Wahrscheinlichkeitsverteilungen*. Dissertation, Mathematisch-Naturwissenschaftliche Fakultät, Rheinisch-Westfälische Technische Hochschule (RWTH) Aachen.

MENDELL, N.P., THODE, H.C., & FINCH, S.J. (1991): The likelihood ratio test for the two-component normal mixture problem: power and sample-size analysis. *Biometrics* 47, 1143-1148. Correction: 48 (1992) 661.

MENDELL, N.P., FINCH, S.J., and THODE, H.C. (1993): Where is the likelihood ratio test powerful for detecting two-component normal mixtures? *Biometrics* 49, 907-915.

MIZUTA, M. (1995): Asymmetric principal points of symmetric distributions. Proc. 10th Symposium of the Japanese Society of Computational Statistics, 189-196 (in Japanese).

MIZUTA, M. (1998): Two principal points of symmetric distributions. In: A. Rizzi, M. Vichi, H.-H. Bock (eds.): *Advances in data science and classification*. Springer Verlag, Heidelberg, 171-176.

PÄRNA, K. (1991): Clustering in metric spaces: some existence and continuity results for *k*-centers. In: M. Schader (ed.): *Analyzing and modeling data and knowledge*. Springer Verlag, Heidelberg, 85-91.

PESCH, Ch. (1999): *Eigenschaften des gegenüber Ausreißern robusten MCD-Schätzers und Algorithmen zu seiner Berechnung*. Dissertation, Universität Passau,.

POLLARD, D. (1982): A central limit theorem for *k*-means clustering. *Annals of Probability* 10, 919-926.

PÖTZELBERGER, K. (2002): Quantization of models: Local approach and asymptotically optimal partitions. In: K. Jajuga, A. Sokolowski, H.-H. Bock (eds.): *Classification, clustering and data analysis*. Springer Verlag, Heidelberg, 97-105.

PÖTZELBERGER, K., and H. STRASSER (2001): Clustering and quantization by MSP-partitions. *Statistics and Decisions* 19, 331-371.

PUNIN, J.R., KRISHNAMOORTHY, M.S., and M.J. ZAKI (2003): Web usage mining - languages and algorithms. In: M. Schwaiger and O. Opitz (eds.): *Exploratory data analysis in empirical research*. Springer Verlag, Heidelberg, 266-281.

RAMONI, M.F., SEBASTIANI, P., and I.S. KOHANE (2002): Cluster analysis of gene expression dynamics. *Proc. National Academy of Sciences* 99, 9121-9126.

REDNER, R.A. (1981): Note on the consistency of the maximum likelihood estimate for non identifiable distributions. *Annals of Statistics* 9, 225-228.

ROUSSEEUW, P.J. (1983): Multivariate estimation with high breakdown point. In: W. Grossmann, G. Pflug, I. Vincze, W. Werts (eds.): *Mathematical statistics and applications*. 8th Pannonian Symposium. Reidel, Dordrecht, 283-297.

ROUSSEEUW, P.J., KAUFMAN, L., and E. TRAUWAERT (1996): Fuzzy clustering using scatter matrices. *Computational Statistics and Data Analysis* 23, 135-151.

ROUSSEEUW, P.J. and K. VAN DRIESSEN (1999): A fast algorithm for the minimum covariance determinant estimator. *Technometrics* 41, 212-223.

ROUSSEEUW, P.J., and S. VERBOVEN (2001): Robust estimation in very small samples. *Computational Statistics and Data Analysis* 40, 741-758

SHEIKHOLESLAMI, G., CHATTERJEE, S., and A. ZHANG (1998): Wave-Cluster: a multi-resolution clustering approach for very large spatial databases. Proc. 24th Conf. on Very Large Data Bases, 1998.

SHIMIZU, N., MIZUTA, M., and Y. SATO (1997): Some properties of principal points. Proc. 9th Korea and Japan Joint Conference on Statistics: Multivariate Analysis and Computation. Cheju, Korea, 257-262 (in Japanese).

SHIMIZU, N., MIZUTA, M., and Y. SATO (1998): Some properties of principal points. *Japanese Journal of Applied Statistics* 27, 1-16 (in Japanese).

SHIMIZU, N., MIZUTA, M., and Y. SATO (1999): Sufficient conditions for uniqueness of principal points of symmetric univariate distributions. *Bulletin of the Computational Statistics in Japan* 12, 45-53 (in Japanese).

SILVERMAN, B.W. (1981): Using kernel density estimates to investigate multimodality. *J. Royal Statistical Society* B 43, 97-99.

SILVERMAN, B.W. (1983): Some properties of a test for multimodality based on kernel density estimates. In: J.F.C. Kingman and G.E.H. Reuter (Eds.): *Probability, Statistics and Analysis*. Cambridge University Press, Cambridge, UK, 248-259.

SPÄTH, H. (1985): *Cluster dissection and analysis*. Wiley, Chichester.

STRASSER, H. (2000): Towards a statistical theory of optimal quantization. In: W. Gaul, O. Opitz, and M. Schader (eds.): *Data analysis. Scientific modeling and practical application*. Springer-Verlag, Heidelberg, 2000, 369-383.

TARPEY, Th. (1994): Two principal points of symmetric, strongly unimodal distributions. *Statistics and Probability Letters* 20, 253-257.

TARPEY, Th. (1998): Self-consistent patterns for symmetric multivariate distributions. *J. of Classification* 15, 57-79.

TARPEY, Th., Li, L., FLURY, B.D. (1995): Principal points and self-consistent points of elliptical distributions. *Annals of Statistics* 23, 103-112.

THODE, H.C., FINCH, S.J., & MENDELL, N.R. (1988): Simulated percentage points for the null distribution of the likelihood ratio test for a mixture of two normals. *Biometrics* 44, 1195-1201.

TITTERINGTON, D.M. (1990): Some recent research in the analysis of mixture distributions. *Statistics* 21, 619-641.

TITTERINGTON, D.M., A.F.M. SMITH and U.E. MAKOV (1985): *Statistical analysis of finite mixture distributions*. Wiley, New York.

TRUSHKIN, A.V. (1982): Sufficient conditions for uniqueness of a locally optimal quantizer for a class of convex error weighting functions. *IEEE Trans. Information Theory* IT-28, 187-198.

WANG, W., YANG, J., and R. MUNTZ (1997): STING: A statistical information grid approach to spatial data mining. Proc. 23rd Intern. Conf. on Very Large Data Bases (VLDB'97). Morgan Kaufman, 186-195.

WISHART, D. (2003): *k*-means clustering with outlier detection, mixed variables and missing values. In: M. Schwaiger and O. Opitz (eds.): *Exploratory data analysis in empirical research*. Springer Verlag, Heidelberg, 216-226.

WOLFE, J.H. (1971): A Monte-Carlo study of the sampling distribution of the likelihood ratios for mixtures of multinormal distributions. Technical

Bulletin STB, 72-2. U.S. Naval Personnel and Training Research Laboratory, San Diego.

YAMAMOTO, W., and N. SHINOZAKI (2000): On uniqueness of two principal points for univariate location mixtures. *Statistics and Probability Letters* 46, 33-42.

YEUNG, K.Y., HAYNOR, D.R., and W.L. RUZZO (2001): Validating clustering for gene expression data. *Bioinformatics* 17, 309-318.

ZHANG, B., HSU, M., and U. DAYAL (2000): K-harmonic means – a spatial clustering algorithm via boosting. In: J.F. Roddick and K. Hornsby (Eds.): *Temporal, spatial and spatio-temporal data mining (TSDM) 2000*. Springer Verlag, Berlin, 31-45.

ZHANG, T., RAMAKRISHNAN, R., and M. LINVY (1996): BIRCH: An efficient data clustering method for very large databases. Proc. ACM SIGMOID Intern. Conf. on Management of Data. ACM Press, 103-114.

ZHANG, H., YU, C.-Y., SINGER, B., and M. XIONG (2001): Recursive partitioning for tumor classification with gene expression microarray data. *Proc. National Academy of Science* 98, 6730-6735.

ZIEGLER, A., HARTMANN, O., KÖNIG, I.R., and H. SCHÄFER (2003): Statistical genetics – present and future. In: M. Schwaiger and O. Opitz (eds.): *Exploratory data analysis in empirical research*. Springer Verlag, Heidelberg, 401-409.

ZOPPÈ, A. (1997): On uniqueness and symmetry of self-consistent points of univariate continuous distributions. *J. of Classification* 14, 147-158.

STATISTICS IN TRANSITION, October 2002
Vol. 5, No. 5, pp. 759–777

ORTHOGONAL COMPONENTS FOR GROUPED DATA: REVIEW AND APPLICATIONS

W.J. Krzanowski[1]

ABSTRACT

Principal component analysis is a key technique in the analysis of multivariate data, encompassing three possible objectives: description, interpretation and modelling of the data. Orthogonality is a central plank in all three aspects. However, *a-priori* grouping of individuals is not accommodated within the technique. Traditionally, canonical variate analysis is used when data have this structure, but this technique has a number of drawbacks as regards each of the three objectives. Consequently, attempts have been made over the past twenty-five years to uphold the principal component approach while allowing for group structure in the data. The resultant techniques have been published in a wide range of journals, so the main purpose of the present article is to draw this work together in a single review. We also highlight a variety of generic applications in which the techniques have been used. These include discriminant analysis, cluster analysis, distance-based analysis, spatial analysis and industrial process control.

Key words. Eigenvalues; eigenvectors; linear transformations; spectral models; subspace projection.

1. Introduction

Principal component analysis (PCA) is one of the most heavily used of multivariate techniques, and it has found applications in a wide variety of substantive areas (see, e.g., Jackson, 1991, or Jolliffe, 1986). Mathematical development took place throughout the 20th century, and the three prime objectives in most applications of the methodology were identified at approximately evenly-spaced intervals during this period. Pearson (1901) took a geometrical standpoint, using the representation of p variables measured on a sample of n objects as a set of n points in p dimensions, and introduced the technique as one that successively identifies the r-dimensional subspaces of

[1] *Address for correspondence:* School of Mathematical Sciences, University of Exeter, Laver Building, North Park Road, Exeter EX4 4QE.

closest fit to such a set of points for $r = 1, 2, ..., p–1$. This formulation underpins the use of PCA for *description* of the multivariate data. By contrast, Hotelling (1933) took an algebraic approach, and established that Pearson's principal components were also the orthogonal directions in space that successively maximised the variance of the data points. In other words, the *r*-dimensional subspace of closest fit to the data is also the one in which the scatter of the points is maximised, and this feature is central to the use of PCA for *interpreting* the multivariate data via reification (Krzanowski, 2000 p. 54). The third major advance was by Anderson (1963), who took a statistical approach and developed the distributional theory underlying principal components. This unified previous ad-hoc results, and opened up the possibility of PCA being used to *model* a set of multivariate data.

Orthogonality of components plays a key role in each of these objectives. In multivariate description, orthogonal components ensure that subspaces are derived from the original data space by orthogonal projection, thus ensuring that there is no distortion of the original configuration. In multivariate interpretation, having orthogonal components means that identified sources of "important" variation are uncorrelated, while in multivariate modelling orthogonal components provide a simple population dispersion structure. However, *a-priori* grouping of units is not catered for in any of these formulations. Instead, imposition of such grouping structure on the data led to the development of canonical variate analysis (CVA) in the second half of the 20th century. Objectives of this technique broadly parallel those of PCA. Rao (1948) derived the unweighted version as appropriate for defining the r-dimensional subspace in which the total Mahalanobis squared distance between all pairs of groups is maximised, and Ashton *et al* (1957) highlighted this approach as the one providing the best description of between-group differences. Bryan (1951) provided the algebraic details for generating components that successively maximise the between-group relative to the within-group variance, and this permits between-group differences to be interpreted in analogous manner to PCA interpretation. Finally, Campbell (1984) formulated a model-based version of CVA in which the population means are expressed as a function of the population canonical variates, which enables the modelling of data to be extended to situations involving *a-priori* group structure.

However, despite the evident parallels between CVA and PCA, the former technique embodies some fundamental differences from the latter. First, it produces non-orthogonal components, which induce a deformation of the multivariate space if used as the axes of a subspace representation. Second, if there are g groups then the maximum number of components that can be produced by the technique is the smaller of $g - 1$ and p, which means that very few components can be obtained when g is small. Third, there are problems of singularity when either p is very large or n is very small, situations that occur

frequently in applications such as in spectroscopy or chemo metrics that involve automatic data recording devices.

Attention has thus been paid over the past twenty-five years to providing various extensions to PCA that can be used in the presence of group structure in the data and that do not suffer from the above drawbacks. The purpose of this article is to review these developments, and to show how they can be used in such generic areas as discriminant analysis, cluster analysis, distance-based analysis, spatial analysis and process control. The review will be split into the three objectives listed earlier, namely description, interpretation and modelling. First, however, we establish some notation and terminology.

2. Fundamentals

We suppose that the data comprise n observations $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_n$ on a p-element vector $\mathbf{X}$. The sample mean vector will be denoted by $\bar{\mathbf{x}}$, and the sample covariance matrix by $\mathbf{S}$. We have already referred to the representation of this sample as n points in p-dimensional space; in the presence of g a-priori groups or populations, the n observations are divided into g sets ($\mathbf{x}_{ij}$, $i = 1, ..., g$; $j = 1, ..., n_i$) and the cloud of points into g more or less distinct sub-clouds. Group mean vectors will be denoted $\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, ..., \bar{\mathbf{x}}_g$, the between-group covariance matrix $\mathbf{B} = \frac{1}{g-1}\sum_{i=1}^{g} n_i (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^T$, and the within-group covariance matrix $\mathbf{W} = \frac{1}{n-g}\sum_i \sum_j (\bar{\mathbf{x}}_{ij} - \bar{\mathbf{x}}_i)(\bar{\mathbf{x}}_{ij} - \bar{\mathbf{x}}_i)^T$.

Many multivariate techniques involve linear combinations $\mathbf{Y}_i = \mathbf{l}_i^T \mathbf{X}$, i.e. $\mathbf{Y} = \mathbf{L}^T\mathbf{X}$ where $\mathbf{l}_i$ is the ith column of $\mathbf{L}$. In particular, PCA is defined by $\mathbf{Y} = \mathbf{L}^T\mathbf{X}$ where $\mathbf{L}^T\mathbf{S}\mathbf{L} = \mathbf{D}$ (diagonal) and $\mathbf{L}^T\mathbf{L} = \mathbf{I}$
and CVA by
$\mathbf{Y} = \mathbf{L}^T\mathbf{X}$ where $\mathbf{L}^T\mathbf{B}\mathbf{L} = \mathbf{D}$ (diagonal) and $\mathbf{L}^T\mathbf{W}\mathbf{L} = \mathbf{I}$
We consider now other possibilities for such transformations.

3. Interpretation

One problem frequently met is when several groups of invidividuals have had the same variables measured on them, PCA has been conducted on each group separately, and then it is desired to compare the first k components (say) of each group (i.e. to compare $\mathbf{L}_i$, $i = 1, ..., g$). This may occur, for example, if the groups are individuals subjected to different experimental conditions, or if they are observations taken in different environments, or at different times.

The first attempt at tackling this problem was by Krzanowski (1979), who showed that the eigenvectors $\mathbf{b}_i$ (i = 1, ..., k) of $\mathbf{H} = \sum_{j=1}^{g} \mathbf{L}_j \mathbf{L}_j^T$ give the successive components that "agree most closely" with all g sets of PCs. The corresponding eigenvalues are $\lambda_i = \sum_{t=1}^{g} \cos^2 \delta_{ti}$ where $\delta_{ti} = \cos^{-1}\left(\sqrt{\mathbf{b}_i^T \mathbf{L}_t \mathbf{L}_t^T \mathbf{b}_i}\right)$ is the discrepancy between $\mathbf{b}_i$ and $\mathbf{L}_t$. Some limited simulation studies were conducted by Krzanowski (1982) into the null distribution of critical angles produced by the technique, but essentially it has been used descriptively rather than inferentially in applications.

By contrast, Keramidas *et al* (1987) proposed a more inferential method. They first advocated the identification of "well determined" components across all groups, followed by graphical comparison of the latter between the g groups. This comparison could be effected by Gamma probability plots of squared Euclidean distances between each group component and either a pre-specified or a "typical" component, and an extension to comparison of subspaces was outlined. A number of examples illustrated the utility of the technique. However, one evident drawback is that a large number of groups is a necessary condition of successful application of the probability plots (and all the examples had such large numbers). If the number of groups is relatively small, the technique described in section 5.2 below may be more appropriate.

4. Description

Here we consider methods that have been proposed for projecting the data into low-dimensional subspaces, subject to the constraints that the subspaces are defined by orthogonal components and the projection is such as to highlight differences between groups.

Since PCA produces orthogonal components ordered by sample variance, one obvious possibility is to use re-ordered principal components. Instead of ordering them by $d_i = \mathbf{l}_i^T \mathbf{S} \mathbf{l}_i$, order them by

$$e_i = \frac{\mathbf{l}_i^T \mathbf{B} \mathbf{l}_i}{\mathbf{l}_i^T \mathbf{W} \mathbf{l}_i} \text{ or equivalently } \frac{d_i}{\mathbf{l}_i^T \mathbf{W} \mathbf{l}_i}$$

This uses ordinary PCA but arranges the components by the CVA criterion, and was proposed by Krzanowski (1992) following Chang (1983) who suggested Mahalanobis distance as the ordering criterion for the two-group case.

Of course the problem is that the wrong criterion is being optimised, so while the first k re-ordered principal components may be better at representing group differences than the original top k components, they do not necessarily have the right optimality properties. As an approximation to the desired

optimality Yendle & MacFie (1989) proposed their method of discriminant principal components: rescale each variable with the inverse of its pooled within-groups standard deviation and then do PCA of the between-groups covariance matrix from the scaled data. While this goes part of the way, it still ignores the within-groups covariances so will not necessarily produce the best representation. Another possibility, again sub-optimal, is the use of partial least squares (Garthwaite, 1994; Martens & Martens, 2001) which can be implemented in the present context by setting group dummy variables Y_i followed by the multivariate regression formulation of the technique. However, while this method incorporates group information into an orthogonal-component derivation, it does not necessarily do so in an optimal fashion.

The optimal approach is given in the stream of publications dealing with orthogonal canonical variates, i.e. with orthogonal components that maximise between-to-within group variance. This stream was initiated by Foley & Sammon (1975) for the two-group special case, but was taken up for the general case by Okada & Tomita (1985). Their approach was a sequential process: given the first r components $\mathbf{l}_1$, ..., $\mathbf{l}_r$, find a basis $\mathbf{b}_1$, ..., $\mathbf{b}_{p-r}$ for their orthocomplement (e.g. Gram-Schmidt); setting this basis as columns of $\mathbf{P}$, form $\mathbf{W}_r = \mathbf{P}^T\mathbf{W}\mathbf{P}$ and $\mathbf{B}_r = \mathbf{P}^T\mathbf{B}\mathbf{P}$; the iterations $\mathbf{c}_{k+1} = \mathbf{W}_r^{-1}\mathbf{B}_r\mathbf{c}_k$ will then converge to $\mathbf{l}_{r+1} = \mathbf{c}$.

Duchene & Leclercq (1988) organised this process as a matrix eigenvector extraction, and generalised it to allow any mixture of CVA & PCA criteria, while Hamamoto *et al* (1991; 1993) improved the algorithm to optimise discriminant feature selection and conducted a comparison of discriminant performances. However, chemometric and other applications in which p is much larger than n are not amenable to these methods because of problems with inverting $\mathbf{W}$. Krzanowski (1995) proposed handling the case of singular $\mathbf{W}$ by initial projection into a subspace, optimisation in this subspace, and then back-projection into the full space. Kiers (1997) provided improved optimisation algorithms which permit simultaneous rather than sequential derivation of components, while Kiers & Krzanowski (2000) suggested an extension to obtain projections accentuating *extreme* groups. This involves defining $\mathbf{B}_i$ to be the between-group matrix for the two-group case of group i and the union of all other groups, and then maximising over $\mathbf{l}$ and i either $\mathbf{l}^T\mathbf{B}_i\mathbf{l}$ or $\dfrac{\mathbf{l}_i^T\mathbf{B}\mathbf{l}_i}{\mathbf{l}_i^T\mathbf{W}\mathbf{l}_i}$ subject to orthogonal vectors $\mathbf{l}$.

To illustrate some of the above techniques, consider a set of data in which 24 food samples were rated on 31 sensory characters by trained judges (Ian Wilson, personal communication). The 24 samples were divided into 4 groups: group 1 comprised 6 reformed meats, group 2 contained 5 types of sausage, group 3 had 7 whole meats and group 4 was made up of 6 varieties of beefburger. Note that here p is greater than $n - 4$, so the data produced a singular $\mathbf{W}$ and orthogonal canonical variates had to be derived using the method of Krzanowski (1995).

Figure 1 shows the 24 samples plotted against the first two (re-ordered) principal components. Only group 3 seems to be reasonably compact, while the other groups are all intermingled. Figure 2 shows the 24 samples plotted against the first two partial least squares components. The samples in group 3 have become more tightly clustered, but the other three groups are still quite intermingled. Figure 3 shows the 24 samples plotted against the first two orthogonal canonical variates. Groups 2 and 3 are now seen to be very tightly clustered and well separated from the other samples, while groups 1 and 4 (reformed meats and beefburgers!) are relatively indistinguishable.

Figure 1.

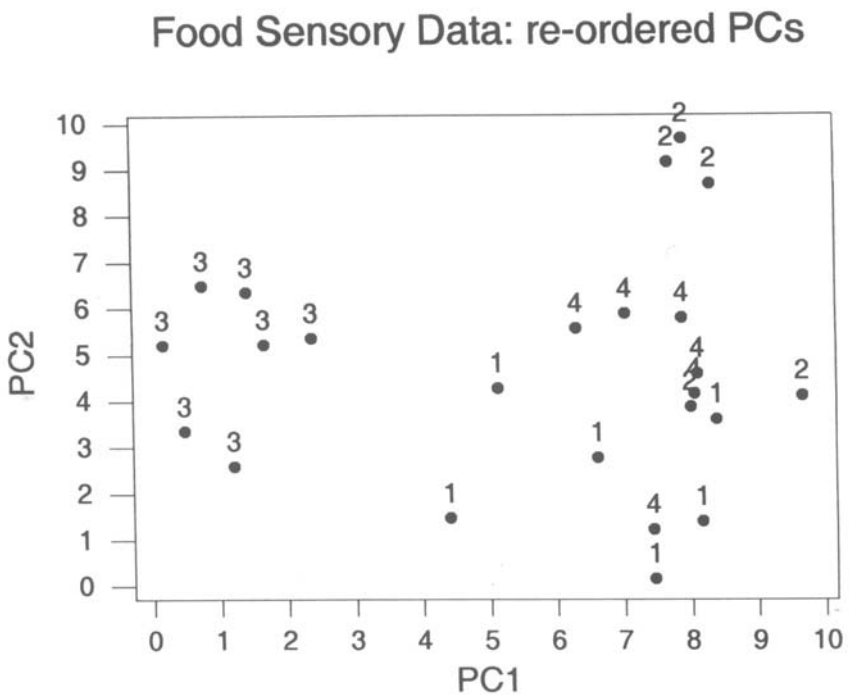


Figure 2.

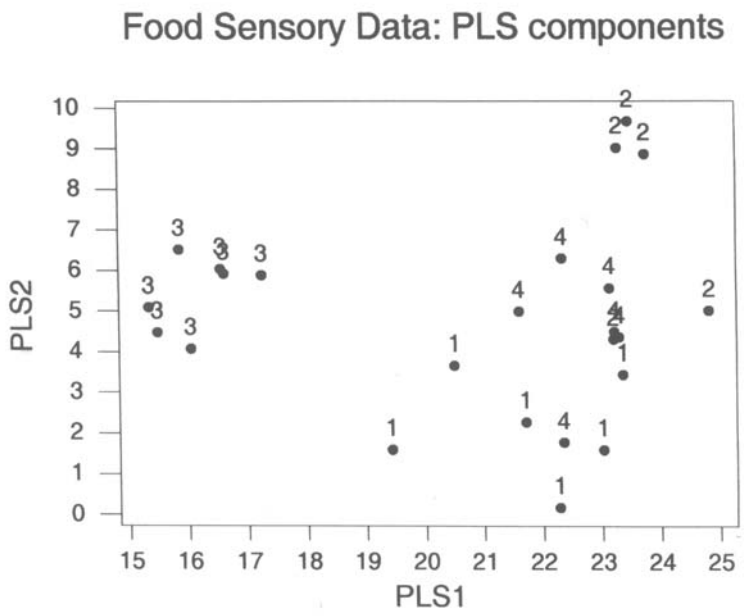
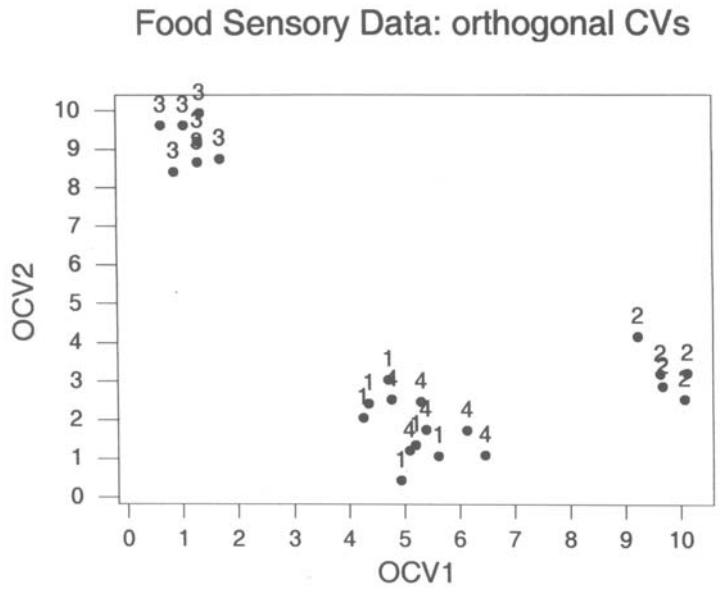

Food Sensory Data: PLS components
PLS2
PLS1
0
1
2
3
4
5
6
7
8
9
10
15
16
17
18
19
20
21
22
23
24
25

Figure 3.


Food Sensory Data: orthogonal CVs
OCV2
OCV1
0
1
2
3
4
5
6
7
8
9
10

4.1 Application: projection pursuit clustering

Bock (1986) and Heiser & Groenen (1997) have established connections between multidimensional scaling (MDS) and cluster analysis, while Bock (1987) has proposed the following clustering method in MDS by projection pursuit. Let I(A,X) be the projection index, as a function of the *d*-dimensional projection A and the clustering X. The steps of the algorithm are:

1. Produce an initial clustering of points.
2. Optimise I over A for fixed X.
3. Optimise I over X for fixed A.
4. Return to 2 until there is no further improvement in I.

Bock effectively used I = trace($\mathbf{L}^T\mathbf{B}_c\mathbf{L}$), $\mathbf{L}^T\mathbf{L} = \mathbf{I}$, where the subscript *c* denotes calculation with regard to current clustering X. Bolton & Krzanowski (2003) investigate several indices based on orthogonal canonical variates, including:

$$I_1 = \sum_{i=1}^{d} \frac{\mathbf{l}_i^T\mathbf{B}_c\mathbf{l}_i}{\mathbf{l}_i^T\mathbf{W}_c\mathbf{l}_i} \;\&\; \mathbf{l}_i^T\mathbf{l}_i = \delta_{ij}$$

$$I_2 = \frac{\sum \mathbf{l}_i^T\mathbf{B}_c\mathbf{l}_i}{\sum \mathbf{l}_i^T\mathbf{W}_c\mathbf{l}_i} \;\&\; \mathbf{l}_i^T\mathbf{l}_i = \delta_{ij}$$

Several examples show that these indices perform well in separating groups that are known to be present. Note however that I and I_2 coincide when the data are initially sphered.

5. Modelling

The first attempt at modelling principal components in the presence of groups was by Flury (1984), who proposed the common principal component (CPC) model: $\mathbf{X}$ has a $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ distribution in population π_i (i = 1, ..., g) where $\boldsymbol{\Sigma}_i = \boldsymbol{\Lambda}\boldsymbol{\Delta}_i\boldsymbol{\Lambda}^T$ and $\boldsymbol{\Lambda}\boldsymbol{\Delta}_i\boldsymbol{\Lambda}^T = \mathbf{I}$. Here $\boldsymbol{\Lambda}$ contains a set of eigenvectors common to all populations, while the $\boldsymbol{\Delta}_i$ are population-specific diagonal matrices of eigenvalues.

Given samples of sizes n_i from populations π_i, with sample means $\bar{\mathbf{x}}_i$ and covariance matrices $\mathbf{S}_i$, the first step in an analysis is to obtain estimates $\mathbf{L}$, $\mathbf{D}_i$ of $\boldsymbol{\Lambda}$, $\boldsymbol{\Delta}_i$ (i = 1, ..., g). There are two possible approaches:

- Maximum likelihood [F-G algorithm; Flury & Constantine (1985)], assuming normality of populations. Here $\mathbf{L}$ is chosen to minimise $\prod_{i=1}^{g}\left[\frac{|\mathrm{diag}\mathbf{L}^T\mathbf{S}_i\mathbf{L}|}{|\mathbf{S}_i|}\right]^{n_i}$ and then $\mathbf{D}_i$ is set equal to $\mathbf{L}^T\mathbf{S}_i\mathbf{L}$.

- Least squares [Clarkson (1988)], if no distributional assumptions are made. Here **L** is chosen to minimise $\sum_{i=1}^{g} n_i \left[\sum\sum_{j\neq 1}\left(\mathbf{L}^T\mathbf{S}_i\mathbf{L}\right)\right]$ and then $\mathbf{D}_i$ is set equal to $\mathbf{L}^T\mathbf{S}_i\mathbf{L}$..

With either method, we would then use the transformation $\mathbf{Y} = \mathbf{L}^T\mathbf{X}$ where $\mathbf{L}^T\mathbf{S}_i\mathbf{L} \approx \text{diag}(\mathbf{D}_i)$ and $\mathbf{L}^T\mathbf{L} = \mathbf{I}$. Adequacy of fit of this model can be checked either informally or formally:

- Informal: inspect the correlation matrices $\mathbf{R}_i = \mathbf{E}_i^{-1/2}\mathbf{L}^T\mathbf{S}_i\mathbf{L}\mathbf{E}_i^{-1/2}$ where $\mathbf{E}_i = \text{diag}\mathbf{D}_i$. All off-diagonal elements should be "close to" zero if the model is adequate.

  Formal: compute the likelihood-ratio test statistic $T = \sum_{i=1}^{g} n_i \ln\left|\mathbf{S}_i^{-1}\left(\mathbf{L}^T\mathbf{D}_i\mathbf{L}\right)\right|$. This has an asymptotic $\chi^2_{p(p-1)(g-1)/2}$ distribution if the model is adequate.

Flury (1988) gives many substantive applications (anthropometry, biometry, finance etc); see also Boik (2002). Flury (1986) and Yuan & Bentler (1994) provide some asymptotic distribution theory, while Reyment (1997) applies the model to logratios in geological compositional data.

An alternative, more general, formulation of this model has been given by ten Berge (1986) and Kiers & ten Berge (1994). Let:

Z be the ($n \times p$) data matrix ("standard scores"),

L be the ($p \times k$) matrix of component weights, and

P be the ($p \times k$) "pattern" matrix used to reconstruct **Z** from columns of the component scores matrix **ZL**. Then PCA is given by the **P** and **L** which minimise

$$f(\mathbf{L},\mathbf{P}) = \left\|\mathbf{Z} - \mathbf{Z}\mathbf{L}\mathbf{P}^T\right\|^2$$

and for g groups we can generalise the optimality criterion to

$$f(\mathbf{L}_1,...,\mathbf{L}_g;\mathbf{P}_1,...,\mathbf{P}_g) = \sum_{i=1}^{g}\left\|\mathbf{Z}_i - \mathbf{Z}_i\mathbf{L}_i\mathbf{P}_i^T\right\|^2$$

Optimisation of this criterion is effected by alternating least squares, and the solutions can be subject to constraints. Four possibilities exist:

1. (Unconstrained) $f^*_{sep} = \min f$ over separate $\mathbf{L}_i$, $\mathbf{Z}_i$ $\forall_i$;
2. (SCA–L) $f^*_{\mathbf{L}} = \min f$ subject to $\mathbf{L}_i = \mathbf{L}$ $\forall_i$;
3. (SCA–S) $f^*_{\mathbf{S}} = \min f$ subject to $\mathbf{S}_i = \left(\mathbf{Z}_i^T\mathbf{Z}_i\right)\mathbf{L}_i \forall_i$;
4. (SCA–P) $f^*_{\mathbf{P}} = \min f$ subject to $\mathbf{P}_i = \mathbf{P}$ $\forall_i$.

The Flury CPC model is equivalent to SCA–L.

5.1. Applications of the CPC model

5.1.1. (i) Discriminant analysis

Standard (two-group) discriminant analysis theory (Krzanowski, 2000) says that we should allocate **x** to one of $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ according as

$$Q(\mathbf{x}) = -\frac{1}{2}\mathbf{x}^T\left(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1}\right)\mathbf{x} + \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2$$

is ≥ or < a constant (which depends on the $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$.

Given training samples from each population, we estimate $\boldsymbol{\mu}_i$ by the sample means. Estimation of $\boldsymbol{\Sigma}_i$ can take various routes:

1. (Unconstrained) $\hat{\boldsymbol{\Sigma}}_i = \mathbf{S}_i$ for i = 1, 2 (yielding the usual quadratic discriminant function).
2. (CPC) $\hat{\boldsymbol{\Sigma}}_i = \mathbf{L}\mathbf{D}_i\mathbf{L}^T$ for i = 1, 2.
3. (Proportionality) Assuming $\boldsymbol{\Sigma}_2 = \rho\boldsymbol{\Sigma}_1$ implies that $\boldsymbol{\Delta}_2 = \rho\boldsymbol{\Delta}_1$ and leads to constrained CPC (Flury, 1988).
4. (Equality) $\hat{\boldsymbol{\Sigma}}_i = \mathbf{W}$ for both i (yielding the usual linear discriminant function).

It has been shown that parsimonious models are often best, even when they are not "true". In particular, model 3 works well. [Flury & Schmid (1992); Flury *et al* (1994)]

5.1.2. (ii) Generalising CVA

Canonical variate analysis requires the assumption of equal population dispersion matrices $\boldsymbol{\Sigma}_i$. The technique can be generalised by using the CPC model, and exploiting the equivalence between CVA and MDS based on Mahalanobis distance (Krzanowski, 2000, p302). For this we need to derive a distance between populations that follow the CPC model. Several such distances are available:

- Hellinger/Matusita distance:

  The affinity ρ_{ij} between $f_i(\mathbf{x})$ and $f_j(\mathbf{x})$ is given by $\int\sqrt{f_i(\mathbf{x})f_j(\mathbf{x})}d\mathbf{x}$, and from this we can then define the distance δ_{ij} between the distributions by any of $\sqrt{2(1-\rho_{ij})}$, $\cos^{-1}\rho_{ij}$ or $-\ln\rho_{ij}$ (Gower, 1967). For the CPC model we have (Krzanowski, 1990):

$$\rho_{ij} = 2^{p/2} \frac{\left|\mathbf{D}_i \mathbf{D}_j\right|^{1/4}}{\left|\mathbf{D}_i + \mathbf{D}_j\right|^{1/2}} \exp\left[-\frac{1}{4}\left(\overline{\mathbf{y}}_i - \overline{\mathbf{y}}_j\right)^T \left(\mathbf{D}_i + \mathbf{D}_j\right)\left(\overline{\mathbf{y}}_i - \overline{\mathbf{y}}_j\right) \right]$$

where $\overline{\mathbf{y}}_i = \mathbf{L}^T \overline{\mathbf{x}}_i$. An application of this distance to classification of orchids has been given by Tyteca & Dufrêne (1993).

- Rao distance:

  The distance between $f(\mathbf{x}|\boldsymbol{\theta}_1)$ and $f(\mathbf{x}|\boldsymbol{\theta}_2)$ is defined to be the geodesic distance $s(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ with respect to the metric

$$ds^2 \sum\sum g_{ij}(\boldsymbol{\theta}) d\theta_i d\theta_j$$

  in the parameter space $\boldsymbol{\Theta}$, where

$$g_{ij}(\boldsymbol{\theta}) = E\left[\frac{\partial}{\partial \theta_i} \ln f(\mathbf{x} \mid \boldsymbol{\theta}) \frac{\partial}{\partial \theta_j} \ln f(\mathbf{x} \mid \boldsymbol{\theta}) \right]$$

  For the CPC model we find (Krzanowski, 1996):

$$\delta_{ij} = \sqrt{\delta_{ij1}^2 + \delta_{ij2}^2 + \ldots + \delta_{ijp}^2}$$

  where

$$\delta_{ijs} = \left(\sqrt{2}\right) \cosh^{-1}\left[\frac{\left(\overline{\mathrm{y}}_{\mathrm{is}} - \overline{\mathrm{y}}_{\mathrm{js}}\right)^2 + 2d_{is} + 2d_{js}}{4\sqrt{d_{is} d_{js}}} \right]$$

Krzanowski (1990) has demonstrated the difference that is obtained between use of standard canonical variate analysis and MDS on Matusita distances for a data set (Venezuelan students) in which the hypothesis of equal population dispersion matrices is rejected while the common principal component model provides an adequate fit. Applying MDS with Rao's distance to this data set produces almost exactly the same picture as MDS with Matusita's distance.

5.1.3 (iii) Spatial analysis

Suppose that $\mathbf{X}(\mathbf{s}) = [X_1(\mathbf{s}), \ldots, X_p(\mathbf{s})]^T$ is a p-valued 2nd-order stationary point process observed at locations $\mathbf{s}$ with:

- pointwise dispersion matrix $\Sigma = \mathrm{cov}(\mathbf{X}(\mathbf{s}))$;
- lag δ dispersion matrix $\boldsymbol{\Omega}(\boldsymbol{\delta}) = \mathrm{cov}(\mathbf{X}(\mathbf{s}), \mathbf{X}(\mathbf{s} + \boldsymbol{\delta}))$;
- semi-variance matrix $\boldsymbol{\Gamma}(\boldsymbol{\delta}) = \frac{1}{2}\mathrm{cov}\left[\mathbf{X}(\mathbf{s} + \boldsymbol{\delta}) - \mathbf{X}(\boldsymbol{\delta})\right] = \boldsymbol{\Sigma} - \boldsymbol{\Omega}(\boldsymbol{\delta})$.

We look for components $Y_i = \mathbf{l}_i^T \mathbf{X}$ which have "good" properties, but how do we define "good"?

Switzer (1985) seeks the Y_i that successively maximise autocorrelations $r_i(\boldsymbol{\delta})$ at a given lag $\boldsymbol{\delta}$, and shows that the coefficients $\mathbf{l}_i$ are solutions of $(\boldsymbol{\Omega}(\boldsymbol{\delta}) - r_i\boldsymbol{\Sigma})\mathbf{l}_i = \mathbf{0}$. The resulting Y_i are mutually uncorrelated, both at a point and at lag $\boldsymbol{\delta}$.

Hence the Y_i will be mutually uncorrelated at a set of lags $\delta_1, ..., \delta_g$ if (Bailey and Krzanowski, 2000):

$(\boldsymbol{\Omega}(\boldsymbol{\delta}_k) - r_i\boldsymbol{\Sigma})\mathbf{l}_i = \mathbf{0}$ for $k = 1, ..., g$,

i.e. $(\boldsymbol{\Gamma}(\boldsymbol{\delta}_k) - (1 - r_i)\boldsymbol{\Sigma})\mathbf{l}_i = \mathbf{0}$ for $k = 1, ..., g$,

i.e. $(\mathbf{Q}(\boldsymbol{\delta}_k) - (1 - r_i)\mathbf{I})\mathbf{m}_i = \mathbf{0}$ for $k = 1, ..., g$,

where $\boldsymbol{\Sigma} = \mathbf{U}\mathbf{F}\mathbf{U}^T$, $\mathbf{Q}(\boldsymbol{\delta}) = \mathbf{F}^{-1/2}\mathbf{U}^T\boldsymbol{\Gamma}(\boldsymbol{\delta})\mathbf{U}\mathbf{F}^{-1/2}$, and $\mathbf{l} = \mathbf{U}\mathbf{m}$.

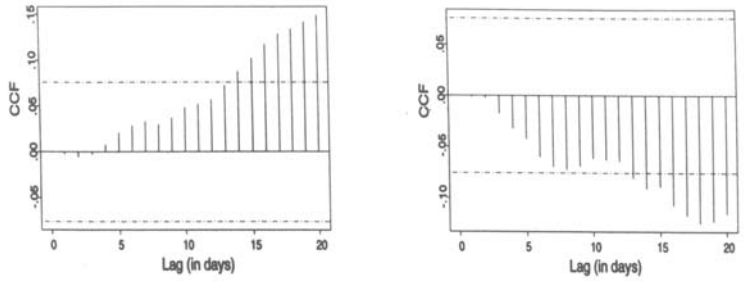
An (approximate) solution is thus given by a CPC analysis of $\mathbf{Q}(\boldsymbol{\delta}_k)$, $k = 1, ..., g$. Further spatial justification of this model can be provided in terms of "co-regionalisation" and "intrinsic correlation" models, while practical questions include choice of lags $\boldsymbol{\delta}_k$ and estimation of matrices. See Bailey and Krzanowski (2000) for details.

5.1.4. (iv) Process control

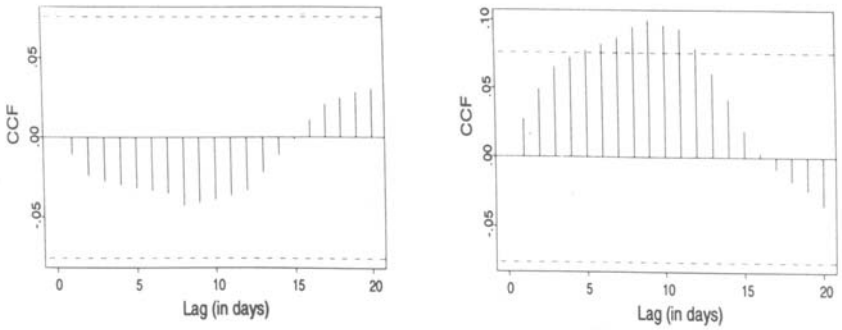
Multivariate control charts based on Hotelling's T^2 are well established, but don't identify *why* a process has gone out of control. One possible remedy might be to conduct a PCA of process data, followed by application of univariate control charts to each component separately, adjusting for autocorrelation and desired overall average run length (ARL). However, it can be shown that such adjustments are incorrect because, although the principal components are "instantaneously" uncorrelated [i.e. $\text{corr}(Y_i(t), Y_j(t)) = 0$ for $i \neq j$], they are not "temporally" uncorrelated [i.e. $\text{corr}(Y_i(t), Y_j(t - \delta)) \neq 0$ for $i \neq j$]. This feature is illustrated in the top part of Figure 4, which shows the cross-correlograms between two pairs of principal components for a six-variable set of process data collected daily over a period of nearly two years from the fluidised catalytic cracking unit of an industrial plant (Dr Phil Jonathan and Julie Badcock, personal communications). The cross-correlations build steadily with increasing lag, and exceed the critical null hypothesis value for all lags beyond a certain point.

Defining $\boldsymbol{\Sigma}$, $\boldsymbol{\Omega}(\delta)$ and $\boldsymbol{\Gamma}(\delta)$ as for the spatial case but now for (1-dim) *temporal* data, better components should be given from a CPC analysis over a range of lags δ_k. This can be seen in the bottom part of Figure 4, where the cross-correlations between two corresponding pairs of CPC components for the fluidised catalytic cracking data stay within the null hypothesis limits everywhere except for a small set of lags. A fuller investigation of performance of this approach is currently under way.

Figure 4.



Cross-correlograms between Principal Components 1 & 2 and 2 & 3



Cross-correlograms between 'temporally uncorrelated' components 1 & 2 and 2 & 3

5.2. Extension to subspace modelling

Let us again suppose that $\mathbf{X}$ has a $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ distribution in population π_i for $i = 1, ..., g$. Then Flury (1986) extended the CPC model to the partial common principal component model by assuming that only q of the components are common to each population, while the remaining $p - q$ are "population specific", i.e. $\boldsymbol{\Sigma}_i = \left(\boldsymbol{\Lambda}^{(c)} \vdots \boldsymbol{\Lambda}_i^{(s)}\right)\boldsymbol{\Delta}_i\left(\boldsymbol{\Lambda}^{(c)} \vdots \boldsymbol{\Lambda}_i^{(s)}\right)^T$ where $\boldsymbol{\Lambda}^{(c)}$ has q columns and the $\boldsymbol{\Lambda}_i^{(s)}$ each have $p - q$.

Approximate maximum likelihood estimates are available from adaptation of the F-G algorithm, and the likelihood-ratio test of this model has the same form as for the full CPC model but degrees of freedom depend on number of common components. A small modification enables a "common subspace" to be identified.

However, the basic problem with Flury's CPC subspaces is that they involve *unordered* eigenvectors which have an arbitrary matching from group to group, rather than forming subspaces based on the top components of each group. Schott (1991) therefore introduced a different model, which assumes that the *first m* pcs of each population lie in the same subspace. This model can be tested by testing whether $\mathbf{P}_{im} = \boldsymbol{\lambda}_{i1}\boldsymbol{\lambda}_{i1}^T + \ldots + \boldsymbol{\lambda}_{im}\boldsymbol{\lambda}_{im}^T$ is the same for $i = 1, \ldots, g$. To do this we need the following calculations:

Form the spectral decomposition of $(n_1\mathbf{S}_1 + \ldots + n_g\mathbf{S}_g)$, partitioning the matrix of eigenvectors so that the first m columns are in $\mathbf{K}_1$ and the remaining columns are in $\mathbf{K}_2$:

i.e. $(n_1\mathbf{S}_1 + \ldots + n_g\mathbf{S}_g) = (\mathbf{K}_1 \vdots \mathbf{K}_2)\mathrm{E}(\mathbf{K}_1 \vdots \mathbf{K}_2)^T$.

Then obtain the sample equivalent of $\mathbf{P}_{im}$ by replacing the $\boldsymbol{\lambda}_{im}$ by their sample equivalents $\mathbf{l}_{im}$:

i.e. $\hat{\mathbf{P}}_{im} = \mathbf{l}_{i1}\mathbf{l}_{i1}^T + \ldots + \mathbf{l}_{im}\mathbf{l}_{im}^T$.

Finally form $\mathbf{t}_i = \mathrm{vec}\left(\mathbf{K}_1^T\hat{\mathbf{P}}_{im}\mathbf{K}_2\right)$; $\bar{\mathbf{t}} = \left(\sum n_i\boldsymbol{\Phi}_i^{-1}\right)^{-1}\left(\sum n_i\boldsymbol{\Phi}_i^{-1}\mathbf{t}_i\right)$; and

$$\boldsymbol{\Phi}_i = \sum_{j=1}^{m}\sum_{s=1}^{p-m}\frac{a_{ij}b_{is}}{\left(a_{ij}-b_{is}\right)^2}\mathbf{q}_{is}\mathbf{q}_{is}^T \otimes \mathbf{p}_{ij}\mathbf{p}_{ij}^T \quad \text{where}$$

a_{ij}, $\mathbf{p}_{ij}$ is the jth eigenvalue/vector of $\mathbf{K}_1\mathbf{S}_i\mathbf{K}_1$, and

b_{ij}, $\mathbf{q}_{ij}$ the jth eigenvalue/vector of $\mathbf{K}_2\mathbf{S}_i\mathbf{K}_2$.

Then the test statistic is $T_{gm} = \sum_{i=1}^{g} n_i\left(\mathbf{t}_i - \bar{\mathbf{t}}\right)^T\boldsymbol{\Phi}_i^{-1}\left(\mathbf{t}_i - \bar{\mathbf{t}}\right)$, and under the hypothesis of common subspaces, T_{gm} has an asymptotic $\chi^2_{(g-1)m(p-m)}$ distribution. Schott(1991) considers also the correlation matrix case and robust covariance estimation, as well as checking on the test performance by simulation.

Schott (1999) extends the results to a test of the hypothesis that the first m pcs in each population lie in a subspace of dimension r for $m < r < t$, where $t = \min(gm, p)$. The case $r = m + 1$ is of particular interest ("almost coincident subspaces"). A sequential procedure is outlined, and examples are given.

5.3. Model generalisations

There have been two noteworthy generalisations of the CPC model in recent years.

i. Banfield and Raftery (1993) introduced the formulation

$$\boldsymbol{\Sigma}_k = a_k \mathbf{L}_k \mathbf{D}_k \mathbf{L}_k^T,$$

where any of a_k, $\mathbf{L}_k$ and $\mathbf{D}_k$ are allowed to be either group-specific or common across groups. Taking all possible combinations, and including the cases $\mathbf{L}_k = \mathbf{I}$ and $\mathbf{L}_k \mathbf{D}_k \mathbf{L}_k^T = \mathbf{I}$, leads to 14 different models. Banfield & Raftery (1993) have used this formulation in normal mixture modelling for cluster analysis, Bensmail & Celeux (1996) for regularization in discriminant analysis, and Bensmail & Bozdogan (2002) for kernel density modelling in nonparametric mixture cluster analysis.

ii. Boik (2002) has provided a more comprehensive spectral decomposition: $\boldsymbol{\Sigma}_k = \mathbf{L}_k \mathbf{D}_k \mathbf{L}_k^T$ where $\mathbf{L}_k = \mathbf{L}_0 \boldsymbol{\Psi}_k$. and $\boldsymbol{\Psi}_k = \otimes_{j=1}^{d_k} \boldsymbol{\Psi}_{t_{kj}}$, for

$\mathbf{L}_0$ ($p \times p$) orthogonal and $\boldsymbol{\Psi}_{t_{kj}}$ ($r_{ij} \times r_{ij}$) orthogonal.

This is a very flexible model, allowing shared eigenspaces without coincident eigenvalues, features common within subsets of groups, and partial commonality within selected groups. Boik (2002) demonstrates its use on a range of examples.

6. Comments

We have traced recent developments of orthogonal components in the presence of group structure under the three headings of interpretation, description and modelling. Most of the work under the first and third of these headings has appeared in the traditional statistical literature and would be viewed as mainstream statistical development by practitioners. By contrast, the work on description has predominantly appeared in journals devoted to pattern recognition, computing or chemometrics, so is perhaps less familiar to statisticians. It is not without various points of debate. For example, some would argue that it is better to focus on uncorrelated scores, as in ordinary canonical variates, rather than on uncorrelated derived variates as in the orthogonal version. However, it is not the aim of this article to enter into such debates, but simply to bring the full range of available computational tools to the attention of the statistician. Orthogonal canonical variates should be placed alongside the ordinary variety as an extra weapon in the analytical armoury, for use whenever circumstances warrant.

Notwithstanding this aspect, it is clear that the most vigorous development in this area has been in terms of common principal component and extended models for grouped multivariate data. This development has evidently not yet fully run its course, and we can look forward to further advances in the future.

REFERENCES

ANDERSON, T. W. (1963). Asymptotic theory for principal component analysis. *Annals of Mathematical Statistics*, 34, 122-48.

ASHTON, E. H., HEALY, M. J. R. and LIPTON, S. (1957). The descriptive use of discriminant functions in Physical Anthropology. *Proceedings of the Royal Society, Series B*, 146, 552-572.

BAILEY, T. C. and KRZANOWSKI, W. J. (2000). Extensions to spatial factor methods with an illustration in Geochemistry. *Mathematical Geology*, 32, 657-82.

BANFIELD, J. D. and RAFTERY, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49, 803-21.

BENSMAIL, H. and BOZDOGAN, H. (2002). E-M algorithm for multivariate kernel mixture-model cluster analysis for mixed data. Presented at IFCS2002 meeting, Krakow, Poland.

BENSMAIL, H. and CELEUX, G. (1996). Regularized Gaussian discriminant analysis through eigenvalue decomposition. *Journal of the American Statistical Association*, 91, 1743-8.

BOCK, H-H. (1986). Multidimensional Scaling in the Framework of Cluster Analysis. In: *Studien Zur Klassifikation*, P. Degens, H-J. Hermes and O. Opitz (eds), pp247-58, Frankfurt: INDEKS-Verlag.

BOCK, H-H. (1987). On the Interface Between Cluster Analysis, Principal Component Analysis, and Multidimensional Scaling. In: *Multivariate Statistical Modeling and Data Analysis*, H. Bozdogan and A. K. Gupta (eds), pp17-34, Dordrecht: Riedel.

BOIK, R. J. (2002). Spectral models for covariance matrices. *Biometrika*, 89, 159-82.

BOLTON, R. J. and KRZANOWSKi, W. J. (2003). Projection pursuit clustering for exploratory data analysis. *Journal of Computational and Graphical Statistics, to appear*.

BRYAN, J. G. (1951). The generalized discriminant function: mathematical foundation and computational routine. Harvard Educational Review, 21, 90-5.

CAMPBELL, N. A. (1984). Canonical variate analysis – a general model formulation. *Australian Journal of Statistics*, 26, 86-96.

CHANG, W. C. (1983). On using principal components before separating a mixture of multivariate normal distributions. *Applied Statistics*, 32, 276-86.

CLARKSON, D. B. (1988). A least squares version of Algorithm AS211: The F-G diagonalisation algorithm. Algorithm ASR74, *Applied Statistics*, 37, 317-21.

DUCHENE, J. and LECLERCQ, S. (1988). An optimal transformation for discriminant analysis and principal component analysis. *IEEE Transactions on PAMI*, 10, 978-83.

FLURY, B. (1984). Common principal components in *k* groups. *Journal of the American Statistical Association*, 79, 892-8.

FLURY, B. (1986). Asymptotic theory for common principal component analysis. *Annals of Statistics*, 14, 418-30.

FLURY, B. (1987). Two generalizations of the common principal component model. *Biometrika*, 74, 59-69.

FLURY, B. (1988). *Common Principal Components and Related Models*. New York: Wiley.

FLURY, B. and CONSTANTINE, G. (1985). The F-G diagonalisation algorithm. Algorithm AS211, *Applied Statistics*, 34, 177-83.

FLURY, B. and SCHMID, M. J. (1992). Quadratic discriminant functions with constraints on the covariance matrices: some asymptotic results. *Jour. Multivariate Analysis*, 40, 244-61.

FLURY, B., SCHMID, M. J. and NARAYANAN, A. (1994). Error rates in quadratic discrimination with constraints on the data matrices. *Journal of Classification*, 11, 101-20.

FOLEY, D. H. and SAMMON, J. W. (1975). An optimal set of discriminant vectors. *IEEE Transactions on Computers*, 24, 281-9.

GARTHWAITE, P. H. (1994). An interpretation of partial least squares. *Journal of the American Statistical Association*, 89, 122-7.

GOWER, J. C. (1967). Multivariate analysis and multidimensional geometry. *The Statistician*, 17, 13-28.

HAMAMOTO, Y., MATSUURA, Y., KANAOKA, T. and TOMITA, S. (1991). Note on the orthonormal discriminant vector method for feature extraction. *Patt. Recog.*, 24, 681-4.

HAMAMOTO, Y., KANAOKA, T. and TOMITA, S. (1993). On a theoretical comparison between the orthonormal discriminant vector method and discriminant analysis. *Pattern Recognition*, 26, 1863-7.

HEISEr, W. J. and GROENEN, P. J. F. (1997). Cluster differences scaling with a within-clusters loss component and a fuzzy successive approximation strategy to avoid local minima. *Psychometrika*, 62, 63-83.

HOTELLING, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24, 417-41.

JACKSON, J. E. (1991). *A User's Guide to Principal Components*. Wiley.

JOLLIFFE, I. T. (1986). *Principal Component Analysis*. Springer-Verlag.

KERAMIDAS, E. M., DEVLIN, S. J. and GNANADESIKAN, R. (1987). A graphical procedure for comparing the principal components of several covariance matrices. *Communications in Statistics – Simulation*, 16, 161-91.

KIERS, H. A. L. (1997). Discrimination by means of components that are orthogonal in the data space. *Journal of Chemometrics*, 11, 533-45.

KIERS, H. A. L. and ten BERGE, J. M. F. (1994). Hierarchical relations between methods for simultaneous component analysis and a technique for rotation to simple simultaneous structure. *British Journal of Mathematical and Statistical Psychology*, 47, 109-26.

KIERS, H. A. L. and KRZANOWSKI, W. J. (2000). Projections distinguishing isolated groups in multivariate data spaces. In: *Data Analysis*, W. Gaul, O. Opitz and M. Schader (eds),Berlin: Springer.

KRZANOWSKI, W. J. (1979). Between-groups comparison of principal components. *Journal of the American Statistical Association*, 74, 703-7, (correction 76, 1022).

KRZANOWSKI, W. J. (1982). Between-groups comparison of principal components – some sampling results. *Jour. Statistical Computation and Simulation*, 15, 141-54.

KRZANOWSKI, W. J. (1990). Between-group analysis with heterogeneous covariance matrices: the common principal component model. *Journal of Classification*, 7, 81-98.

KRZANOWSKI, W. J. (1992). Ranking principal components to reflect group structure. *Journal of Chemometrics*, 6, 97-102.

KRZANOWSKI, W. J. (1995). Orthogonal canonical variates for discrimination and classification. *Journal of Chemometrics*, 9, 509-20.

KRZANOWSKI, W. J. (1996). Rao's distance between normal populations that have common principal components. *Biometrics*, 52, 1467-71.

KRZANOWSKI, W. J. (2000). *Principles of Multivariate Analysis: a User's Perspective*. Oxford University Press.

MARTENS, H. and MARTENS, M. (2001). *Multivariate Analysis of Quality*. Wiley.

OKADA, T. and TOMITA, S. (1985). An optimal orthonormal system for discriminant analysis. *Pattern Recognition*, 18, 139-144.

PEARSON, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science, Sixth Series*, 2, 559-72.

RAO, C. R. (1948). The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society, Series B*, 10, 159-203.

REYMENT, R. A. (1997). Multiple group principal component analysis. *Math Geol*, 29, 1-16.

SCHOTT, J. R. (1991). Some tests for common principal components in several groups. *Biometrika*, 78, 771-7.

SCHOTT, J. R. (1999). Partial common principal component subspaces. *Biometrika*, 86, 899-908.

SWITZER, P. (1985). Min/max autocorrelation factors for multivariate spatial imagery. *Computer Science and Statistics*, Proc. 16th symposium on the Interface, 13-16.

TYTECA, D. and DUFRÊNE, M. (1993). On the use of distances in the taxonomic study of critical plant groups – case studies of Western European Orchidaceae. *Annals of Botany*, 71, 257-77.

YENDLE, P. W. and MacFie, H. J. H. (1989). Discriminant principal component analysis. *Journal of Chemometrics*, 3, 589-600.

YUAN, K-H. and BENTLER, P. M. (1994). Test of linear trend in eigenvalues of k covariance matrices with applications in common principal components analysis. Communications in Statistics – Theory and Methods, 23, 3141-56.

STATISTICS IN TRANSITION, October 2002
Vol. 5, No. 5, pp. 779–785

SOME REMARKS ON THE TASKS OF STATISTICS ON THE VERGE OF THE TWENTY FIRST CENTURY

Czesław Domański[1]

ABSTRACT

This article tries to show the development of statistics as the self-independent branch of science. It was established not before the second quarter of the former century. It has been treated as the method of obtaining information from observed data and as the logic of making decisions in conditions of uncertainty. Statistical knowledge has become more and more precious for representatives of all occupations. Nowadays statistics is understood as a logic, by means of which we can transform data into information.

1. Introduction

Statistics as the self-independent branch of science was established not before the second quarter of the former century. It has been treated as the method of obtaining information from observed data and as the logic of making decisions in conditions of uncertainty (Rao 1994). Statistical knowledge has become more and more precious for representatives of all occupations. The fact that we can find statistics almost everywhere results from necessity of understanding statistics which is more important than understanding any other branch of science.
The etymological definition of statistics can be formulated as follows:

- It is data obtained by certain steps.

Therefore, the most important task is giving answers to the following questions:

- What do data signify?;
- How to use things that are signified by data to specific purposes?

In other words, we must know the kind and scope of information will enable us to solve concrete problem contained in data.

[1] Chair of Statistical Methods Institute of Econometrics and Statistics. Faculty of Economics and Sociology. University of Łódź, Poland.

Claude Shannon, who laid foundations for information theory gave the most logical and concise answer to the question "What is information?" The answer was: "Eliminating uncertainty".

Data create material which allows us to say how far an answer to a given question is correct or how confident about the answer we can be. Data should be transformed in order to eliminate the uncertainty. Knowledge about the size of uncertainty contained in data is the key to make the correct decision. It allows us to compare the consequences of various possibilities and to choose one that is the least harmful.

Nowadays, statistics is understood as a logic, by means of which we can transform data into information. Generally there is a necessity of developing statistics as the methodology of making decisions based on data existing under uncertainty. Significance of information is considerably larger than the experience or technical knowledge, in preparing and carrying out each investment project. We realize this more and more often.

We can assume that the modern statistics is developing as a meta-science. The future of statistics understood in such a way lies in a suitable transmission of statistical ideas to scholars of other branches of knowledge. It will depend on the way of expressing basic problems by the representatives of these branches and most of all on their awareness and understanding of statistics.

Statisticians have to live up to the organisational and conceptual expectations and challenges concentrated around the contents of their branch. Organisational problems are connected with fast conversion of satellite data, meteorological data, capital market data, genetic information and so on. Computer experts try to cope with gathering information but there is still one problem – evaluation of their quality and interpretation of this information.

One very often comes across with the statistical research which is not even consulted with an expert – statistician, not mentioning his participation in these research. There are many possibilities of obtaining wrong data. The research shows that 2/3 of errors which can appear in analytic research are connected with research design, construction of questionnaire and statistical observation (data gathering). If errors connected with the analysis and interpretation of results are noticed early enough they can be corrected. On the other hand we are not able to correct the errors which appeared while designing a project. The main result of the research process is usually a publication. Therefore scientific periodicals should do their best to eliminate research results standing below a certain level. Each statistical publication or publication based on statistical material should be reviewed by an expert – statistician.

Nowadays many papers in which statistical analysis is incorrectly used are published. We often meet the incompatible results used in analogous papers which can be caused by the statistical incompetence of authors.

2. Tasks of statistics

The development of computer science, and most of all the development of electronics and as a consequence the development of computer technology brought the following challenges :

1. The new technical possibilities of calculations, fast access to data, saving huge amount of data on small data carriers, fast copying, complex applications, linking various data bases and so on.
2. Broadening of data basis to include the administrative information. Nowadays law secures the use of administrative files for statistical purposes.
3. Creating merged specification including statistical and administrative data about people, companies, buildings and so on, within national identification number.
4. Decentralizing computer science – every researcher has a personal computer which lets him perform operations that in the past were impossible without costly equipment.
5. Unrestricted possibilities of application of mathematical, statistical, econometric and operation research methods, both classical and nonparametric ones, exact and asymptotic ones, based on theoretical and simulation models.
6. Interpretation of results and their selection and preparing to the decision making process.

More and more often statistics is used as the "source of cognition" and also as "the way of thinking". Statistical thinking needs appropriate preparation which means understanding of concepts, ability of comparing various data sets and knowledge, for example, why two sets of numbers, which refer to the same phenomenon, can show differences. The cases of using individual facts for generalizations have nothing in common with statistical thinking. Unfortunately we witness very often that convincing power of individual, even anecdotic, facts makes on audience bigger impact than precise quantitative analysis.

Obviously we must realize that very few people depend only on one source of knowledge. Relatively few researchers base their methods only on statistical thinking. Even in political and government circles in which historical role of statistics is considerable; the role of statistical thinking is not even satisfactory as we would like it to be. This situation should be changed by wider use of computer technology.

We still are at the early stage of technological revolution which greatly broadens the possibilities of applying statistical data and using them in political debates and many other fields of social activities. This revolution began when the Internet was invented. Statistics that in the past were hardly available and could be found only in publications now are easy to find in a couple of minutes. It can be copied to personal computers and, in a growing numbers of cases; it can be worked on directly.

It is hard to predict all consequences of huge access facilities to official statistical data. One of them will undoubtedly be a biased way of using data and making a lot of mistakes. Soon, it will be of great importance to help unprepared

users to perform data analysis and interpretation. It will also result in bigger pressure on improving the co-operation between data gathering and classification.

The fast development of computer technology will cause easier access to information and as a consequence create new tasks for statistics:

- adaptation of statistics to the user's needs,
- evaluation of data and data interpretation and analysis,
- punctuality of statistical research realisation using the newest methodologies and technologies.

At the beginning of the 21st century we can name three factors crucial to the role of statistics and the directions of its development:

1. The introduction of the model of open and competitive market economy in particular countries.
2. Technological revolution which has taken place in information sector and had a great influence on society transformation.
3. The creation of global "economic village". This process is necessary in market economy and it is possible just due to the technological revolution.

The factors listed above cause huge changes in political, economic and social life. They contribute to the tightening of interdependence between countries in which the local authorities act completely independently from their partners with large difficulties. At the same time we observe radical changes in economic and social life which need general change of thinking and activity style. It becomes necessary to inform, explain and educate people so they could understand the scale of changes and take actively part in them.

Statistics allows to predict directions of society evolution, define the conditions of making decisions by providing reliable and up to date information about actual situation and its changes in various sectors of the economy and social groups.

Authorities of various levels often do not realize that they need statistics to formulate, follow and define their activities. Statistical information is also necessary to estimate the size of social groups needing special care: the poor, the disabled, the old-aged.

Citizens of every country are in need of statistics providing information about their territory and their neighbours. Thanks to statistics people are able to participate actively in democratic process. On the ground of observing economic activity of a city, region or community, statistics allows to define its own economical status and help to set the future strategy.

Market economy can not provide statistical data which is necessary for its own activity, statistics is a kind of infrastructure of market economy but still data gathering is a long lasting and expensive process in which market economy objects do not want to participate. In this situation national and municipal governments are responsible for running statistical system.

Fast changes in society are in general connected with growing uncertainty. As a consequence, authorities, businessmen and citizen groups need more and

more statistical information which is hard to get and interpretation of it often demands an expert-statistician.

3. Statistical teaching programmes

Introduced tasks of statistics can be tackled mainly by appropriate teaching programmes on various teaching levels.

We can assume that statistical knowledge has become necessary both as an element of general education of all active people and professional preparation of even larger group of future specialists.

Bessaut and Mac Phearson propose statistics to be lectured in technical schools with reference to concrete applications areas which are fundamental from the point of view of prospective employers needs in the conditions of the information society.

They distinguish 5 most important blocks:

1. Mathematical statistics: data gathering, data analysis, theory (for example, elements of probability, confidence intervals, correlation and regression analysis, statistical tests).
2. Statistics not basing on mathematics: communication, designing, design management.
3. Statistical computer science: text transformation, data transformation, statistical calculations.
4. Mathematical basis: differential and integral calculus and linear algebra.
5. Statistics in different disciplines – knowledge and experience in data interpretation in context of particular applications fields.

These proposals confirm a widespread idea that statistics teaching can be effective only when ensures getting comprehensive knowledge from lots of branches of science. Statistics can not be treated as "the set of methods which are incoherent with one another". These methods should be expressed by their connection with variety of applications.

Varieties in statistical lecturers scientific profile and differentiation of students interests influence the contents and methods of teaching. We can observe particularly three kinds of scenarios:

i. Statistical concepts and methods are lectured according to standards of statistics course structure with an illustration of application of particular formulas in solving particular practical tasks.
ii. We must concentrate almost only on explaining statistical notions and the ability of using them in particular branches of science with formulas and calculations methods explained in a pretty limited way. We emphasise on explaining the aims of statistics and interpretation of gained results.
iii. We mainly stress on informative-analytic aspects concerning available computer programming used to statistical aims with a small degree of statistical theory and formulas. We concentrate students attention on the

methods of data management and on the realization of statistical procedures with the use of accessible software.

Scenarios presented above contain the most important issues. The graduated students should get to know them and in the same time they should gain appropriate abilities of statistical reasoning. The problem is usually connected with the evaluation of teaching skills in statistics which is connected with all teaching process.

4. Final remarks

Statisticians must cope with challenges connected with all the stages of statistical research and data gathering, data basis organisation, data analysis, presentation, interpretation and publication.

Conceptual problems include new ways of thinking, new methods of treatment of our art which will demand new methodologies with participation of the newest computer technology.

We must solve the problem how to explain incomprehensive difficult methodological problems formulated by mysterious mathematical symbols which are unnecessary to economist, sociologist, biologist, engineer, doctor, journalist, businessman, member of Parliament, government and municipality representative, government agency representative.

It is necessary to change our educational programs. We must construct programs which at the same time stimulate creativity and development of basic abilities of gathering and using information.

Realization of these programs should provide the basic mechanisms and habits of permanent schooling which is necessary in the 21st century. Statistics is unusual branch of science. It is comprehensive and universal. There are not any disciplines which interfere in other branches of science so widely and so often as statistics does. At the same time statistics as a branch cannot, in principle, exist by itself because its nature is to serve to other branches.

We need "statistics in society". We now see the necessity of promoting statistics among citizens. We should notice their needs, they expect it from us. We can ask: are we with them?, in other words: do representatives who serve society (i.e. civil servants, managers, consultants, experts, scientists and others) acquire statistical methods as the integral part of their own affairs or do they only dodge?

We must be adapted to new realities and be there before others "apparent statisticians".

We must work out not only new standards of our discipline but also find new connections with government, local government, industry, science world and other disciplines. We are becoming the information society, although some regress can be seen in the drop of the national production of computer programs, but without statisticians this regress can be even faster because huge data basis can become completely useless.

REFERENCES

RAO, C. R. (1994): Statystyka i prawda, PWN, Warszawa.

STATISTICS IN TRANSITION, October 2002
Vol. 5, No. 5, pp. 787–802

SOME NOTES ON THE SELECTION OF NORMALISATION OF DIAGNOSTIC VARIABLES

Aleksander Zeliaś[1]

ABSTRACT

The paper presents the discussion on the selection of normalisation of diagnostic variables. In the paper a few methods of normalisation of diagnostic variables are presented. The most important methods are: standardisation, unitarisation, quotient transformation, rank method. The statistical properties of the methods discussed by author are studied. The selection of a relevant normalisation formula is also discussed.

Ke ywords: diagnostic variables, normalisation, quotient transformation, rank method, standardisation, unitarisation.

1. Introduction

The purpose of the article is to present selected methods of normalisation of diagnostic variables. The following are the most frequently used methods of normalisation of characteristics: standardisation, unitarisation, and quotient transformation rank method.

Normalisation of variables facilitates comparative analysis of multi-characteristic economic entities with respect of levels of variables used as assessment criteria of a given complex phenomenon. The selection of a relevant normalisation formula is one of the keys to the success of constructing aggregated variables. The researcher's awareness of his goals and awareness of the advantages and disadvantages of individual transformation formulas may make easier by a profound knowledge of the phenomenon.

This article has the following layout: chapter two provides a summary of complex phenomena which is indispensable for further analysis, chapter three deals with the main selection criteria underpinning the selection of variables and chapter four presents the descriptive approach and the stochastic approach as used

[1] Cracow University of Economics, Department of Statistics, 27, Rakowicka Street, 31-510 Krakow, Poland.

in normalisation processes of diagnostic characteristics. In chapter five the author considers normalisation methods of diagnostic variables most frequently encountered in pertinent literature. The last chapter recapitulates the key issues.

2. Complex phenomena in taxonomic research

In conducting taxonomic research into qualitative phenomena one must compare multi-characteristic objects and order the set of characteristics available. The goal is to arrange the objects in a linear fashion with respect to their diagnostic variables. The object group may include countries, regions, administrative units (*gminas (communities), poviats, voivodships*), enterprises, households etc. These objects can be compared, one to another, by using an available set of diagnostic variables (characteristics) which typify these objects because of the development of a selected qualitative phenomenon which, rather than directly, can be researched and assessed indirectly (as it cannot be measured directly) e.g. economic and social development of countries, regions (usually *voivodships*), management quality, financial standing of enterprises, product analysis in the market, assessment of the effectiveness of promotional activities, level of environmental degradation, the standard of living of the population, level of poverty in society, level of agricultural production (plants and animals).

In all of the above-mentioned situations, diagnostic variables characterising a given phenomenon must be subjected to normalisation by means of one selected normalisation method. An aggregate variable becomes the main criterion of a procedure of ordering multi-characteristic objects in a linear manner.

An aggregated variable is developed in the following manner: 1) purpose and scope of research are determined; 2) data are collected; 3) diagnostic characteristics are selected; 4) variables are normalised, 5) variables are aggregated.

3. Principal criteria used to select diagnostic variables

Aggregate characterisation of economic phenomena relies on the so-called aggregated variables (aggregated, taxonomic measures of development) The replacement of a set of many diagnostic variables of a descriptive character, e.g. variables describing the population's standard of living, with an aggregated variable permits not only the reduction of their overall number but also full elimination of colinearity (their excessive correlation). In dividing objects into similar type groups high correlation of variables is undesirable.

Multi-dimensional object sets are characterised by:

$$Q = \{Q_1, ..., Q_m\}, \tag{1}$$

where m is the number of analysed objects whilst a set of diagnostic variables is:

$$X = \{X_1, ..., X_k\}, \tag{2}$$

where:

k – number of analysed variables, assuming that $m >> k$.

The variables $X_{1, ..., }X_k$ allow distinguishing amongst objects and some do so better than others. Moreover, these variables will often possess different number crunching and various ranges of variation, which often precludes their direct comparison.

The knowledge of the object of research is a precondition of appropriate selection of diagnostic variables X_1, ..., X_k. Knowledge of the object of research and even a researcher's intuition are essential here. Variables X_1, ..., X_k under consideration must arise out of clear-cut pertinent links with the qualitative phenomenon under consideration. In the event that there are no adequate theories, one could use e.g. the opinion of a competent team of experts.

In order to divide a given set of objects Q_1, ..., Q_m into separate and relatively uniform sub-sets composed of objects with a similar value of variablesX_1, ..., X_k , we treat objects Q_1, ..., Q_m as vectors whose coordinates are values of variables reached by the vectors. Thus we have:

$$\mathbf{Q}_i = [x_{i1}\ x_{i2} \ldots x_{ik}], \tag{3}$$

where:

x_{ij} ($i = 1, ..., m; j = 1, ..., k$) – value of j variable in an i multi-dimensional object.

A set of objects Q_1, ..., Q_m described in the above manner can be presented in the form of the so-called observation matrix:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ x_{m1} & x_{m2} & \cdots & x_{mk} \end{bmatrix}. \tag{4}$$

Having defined its completeness it becomes a classification base of a set of multi-characteristic objects Q_1, ..., Q_m. In matrix (4) objects are researched, which we treat as vectors defined by means of formula (3), whilst the columns contain values of diagnostic variables X_1, ..., X_k.

It should be added that the capacity of the database presented as a matrix (4) is determined by the product of the dimensions of the matrix: $m \times k$, where: m – number of objects, k – number of diagnostic variables.

It is worth noting that too many diagnostic variables X_1, ..., X_k in a set will often impede or even prevent classification of multi-characteristic objects. This is

the reason why a pre-selected set of variables X_1, ..., X_k. must be subjected to further selection in terms of:

1) criteria of assessment of formal and pertinent qualities of variables,
2) criteria of the value of information provided by variables.

The above implies that in selecting diagnostic variables one should rely not only on pertinent arguments, but also statistical considerations (Hellwig, 1969; Zelias, 1982; Pociecha, 1996).

Whilst analysing diagnostic variables in terms of their formal and pertinent aspects, one should distinguish in the initial set of variables X_1, ..., X_k. at least three distinct sub-sets: 1) stimulants (*S*), 2) destimulants (*D*), 3) nominants (*N*). Stimulants are variables whose high values permit classifying a given object as superior from the point of view of an aggregated variable. With de-stimulants the opposite holds true, i.e. high values justify classifying an object as being inferior. The third diagnostic variable, which appears rather infrequently in empirical research is called the nominant. The nominant is a variable, which has the most favourable nominal value in terms of the assessment of objects, e.g. share of investment in GDP or population rise, which in some countries may be stimulants, and in others de-stimulants of growth. Variable nominats are increasing/decreasing functions, so their absolute values increasing to a nominal level have a positive impact on the assessment of a phenomenon, whilst further increase above the nominal level has a negative impact on the same. It is obvious that in the practical application of taxonomic methods one should focus on both stimulants and de-stimulants, and eliminate neutral variables, which are in no way connected with a researched complex phenomenon. The character of diagnostic variables should be, in principle, determined based on pertinent criteria. In the event that there is no applicable theory, one can make use of statistical methods based on correlation computation or the opinion of a team of competent experts.

Further research necessitates the transformation of de-stimulants into stimulants. For example in terms of the level and standard of living, the number of people per one room is a de-stimulant. If we compute the opposite of that indicator we will have a new variable which will be a stimulant and will mean the number of rooms per one person.

It should also be stated that one must always prefer such a transformation of variable de-stimulants into variable stimulants, which will permit one to attribute, transformed variables to certain economic interpretations.

Given that users of research data have various needs and expectations, one can analyse complex qualitative phenomena by using: 1) the statistics-based approach; 2) the dynamic approach. In the case of the statistics-based approach, the phenomenon under analysis is subjected to an analysis over a given period (moment). The required statistical data form a two-dimensional matrix $\boldsymbol{X}$, which is computed based on formula (4). In the dynamic approach, the phenomenon researched is analysed in objects, and additionally over time. The acceptance of statistical data representing numerous consecutive periods means that the data will

make a three-dimensional data matrix, which will be defined by the following formula:

$$\boldsymbol{X}^{(t)} = \left[x_{ij}^{(t)} \right] \qquad \begin{pmatrix} i = 1,\ldots,m \\ j = 1,\ldots,k \\ t = 1,\ldots,n \end{pmatrix}, \tag{5}$$

where n means the number of researched periods.

The product of the dimensions of the matrix determines the above means that the capacity of a database presented as matrix (5): $m \times k \times n$, where n represents the number of analysed periods.

4. Descriptive approach and stochastic approach in normalisation of diagnostic variables

A qualitative phenomenon can be analysed based on:
1) the descriptive approach (deterministic);
2) the stochastic approach.

It is assumed that in the stochastic approach, a set of objects is composed of a randomly selected sample of the population. Such an approach is called stochastic as it presupposes that researched variables describing analysed objects are selected at random, which entails a need to introduce some of the concepts used in mathematical statistics (classic and frequentist probability, expected value, variance, estimators, assessments, assessment errors, significance tests, etc). In many empirical tests this approach cannot, as a rule, be justified by anything. Neither do we know what practical sense the results obtained from numerical calculations make.

In the descriptive approach there is no mention of the randomness of the sample. Collected objects are all objects that are available. Thus the variables are not random ones, but merely regular variables in the true sense of the word.

The former approach is used chiefly in natural sciences that are in biology, chemistry, and physics. This is connected with the possibility of repeating the scientific experiment under the very same conditions. In the second approach one does not have any opportunity to research into the process of complex economic phenomena in artificial conditions. The conditions under which the numerical data were obtained are impossible to re-create here: the data derive chiefly from statistical reporting. In the course of economic research, frequently a set of business entities is taken into consideration, together with the entire population (set of *voivodships, poviats* or *gminas*) making the research exhaustive (comprehensive). Considering the above a descriptive approach is more justifiable. The author believes that the descriptive approach can be applied more widely than the stochastic approach to do research into economic phenomena .

5. Normalisation methods of diagnostic variables

From the point of view of taxonomic research it is important to ensure that the final diagnostic variables are comparable. This means, amongst others, that it is necessary to strip variables of their natural units, through which diagnostic characteristics are expressed and bring the variables to a state when they lend themselves to comparison, which implies smoothing of the range of variability of the characteristics. In order to achieve this, use is made of methods of normalisation of diagnostic variables measured on an interval and ratio level of measurement, whose general formula can be expressed as follows[1]:

$$X' = \frac{X - a}{b} = \beta X + \alpha \quad \text{when } X \in S,\ b \neq 0, \tag{6}$$

where:

X – diagnostic variables ($X \in S$),
X' – transformed variables,
a, b – normalisation parameters,
$\beta = \frac{1}{b}$ – coefficient with the diagnostic variable X,
$\alpha = -\frac{a}{b}$ – constant.

The transformed variable X' has the following descriptive parameters:

$$\bar{x}' = \frac{\bar{x} - a}{b}, \tag{7}$$

$$s_{x'} = \frac{s_x}{|b|}, \tag{8}$$

$$V_{x'} = \frac{s}{|\bar{x} - a|}, \tag{9}$$

where a and b are defined in the same way as in formula (6). The arithmetic mean $\bar{x}'$ of a transformed variable depends on the arithmetic mean of variable X $(\bar{x})$ and normalisation parameters a and b. On the other hand, standard deviation s_x depends on the standard deviation of the real variable X (s_x) and the value of b,

[1] We assume that a set of diagnostic variables describing economic entities is a set of stimulants ($X \in S$). If we have de-stimulants, ($X \in D$), we use the transformation: X' = ($a - X$):b, where parameters a and b are defined in the same way as in formula (6). We also assume that the set of values of diagnostic variables there are no outstanding values (rare or unrelated).

whilst the variation coefficient $V_{x'}$ depends on the arithmetic mean and standard deviation of variable X and the value of parameter a.

The most commonly used normalisation methods of diagnostic characteristics include:

1) Standardisation by two methods, when parameter a is equal to the arithmetic mean of variable X $\left(a = \bar{x}\right)$,

   b – standard deviation ($b = s_x$) and $a = 0$, $b = s_x$.

   By applying the above assumptions to formula (6) we will obtain:

$$x_i' = \frac{x_i - \bar{x}}{s_x}, \quad s_x > 0 \quad (i = 1, ..., m), \tag{10}$$

$$x_i' = \frac{x_i}{s_x}, \quad s_x > 0. \tag{11}$$

2) Unitarisation, when a equals zero, average value, lowest and highest value, b – range of standard variable $X(R_x = \max_i x_i - \min_i x_i)$.

   Below are selected normalising formulas based on unitarisation methods:

$$x_i' = \frac{x_i}{R_x} \quad (i = 1, ..., m), \tag{12}$$

$$x_i' = \frac{x_i - \bar{x}}{R_x}, \tag{13}$$

$$x_i' = \frac{x_i - \min_i x_i}{R_x}, \tag{14}$$

$$x_i' = \frac{x_i - \max_i x_i}{R_x}. \tag{15}$$

3) Quotient transformation, when $a = 0$, b is any number d=other than the value of range R_x (b is most often the arithmetic mean of variable X, minimum value of the variable, sum of the realisations of the variable one by one, sum of squares of variable measurements, root of sum of squares of observations.

The above means that:

$$x_i' = \frac{x_i}{\max\limits_i x_i}, \quad \max_i x_i \neq 0, \quad (i = 1, ..., m), \tag{16}$$

$$x_i' = \frac{x_i}{\min\limits_i x_i}, \quad \min_i x_i \neq 0, \tag{17}$$

$$x_i' = \frac{x_i}{\bar{x}}, \quad \bar{x} \neq 0, \tag{18}$$

$$x_i' = \frac{x_i}{\sum\limits_{i=1}^{m} x_i}, \quad \sum_{i=1}^{m} x_i \neq 0, \tag{19}$$

$$x_i' = \frac{x_i}{\sum\limits_{i=1}^{m} x_i^2}, \tag{20}$$

$$x_i' = \frac{x_i}{\left(\sum\limits_{i=1}^{m} x_i^2\right)^{1/2}}. \tag{21}$$

It is easy to note that in standardisation methods, the coefficient β in equation (6) is the opposite of standard deviation of normalisation variable X ($\beta = 1/s_x$), in methods relying on unitarisation, coefficient $\beta = 1/ R_x$, where R_x is a range of variable X, whilst in methods defined by means of quotient transformations, the coefficient is the opposite of the base of normalisation of variable X. Additionally, in methods relying on a_quotient transformation, the constant α of formula (6) normally equals zero.

In selecting a normalisation procedure one must remember that in the case of standardisation carried out by means of formula (10), when parameter a equals the arithmetic mean of variable X, b equals standard deviation, transformed variable X' has an average totalling zero $\left(\bar{x}' = 0\right)$ and standard deviation totalling one $\left(s_{x'} = 1\right)$. Thus not only the average value but also variation is made uniform. This eliminates variation as a base of differentiation of economic entities. The appearance of negative values of transformed variable X' is another important result.

It is easy to note that the transformation of variable X into variable X' based on formula (10) produces wide ranging variation bands for each normalised diagnostic variable. When a diagnostic characteristic is a stimulant, the limits of the variation band of a normalised characteristic can be expressed by means of the following formula:

$$x_i' = \in \left\langle \frac{\min\limits_i x_i - \bar{x}}{s_x}, \frac{\max\limits_i x_i - \bar{x}}{s_x} \right\rangle, \; X \in S. \tag{22}$$

The range of this interval is defined by four characteristics of variable X: $\bar{x}, s_x, \min\limits_i x_i$ and $\max\limits_i x_i$.

The standardisation method relying on formula (11), where $a = 0$ i $b = s_{x,}$ is characterised by average normalised values, $\bar{x}' = \bar{x} / s_x$ and a variation constant of these values, $s_{x'}^2 = 1$. Standard deviation is of course $s_{x'} = 1$.

Use of formula (11) produces normalised variable X' fitting into the interval:

$$x_i' \in \left\langle \frac{\min\limits_i x_i}{s_x}, \frac{\max\limits_i x_i}{s_x} \right\rangle, X \in S. \tag{23}$$

In that case, the limits of variation of the interval of normalised variable X' are not constant but variable and depend on three parameters $s_x, \min\limits_i x_i$ and $\max\limits_i x_i$.

It is worth noting that in this case on the one hand there is nothing to prevent normalisation of diagnostic characteristics, which are negative, positive and zero values. We should also note that normalisation will produce the following values: zero for $x_i = 0$, positive for $x_i > 0$ and negative for $x_i < 0$.

Use of the unitarisation methods will cause the range of a normalised characteristic X' to be constant and amount to one in all four formulas $(a = 0, \; a = \bar{x}, \; a = \min\limits_i x_i, \; a = \max\limits_i x_i, \; b = R_x)$.

Use of parameter a at its highest value will cause the transformed variable to have non-positive values.

Let us also note that a variable normalised in keeping with formula (12) is:

$$x_i' \in \left\langle \frac{\min\limits_i x_i}{R_x}, \frac{\max\limits_i x_i}{R_x} \right\rangle, \tag{24}$$

$$R_{x'} = 1. \tag{25}$$

(24) and (25) indicate that variable X' normalised in keeping with formula (12) has a constant range equalling a unit, whilst the floor and ceiling of the range of normalised variable may have different location on the real axis.

In respect of formula (13), the values of X' normalised variable fit into the following range

$$x_i' \in \left\langle \frac{\min\limits_i x_i - \bar{x}}{R_x}, \frac{\max\limits_i x_i - \bar{x}}{R_x} \right\rangle \tag{26}$$

and:

$$R_{x'} = 1. \tag{27}$$

On the other hand, quotient transformation is made in keeping with formula (6), in which coefficient $\beta = 1/b$, and b is the so-called base of variable X normalisation, and $\alpha = 0$. Quotient transformations so defined (cf. formulas (16)–(21)) satisfy the following recommendations: 1) additivity requirement (laying down formal basis for the conduct of basic arithmetic operations/activities in sets of primary values of variables with different number crunchiness), 2) non-negativity requirement (all realisations of variables are non-negative). On the other hand, they do not satisfy the requirement of a constant interval of variation of normalised values in terms of a constant range and constancy of extreme values, e.g. one often attempts to ensure that normalised values of diagnostic variables fit into a range of <0, 1>.

In empirical research, the most often used normalisation base of variable X is its arithmetic mean $(b = \bar{x})$. So transformed variables have the following four desirable qualities:

1) they are comparable (additivity requirement), as measurement units have been eliminated from the different number crunching variables;
2) they have varied variances;
3) the arithmetic mean of each equals the number of researched objects;

By applying formula (18), we will have:

$$x_i' \in \left\langle \frac{\min\limits_i x_i}{\bar{x}}, \frac{\max\limits_i x_i}{\bar{x}} \right\rangle, \; X \in S. \tag{28}$$

As (28) indicates, one cannot assume that the lower and upper limits of dispersion ranges are equal for all variables included in the set of diagnostic characteristics.

We should further note that the normalisation of variables $X_1, ..., X_k$ by reference /by determining a reference system, in terms of computations, is markedly simpler than standardisation, further helping one to avoid the inconvenience of having to deal with negative values of standardised variables $X'_1,...,X'_k$. Normalisation based on formula (6), when $a = 0$, $b = \bar{x}$, allows, in contrast with standardisation, which causes each variable to have equal influence on the results of research, retaining varied variances of variables, hence giving them varied importance $V_{x'_1}=V_{x_1}=\cdots=V_{x'_k}=V_{x_k}$, where $V_{x'_j}$ and V_{x_j} (j = 1,..., k) are coefficients of variance of variables X'_j and X_j (j = 1,...,k).

Table 1. Distribution of variable values after normalisation

Normalisation formula	Arithmetic mean of transformed variable (X')	Standard deviation of transformed variable (X')	Range for normalised variables X'
$(x_i - \bar{x})s_x$	0	1	R_x / s_x
x_i / s_x	$\bar{x} / s_x$	1	R_x / s_x
x_i / R_x	$\bar{x} / R_x$	s_x / R_x	1
$(x_i - \bar{x}) / R_x$	0	s_x / R_x	1
$(x_i - \min_i x_i) / R_x$	$(\bar{x}_i - \min_i x_i) / R_x$	s_x / R_x	1
$x_i / \max_i x_i$	$\bar{x} / \max_i x_i$	$s_x / \max_i x_i$	$R_x / \max_i x_i$
$x_i / \bar{x}$	1	$s_x / \bar{x}$	$R_x / \bar{x}$
$x_i \Big/ \sum_{i=1}^{m} x_i$	$1 / n$	$s_x \Big/ \sum_{i=1}^{m} x_i$	$R_x \Big/ \sum_{i=1}^{m} x_i$
$x_i \Big/ \sqrt{\sum_{i=1}^{m} x_i^2}$	$\bar{x} \Big/ \sqrt{\sum_{i=1}^{m} x_i^2}$	$s_x \Big/ \sqrt{\sum_{i=1}^{m} x_i^2}$	$R_x \Big/ \sqrt{\sum_{i=1}^{m} x_i^2}$

Source: M. Walesiak, 2002, p. 21.

Generally, all of the normalisation formulas presented so far, which are linear transformations of observations of each variable, maintain skewness and kurtosis of the variables distribution (see: M. Walesiak, 2002, p. 20). Additionally, all normalisation formulas retain an unchanged value of the linear correlation coefficient in respect of each pair of variables.

The value of characteristics $\bar{x}', s_{x'}$ i $R_{x''}$ for selected normalisation procedures is presented in table 1.

Table 1 indicates that normalisation formulas (unitarisation, quotient transformation with a base of normalisation equalling to a range) guarantee normalised values of variables variance and also constant range for all variables.

Standardisation and transformation with a normalisation base equal to standard deviation enforce uniformity of the values of all variables in terms of variance.

If we take the quotient transformation with a normalisation base equal to the maximum and the root of the sum of squares of observations, then the normalised values of variables will ensure varied variance, arithmetic mean and range.

Moreover, quotient transformation with a normalisation base equal to the sum and arithmetic mean ensure for normalised values of variables verse/varied variance, range and constant arithmetic mean for all variables.

The last group of normalisation procedures of diagnostic variables includes the rank method procedures. It is generally used to arrange objects based on qualitative variables measured on a scale (these are situations that are very frequent in marketing research). Variables measured on an ordinal level of measurement are subjected to a rank procedure, so that each i realisation of variable X in a series of its values arranged (non-decreasingly or non-increasingly) is assigned an adequate rank. If we consider m elements, then the "best object" is ranked m, whilst the "worse object" is ranked 1.

The appropriate ranks for m objects will be:

$$x_i' = \begin{cases} 1 & \text{dla} \quad \min\limits_i x_i \\ & \qquad (i = 1,...,m) \\ m & \text{dla} \quad \max\limits_i x_i \end{cases} \tag{29}$$

It is worth noting that $x_i' \in <1,\ m>$ $(i = 1,...,m)$, and the range of so normalised a variable equals the number of objects less one ($R_{x'} = m - 1$). The range is constant for all variables normalised by means of the rank method.

Ranking satisfies the additivity requirement, as it makes disparate characteristics comparable (normalised values of variables are numbers without crunching) and the non-negativity requirement (all normalised values of variables are non-negative).

When we deal with variables measured on powered scales, i.e. interval and ratio ones, we normally use standardisation, unitarisation and quotient transformation.

It is worth noting that the variables measured on a interval and ratio scale can be processed with the a rank method ,which explains the ease with which they transit from a powered to a weaker scale, however such a procedure must

invariably be associated with loss of statistical data, which manifests itself, amongst others, in restrictions on use of various statistical methods.

One must also consider a case where variables can be positive, zero and negative. This often happens in the course of analysing the financial standing of companies, banks and other business ventures. This implies that one must select a normalisation method, which transforms real variables of any value ($x_i \in R$, where R is a set of real numbers).

Using one of the normalisation formulas (10)–(21), one obtains the following normalised matrix:

$$\mathbf{X}' = \begin{bmatrix} x'_{11} & x'_{12} & \cdots & x'_{1k} \\ x'_{21} & x'_{22} & \cdots & x'_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ x'_{m1} & x'_{m2} & \cdots & x'_{mk} \end{bmatrix}, \tag{30}$$

where x'_{ij} $(i = 1,...,m;\ j = 1,..., k)$ – normalisation of the value of j variable in i object , which are directly comparable.

6. Conclusions

The selection of a proper formula to normalise diagnostic variables is one of the key issues involved in the construction of aggregated variables. In making such a choice one may benefit from a profound knowledge of the phenomenon under analysis, one's goal and cognisance of various transformation formulas. However, it is more important than ever to take into consideration economic analysis, especially when one can encounter a theory that presents the characteristics defining a complex phenomenon clearly enough.

One should also acknowledge that the results of linear ordering of objects depend also on the selection of a proper formula normalising diagnostic variables. Researches who make use of multi-dimensional comparative analysis methods in carrying out their empirical work, highly value those formulas that guarantee stable or nearly stable areas of variance of normalised variables. Likewise, they frequently make use of the constancy of parameters characterising normalised variables.

Selecting a normalising transformation one must bear in mind that normalisation has an impact on the results of linear ordering of objects by effecting the relationships between descriptive parameters of real and transformed diagnostic variables (cf. formulas (7)–(9)). Furthermore, there are no grounds to believe that one can design a universal normalisation formula, as each formula satisfies only selected criteria (e.g. deprivation of number crunching, which express diagnostic characteristics, varied variation measured by means of e.g.

standard deviation, constancy of the range of all normalised characteristics, non-negativity of normalised qualities, possibility of normalising characteristics adopting both positive and negative values or only negative ones).

To conclude, the author would like to emphasise that normalisation of diagnostic characteristics is extremely important and absorbing, the more so that it facilitates understanding of the interdependencies amongst problems and their complexity, not to mention the fact that it allows spotting threats early on in the course of research and taking adequate preventive measures.

REFERENCES

BARTOSIEWICZ, S. "Propozycja metody tworzenia zmiennych syntetycznych", /Proposed Methods of Developing Aggregate Variables/, *Prace Naukowe Akademii Ekonomicznej we Wrocławiu*, 84., 1976.

BORYS, T. "Metody normowania cech w statystycznych badaniach porównawczych", /Methods of Normalisation of Characteristics in Comparative Statistical Research/, *Przegląd Statystyczny*, vol. 2, 1978.

CIEŚLAK, M. "Taksonomiczna procedura programowania rozwoju gospodarczego i określania zapotrzebowania na kadry kwalifikowane", /Taxonomic Programming Procedure of Economic Growth and Determining Management Demand/, *Przegląd Statystyczny*, vol.1, 1974.

GAŁUSZKA, B. "O metodzie szacowania brakujących danych przekrojowych", /On Method of Estimating Missing Comprehensive Data/, *Przegląd Statystyczny*, vol. 2, 1992.

GATNAR, E. Symboliczne metody klasyfikacji danych, /Symbolic Methods of Data Classification/, Warszawa: Wydawnictwo Naukowe PWN, 1998.

GRABIŃSKI, T. "Wielowymiarowa analiza porównawcza w badaniach dynamiki zjawisk gospodarczych", /Multi-dimensional Comparative Analysis in Research into Dynamics of Economic Events/, Akademia Ekonomiczna w Krakowie. *Zeszyty Naukowe*, *seria specjalna: Monografie,* 61, Kraków, 1984.

GRABIŃSKi, T. (1992), Metody taksonometrii */Taxonometric Methods/, Kraków: Akademia Ekonomiczna* .

GRABIŃSKI, T., S, WYDYMUS, A. ZELIAŚ, Z badań nad metodami szacowania brakujących informacji, /On Research into Methods of Estimating Missing Information/, Kraków: Akademia Ekonomiczna w Krakowie: *Zeszyty Naukowe*, 114, 1979.

GRABIŃSKI, T., S. WYDYMUS, A. ZELIAŚ, Metody taksonomii numerycznej w modelowaniu zjawisk społeczno-gospodarczych, /Numerical Taxonomic Methods in Modelling Social and Economic Events /, ed. A. Zeliaś, Warszawa: PWN, 1989.

HELLWIG, Z., Zastosowanie metody taksonomicznej do typologicznego podziału krajów ze względu na poziom ich rozwoju oraz zasoby i strukturę wykwalifikowanych kadr, /Use of the Taxonomic Method for Typological Classification of Countries in Terms of Their Development, Resources and Structure of Professional Qualifcations/, *Przegląd Statystyczny*, vol. 4., 1968.

HELLWIG, Z., Problem optymalnego wyboru predykant, /Problems of Optimum Selection of Predykant/, *Przegląd Statystyczny*, vol. 1, 1969.

JAJUGA, K., Statystyczna analiza wielowymiarowa */Multi-dimensional Statistical Analysis/,* Warszawa: Wydawnictwo Naukowe PWN, 1993.

KORDOS, J., Jakość danych statystycznych, /Quality of Statistical Data/, Warszawa: PWE, Warszawa, 1988.

KUKUŁA, K., Metoda unitaryzacji zerowanej, /Zero Unitarisation Method/, Warszawa: Wydawnictwo Naukowe PWN, 2000.

NOWAK, E., Problemy doboru zmiennych do modelu ekonometrycznego, /Problems Involved in Selecting Variables for Economic Models/, Warszawa: PWN, 1984.

NOWAK, E., Metody taksonomiczne w klasyfikacji obiektów społeczno-gospodarczych, /Taxonomic Methods in Classifying Social and Economic Objects/, Warszawa: PWE, 1990.

NOWAK, E., Problem informacji w modelowaniu ekonometrycznym, /Information Problems in Econometric Modelling/,Warszawa: PWN, 1990.

POCIECHA, J., Metody statystyczne w badaniach marketingowych, /Statistical Methods in Marketing Research/, Warszawa:PWN, 1996.

POCIECHA, J., B. PODOLEC, A. SOKOŁOWSKI, K. ZAJĄC., Metody taksonomiczne w badaniach społeczno-ekonomicznych, /Taxonomic Methods in Social and Economic Research/, Warszawa: PWN, 1988.

STRAHL, D., Propozycja konstrukcji miary syntetycznej, /Proposal Regarding Construction of An Aggregate Variable/, *Przegląd Statystyczny*, vol. 2, 1978.

WALESIAK, M., Metody analizy danych marketingowych, /Analysis Methods of Marketing Data/, Warszawa: Wydawnictwo Naukowe PWN, 1996.

WALESIAK, M., Uogólniona miara odległości w statystycznej analizie wielowymiarowej, /Generalised Measure of Distance in Multidimensional

Statistical Analysis/, Wrocław: Wydawnictwo Akademii Ekonomicznej im. Oskara Langego we Wrocławiu, 2002.

ZELIAŚ, A., "Kilka uwag o kryteriach doboru zmiennych w modelach ekonometrycznych", /Some Comments on Selection Criteria of Variables in Econometric Models/ *Folia Oeconomica Cracoviensia*, vol. 24, 1982.

ZELIAŚ A, ed. Taksonomiczna analiza przestrzennego zróżnicowania poziomu życia w Polsce w ujęciu dynamicznym. /Taxonomic Analysis of Spatial Variance of Living Standards in the Light of Dynamic Approach/, Kraków: Wydawnictwo Akademii Ekonomicznej w Krakowie, 2000.

STATISTICS IN TRANSITION, October 2002
Vol. 5, No. 5, pp. 803–821

COMBINATORIAL SEARCH IN MULTIVARIATE STATISTICS

David Banks[1] **and Robert T. Olszewski**[2]

ABSTRACT

Modern computational statistics often requires extensive combinatorial search. But search is expensive, not just in terms of time, but also in terms of the penalty one pays for multiple testing associated with model selection. Therefore it is important for modern statistical procedures to search in the most efficient ways possible. This paper describes several methods for smart combinatorial search, and shows how they can apply to problems in multivariate regression, cluster analysis, and multidimensional scaling.

Keywords: cluster analysis; curse of dimensionality; Gray code; multidimensional scaling; multiple testing; nonparametric regression.

1. Introduction

Modern multivariate analysis faces the curse of dimensionality. This was first coined by Richard Bellman, a mathematician, in the context of numerical approximation (Bellman, 1961, p. 106) The curse of dimensionality implies that when p, the number of variables per observation, is large then inference is more difficult. And the level of difficulty increases very rapidly with p.

In modern computational statistics, there are several ways to show that as the number of explanatory variables increases, the problem of structure discovery becomes harder. This is closely related to the problem of variable selection in model fitting. Classical statistical methods, such as multiple linear regression analysis, avoided this problem by making the strong model assumption that the mathematical relationship between the response variable and the explanatory variables was linear. But most statisticians now prefer alternative analytic approaches, since the linearity assumption is usually wrong for the kinds of applications encountered in practice. This means they must attempt to find ways to mitigate the curse of dimensionality.

[1] CBER, U.S. Food and Drug Administration 1401 Rockville, MD 20850 banksd@cber.fda.gov.
[2] Center for Biomedical Informatics University of Pittsburgh, Pittsburgh PA 15213.

There are three nearly equivalent ways to describe the curse of dimensionality, and each provides a usefully different perspective on the problem:

- For large p, nearly all datasets are too small.
- For large p, nearly all datasets are multicollinear (or concurve, the nonparametric generalization of multicollinearity).
- The number of structural functions to consider grows quickly (faster than exponentially) with p.

These problems are reduced if data are collected using an appropriately dispersed statistical design, such as Latin hypercube sampling (cf. Stein, 1987). However, except for simulation experiments, this is difficult to achieve.

This paper focuses upon the third formulation of the curse. For specificity, consider the case of regression analysis; here one wants to find structure that predicts the value of the response variable Y from a vector of explanatory variables $\boldsymbol{X}$ in $\mathbb{R}^p$, the p-dimensional Euclidean space. In this case one must consider various functional forms for models that relate the $\boldsymbol{X}$ to the Y; so consider the problem of fitting a polynomial model of degree less than or equal to 2.

The important thing to note is that, as the third version of the curse expresses, there is an explosion in the number of possible models. When p=1 (i.e., a single explanatory variable), there are seven possible regression models:

$$Y = \beta_0 + \varepsilon$$
$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon$$
$$Y = \beta_1 X_1 + \varepsilon$$
$$Y = \beta_0 + \beta_2 X_1^2 + \varepsilon$$
$$Y = \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon$$
$$Y = \beta_0 + \beta_2 X_1^2 + \varepsilon$$

where ε denotes noise in the observation of Y. For p=2 there are 63 models to consider (including interaction terms of the form $X_1 X_2$), and simple combinatorics shows that, for general p, there are $2^{[1 + p + p(p+1)/2]} - 1$ models of degree 2 or less. Since real-world applications usually explore much more complicated functional relationships than low-degree polynomials, modern computational statisticians need vast quantities of data to discriminate among the many possibilities when p is large (or even moderate).

When fitting many models, it is necessary to make repeated significance tests to determine which mathematical terms add substantial explanation. But if one makes repeated significance tests, it is well known that a certain proportion of the tests will be misleading; in particular, one will include terms that are not useful. One approach to adjust for this is to use Bonferroni correction, or other procedures that control for multiple testing (cf. Westfall and Young, 1993). But these methods typically tend to be very conservative, and one pays an especially

high price when the tests are dependent, as happens when comparing models that incorporate very similar sets of explanatory variables.

It used to be popular to avoid multiple testing and just include all of the explanatory variables, and in the early days of data mining many naive computer scientists did so. But statisticians knew that this violated the Principle of Parsimony and led to inaccurate prediction. If one does not severely limit the number of variables (and the numbers transformations of the variables, and the number of interaction terms between variables) in the final model, then one ends up *overfitting* the data. Overfit happens when the chosen model describes the random noise as well as the true signal.

If one is interested only in prediction, the curse of dimensionality is less of a problem for future data whose explanatory variables have values close to those observed in the past. But an unobvious consequence of large values of p is that nearly all of the new observation vectors tend to be far from those previously seen. Furthermore, if one needs to go beyond simple prediction in order to develop uncertainty statements or interpretable models, then the curse of dimensionality can be an insurmountable obstacle. Usually the most one can achieve is local interpretability, and that happens only where data are locally dense. For more detailed discussions of the curse, readers should consult Hastie and Tibshirani (1990) and Scott and Wand (1991).

In order to minimize the problem of overfit that results from the conservatism of multiple testing in the context of model selection, this paper considers various strategies for efficient combinatorial search. The heuristic behind this approach is that if one can minimize the number of models that are tested, then one can both improve computational speed and prevent the inclusion of spurious variables or false structure. The issues involved with efficient combinatorial search are illustrated with applications to problems in multivariate nonparametric regression, cluster analysis, and multidimensional scaling. The methodology relies upon combinatorial algorithms for enumerating binary sequences, listing permutations, and identifying all possible subsets of fixed size from a given set of elements.

2. Nonparametric Multivariate Regression

A central problem in modern computer-intensive statistics and data mining is finding good variable selection techniques for nonparametric (or flexible) multivariate regression. Suppose one wants to perform nonparametric multiple regression with p explanatory variables. Then there are 2^p possible sets of variables that might be fed into the software that fits the model. Depending upon the application, there are many different kinds of software, such as MARS (Multivariate Adaptive Regression Splines; see Friedman, 1991), ACE (Alternating Conditional Expectations; see Breiman and Friedman, 1985), neural nets (cf. Bock, 1998 or Lebart, 1998 for applications to cluster analysis and

discriminant analysis, respectively) and so forth. The key to the success of a particular piece of software in a particular application often depends on how well that software is able to search among the possible variable combinations to find the best one to use in model fitting.

Even for relatively small values of p, it is computationally infeasible to fit and examine all of the possible subsets of variables. And the available methodology for correction in multiple testing is so conservative that essentially no set of variables will meet adequate levels of significance, which is why most of the modern software packages for computer-intensive statistics do not control for the number of tests that are made.

Traditional methods for model selection in multiple regression analysis include forward selection, backwards elimination, and stepwise linear regression (cf. Weisberg, 1985, Chapter 8). None of these attempts a full search of the space of possible models, and each executes a greedy search according to a specified testing procedure. Unfortunately, these methods break down for large values of p, both in terms of mean integrated squared error (cf. Banks, Olszewski and Maxion, 2002) and in terms of multiple testing correction for accurate selection of variables (cf. Pinsker, Kipnis and Grechanovsky, 1987). For example, an extensive set of simulations in the case of stepwise selection for multiple linear regression when the true model is linear finds that, when using standard defaults in common packages, each spurious variable has almost a 15% chance of being falsely included, independently of whether other spurious variables appear.

2.1. Experimental Design Methodology

Instead of doing greedy search to find adequate multiple linear models, it is possible to use ideas from experimental design. A version of this strategy was independently proposed by Clyde (1999) in the context of model averaging.

To implement this approach, one can view each explanatory variable as a factor in an experimental design. The factors all have two levels, corresponding to whether or not the explanatory variable is included in the model. This enables one to perform a 2^{p-k} fractional factorial experiment in which one fits a multiple regression model to the included variables and records some prechosen measure of goodness-of-fit. Obviously, one takes k to be sufficiently large that it is possible to perform the computations in a reasonable amount of time and also to limit the effect of multiple testing. But since this is a computer experiment and one can fit models very quickly, this still enables one to collect a great many observations on different models.

There are several possible measures of goodness-of-fit. Most were developed in the context of multiple linear regression analysis, but they can be extended (with varying degrees of success) to nonparametric problems. Standard choices include:

- R^2, the proportion of variance in the observations that is explained by the model.

- Adjusted R^2, the proportion of variance in the observations that is explained by the model, but with an adjustment to account for the number of variables in the model.
- Mallows' C_p, a measure of predictive accuracy that takes account of the number of terms in the model.
- MISE, the mean integrated squared error of the fitted model over a given region (often the hyperrectangle defined by the minimum and maximum values taken by each explanatory variable used in the model.

The statistical properties of the first three measures are discussed in Weisberg (1985, p. 185-190); a good discussion of the MISE is given in Scott (1992, section 2.4).

We recommend using the square root of the adjusted R^2. This transformation of the measure of fit appears to stabilize the variance in the observations, thereby supporting use of analysis of variance techniques and response surface methodology in the search for the best-fitting model (cf. Myers, 1999). These techniques perform best when the observations from the computer experiment have an approximate normal distribution with common variance.

In this framework, it is straightforward to use analysis of variance to examine which factors and factor combinations have a significant influence on the observations. Significant main effects correspond to explanatory variables that contribute on their own. Significant interaction terms correspond to subsets of variables whose joint inclusion in the model provides explanation, or whose joint inclusion lowers the measure of fit through redundant explanation. If one is performing multiple linear regression, then these results are implicit in the standard tests of significance for coefficients and the combinatorial search procedure adds no new insight; however, if one is performing a modern computer-intensive nonparametric regression, then the methods described here enable better search for influential variables.

Based on the results of the designed experiment, one can ultimately find and fit the model that includes all variables corresponding to significant main effects or interactions. And the factorial design reduces the penalty one pays for multiple testing, as compared to exhaustive search or other less-efficient searches.

2.2. Gray Code Search

Factorial designs are not the only way to search. A different approach, based on combinatorial properties of binary sequences, may offer even more improvement.

For multiple regression, the 2^p possible models can be identified with the 2^p vertices of the unit hypercube in p dimensions. The (0, 0,…,0) vertex corresponds to the model with all variables excluded (one just fits the average of the response variable), whereas the (1,1, …l) model is the regression in which all variables are included. From this perspective, a clever search of vertices of the hypercube would be an attractive search algorithm for finding a good regression model.

A Hamiltonian circuit of the unit hypercube is a traversal that reaches each vertex exactly once (and necessarily, it never travels the same edge twice). There are many possible Hamiltonian circuits - the exact number is not known in general, even asymptotically (cf. Gilbert, 1958). From the standpoint of model search, one wants a Hamiltonian circuit that has desirable properties of symmetry, and which treats all vertices in the same way (i.e., there should be no distinguished vertex, or model, in the circuit).

If it were possible to traverse the edges of the hypercube visiting each vertex just once and maintaining a certain balance in the process, then the sequence of visited vertices would form the basis for an efficient search algorithm.

The Gray code (1939) is a procedure for listing the vertices of the hypercube in such a way that there is no repetition, each vertex is one edge from the previous vertex, and all vertices in a neighborhood are explored before moving on to a new neighborhood. Wilf (1989) describes the mathematical theory and properties of the Gray code system.

To explain the Gray code algorithm, consider the case of four explanatory variables, or the unit hypercube in $|^4$. Table 1 shows the rank of the vertex in the Gray code traversal, the binary digit representation of the rank, and the bit string of the visited vertex on the hypercube:

Table 1. Gray code vertex rank, binary rank, and vertex string

0	0000	0000	8	1000	1100
1	0001	0001	9	1001	1101
2	0010	0011	10	1010	1111
3	0011	0010	11	1011	1110
4	0100	0110	12	1100	1010
5	0101	0111	13	1101	1011
6	0110	0101	14	1110	1001
7	0111	0100	15	1111	1000

The Gray code has several kinds of subtle balance. For example, it can be generated by reflection and recursion. Let L_p be the list of all possible binary bit strings of length p, arranged in the Gray code order. Then generate the first half of L_{p+1} by writing a zero in front of each element in the list L_p. For the second half of L_{p+1}, write L_p in reverse order, and then prefix each element with a one. By concatenating the lists, one obtains L_{p+1} (the reader should observe this pattern in Table 1).

Suppose one prefixes each Gray code string with an infinite number of zeroes. This makes it possible to consider the numbers corresponding to the Gray code strings as an infinite series:

0, 1, 3, 2, 6, 7, 5, 4,12, 13, 15, 14, 10, 11, 9, 8, …

Note that each number in the sequence is relatively close to its neighbors. A theorem due to Yuen (1974) shows that two strings in the Gray code whose Hamming distance is at least d must have ranks that differ by at least $[2d/3]$ (here $[\cdot]$ is the nearest-integer function), and this provides the greatest possible separation. This means that the traversal explores models locally and exhaustively, rather than swooping back after a distant excursion.

From our perspective, the key point from Yuen's theorem is that if starts at an arbitrary model, then goes a large number of steps in the Gray code traversal, one ends up at a vertex corresponding to a model that is very different from the starting point. This property suggests that by taking every dth step, for d large, and then testing the corresponding model, one is performing a thorough search of the set of possible models.

Wilf gives a theorem that is useful in rapidly generating the Gray code strings that correspond to particular ranks:

Theorem 1: Let $m = 3\ a_i\ 2^i$ be the representation of integer m in binary notation. Let $\ldots b_3\ b_2\ b_1\ b_0$ be the string for the vertex of rank m in the Gray code. Then

$$b_i = a_i + a_{i+1} \pmod 2$$

for i=0,1,2, … (this solves the *ranking problem* for a Gray code, since it gives the code with rank m in the list).

Proof: (Wilf, 1989, p. 4.) The proof proceeds by induction on p. When p=0, the statement is trivial. Assume now that Theorem 1 is true for all ranks in the Gray code list L_{p-1} of binary sequences of length p ! 1. Consider the string of rank m on the Gray code list L_p of binary sequences of length p. If $m < 2^{p-1}$! 1 then the result is immediate since the string of rank m is the same, except for the prefatory 0.

If $m \geq 2^{p-1}$, then we can write $m' = 2^{p-1} - m$. The theorem holds for the rank m' because $m' < 2^{p-1} - 1$. And the bits in the strings of rank m and rank m' are related by

$$a_i(m) = 1 + a_i(m') \text{ for } i = 0, 1, \ldots, p\text{-}1. \qquad \square$$

This theorem gives us a way to generate the Gray code string for a specified rank, and thus the theorem is sometimes called the "ranking theorem." A theorem that goes the other way, to find the rank of a given code string, solves the unranking problem. To use Theorem 1 to efficiently explore a set of models, suppose one decides to examine only 100 models and then infer the final fit. If there are p explanatory variables, one takes $d = [2^p/100]$, and then finds the Gray code sequence of rank d, $2d$, …, $100d$. Each sequence defines a particular set of variables that may be included or excluded.

In practice, one would examine the 100 model fitting results, probably in terms of the square root of the adjusted R^2, and then home in on the region of the cube that provides good explanation. This enables one to quickly identify the

vertex or bit string corresponding to the set of variables that provides near-optimal explanation. One might make additional Gray code searches in the region of the best results from the first search, and iterate to find the final model. Determining good rules of thumb for practice requires an extensive simulation experiment that would compare the effectiveness of Gray code search procedures, with corrections for multiple testing, against other standard procedures for model selection.

3. Cluster Analysis

Suppose one wants to do a cluster analysis in which each case has many explanatory variables. In this situation, there can be many different cluster structures. Also, the clusters that derive from conventional procedures (i.e., procedures that are not computationally tuned to be robust) tend to be unstable (cf. Fowlkes, Gnanadesikan and Kettenring, 1988).

As an example to focus our attention, consider a market segmentation study using data from supermarket scanners. The data set tells exactly what, and how much, of each product was purchased by each customer over the period of a year. Some items, such as expensive shampoo and nail polish, divide the customers into clusters that can be retroactively interpreted as men and women. Other items, such as diapers and baby food, divide the customers into clusters that correspond to parents with infants and those without. And other products, such as tortilla flour and refried beans, or matzoth meal and kosher salt, or shitake mushrooms and bamboo shoots, are able to cluster people as Hispanic, Jewish, or Chinese, respectively, and thereby show ethnic patterns. In contrast, products such as eggs, butter, salt, and soap are noise variables, and provide little information about demographic cluster structure.

In cluster analysis, we want automatic ways to ignore the noise variables and to focus on those variables that provide information about the cluster structure(s). And when the number of measurements per case is large, then there can easily be multiple cluster structures, as indicated in the supermarket scanner data example.

No clustering method is uniformly superior in all applications. But one approach that seems promising in this context is to look first for cluster structure on a variable-by-variable basis. Clearly, records on the number of purchases of cosmetics or diapers or shitake mushrooms will probably show two very distinct groups; one cluster does not buy these items regularly, but a second cluster buys them routinely. So for many of the most useful variables, single variables can provide good segregation.

It is computationally efficient to do cluster analyses of all of the cases for each of the variables separately. This is a univariate cluster analysis, and thus can be done quite rapidly, and the results can be stored for later use. If we take the market segmentation problem as our guide, then we expect each variable to divide the cases into at most two groups. In other applications, different heuristics would

apply, and the following discussion should be modified accordingly. But for our example problem, the primary result from this first step should be, for each variable, a list of the cases that fall into one cluster, a list of the cases that fall into a second cluster, and a measure of how well-separated the clusters are.

Given those lists, the next step is to decide how closely two variables agree on the cluster structure. For any pair of variables, there are several possible measures of how closely their two lists agree. For example, one can construct a simple contingency table in which the rows represent the two lists from the first variable and the columns represent the two lists from the second variable. The entries in each cell correspond to the number of cases that are in each cross-classification. Then any measure of deviation from independence, such as the traditional Pearson goodness-of-fit statistic, indicates significant agreement on the cluster membership of the cases between the lists generated by Variable 1 and Variable 2.

Variables that show strong dependence would probably include nail polish, emery boards, and aftershave lotion. Variables that would show independence probably include nail polish, kosher hotdogs, and diapers. In order to highlight the cluster structure and enable interpretation, it is helpful if the variables that provide similar lists are grouped together for the analyst, while those variables that have little cluster structure or very different cluster structure are grouped apart.

One approach that has been taken in the cluster analysis of such data is that described by Wilhelm, Wegman and Symanzik (1999). They describe the use of parallel coordinate plots to find, visualize, and interpret cluster structure. Parallel coordinate plots were proposed by Inselberger (1981) and provide an appealing alternative to Cartesian coordinate systems for visualizing high-dimensional structure. Essentially, instead of having the graphical image show axes meeting at right angles (which obviously breaks down when $p > 3$), parallel coordinate plots align each axis in parallel. The point that represents an observation in Cartesian systems becomes a line, linking the values of the observation on each axis.

Figure 1 shows a parallel coordinate plot of eight cases, with seven variables for each case. For each case, a line links the seven observations on that case. Note that in terms of variables X_1, X_2, and X_3, the graph shows strong cluster structure, with cases corresponding to circles and squares separating very cleanly on each axis. Similarly, for variables X_4, X_5, and X_6, the cases no longer cluster in terms of squares and circles, but they do cluster strongly in terms of color, black or white, for each axis. Finally, for variable X_7, there does not seem to be any cluster structure at all. In terms of the market segmentation example, one can think of the circles and squares as being men and women, and the black and white as denoting parents and non-parents.

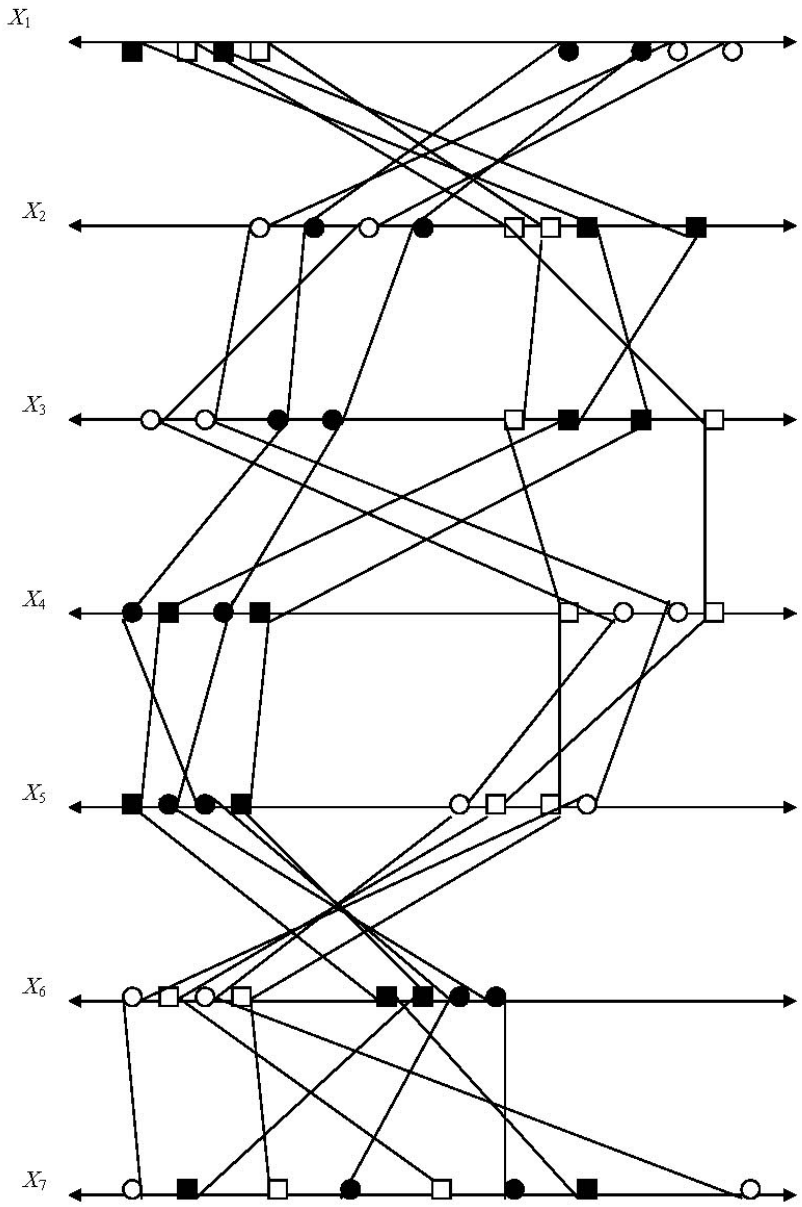
The problem of performing cluster analysis using parallel coordinate plots can be approached in many ways. But Figure 1 shows that it is very helpful, in terms of both cluster discovery and interpretation, if the axes that capture the same kind of cluster structure are arranged to be adjacent. Otherwise, the visual

signal is hard to detect in large datasets with many different variables. If one can find an arrangement that puts variables that contribute to common clustering structure together, this aids understanding. This can be viewed as a special kind of feature selection-instead of selecting features that contribute to structure, and leaving out features that are essentially noise, one must find the features that contribute to different kinds of cluster structure, and separate those out from the features that are essentially noise.

To achieve this arrangement, one must search the space of all possible permutations of the axes. If this can be done efficiently, then one can reduce computational time and also minimize the number of statistical tests that are performed during the search, thereby limiting the impact of multiple testing on the number of false findings of significance. In the example we have discussed, based upon cluster-membership agreement for each variable, it is possible to use direct measures of agreement to find candidate axes for adjacent representation. But in more general problems, where multivariate behavior is important for cluster structure, the full power of search over the set of all possible arrangements of the axes becomes vital.

There are many ways to enumerate the list of all possible permutations variables; this is equivalent to listing the permutations of the indices 1, …, p. From a computational standpoint, it is appealing if the algorithm produces a sequence of permutations with the property that permutations that are near each other in the list are also near in terms of some appropriate metric on the space of permutations. The advantages from such a property are very similar to those that arise in the context of efficient search of binary vectors via Gray code enumeration.

Figure 1. Parallel coordinate plot with two cluster structures



Although there does not yet seem to be an analogue of Yuen's theorem for permutations, the procedure that best imitates the Gray code properties is the Steinhaus-Johnson-Trotter algorithm (cf. Trotter, 1962 and Johnson, 1963). The key feature is that at each step, exactly two elements are interchanged and these two elements are adjacent to each other in the previous permutation. The algorithm can be defined recursively: Suppose one has the sequence of Steinhaus-Johnson-Trotter permutations of p-1 integers. Then one can construct the sequence for p integers by inserting p into each permutation in all possible ways, beginning at the right of the first permutation and moving to the left, then sweeping from right to left, left to right, and so forth, always in alternation. This inductive description is relatively transparent to explain, but no induction is needed to actually generate the sequence; code for such generation is available in many places (cf. Niejenhuis and Wilf, p. 59-60)

Table 2 shows how the induction argument works. The first permutation of four elements is 1,2,3,4 and the last is 2,1,3,4. The first column is the Steinhaus-Johnson-Trotter sequence of permutations for three elements, from first to last. The following four sequences in each row are given by inserting the "4" into the initial triplet at the correct places.

Table 2. Steinhaus-Johnson-Trotter permutations of four elements

1,2,3	gives	1,2,3,4	1,2,4,3	1,4,2,3	4,1,2,3
1,3,2	gives	4,1,3,2	1,4,3,2	1,3,4,2	1,3,2,4
3,1,2	gives	3,1,2,4	3,1,4,2	3,4,1,2	4,3,1,2
3,2,1	gives	4,3,2,1	3,4,2,1	3,2,4,1	3,2,1,4
2,3,1	gives	4,2,3,1	2,4,3,1	2,3,4,1	2,3,1,4
2,1,3	gives	4,2,1,3	2,4,1,3	2,1,4,3	2,1,3,4

In practice, one doesn't want to enumerate all of the permutations. Instead, one wants to study the adequacy of the clustering produced by different permutations. Suppose one intends to examine only 100 permutations at first and then search more carefully near those permutations that have the best statistical properties. To do this, one takes d=[p!/100] and then finds the permutation corresponding to rank d, 2d, …, 100d in the sequence. This is the same kind of search strategy used in Section 2.2.

As before, this kind of search requires a quick algorithm for solving the ranking problem. That is, for a given integer between 1 and p!, one needs to be able to quickly find the corresponding permutation in the Gray code list. This is given in many places (cf. Stanton and White, 1986, p. 73). Using this, one can generate the permutations corresponding to, say, ranks d, 2d, …, 100d and assess how well they segregate the variables that determine the cluster structure(s).

To assess the value of a particular sequence of axes, one can use any index that increases with increasing cluster structure on each axis and also increases

with agreement on the cluster membership between adjacent axes. Many indexes are possible, and it is necessary to perform a simulation experiment to determine which ones work best in practice and to provide heuristics for implementing search strategies of the kind described in this section.

4. Multidimensional Scaling

Multidimensional scaling is a procedure that takes a proximity matrix between cases, and then finds locations for the cases in a low-dimensional space that conform, as closely as possible, to the given inter-case proximities. In practice, the proximity matrix is usually measured with error, and one represents the cases by estimating locations in a two-dimensional plot that respect, as closely as possible, the orderings implied by the elements of the proximity matrix. Wish and Carroll (1982) and Kruskal (1975) provide readable introductions to theory behind multidimensional scaling.

To make the multidimensional scaling problem concrete, suppose one has a list of United States cities and knows the time required for driving between each pair of cities. The inter-city proximity matrix contains the drive time, and the multidimensional scaling solution looks very much like the map of the United States (in some random orientation, since compass points are irrelevant to the algorithm). Minor distortions in the map distances are caused when mountain ranges force the roadways to take less direct paths, or when different routes have roads with substantially different speed limits.

Now suppose that instead of driving time one has a matrix of flight time on commercial aviation. For this proximity matrix, the multidimensional scaling solution is highly distorted because United States carriers use the hub-and-spoke system. Passengers traveling from a small city to a distant small city must generally change airplanes at an intermediate large city. As a consequence, the two-dimensional map built from such data tends to place the large cities (i.e., the hub airports) in a central circle, and then the small cities are splayed around on the outer half of a circle centered on their nearest hub airport. The air travel solution tends to have more "stress" than road travel solution, since there is more intrinsic incompatibility within the proximity matrix used for flying.

Stress is an important concept in multidimensional scaling, and it will drive much of the subsequent discussion. Suppose the proximity matrix contains values r_{ij}, those being interpretable as the approximate distance between case i and case j, and let d_{ij} represent the fitted distances between case i and case j in a Euclidean space of fixed dimension. (the r_{ij} will not, in general, satisfy the properties of a true metric, but the d_{ij} must). Let $f(\cdot)$ be the monotone deformation function used by the multidimensional scaling algorithm to bring the proximities as near to the fitted distances as possible. Then the raw stress value of the configuration is given by:

$$\Phi = \varphi \left[d_{ij} - f(r_{ij}) \right]^2$$

where Φ can be viewed as a measure of the lack of fit. There are several other measures that may be used instead of stress, but most of these are quite similar in spirit and typically are based on sums of squares, perhaps weighted or penalized in some way. See Borg and Groenen (1997) for details.

In practice, multidimensional scaling is highly sensitive to outliers. For example, suppose one is building a proximity matrix from driving times, but that along one leg of the trip, the driver has a breakdown and must wait a day while the automobile is fixed. This aberration in the data introduces enormous stress in the multidimensional scaling problem and frequently distorts the answer; it may happen that the resulting low-dimensional map no longer resembles the United States at all, and that crucial interpretability is lost. More commonly, a few large, aberrant values lead researchers to falsely conclude that the appropriate dimension for the multidimensional scaling solution is larger than it should be; the outlier causes the corresponding case to sit on top of a "mountain" in $\mathbb{R}^3$ rather than lie in a planar map in $\mathbb{R}^2$.

To solve the problem of significant distortion introduced by a small number of noisy observations, one must seek a robust solution. In the spirit of *S*-estimators (cf. Rousseeuw and Yohai, 1984), we seek to find a relatively small fraction of the cases which can be excluded from the analysis, but whose exclusion enables much lower-stress solutions. That search entails looking at all subsets of k cases out of n, and then applying the multidimensional scaling algorithm to the $n \,!\, k$ cases retained for analysis. The subset whose exclusion provides the least-stress solution provides a robustification of multidimensional scaling that is resistant to outliers. (A variant on this strategy is not to eliminate cases, but only to eliminate distances; since there are many more inter-case distances than there are cases, this is a harder problem.)

We need to have the analogue of a Gray code for enumerating all possible subsets. It turns out that there are several possibilities (cf. Wilf, 1989, p. 10). But some of these possibilities do not exist for all values of n and k, and others add restrictions (such as monotonicity in the case labels) that do not seem statistically relevant. Therefore we prefer the revolving door algorithm for generating the list A(n,k), which consists of all subsets of size k from a set n, arranged so that only one element changes between each adjacent subset in the sequence.

There is a convenient trick for producing these lists recursively. Given the lists A(n-1,k) and A(n-1,k-1), one can generate A(n,k) by first writing the list A(n-1,k) and then appending, in reverse order, the list A(n-1,k-1) with n added as the last element in each subset. The reader can verify this algorithm in Table 3.

Table 3. Revolving door subsets

Sets of 3 from 5	Sets of 3 from 4	Sets of 2 from 4
A(5,3)	**A(4,3)**	**A(4,2)**

1,2,3	1,2,3	1,2
1,3,4	1,3,4	2,3
2,3,4	2,3,4	1,3
1,2,4	1,2,4	3,4
1,4,5		2,4
2,4,5		1,4
3,4,5		
1,3,5		
2,3,5		
1,2,5		

In order to mimic the search procedures described previously, one needs a solution to the ranking problem for subsets. This enables one to pick and generate the subset having a particular rank on the Gray code list. Unfortunately, the authors are not aware of any solution to this problem in the current literature, but we are not combinatorial mathematicians and we invite correction for others more expert than ourselves.

If no solution to the ranking problem is available, then the strategy can still be successful in small problems, or in problems for which extensive preparatory calculation is possible. This is because there is a fast algorithm to produce the revolving-door sequence that does not rely upon the recursive definition (cf. Nijenhuis and Wilf, 1978, p. 33-35). Therefore one can quickly enumerate the list, stopping only to store subsets that are far apart in the Gray code ordering. Those are the subsets that are used for applying the multidimensional scaling routine, and whose corresponding stresses will be used to identify cases that contribute largely to lack of fit in low dimensions.

Each multidimensional scaling run provides information on the stress associated with a particular subset of the cases. Cases that are always associated with high-stress solutions should be removed as outliers. Cases that give low-stress solutions should be retained. Often one wants to plot a curve of the stress against the number of dimensions in the fit, in order to decide when the scaling algorithm has reached the point of diminishing returns with respect to dimension.

5. Conclusions

This paper argues that combinatorial search issues are playing an increasingly significant role in modern computational statistics. Such searches arise in various ways, including:

- Strategies for model selection, where the number of variables is too large to allow enumeration and testing of each model.

- Strategies for seriation, to enhance visual displays or to group objects according to some measure of similarity.
- Strategies for identifying and removing outliers in complex, computer-intensive analyses.

Besides these cases, as treated in the present paper, it is easy to imagine that one would want to find efficient ways to search the set of rooted binary trees with *n* terminal vertices, for purposes of phylogeny (cf. Gordon, 1986, or Lapointe and Cucumel, 1997) or for CART analysis (cf. Shannon and Banks, 1997) or for cluster analysis (cf. Banks and Constantine, 1998).

Our contribution to the search problem is to note that Gray codes have attractive features as a guide to efficient search. The traditional Gray code was applied to binary sequences, and listed the sequences in such a way that sequences that were near on the list differed from each other only in small, controlled ways. Subsequent researchers have found many other Gray code schemes for different kinds of combinatorial objects, such as permutations, subsets of fixed size, rooted trees, partitions, and so forth. The advantage of the Gray code list is that one can take large steps along the list, extracting a sequence or permutation or subset or tree at each step, and then do the computationally intense analysis on only the extracted objects. Since the objects are well separated in their corresponding Gray codes, then the objects are all mutually very different. This ensures that the search covers the space of possible objects more thoroughly than a simple random search would do.

In order to make this kind of search strategy work, it is very desirable to have a solution to the ranking problem. This means that for a given positive integer, one is able to generate the object that has that rank on the Gray code list. Without a ranking theorem, it is usually necessary to generate the entire Gray code list, and then sort out the objects that are a fixed and large distance apart on the list. But for most problems, full enumeration quickly becomes infeasible, because the number of combinatorial objects tends to grow very quickly.

Our broad strategy is to use the Gray code search to generate the objects, then perform whatever statistical analysis is appropriate for each object, and then compare those results to find out which regions of the Gray code list tend to optimize some criterion, such as goodness-of-fit, or similar cluster structure on adjacent axes, or minimum stress in a two-dimensional scaling representation. The exact approach depends upon the situation, and an iterative search-and-refine exploration is likely to produce better results than a single search.

As the next step in this research program, we intend to perform a number of simulation experiments. These experiments will focus upon the variable selection (or feature selection) problem described in Section 2. Those experiments should provide comparative insight on the value of Gray code search versus selection based upon the results of fractional factorial designs versus selection based upon conventional regression methodologies.

REFERENCES

BANKS, D.L., and CONSTANTINE, G.M. (1998). "Metric models for random graphs," *Journal of Classification*, 15, 199-224.

BANKS, D.L., OLSZEWSKI, R.T., and MAXION, R. (2002). "Comparing methods of multivariate nonparametric regression," to appear in *Communications in Statistics: Simulation and Computation*.

BELLMAN, R.E. (1961). *Adaptive Control Processes*. Princeton University Press, Princeton, N.J.

BOCK, H.-H. (1998). "Clustering and neural networks." In *Advances in Data Science and Classification: Proceedings of the 6th Conference of the International Federation of Classification Societies*, A. Rizzi, M. Vichi, and H.-H. Bock, eds. Springer-Verlag, pp. 265-277.

BORG, I. and GROENEN, P.J.F. (1997). *Modern Multidimensional Scaling*. Springer-Verlag, New York.

BREIMAN, L. and FRIEDMAN, J.H. (1985), "Estimating optimal transformations for multiple regression and correlation," *Journal of the American Statistical Association*, 80, 580-598.

BREIMAN, L., FRIEDMAN, J., OLSHEN, R.A. and STONE, C. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, CA.

CLYDE, M. (1999). "Bayesian model averaging and model search strategies (with discussion)." In *Bayesian Statistics 6.* J.M. Bernardo, A.P. Dawid, J.O. Berger, and A.F.M. Smith Eds. Oxford University Press, pp 157-185.

FOWLKES, E. B., GNANADESIKAN, R., and KETTENRING, J.R. (1988), "Variable selection in clustering," *Journal of Classification*, 5, 205-228.

FRIEDMAN, J.H. (1991). "Multivariate additive regression splines," *Annals of Statistics*, 19, 1-66.

GILBERT, E.N. (1958). "Gray codes and paths on the n-cube," *Bell System Technical Journal*, 37, 815-826.

GORDON, A.D. (1986). "Consensus supertrees: The synthesis of rooted trees containing overlapping sets of labeled leaves," *Journal of Classification*, 3, 335-348.

GRAY, F. (1953). "Pulse code communication." United States Patent 2,632,058, granted March 17, 1953.

HASTIE, T.J., and TIBSHIRANI, R.J. (1990). *Generalized Additive Models*. Chapman and Hall, New York.

INSELBERG, A. (1985). "Intelligent instrumentation and process control," *Proceedings of the Second Conference on Artificial Intelligence*, 302.

INSELBERG, A., and DIMSDALE, B. (1988), "Visualizing multi-dimensional geometry with parallel coordinates," *Computer Science and Statistics: Proceedings of the 20th Symposium on the Interface*, 115-120.

JOHNSON, S.M. (1963). "Generation of permutations by adjacent transpositions." Mathematics of Computation, 17, 282-285.

KRUSKAL, J. B. (1975). "Locally linear and locally isometric mapping." In *Theory, Methods, and Applications of Multidimensional Scaling and Related Techniques*. National Science Foundation, Washington, DC, pp. 27--33.

LAPOINTE, F.-J., and CUCUMEL, G. (1997). "The average consensus procedure: Combination of weighted trees containing identical or overlapping sets of taxa, *Systematic Biology*, 46, 306-312.

LEBART, L. (1998), "Correspondence analysis, discrimination, and neural networks'." In *Data Science, Classification and Related Methods: Proceedings of the Fifth Conference of the International Federation of Classification Societies*, pp. 423-430.

MYERS, R.H. (1999), "Response surface methodology Current status and future directions," *Journal of Quality Technology*, 31, 30-44.

NIJENHUIS, A. and WILF, H. (1978). *Combinatorial Algorithms for Computers and Calculators*, 2nd ed. Academic Press, London, U.K.

PINSKER, I. SH., KIPNIS, V., and GRECHANOVSKY, E. (1987), "The use of conditional cutoffs in a forward selection procedure," *Communications in Statistics, Part A - Theory and Methods*, 16, 2227-2241.

ROUSSEEUW, P. and YOHAI, V. (1984). "Robust regression by means of *S*-estimators," in *Robust and Nonlinear Time Series Analysis*, Lecture Notes in Statistics No. 26, ed. by J. Franke, W. Härdle, and D. Martin. Springer-Verlag, Berlin, pp. 256-272.

SAKATA, S. and WHITE, H. (1998). "Breakdown points," in *Encyclopedia of Statistical Sciences, Update Volume 3*, ed. by S. Kotz, C. B. Read, and D. Banks. John Wiley and Sons, New York, pp. 84-89.

SCOTT, D.W. (1992), *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, New York, NY.

SCOTT, D. and Wand, M.P. (1991). "Feasibility of multivariate density estimates." *Biometrika*, 78, 197-205.

SHANNON, W., and BANKS, D.L. (1997). "An MLE strategy for combining CART models," *Statistics in Medicine*, 29, 540-544.

STANTON, D., and WHITE, S. (1986). *Constructive Combinatorics*. SpringerVerlag, New York, NY.

STEIN, M. (1987), "Large sample properties of simulations using Latin hypercube sampling," *Technometrics*, 29, 143-151.

TROTTER, H.F. (1962). PERM (Algorithm 115). *Communications of the ACM*, 5, 434-435.

WEISBERG, S. (1985). *Applied Linear Regression*. John Wiley & Sons, New York, N.Y.

WESTFALL, P.H., and YOUNG, S.S. (1993). *Resampling-based multiple testing: examples and methods for P-value adjustment.* John Wiley & Sons, New York, N.Y.

WILF, H.S. (1989). Combinatorial Algorithms: An Update. CBMS-NSF 55, Society for Industrial and Applied Mathematics, Philadelphia, PA.

WISH, M., and CARROLL, J.D. (1982), "Multidimensional scaling and its applications", *Handbook of statistics 2: Classification, Pattern Recognition and Reduction of Dimensionality*, 317-45.

WILHELM, A.F.X, WEGMAN, E.J., and SYMANZIK, J (1999). "Visual clustering and classification: The Oronsay particle size data set revisited," *Computational Statistics*, 14, 109-146.

YUEN, C.K. (1974). "The separability of Gray code," IEEE Transactions on Information Theory, p. 688.

STATISTICS IN TRANSITION, October 2002
Vol. 5, No. 5, pp. 823–835

SIMULTANEOUS VISUALIZATION OF TWO TRANSITION TABLES

Michael Greenacre[1] and José G. Clavel[2]

ABSTRACT

The case of two transition matrices of frequencies is considered, that is two square asymmetric matrices of counts where the rows and columns of the matrices are the same objects observed at three different time points. Alternative ways of jointly visualizing the tables are considered. We generalize an existing idea, where a square matrix is decomposed into symmetric and skew-symmetric parts, to the case of two matrices. This approach leads to a decomposition into four components: (1) average symmetric, (2) average skew-symmetric, (3) symmetric difference from average, and (4) skew-symmetric difference from average. The method is illustrated with data from a study of changing values over three generations.

***Keywords*:** Correspondence analysis, matrix decomposition, skew-symmetry, transition matrices.

1. Introduction

We consider the data in Tables 1 and 2 (Garcia *et al.* 1997) concerning the values that three generations recognized as the most important. The data collection took place in the School of Economics at the University of Murcia in Spain. A sample of 129 students was asked to obtain the information from their families. The ten values from which they – and their parents and grandparents – could choose were: honesty, abbreviated as *ho*, family *fa*, culture *cu*, responsibility *re*, happiness *ha*, solidarity *so*, freedom *fr*, loyalty *lo*, industriousness *in* and tolerance *to*. The data are reported in the form of cross-tabulations between grandparents and parents and between parents and students.

[1] Facultat de Ciències Econòmiques i Empresarials ,Universitat Pompeu Fabra, Ramon Trias Fargas, 23–27, E-08005 Barcelona, Spain, e-mail: michael@upf.es.
[2] Facultad de Economía y Empresa, Dpt. Métodos Cuantitativos para la Economía, Universidad de Murcia, Campus de Espinardo, E-30100 Murcia, Spain, e-mail: jjgarvel@um.es.

For example, from Table 1 we can see that, of the 35 grandparents that chose honesty as the most important value, six of the associated parents chose honesty, 9 chose family, 1 chose culture, and so on.

In the case of a single transition table of this kind, Greenacre (2000) proposed separate analyses of the symmetric and skew-symmetric parts of the table, based on an idea of Constantine & Gower (1978) and Gower (1980). He showed how this idea could be implemented in the correspondence analysis framework by setting up a block matrix of the table reproduced twice down the diagonal and in transposed form in the off-diagonal positions. That is, if **N** is the transition matrix, the block matrix analyzed is:

$$\begin{bmatrix} \mathbf{N} & \mathbf{N}^T \\ \mathbf{N}^T & \mathbf{N} \end{bmatrix} \tag{1}$$

As shown by Greenacre (2000), the simple correspondence analysis of this block matrix yields exactly the symmetric and skew-symmetric analyses in one joint analysis, with the principal axes of the two analyses appearing interleaved in the solution and in order of importance.

The purpose of this paper is to go one step further by analyzing two square tables simultaneously in the same style. These two tables could be mobility tables, for example changing professions from grandfather to father and from father to son. Alternatively, the two tables could arise as a subdivision of a table according to a binary variable such as gender (male/female). In the former example we would be interested in comparing the transition in the first generational change with the second, in the latter example we would be interested in comparing the difference in the transitions between males and females.

In Section 2 we give a brief technical summary of correspondence analysis and how it can be applied to the case of a single square matrix as well as to a pair of matrices. In Section 3 we show the results of applying this methodology to the data in Tables 1 and 2. We conclude with some discussion in Section 4.

Table 1. Changing values from grandparents to parents

GP to P	HO	FA	CU	RE	HA	SO	FR	LO	IN	TO	
ho	6	9	1	3	3	5	3	2	1	2	35
fa	9	5	0	4	5	4	0	0	2	3	32
cu	0	0	0	1	1	0	0	0	0	0	2
re	0	0	0	0	1	2	1	0	0	0	4
ha	5	3	1	2	3	1	0	0	0	0	15
so	2	1	0	1	0	0	1	1	0	0	6
fr	3	2	0	0	3	1	3	0	1	1	14
lo	0	0	0	2	2	3	1	0	1	0	9
in	0	0	1	1	0	1	0	3	0	0	6
to	1	0	0	1	1	1	0	2	0	0	6

Table 2. Changing values from parents to students

P to S	HO	FA	CU	RE	HA	SO	FR	LO	IN	TO	
ho	5	0	0	5	9	0	6	1	0	0	26
fa	2	1	0	0	10	1	3	3	0	0	20
cu	0	0	0	0	0	1	2	0	0	0	3
re	0	0	1	0	4	0	6	3	0	1	15
ha	1	1	0	0	12	0	4	1	0	0	19
so	2	0	0	1	4	9	2	0	0	0	18
fr	1	0	1	0	1	0	4	1	0	1	9
lo	0	0	0	0	1	0	3	2	0	2	8
in	0	1	0	0	0	1	2	0	1	0	5
to	1	1	0	1	2	0	1	0	0	0	6

2. Methodology

2.1. Correspondence analysis

Correspondence analysis (CA) can be defined as a method for weighted least-squares approximation of a matrix of counts. In general, suppose that the data matrix **N** has been divided by its grand total n to obtain $\mathbf{P} = \mathbf{N}/n$, called the *correspondence matrix*. Suppose **P** has row and column sums **r** and **c** respectively, and that $\mathbf{D}_r$ and $\mathbf{D}_c$ are diagonal matrices with the elements of **r** and **c** on the diagonal. Thus when **N** is a contingency table, **P** is the (sample) discrete bivariate distribution and **r** and **c** the marginal distributions. CA can be defined as the reduced-rank matrix approximation of **P** by weighted least squares, minimizing the following expression:

$$\text{trace}\left[\mathbf{D}_r^{-1}\left(\mathbf{P}-\hat{\mathbf{P}}\right)\mathbf{D}_c^{-1}\left(\mathbf{P}-\hat{\mathbf{P}}\right)^T\right] = \sum_i\sum_j \frac{\left(p_{ij}-\hat{p}_{ij}\right)^2}{r_i c_j} \qquad (2)$$

for a matrix $\hat{\mathbf{P}}$ of given reduced rank. We know that the best rank 1 approximation is given by $\hat{\mathbf{P}} = \mathbf{r}\mathbf{c}^T$, called the *trivial solution*, so that we can equivalently consider the approximation of the centred matrix $\mathbf{P} - \mathbf{r}\mathbf{c}^T$. The solution for any low rank is given by the singular value decomposition (SVD) of the matrix of standardized residuals $\mathbf{D}_r^{-1/2}\left(\mathbf{P} - \mathbf{r}\mathbf{c}^T\right)\mathbf{D}_c^{-1/2}$ (see, for example, Blasius and Greenacre, 1994):

$$\mathbf{D}_r^{-1/2}\left(\mathbf{P} - \mathbf{r}\mathbf{c}^T\right)\mathbf{D}_c^{-1/2} = \mathbf{U}\mathbf{D}_a\mathbf{V}^t \qquad \text{where } \mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I} \quad (3)$$

For constructing CA maps, the *principal coordinates* of the row and column points are given by $\mathbf{F} = \mathbf{D}_r^{-1/2}\mathbf{U}\mathbf{D}_a$ and $\mathbf{G} = \mathbf{D}_c^{-1/2}\mathbf{V}\mathbf{D}_a$ respectively. For example, to plot the rows and columns in two dimensions, the rank 2 solution

given by the first two columns of **F** and **G** are used. The resulting plot is called the *symmetric map*, as opposed to other so-called asymmetric maps, described in the following (see also Greenacre 1984, 1993).

An alternative way of defining CA is as a weighted least-squares approximation of the row or column *profiles* of the table. A profile is a row or column of the matrix divided by its corresponding sum. For example, the row profiles are the rows of the matrix $\mathbf{D}_r^{-1}\mathbf{P}$, in which case CA can be defined as the approximation of the row profiles by points in a low-dimensional subspace. Distances and scalar products in the space are computed using the *chi-square metric*, a weighted Euclidean metric using $\mathbf{D}_c^{-1}$ as the weighting matrix. Furthermore, the row profiles are weighted by the respective elements of **r**, called the *row masses*. The objective function in this case is:

$$\text{trace}\left[\mathbf{D}_r\left(\mathbf{D}_r^{-1}\mathbf{P}-\hat{\mathbf{Q}}\right)\mathbf{D}_c^{-1}\left(\mathbf{D}_r^{-1}\mathbf{P}-\hat{\mathbf{Q}}\right)^T\right]=\sum_i r_i\sum_j\frac{\left(p_{ij}/r_i-\hat{q}_{ij}\right)^2}{c_j} \qquad (4)$$

Again we have a trivial solution because it turns out that the row vector $\mathbf{c}^T$ comes closest to all the row profiles in terms of weighted least sum-of-squared distances, so that it is equivalent to approximate the centred profiles $\mathbf{D}_r^{-1}\mathbf{P}-\mathbf{1}\mathbf{c}^T$. Again this problem is solved using the SVD of the matrix $\mathbf{D}_r^{1/2}(\mathbf{D}_r^{-1}\mathbf{P}-\mathbf{1}\mathbf{c}^T)\mathbf{D}_c^{-1/2}$, which is identical to the matrix of standardized residuals decomposed previously, so the solution is as before. Because we think of the matrix as a set of rows, it is often convenient to visualize the results using the *asymmetric map*. In this case the row profiles would be plotted using principal coordinates **F** as before, but the columns would be in rescaled coordinates called *standard coordinates* $\mathbf{Y} = \mathbf{D}_c^{-1/2}\mathbf{V} = \mathbf{G}\mathbf{D}_a^{-1}$. These column points are the projections of the unit vectors onto the optimal subspace, and together **F** and **Y** constitute a biplot of the frequency table (Greenacre, 1992).

In both definitions of CA the *total inertia* of the table, a measure of the table's total variation, is equal to the weighted sum-of-squares of the centred matrix being approximated:

$$\text{total inertia}=\sum_i\sum_j\left(p_{ij}-r_ic_j\right)^2/\left(r_ic_j\right) \qquad (5)$$

The inertia accounted for by the rank K^* solution (or K^* -dimensional solution) is equal to the weighted sum-of-squares of the matrix approximation, which is equal to $\sum_{k=1}^{K^*}a_k^2$. The minimum value of (2) (or (4)), which is the residual inertia not accounted for, is equal to the remaining sum-of-squared singular values: $\sum_{k=K^*+1}^{K}a_k^2$.

The special case considered here is the application of CA to square tables where the rows and columns refer to the same set of objects. For a transition table

where the rows refer to the first time point and the columns the second (e.g., father and son), the row profiles are the relative frequencies of change from time one to time two.

2.2. Analysing concatenated or stacked tables

One way of visualizing two data matrices jointly, is to concatenate the matrices side by side, or to stack one on top of another. Since our interpretation of the transition matrices is in terms of row profiles, the preferred way of combining the tables in this case would be to stack them. CA is then applied to the concatenated table (Blasius 1994, Greenacre 1994). A common problem with the CA of such tables is, as pointed out by Greenacre (2000), the predominant role played by the diagonal of the table, in fact by the *symmetric part* of the table as a whole.

2.3. Decomposition into symmetric and skew-symmetric components

Greenacre (2000) adapted the ideas of Constantine & Gower (1978) and Gower (1980) to the decomposition of the correspondence matrix into symmetric and skew-symmetric components respectively, as follows:

$$\mathbf{P} = \mathbf{S} + \mathbf{T} \tag{6}$$

Tyo solve the centring problem, Greenacre considered the average of the row and column margins $\mathbf{w} = \frac{1}{2}(\mathbf{r} + \mathbf{c})$ as the centre, so that the decomposition is actually:

$$\mathbf{P} - \mathbf{w}\mathbf{w}^T = \mathbf{S} - \mathbf{w}\mathbf{w}^T + \mathbf{T} \tag{7}$$

with the metric also being defined as $\mathbf{D}_w^{-1}$. The corresponding decomposition of inertia is thus

$$\sum_i \sum_j \left(p_{ij} - w_i w_j\right)^2 / \left(w_i w_j\right) = \sum_i \sum_j \left(s_{ij} - w_i w_j\right)^2 / \left(w_i w_j\right) + \sum_i \sum_j t_{ij}^2 \left(w_i w_j\right) \tag{8}$$

A convenient way to perform the **CA** on the separate matrix components was shown to be the simple **CA** algorithm applied to the block matrix:

$$\tilde{\mathbf{N}} = \begin{bmatrix} \mathbf{N} & \mathbf{N}^T \\ \mathbf{N}^T & \mathbf{N} \end{bmatrix} \tag{9}$$

For a $p \times p$ matrix **N**, the CA of the block matrix yields $2p - 1$ dimensions, $p - 1$ of which coincide with the symmetric part of **N**. These dimensions correspond to coordinate matrices which have vectors of coordinates reproduced twice, i.e. of the form:

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f} \end{bmatrix}$$

Clearly only one block of coordinates needs to be used for plotting, so that there is one set of points displayed.

The other p dimensions (or $p - 1$ if p is an odd number) correspond to pairs of equal singular values (i.e., equal principal inertias), and are the so-called *bimensions* of the skew-symmetric component, with pairs of row and column principal coordinate vectors of the form:

$$\text{rows: } \begin{bmatrix} \mathbf{f}_1 & \mathbf{f}_2 \\ \mathbf{f}_1 & -\mathbf{f}_2 \end{bmatrix}, \text{ columns : } \begin{bmatrix} \mathbf{f}_2 & -\mathbf{f}_1 \\ \mathbf{f}_2 & \mathbf{f}_1 \end{bmatrix}$$

Each block of coordinates is a 90 degree rotation of the other. To apply the usual rules of interpretation it would be necessary to plot row and column points. But, as before, only one set of points needs to be plotted, say the first block of row points, since the 90 degree relationship means that we can interpret the areas of triangles as estimates of the elements of the skew-symmetric matrix (see Greenacre, 2000, and the application in Section 3).

In practice we map the first two dimensions of the symmetric part and the first two dimensions of the skew-symmetric part, in separate maps. Thus in terms of parsimony, the symmetric/skew-symmetric decomposition has the same number of displayed points as an ordinary CA of the original table which would have two sets of points in one map.

2.4. A pair of matched transition tables

We can generalize the above ideas to the case of two transition tables which are matched in the sense that they have the same row and column labels, as in the present application. The idea is as follows. Suppose the two tables are **M** and **N** respectively, each $p \times p$ and each crosstabulating the same set of n individuals. We set up the following $4p \times 4p$ block matrix:

$$\begin{bmatrix} \mathbf{M} & \mathbf{N} & \mathbf{M}^T & \mathbf{N}^T \\ \mathbf{N} & \mathbf{M} & \mathbf{N}^T & \mathbf{M}^T \\ \mathbf{M}^T & \mathbf{N}^T & \mathbf{M} & \mathbf{N} \\ \mathbf{N}^T & \mathbf{M}^T & \mathbf{N} & \mathbf{M} \end{bmatrix} \tag{10}$$

This matrix has a 2×2 block pattern, where the matrix

$$\begin{bmatrix} \mathbf{M} & \mathbf{N} \\ \mathbf{N} & \mathbf{M} \end{bmatrix}$$

plays the role of the single matrix we had before in (9). So the CA of (10) yields the analyses of the symmetric part:

$$\begin{bmatrix} \mathbf{M} & \mathbf{N} \\ \mathbf{N} & \mathbf{M} \end{bmatrix} + \begin{bmatrix} \mathbf{M}^T & \mathbf{N}^T \\ \mathbf{N}^T & \mathbf{M}^T \end{bmatrix} = \begin{bmatrix} \mathbf{M}+\mathbf{M}^T & \mathbf{N}+\mathbf{N}^T \\ \mathbf{N}+\mathbf{N}^T & \mathbf{M}+\mathbf{M}^T \end{bmatrix} \tag{11}$$

and the skew-symmetric part:

$$\begin{bmatrix} \mathbf{M} & \mathbf{N} \\ \mathbf{N} & \mathbf{M} \end{bmatrix} - \begin{bmatrix} \mathbf{M}^T & \mathbf{N}^T \\ \mathbf{N}^T & \mathbf{M}^T \end{bmatrix} = \begin{bmatrix} \mathbf{M}-\mathbf{M}^T & \mathbf{N}-\mathbf{N}^T \\ \mathbf{N}-\mathbf{N}^T & \mathbf{M}-\mathbf{M}^T \end{bmatrix} \tag{12}$$

(in both cases we omit the division by 2 of the sum and the difference, which in any case does not affect the correspondence analysis).

Now (11) and (12) are themselves block matrices, and Greenacre (2001) has shown the similar, but more general, result that the analysis of any block matrix of the form:

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B} & \mathbf{A} \end{bmatrix}$$

yields an analysis of the sum $\mathbf{A} + \mathbf{B}$ and the difference $\mathbf{A} - \mathbf{B}$. Hence the analyses of (11) and (12) yield two components each:

$$(\mathbf{M} + \mathbf{M}^T) + (\mathbf{N} + \mathbf{N}^T) \text{ and } (\mathbf{M} + \mathbf{M}^T) - (\mathbf{N} + \mathbf{N}^T)$$

and

$$(\mathbf{M} - \mathbf{M}^T) + (\mathbf{N} - \mathbf{N}^T) \text{ and } (\mathbf{M} - \mathbf{M}^T) - (\mathbf{N} - \mathbf{N}^T)$$

which gives a total of four components altogether:

1. $(\mathbf{M} + \mathbf{M}^T) + (\mathbf{N} + \mathbf{N}^T)$ (average symmetric part)
2. $(\mathbf{M} + \mathbf{M}^T) - (\mathbf{N} + \mathbf{N}^T)$ (difference in the symmetric parts)
3. $(\mathbf{M} - \mathbf{M}^T) + (\mathbf{N} - \mathbf{N}^T)$ (average skew-symmetric part)
4. $(\mathbf{M} - \mathbf{M}^T) - (\mathbf{N} - \mathbf{N}^T)$ (difference in the skew-symmetric parts)

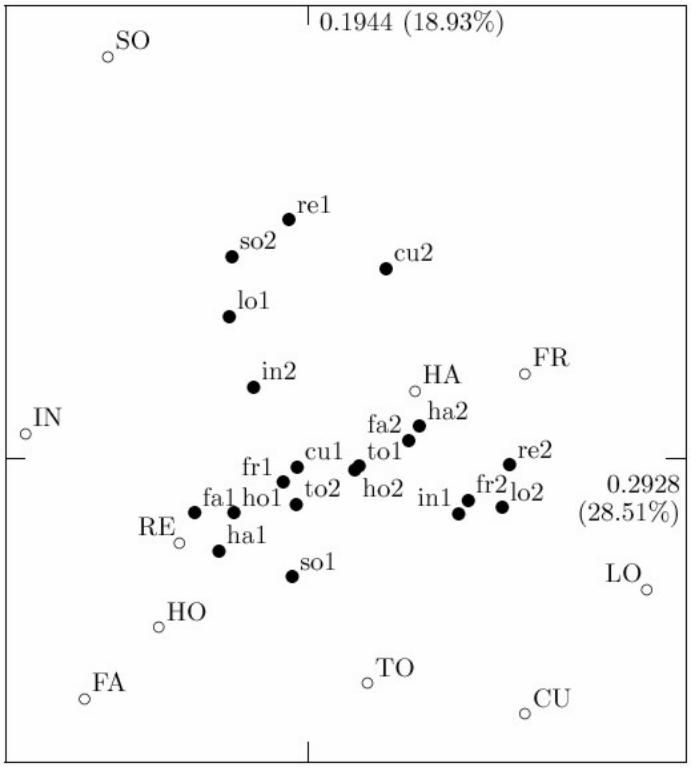
The application will demonstrate how these components appear in the solution in practice.

3. Results

We now look at the results of this methodology applied to the data of Tables 1 and 2. Both transition tables are rather sparse because of the low sample size relative to the number of cells in the tables, so we should be careful in checking that the features that we observe in the maps actually exist in the original data. One way of ensuring correct interpretation of the maps is to look at the contributions to the inertia of each axis, and to restrict our interpretation to those points which make large contributions.

The CA of the concatenated tables is shown in Figure 1. Rows from the first transition are indicated by 1 and from the second by 2. There are three important groups of "receiving" values: solidarity (SO) at top left, friendship (FR) and loyalty (LO) on the right, and family (FA) and honesty (HO) at bottom left. Many of the changes between the first to the second transition are towards the right of the display, that is towards the "receiving" values of friendship and loyalty, in particular the values fr, lo, re, fa all move from the left, often from top left, towards these values. This means that in the parent–to–student transition there are changes from other values towards friendship and loyalty. The points so1 and so2 show a change towards the solidarity value, that is, grandparents with solidarity values changed to other values, whereas parents with solidarity value have many children with this value too.

Figure 1. Asymmetric map of stacked table of both transition



The results of the joint analysis of the two tables are given in Table 3, and two of the four possible maps are given in Figures 2 and 3 as examples of the interpretation. First notice the patterns of the coordinates shown in Table 3. The classification is decided as follows: apply the signs to the matrices **M**, **N**, $\mathbf{M}^T$, $\mathbf{N}^T$ (in the order as they appear in the first row or column of the block matrix). Thus

dimensions 1 and 3, for example, with sign pattern + – + –, correspond to the component $\mathbf{M} - \mathbf{N} + \mathbf{M}^T - \mathbf{N}^T = (\mathbf{M} + \mathbf{M}^T) - (\mathbf{N} + \mathbf{N}^T)$, that is the difference in the symmetric parts. Accumulating inertias and percentages of inertia for each of the components we obtain in decreasing order: difference in symmetric parts 0.4746 (35.6%), difference in skew-symmetric parts 0.3407 (25.6%), average of symmetric parts 0.2741 (20.6%) and average of skew-symmetric parts 0.2425 (18.2%).

Figure 2. Dimensions 1 and 3 of block tables of both transitions: difference between the symmetric parts

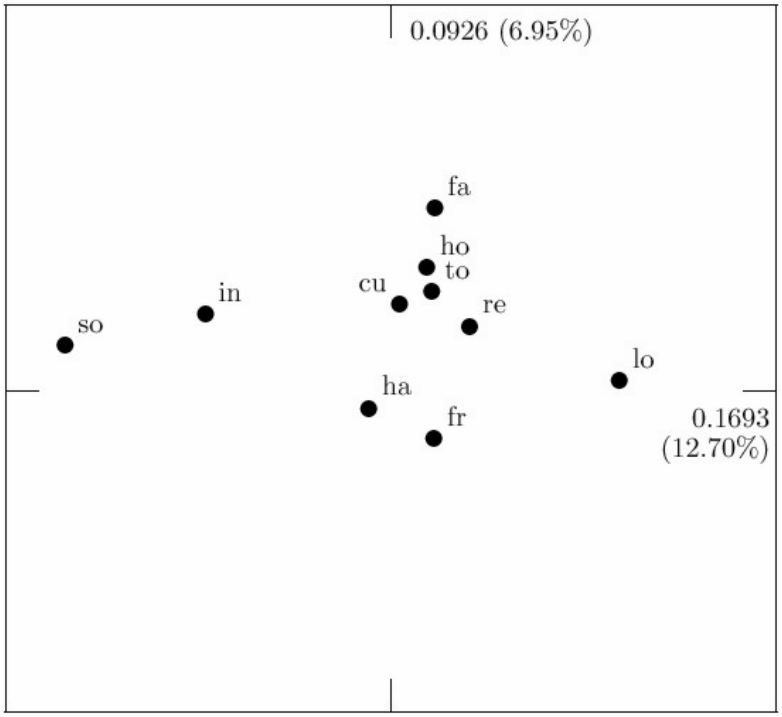


Figure 2 shows dimensions 1 and 3, accounting for an inertia of 0.2619, which is 55.2% of the inertia of the difference in symmetric parts (the percentages shown in Figure 2 are relative to the total inertia of the four components). Figure 3 shows dimensions 4 and 5, accounting for an inertia of 0.1778, which is 64.9% of the inertia of the average skew-symmetric part.

Figure 3. Dimensions 4 and 5 of block tables of both transitions: average skew-symmetric part

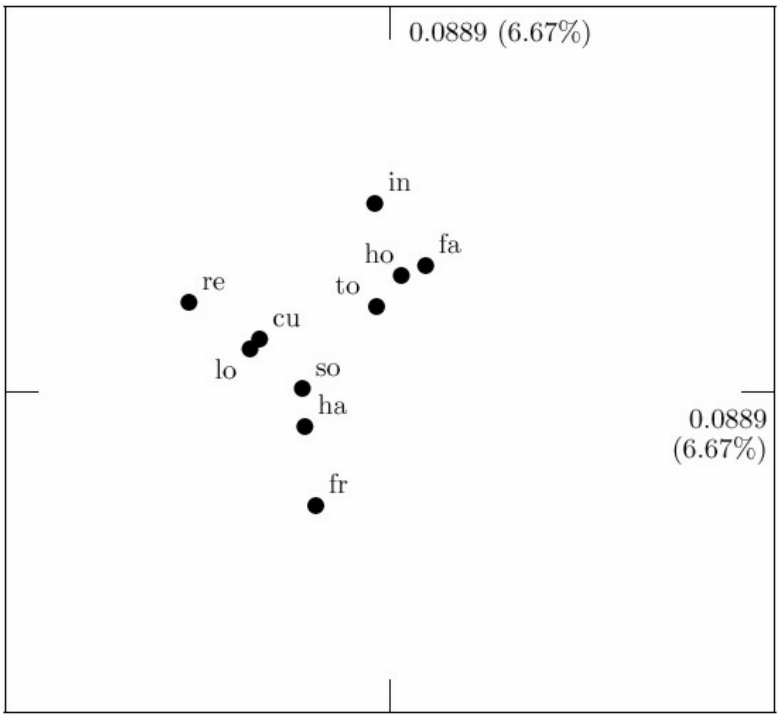


Table 3. Percentages of inertia and patterns of coordinates (ave S = average of symmetric parts, diff S = difference between symmetric parts, ave SS = average of skew-symmetric parts, diff SS = difference between skew-symmetric parts)

Dimension	Inertia	Percentage	Pattern	Classification
1	0,16928	12,70%	+ – + –	diff S
2	0,11057	8,30%	+ + + +	ave S
3	0,09262	6,95%	+ – + –	diff S
4	0,08886	6,67%	+ + – –	ave SS
5	0,08886	6,67%	+ + – –	ave SS
6	0,08147	6,11%	+ – – +	diff SS
7	0,08147	6,11%	+ – – +	diff SS
8	0,07913	5,94%	+ – + –	diff S
9	0,06430	4,83%	+ + + +	ave S
10	0,06286	4,72%	+ – – +	diff SS
11	0,06286	4,72%	+ – – +	diff SS
12	0,04614	3,46%	+ – + –	diff S
13	0,03335	2,50%	+ – + –	diff S
14	0,03311	2,48%	+ + + +	ave S
15	0,03235	2,43%	+ – + –	diff S
16	0,02258	1,69%	+ + + +	ave S

Dimension	Inertia	Percentage	Pattern	Classification
17	0,02131	1,60%	+ + – –	ave SS
18	0,02131	1,60%	+ + – –	ave SS
19	0,02124	1,59%	+ + + +	ave S
20	0,01755	1,32%	+ – – +	diff SS
21	0,01755	1,32%	+ – – +	diff SS
22	0,01637	1,23%	+ – + –	diff S
23	0,01050	0,79%	+ + – –	ave SS
24	0,01050	0,79%	+ + – –	ave SS
25	0,00958	0,72%	+ + + +	ave S
26	0,00845	0,63%	+ – – +	diff SS
27	0,00845	0,63%	+ – – +	diff SS
28	0,00623	0,47%	+ + + +	ave S
29	0,00525	0,39%	+ + + +	ave S
30	0,00478	0,36%	+ – + –	diff S
31	0,00184	0,14%	+ + + +	ave S
32	0,00059	0,04%	+ + – –	ave SS
33	0,00059	0,04%	+ + – –	ave SS
34	0,00035	0,03%	+ – + –	diff S
35	0,00021	0,02%	+ – + –	diff S
36	0,00002	0,00%	+ – – +	diff SS
37	0,00002	0,00%	+ – – +	diff SS
38	0,00000	0,00%	+ + – –	ave SS
39	0,00000	0,00%	+ + – –	ave SS

The interpretation of these maps is not easy, since each has a different style of interpretation. In Figure 2 we are interpreting differences between symmetric parts of the two matrices. The fact that loyalty and solidarity oppose each other corresponds to the “popposite” movements of these two values in Figure 1, each of these values is being reinforced from one transition to the next. In Figure 3, it is areas of triangles that have to be interpreted and the direction of flow is in this case anti-clockwise. For example, the positions of industriousness and freedom are far from the origin and make triangles with large areas with several other values (the areas are of triangles formed by pairs of points and the origin). Their positions show that flows are taking place away from industriousness (e.g., to loyalty), and from some values (e.g., responsibility) to freedom, which can be checked in the data. The frequencies are, however, quite low in general in these tables and the features being visualized are admittedly subtle.

4. Discussion and conclusions

We have shown two ways of tackling the analysis and visualization of a pair of transition tables. When the tables are dominated by their symmetric parts, then

the decomposition into symmetric and skew-symmetric parts makes sense, so that we can interpret the skew-symmetry without the overwhelming influence of the symmetric part.

In the joint analysis of two tables, there are four components of interest: the average symmetric part, the average skew-symmetric part, the difference between the symmetric parts and the difference between the two skew-symmetric parts. This strategy functions optimally when the individual symmetric parts are strong, but not necessarily similar. The total inertia of the two tables is distributed over the four components and this can be useful in quantifying the amount of variance attributable to these four sources. The analysis may be executed by applying a regular correspondence analysis to the matrices set up in a block table where the table and its transpose are included four times in a pattern where no matrix is repeated twice in the same row or column. The map of each component involves only one set of points at a time, with the exception of the difference in symmetric parts where inverse dimensions are possible and where the row and column points are reflections of each other, thereby allowing reconstruction of negative differences on the diagonal of the symmetric matrix.

An alternative and more common way to analyze the tables jointly is a simple stacking of the tables, leading to two points for each row and one point for each column. Our initial experience with these different approaches is that the four-component approach is useful when the symmetric parts of the tables are strong. In all the usual approaches the deviations from symmetry, which are the interesting flows in the tables, would be masked and difficult to see. On the other hand, the four-component model is more complicated to interpret and different rules of interpretation apply to the maps of different components. Applying this method to the real data in this paper, where the tables were quite sparse in data, was particularly difficult. The more regular approach of stacking the tables is useful when the overall differences between the tables is strong, but will not depict the flows accurately in a planar map when the tables have strong symmetric components.

Acknowledgements

This research is partly supported by Spanish Ministry of Science and Technology, grant BFM 2000-1064.

REFERENCES

BLASIUS, J. (1994), “Correspondence analysis in social science research,” in M.J. Greenacre and J. Blasius (eds.), *Correspondence Analysis in the Social Sciences*. London: Academic Press, pp. 23-52.

BLASIUS, J. & GREENACRE, M.J. (1994), "Computation of correspondence analysis," in M.J. Greenacre and J. Blasius (ed.), *Correspondence Analysis in the Social Sciences*. London: Academic Press, pp. 53-78.

CONSTANTINE, A.G. and GOWER, J.C. (1978), "Graphical representation of asymmetry," *Applied Statistics*, 27, 297-304.

GARCÍA, J.G., FRUTOS, I.P. & CLAVEL, J.G. (1997), "Una cadena de Markov para el estudio de los valores sociales," *Cuadernos de Realidades Sociales*, 49-50, 207-229.

GOWER, J.C. (1980), "Problems in interpreting asymmetrical relationships," in F.A. Bisby, J.G. Vaughan and C.A. Wright (eds.), *Chemosystematics: Principles and Practice*. London: Academic Press, pp. 399-409.

GREENACRE, M.J. (1984), *Theory and Applications of Correspondence Analysis*, London: Academic Press.

GREENACRE, M.J. (1992), "Biplots in correspondence analysis," *Journal of Applied Statistics*, 20, 251-269.

GREENACRE, M.J. (1993), *Correspondence Analysis in Practice*, London: Academic Press.

GREENACRE, M.J. (1994), "Multiple and joint correspondence analysis," in M.J. Greenacre and J. Blasius (ed.), *Correspondence Analysis in the Social Sciences*. London: Academic Press, pp. 141-161.

GREENACRE, M.J. (2000), "Correspondence analysis of square asymmetric matrices," *Applied Statistics*, 49, 297-310.

GREENACRE, M.J. (2001), "Analysis of Matched Matrices," Working Paper 539, Department of Economics and Business, Universitat Pompeu Fabra, Barcelona, downloadable at http://www.econ.upf.es/cgi-bin/onepaper?539, submitted for publication.

STATISTICS IN TRANSITION, October 2002
Vol. 5, No. 5, pp. 837–842

WHAT IS DATA MINING?

Eugeniusz Gatnar[1]

ABSTRACT

The paper is devoted to discussion of the notion of Data Mining. The idea of finding patterns in large data sets has been given various names, but the one of Mining in Data became the most popular. Data Mining uses statistical algorithms to discover patterns and regularities (or "knowledge") in data, therefore its relation to statistics is also presented and some methods are discussed. Then several interesting applications of Data Mining methods are presented and at last two still open problems are pointed out.

Keywords: Data Mining, Statistics, Exploratory data analysis, Statistical learning.

1. Introduction

The fast growth of amount of data stored in easy-to-access databases and data warehouses has created a need for a new generation of tools for automated and intelligent database management (Gatnar, 1996). The notion of finding useful patterns in data has been given various names, including knowledge discovery in databases, Data Mining, knowledge extraction, information discovery, information harvesting, etc.

The term "Data Mining" is mostly used by statisticians and data analysts and refers to the application of algorithms for extracting patterns from data or learning from data. The goal of this learning is to understand what the data says.

To infer information from a database two techniques are used:

- *deduction* – when the information is a logical consequence of the data in the database,
- *induction* – when the information is generalised from the data in the database. The general statements about properties of observed objects are called "knowledge".

[1] Institute of Statistics, Katowice University of Economics, ul. Bogucicka 14, 40-226 Katowice, Poland.

The most important for the induction is the selection of the most plausible rules and regularities. The regularities are represented by a simplification of the system described by the data are called "a model".

The creation of such a model is *inductive learning* and the automation of inductive learning processes has been researched in Machine Learning, the subfield of Artificial Intelligence. Data Mining is a special case of Machine Learning where the system is observed through a large database.

On the other hand, Data Mining is often seen as the next step beyond online analytical processing (OLAP) and the next step beyond exploratory data analysis (EDA).

But what is, in fact, Data Mining?

The simplest definition says that Data Mining uses statistical algorithms to discover patterns in data. But many other definitions can be found in the literature, e.g.

- *Fayyad*: "Data Mining is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data".
- *Zekulin*: "Data Mining is the process of extracting previously unknown, comprehensible, and actionable information from large databases and using it to make crucial business decisions".
- *Ferruzza*: "Data Mining is a set of methods used in the knowledge discovery process to distinguish previously unknown relationships and patterns within data".
- *John*: "Data mining is the process of discovering advantageous patterns in data".

2. Data Mining methods

Most Data Mining methods are based on concepts from Machine Learning, pattern recognition and statistics. Their goal is to make prediction or/and give description. Prediction involves using some variables to predict unknown values (e.g. future values) of other variables while description focuses on finding interpretable patterns describing the data.

Data Mining uses methods that can sift through the data in search of frequently occurring patterns, can detect trends, produce generalisations about the data, etc. These tools can discover these types of information with very little guidance from the user.

Data Mining has mostly at least three major components: *classification*, *association rules* and *sequence analysis*.

In *classification* a database is analysed and a set of rules which can be used to classify future data is generated. It allows to find rules that partition the data into several predefined classes.

An *association rule* is a rule that implies certain association relationships among a set of objects in a database. In this process association rules at multiple

levels of abstraction from the relevant set(s) of data in a database are discovered. Mining association rules may require searching large relational database that is quite costly in processing.

In *sequential analysis*, patterns that occur in sequence are discovered. This deals with data that appear in separate transactions. For example: if a customer buys item X in the first week of the month, then he buys item Y in the second week etc.

3. Data Mining products

As Friedman (1997) pointed out, perhaps the largest profits are made by selling tools to the data miners, rather than in doing the actual mining. This is because very large databases must be stored and quickly accessed, and computationally intensive methodology is applied to these data. This requires massive amounts of disk space and fast computers with large RAM memories. Therefore Data Mining opens new markets for such hardware.

Examples of some current Data Mining products are given in table 1.

Table 1. Examples of Data Mining software packages

Company	**Product**
IBM	Intelligent Miner
Tandem	Relational Data Miner
Angoss Software	***KnowledgeSEEKER***
Thinking Machines Corporation	***DarwinTM***
NeoVista Software	ASIC
ISL Decision Systems, Inc.	Clementine
DataMind Corporation	DataMind Data Cruncher
Silicon Graphics	MineSet
California Scientiřc Software	BrainMaker
WizSoft Corporation	WizWhy
Lockheed Corporation	Recon
SAS Corporation	SAS Enterprise Miner

4. Data Mining and statistics

The idea of *learning from data* has been around for a long time. This is also the core idea in statistics (statistical learning). The learning problem can be considered as either supervised or unsupervised.

In the supervised learning the goal is to predict the value of the variable “y” (outcome), based on a number of other variables (predictors). As a result, the prediction model or learner is built. It will predict the outcome for new, unseen objects.

In the unsupervised learning no outcome variable is available. Therefore its goal is to describe associations and patterns among the data.

Although Data Mining has its origins outside statistics it uses many statistical procedures, for example: classification and regression trees (CART, CHAID), rule induction (AQ, CN2), nearest neighbors, clustering methods, association rules, feature extraction, data visualisation, etc.

Some Data Mining software (table 1) include also procedures of neural networks, bayesian belief networks, genetic algorithms, self-organizing maps, and neuro-fuzzy systems.

Unfortunately, many statistical methods as: hypothesis testing, ANOVA, MANOVA, linear regression, discriminant analysis, logistic regression, GLM, canonical correlation, principal component analysis and factor analysis are not offered by almost any of the Data Mining packages presented in table 1.

The sampling methodology is not used in Data Mining applications either, although it could improve accuracy while mitigating computational requirements. Computationally intense procedure operating on a subset of the data may in fact improve accuracy better than a less sophisticated one using the entire database.

5. Why use Data Mining ?

Data Mining is quickly becoming a necessity, especially for those who must analyse data warehouses containing hundreds of gigabytes or terabytes of information.

Banks use Data Mining to identify their most profitable credit-card customers or their highest-risk loan applicants. They also seek to prevent fraud by using a technique called "deviation detection", which finds events that are outside the norm.

Some companies use Data Mining to study how to retain customers, separate profitable customers from unprofitable ones, uncover fraud, sell existing customers new products, and understand why some customers leave. To find a model to identify who are the most profitable customers and to see what the impact would be if they lost those customers (Groth, 1998).

Data Mining techniques help companies, particularly those in banking and finance, to build an accurate customer profile based on consumer behaviour. For example, if a customer uses automated teller machines (ATM) more often than going to a bank and doing transactions with a teller, the bank may offer her/him more ATM services or offer the bank's online service (Berry and Linoff, 1997).

Many successful Data Mining applications are well known. The following list presents some of them:

- identifying buying behaviour patterns from customers,
- finding associations among customer demographic characteristics,
- predicting which customers will respond to mailing,

- identifying profitable and "loyal" customers,
- detecting patterns of fraudulent credit card usage,
- predicting customers that are likely to change their credit card affiliation,
- determining credit card spending by customer groups,
- identifying stocks trading rules from historical market data,
- increasing Web site profitability,
- increasing store traffic and optimising layouts for increased sales,
- determining the distribution schedules among outlets,
- analysing loading patterns,
- determining which medical procedures are claimed together,
- characterising patient behaviour to predict office visits,
- identifying successful medical therapies for different illnesses,

and many more ...

6. Conclusions

Data Mining enables to discover hidden patterns and relationships in large amounts of data. It solves a paradox: the more data you have, the more difficult and time-consuming it is to analyse and draw meaning from it.

Data Mining uses powerful statistical techniques to quickly explore millions of data records, identifying the most valuable and usable information, etc.

As pointed out above, most of the tools and techniques used for Data Mining came from pattern recognition, statistics and database management theory. But there are still two unsolved problems:

- most of the traditional Data Mining techniques failed because of the size of the database. New techniques will have to be developed to process huge data sets. Some of the newly proposed parallel algorithms are now beginning to look into this.
- Data Mining algorithms assume the data to be noise-free. As a result, the most time consuming part of the data analysis becomes data pre-processing. Noisy data and pre-processing this data may take up more time than the statistical algorithm execution time.

REFERENCES

BERRY, M.J. and LINOFF, G. (1997): *Data Mining Techniques*. John Wiley & Sons, New York.

FRIEDMAN, J. H. (1997): Data Mining and Statistics: What's the Connection ? *The 29th Symposium on the Interface*, Houston, TX.

GATNAR, E. (1996): Statistical Software in Transition to Inteligent Systems. *Statistics in Transition, 2*, No 7, 1165-1174.

GATNAR, E. (1997): Data Mining and statistical data analysis. *Statistical Revue, 2*, 309-316 (in Polish).

GROTH, R. (1998): *Data Mining*. Prentice Hall, Upper Saddle River, NJ.

STATISTICS IN TRANSITION, October 2002
Vol. 5, No. 5, pp. 843–856

IMPACT OF LATENT CLASS CLUSTERING OF NSF DOCTORAL SURVEY DATA ON ADJUSTED RAND INDEX VALUES

Michael D. Larsen[1]

ABSTRACT

Latent class analysis is used to form clusters based on multivariate categorical data. The adjusted Rand Index is used to compare the degree to which the resulting clusters correspond to the separation of subjects into females and males. The data considered here are from the U.S. National Science Foundation's 1997 Survey of Doctoral Recipients. The subset of respondents studied received Ph.D.'s between 1990 and 1996 in either than physical or biological sciences or in engineering and work at higher educational institutions. Latent class analysis identifies interesting subgroups of women and men based simultaneously on several categorical variables related to limitations on searching for a career path job, work activities, and family and career status. Simulation is used to evaluate the sampling distribution of the adjusted Rand Index. The latent class cluster solutions, although interesting, generally do not increase the values of the adjusted Rand Index.

Keywords: BIC; Classification; Comparing Partitions; Exploratory Data Analysis; External Criterion; Information Technology.

1. Introduction

The National Science Foundation's Scientists and Engineers Statistical Data System (SESTAT) database (NSF 99-337) is created from biennial surveys that are representative of U.S. scientists and engineers. The surveys are large and of high quality (see NSF 99-337 for coverage limitations). Several qualitative

[1] University of Chicago, USA.

questions focus on desired work activities, adequacy of doctoral training, job search resources, limitations on job search, and work activities.

Latent class analysis is useful when the population under study is considered to be composed of distinct subpopulations, but the identification of members of these populations is difficult or not possible. If the classes have sufficiently different response patterns, then the classes can be identified by their members' patterns of responses. The classes then can be compared across groups in terms of known, demographic divisions such as sex. Latent class analysis will be applied using attitudinal and other variables reported by 1997 Survey of Doctoral Recipients (SDR) respondents who received PhDs in physical and biological sciences and engineering between 1990 and 1996 and work at educational institutions.

Detailed information on educational choices (e.g., Etzkowitz et al 1994; Farmer et al 1999; AAUW 2000) and hiring decisions (e.g., Davison and Burke 2000; Darity and Mason 1998; Top 1991) are not available in SESTAT. These issues are not examined here.

The adjusted Rand Index (ARI; Hubert and Arabie 1985) is one measure of the correspondence between two partitions of a finite set of objects. This index is calculated for various sets of partitions of the data. The first partitions are created by responses to individual categorical variables and by sex of the respondents. The second partitions are created by latent class clusters and again by sex of the respondents. The observed value of the ARI is compared to a simulated sampling distribution in order to assess the extremeness of its value.

Section 2 presents basic ideas of latent class analysis. Section 3 reviews that adjusted Rand Index for comparing partitions and defines the simulation procedure. Section 4 compares women and men on variables used in the latent class analysis. Section 5 describes latent class results and the impact of clustering on the adjusted Rand Index. Section 6 is a summary.

2. Latent Class Analysis

An observation y_i (possibly multivariate) from a finite mixture distribution with G classes has probability density

$$p(y_i \mid \pi, \vartheta) = \sum_{g=1}^{G} \pi_g \, p_g(y_i \mid \vartheta_g), \qquad (1)$$

where π_g ($\sum_{g=1}^{G} \pi_g = 1$), p_g, and ϑ_g are the proportion, the density of observations, and the distributional parameters, respectively, in class g, and $\pi = (\pi_1, \ldots, \pi_G)$ and $\vartheta = (\vartheta 1, \ldots, \vartheta G)$. The likelihood function for π and ϑ

based on n independent observations $y=(y_1, ..., y_n)$ is a product of expression (1) with index i=1, ..., n.

With discrete outcomes, the observed data can be presented as a table of counts of the number of cases in each cell. Let the table have L cells based on K variables. Variable y_i records the cell membership of case i. The density $p_g(* |*)$ from expression (1) for each case in cell l is $\pi_{l|g}$, the probability of being in cell l for a case arising from class g. The parameters $\vartheta_g = (\pi_{1|g}, ..., \pi_{L|g})$ usually are related to one another through a log linear model of dimension less than L. The mixture density for n cases cross-classified into a table with L cells can be written from expression (1) as

$$p(y \mid \pi, \vartheta) = \prod_{i=1}^{n} \left(\sum_{g=1}^{G} \pi_g \left[\prod_{l=1}^{L} \pi_{l|g}^{\,I\{y_i = l\}} \right] \right) \qquad (2)$$

$$= \prod_{l=1}^{L} \left(\sum_{g=1}^{G} \pi_g \, \pi_{l|g} \right)^{n_l}, \qquad (3)$$

where $I\{y_i = l\} = 1$ if case i is in cell l and 0 otherwise, and $n_l = \sum_{i=1}^{n} I\{y_i = l\}$ equals the number of cases in cell l ($n= \sum_{l=1}^{L} n_l$).

Probabilities of class membership can be computed using Bayes' Theorem. Let z_{ig} equal 1 if case i is from mixture class g and 0 otherwise, then the probability of case i being in class g is

$$p(z_{ig} =1 \mid y_i, \pi, \vartheta) = \pi_g \, p_g(y_i \mid \vartheta_g) / \left(\sum_{h=1}^{G} \pi_h \, p_h(y_i \mid \vartheta_h) \right) \qquad (4)$$

In the discrete case, (3) depends only on cell membership and is $\pi_{g|l} = \pi_g \, \pi_{l|g} / \sum_{h=1}^{G} \pi_h \, \pi_{l|h}$. Clusters can be formed by assigning observations to the class for which it has the highest probability of membership.

Let n_{lg} be the number of cases in cell l and class g. The mixture classes can be thought of as being associated with subtables of counts $\mathbf{n}_g = \{ n_{lg}, l=1,...,L \}$, g=1,...,G, which when combined yield the observed table, $\mathbf{n} = (\mathbf{n}_1, ..., \mathbf{n}_L) = (\sum_{g=1}^{G} n_{lg}, l=1, ..., L)$. If the latent indicators (and hence counts) were known, the joint density for y and z, where z is a n x G matrix with entries z_{ig}, would be

$$p(y, z|\ \pi,\ \vartheta) \ = \prod_{i=1}^{n} \prod_{g=1}^{G} (\pi_g\ [\prod_{l=1}^{L} \pi_{l|g}^{\ I\{y_i = l\}}\ ])^{z_{ig}} \quad (5)$$

$$= \prod_{l=1}^{L} \prod_{g=1}^{G} (\pi_g\ \pi_{l|g}\ )^{n_{lg}}. \quad (6)$$

One special example of mixture models for discrete data is the latent class (Haberman 1974 and 1979, Goodman 1974). In this model the chance of having a certain level for field k is assumed to be independent of the levels for other other fields of information.

Maximum likelihood estimation for latent class models can be accomplished with the EM algorithm (Dempster, Laird, and Rubin 1977). See also McLachlan and Peel (2000) and McCutcheon (1987).

Alternative models allow interactions between fields of information within a mixture class. For instance, the density in class g can be defined by a log linear model on the expected counts (or, equivalently, on the probabilities $\vartheta_g = (\pi_{1|g}, ..., \pi_{L|g})$ in the cells of its subtable $\mathbf{n}_g$. The (possibly different) log linear models across the classes can be specified by the sets of variables that interact within that class. Mixture models with log linear interactions within classes have been used by Becker and Yang (1998), Larsen and Rubin (2001), and references therein. See Hagenaars and McCutcheon (2002) for other recent developments. These models are not considered in this paper.

The number of classes G is selected here by minimizing the Bayesian Information Criterion (BIC) measure of fit and complexity (see, e.g., McLachlan and Peel 2000, pages 209-210). In the case of mixture models, the value of BIC is calculated as $-2 \log \Lambda + d \log(n)$, where Λ is the log likelihood (sum of log of (1) over i=1,...,n), d is the number of parameters in the model, and n is the number of observations. The value of $-2 \log \Lambda$ decreases, whereas the d increases, as G increases. The value of G that yields the smallest value of BIC is selected.

3. Adjusted Rand Index

The adjusted Rand Index (ARI; Hubert and Arabie 1985) measures the correspondence between two partitions of a finite set of objects. In the application, one partition will be sex (female, male) and the other will be clusters determined by estimated latent class models. The ARI is a modification of the Rand Index (1971). Rand (1971) addressed the issue of interpreting the results of clustering data by proposing a criterion based on pairs of points and how they are clustered under different groupings. If there are i=1,..., r clusters according to method one and j=1,..., c clusters according to method two, then let n_{ij} be the number of objects smultaneously in method one's cluster i and method two's

cluster j. If n is the total number of points, then $\binom{n}{2}$ is the number of (unordered) pairs of observations. The Rand Index is the counts of the number of pairs that are the same clusters in both methods plus the number of pairs that are in different clusters in both methods divided by the total number of pairs. The statistic can be expressed as

$$[\binom{n}{2} - \frac{1}{2}(\sum_{i=1}^{r} n_{i+}^{2} + \sum_{j=1}^{c} n_{+j}^{2}) + \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}^{2}] / \binom{n}{2} \qquad (7)$$

where a + subscript means summation over the relevant index.

Hubert and Arabie (1985) proposed an adjustment of the Rand Index (7) to account for chance agreement (same cluster both methods or different clusters both methods). Specifically, Hubert and Arabie (1985) use a generalized hyper geometric distribution as a null model and compute the expected value of the Rand Index. They then define the adjusted Rand Index (ARI) as (Index - Expected Index)/(Maximum Index - Expected Index). See Hubert and Arabie (1985) for an expression of ARI in terms of the counts n_{ij}, i=1, ..., r, j=1, ..., c. Milligan and Cooper (1986), in a simulation study, found that the ARI performed best out of five external criteria in evaluating the recovery of cluster structure. The ARI is selected for use based on its properties and these simulation results.

Values near 1 indicate that pairs of observations tend to be located in clusters together or separately in both clustering methods. For large sample sizes, however, the ARI can be quite small (near zero) or even negative. As log odds ratio can be near 0 but still be statistically significant, small ARI values can be statistically significant. In the application, latent class solutions with different numbers of classes also will be used.

Simulation will be used to judge the significance of the degree of recovery of the classes as measured by the ARI. Under the assumption that the observations are assigned to clusters independently by the two methods, the distribution of the ARI is simulated by drawing samples from a hyper geometric distribution. One thousand samples are generated under the null model. The reported tail area is the fraction of samples that have ARI values greater than or equal to the value for the survey data. The method will be applied to clusters defined by individual categorical variables and then to clusters produced by latent class analysis.

4. Description of the Sample

The data analyzed are from the 1997 Survey of Doctoral Recipients (SDR). The SDR 1997 is nationally representative survey of scientists and engineers. See NSF 99-337 (1999) for details on coverage and survey weight information.

Respondents received PhDs in three areas: (1) biological and other life sciences, (2) mathematical and computer sciences, chemistry except for biological chemistry, physics and astronomy, and other physical and related sciences, and (3) electrical engineering, electronics, and communication, chemical, civil, mechanical, and other engineering. Attention is restricted to respondents who recently (1990-1996) received a Ph.D., work at educational institutions beyond the secondary level, and have career path jobs. Of the 1756 in biology and life sciences, 50 percent are female. Of the 824 in mathematical and physical sciences, 33 percent are female. Twenty-four percent of the engineers are women. See Larsen (2002) for more information on this subset of the SDR.

There are large differences between the percent female responding in particular ways to several questions for all three groups. There also are large differences between the three groups. Desired work before beginning the Ph.D., adequacy of doctoral preparation, job search resources, limits on job seeking, work activities, and some other variables are described in this section. When clusters defined by sex and by single categorical variables are compared, simulated ARI tail areas generally are similar to P-values from traditional tests of significance. See Larsen (2002) for more additional tables on these comparisons.

4.1. Desired post-Ph.D. Work

Respondents were asked whether they recalled desiring post-Ph.D. work involving teaching, research, management/administration, professional practice, or other activities at the start of their doctoral program.

A desire to do research is relatively more popular with the males than with the females among the biologists (two-sample test of equal proportions, z= -2.45, P-value= .01; adjusted Rand Index= 0.0012, tail area = .01) and the physical scientists (z= -1.93, P-value=0.05; ARI = 0.0205, tail area = .04). A desire to manage/administer was relatively more popular with females than with males (z=2.95, P-value<.01; ARI = 0.0026, tail area = .001). For the three areas, the desire to teach was relatively stronger for males, but not statistically significantly so. Other differences were not significant or consistent across the three areas. Thus, there are some significant and non negligible differences between female and male desires in this select group.

4.2. Adequacy of Doctoral Program Training

Respondents were asked to rate the adequacy of their doctoral training in terms of general problem solving, oral communication, teaching, collaboration and team work, quantitative, writing, computer, and management or administrative skills, subject matter knowledge, and research integrity. Among the biologists, women were relatively more likely to say they had very adequate preparation in communication (z=2.73, P-value=.01; ARI=0.002, tail area= .015) and ethics (z=2.48, P-value=.01; ARI=0.039, tail area= .039), but relatively less likely to say this about computer skills (z= -2.58, P-value=.01; ARI =.002, tail

area = .021). Women in physical sciences were relatively less likely to say they had very adequate general problem solving skills (z= -3.38, P-value<.01; ARI=0.038, tail area = .001). Female and male engineers did not have significant differences in reporting very adequate versus not as adequate preparation in the eleven areas.

4.3. Use of Job Search Resources

Respondents who were holding, who had held, or who had sought a career path job after Ph.D. were asked whether or not they used any of ten job seeking resources. Most differences in the percent female between those who did and did not use specific resources were nonsignificant. Further, the pattern of usage was not consistent across the three discipline areas. Among the biologists and life scientists, however, women were significantly relatively less likely than were men to use faculty or advisors (z= -2.22, P-value=.03; ARI=0.004, tail area=.015) and to use electronic postings (z= -4.58, P-value<.01; ARI=0.006, tail area= .004). Women in the physical sciences were also significantly relatively less likely to use faculty or advisors (z= -2.30, P-value=.02; ARI = 0.030, tail area = .014), but significantly relatively more likely to use newspapers (z=3.24, P-value<.01; ARI = 0.043, tail area=.001) than were men.

4.4. Limitations on Career Path Job Search

Respondents were asked whether or not their career path job search was limited by five factors. Women are much more relatively likely to say they are limited by family responsibilities, a spouse's career or employment, and a desire to not relocate or move. They are somewhat less likely to say they are limited by the unavailability of a suitable job. Table 1 provides by field the percent female in three categories of responses, results of chi-square tests of homogeneity of proportions, and ARI results.

Table 1. Percentage female in three categories of responses, results of chi-square tests of homogeneity of proportions for five questions for each of three disciplines, and adjusted Rand Index values and tail areas.*

	A great deal		Somewhat		Not much or not at all		Chi squared test		ARI evaluation	
Reason/Discipline	N	%F	N	%F	N	%F	X^2	P-va-	ARI	Tail

								lue	value	area
Family Resp.										
Biology	235	57	462	46	690	43	13.4	.00	.0013	.10
Phys. Sci.	83	43	195	35	389	29	7.3	.03	.0193	.02
Engineering	46	35	107	21	185	20	4.9	.09	.0030	.37
Spouse's job										
Biology	288	67	403	45	557	34	80.3	.00	.0152	.00
Phys. Sci.	129	51	160	38	307	22	36.6	.00	.0580	.00
Engineering	48	48	77	21	171	15	23.5	.00	.0466	.02
Debt burden										
Biology	75	53	204	42	1006	43	3.3	.20	-.0126	.84
Phys. Sci.	21	43	60	28	521	29	1.9	.39	-.0171	.56
Engineering	11	9	33	18	265	23	1.5	.47	-.0337	.95
Desire not to move										
Biology	251	66	366	48	868	40	53.1	.00	.0176	.00
Phys. Sci.	88	44	158	39	454	27	13.9	.00	.0360	.00
Engineering	36	42	87	24	216	19	8.7	.01	.0327	.05
No suitable job										
Biology	306	41	340	43	733	47	4.5	.10	.0050	.00
Phys. Sci.	201	24	218	34	271	36	7.6	.02	.0072	.03
Engineering	83	16	103	18	157	25	3.7	.16	.0085	.04

* Individuals answering ``Not Applicable'' are excluded in each line separately.

4.5. Work Activities and Other Variables

In all three disciplines, men are relatively more likely than women to spend at least ten percent of their time on applied research and design (significantly so for physical scientists and engineers) and basic research (significantly so for biologists and physical scientists). However, females are relatively more likely than males to spend time teaching (significantly so for biologists and physical scientists). This could be viewed as surprising since women were relatively less likely to say that they had desired to teach. Many of the other activities are performed less often in these fields by these respondents and do not show consistent patterns or significant differences.

Other demographic, job related, and education variables show significant differences for these respondents between males and females. Women are significantly relatively less likely than men to have Post doctoral positions (``Postdocs''), to be married and have children living at home (for biologists and engineers), to have government support, and be doing supervisory work. Women are significantly relatively more likely to be licensed in their occupation (for biologists), to be of a minority race (for biologists and physical scientists), to have attended a work-related workshop, seminar or other training activity, and to have memberships in more than two professional organizations (significant for biologists). Other variables have inconsistent or insignificant results.

5. Latent Class Results

Latent class models were fit to subsets of variables described in the previous section. Due to the differences in responses in the three discipline areas, latent class models will be fit separately within the areas. The number of classes in each case is chosen using BIC. The latent class results are described briefly; see Larsen (2002) for more complete descriptions. The association of the latent class results with groups defined by sex is compared using the ARI. Many of the classes are interesting in terms of their compositions of females and males.

5.1. Desired post-Ph.D. Work

Models with two and with three latent classes were fit to the four binary variables describing desired work before beginning the doctoral program. According to the BIC criterion, two classes are preferred for the physical scientists and engineers, but three classes are needed to fit the biologists. Throughout this work, the sample of biologists is the largest and often requires bigger models.

The latent class models do not separate females from males any better than do individual variables as judged by the adjusted Rand Index and its simulated tail area and by Chi square test of homogeneity of proportions. The five individual variables only weakly separated females and males. Sex is not used directly when forming the latent classes. Despite this, the patterns found do seem to make sense for this highly select group of respondents.

5.2. Adequacy of Doctoral Program Training

When latent class models are fit to the eleven variables describing the adequacy of Ph.D. programs, three classes are chosen to fit the physical scientists and engineers and four classes are chosen to fit the biologists. All three groups have classes reporting high, medium, and low levels of adequacy. In no case do the clusters formed by the latent class models distinguish females from males any better than do the most significant variables in section 4. The ARI values for the latent class results are lower and less significant than the highest values for individual variables.

5.3. Use of Job Search Resources

The latent class models that were selected to fit the data in the three disciplines on job seeking resources contain four classes. Similar classes are found in all three disciplines. One class used many resources. Representation of females in this class was higher than average in the physical sciences, but lower than average among the biologists and engineers.

A second class often used four or five of the resources, especially faculty or advisors, professional meetings, electronic postings, professional journals, and informal channels. Representation of women in this type of class was lower than average in all three disciplines.

A third class usually used two resources, faculty or advisors and informal channels, and occasionally others. A fourth class also used selected resources and varied somewhat more across disciplines. Representation of women in these two types of classes were slightly higher than average in all three disciplines.

The association with sex is not any stronger for the biologists and physical scientists, but it is stronger, according the ARI tail area, for engineers. The largest ARI value occurs on the question of use of professional journals (ARI=0.021, tail area = 0.22). For the four classes, the ARI value is 0.014, but the corresponding tail area is 0.02. The reason that the ARI value for the latent class model can decrease but still be significant is because the cross classification has more cells, which has a large effect on tests of significance.

5.4. Limitations on Career Path Job Search

Latent class models were fit to the responses from individuals in the three disciplines to the questions about limits on career path job searches (great deal, somewhat, or not at all). As with previous questions on adequacy of doctoral preparation and job seeking resources, a similar pattern of classes was observed in the three discipline areas. A three class model was chosen for the biologists and for the physical scientists, whereas a two class model was selected for the engineers. One class in the three class models has high probabilities of having a great deal of difficulty with all five areas. A second class has some limitations based on three issues: family responsibilities, spousal employment, and a desire to not move out of the area. A third class tends not to have limitations related to these three issues. The two engineer classes were similar to the first class for the other groups and a combination of their other two classes.

There is a strong association of sex with the latent classes. Simulated values of the ARI were never higher than for the survey data for biologists and physical sciences and higher only 26 times out of 1000 for the engineers. It is not surprising that the latent classes highlight differences between the sexes given the strong association of responses to individual variables with sex, as was presented in section 4. However, the multivariate analysis seems to provide a different picture than would be expected based on table 1. Women are over represented when asked about limits, except for not finding a suitable job. Women, on the contrary, are underrepresented in the class that reports multiple limits (class 1). This could be happening because one subset of women report several limits, but other women report very few limits. The set of women with limits report limits based on family, spouse, and a desire not to relocate. The set of women without limits reports few of these limits and reports that finding a suitable job is not a great problem.

Table 2. Latent class parameter estimates – proportions in the four classes, conditional probabilities – for physical scientists based on limitation on seeking a career path job

	Physical Scientists								
	Class 1			Class 2			Class 3		
Class Proportion	0.62			0.16			0.22		
	GD*	SW	NM	GD	SW	NM	GD	SW	NM
Family responsibilities	.92	.08	.00	.15	.85	.00	.24	.29	.47
Spouse's career or employment	.85	.11	.05	.24	.74	.02	.30	.10	.59
Debt burden	.90	.07	.03	.91	.05	.03	.87	.11	.02
Desire not to relocate or move	.83	.11	.06	.56	.44	.00	.39	.28	.33
Suitable job not available	.49	.24	.26	.35	.33	.32	.46	.34	.20
Percent Female	28%			35%			48%		

* GD = A great deal. SW = Somewhat. NM = Not much or not at all.

5.5. Work Activities

Latent class models were fit to the seven most frequently cited work activities variables. The data for the biologists requires four classes, whereas the physical sciences three and the engineers two. In all three disciplines, there is a strong, significant association of the latent classes with sex. Further, the distinctions between latent classes largely are determined by amounts of applied research, basic research, management, and teaching. Discussion can be found in Larsen (2002).

5.6. Other Variables

Latent class models were fit to eleven of the indicator variables (all except for minority status) from the other variables. A model with six classes was selected for the biologists. The association of the latent classes with sex is extremely significant (chi-square statistic = 75, degrees of freedom = 5, P-value < .01; ARI = 0.0085, tail area < .01). Based on these data, as described in Larsen (2002), there appear to be subgroups of males and of females among the biologists having very different life and work experiences.

The physical scientists require five latent classes for these variables. One class that is 44% female tends to not be married and not to have kids, not to have a postdoc, and to have received workshop/job training. One class that is 24% female tends to have similar family situations, but to have postdocs. Again the classes separate based on family and having postdoctoral positions. Three classes fit the engineers. A similar split based on family and having a postdoctoral position is again important.

For these two fields, however, the association of the latent classes with sex is not clear. For the physical scientists, a chi-squared Statistic has a value of 26, which on four degrees of freedom has a P-value less than 0.01. On the other hand, the ARI value is 0.0030 with a simulated tail area of 0.080. For the engineers, the Chi squared statistic is 11 with on 2 degrees of freedom is a P-value less than

0.01. The ARI value actually is negative, which produces a tail area around 0.70. Thus, the results seem to diverge.

6. Summary

Females and males with recent Ph.D.'s in three discipline areas (biology and life sciences, physical sciences, and engineering) working at educational institutions beyond the secondary level were compared on various training, job search, work activity, and other variables. The selection of an apparently homogeneous group to study should have eliminated many important background differences between the women and men.

There are a few significant differences between women and men in desired work activities, job search resources, and adequacy of doctoral training. Differences varied across disciplines. The fact that women in biology felt on average less adequately trained in computing, whereas women in physical sciences felt less adequately trained in general problem solving could have important discipline-specific implications. A future study could try to explain why there were differences in job search resources.

There are many large, significant differences in limitations when searching for a job, work activities, and family and career status. Latent class analysis identified a subgroup of women in each field who has many job search limitations and another subgroup that does not. Latent class analysis identified subgroups that perform different combinations of work activities. The combinations of work activities were more strongly associated with sex than were the individual activities. Latent class analysis found several subgroups in the discipline areas that differ from one another in terms of postdoc status, marriage and family status, and professional activities. It would be interesting to undertake a longitudinal study of factors predictive of long-term success in these disciplines.

The latent classes generally were not as well associated with sex as were the best individual variables according to traditional significance tests and according to the adjusted Rand Index. The simulation of tail areas for the ARI generally agreed with P-values from significance tests. A couple of times, however, results were more or less significant when using the ARI. Further work is needed to understand the different performance of these procedures.

Analysis will be performed on future waves of SESTAT data. Work also needs to be done comparing the experience of minority groups to other groups in the sciences and in the Information Technology workforce.

Acknowledgments

The author would like to acknowledge helpful discussions with Elaine Zanutto (Wharton School, University of Pennsylvania) and Chen Yang (graduate student, University of Chicago, Department of Statistics) and support from the

U.S. National Science Foundation (grant EIA-0089930). The author also would like to thank the organizers of the International Federation of Classification Societies conference in Cracow, Poland, and the editors of *Statistics in Transition*.

REFERENCES

AAUW Educational Foundation Commission on Technology, Gender, and Teacher Education. (2000), *Tech-Savvy: Educating Girls in the New Computer Age*, Washington, D.C.: American Association of University Women Educational Foundation.

BECKER, M. P., and YANG, I. (1998), "Latent Class Marginal Models for Cross-Classifications of Counts", *Sociological Methodology*, 28, 293-325.

DAVISON, H.K., and BURKE, M.J. (2000), "Sex Discrimination in Simulated Employment Contexts: A Meta-analytic Investigation," *Journal of Vocational Behavior*, 56, 225-248.

DARITY, W.A., and MASON, P.L. (1998), "Evidence on Discrimination in Employment: Codes of Color, Codes of Gender," *The Journal of Economic Perspectives*, 12, 63-90.

DEMPSTER, A.P., LAIRD, N.M., and RUBIN, D.B., (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm", *Journal of the Royal Statistical Society*, Series B, 39, 1-38.

ETZKOWITZ, H., KEMELGOR, C., NEUSCHATZ, M., UZZI, B., and ALONZO, J. (1994), ``The Paradox of Critical Mass for Women in Science," *Science*, 266, 51-54.

FARMER, H.S., WARDROP, J.L., and ROTELLA, S.C. (1999), "Antecedent Factors Differentiating Women and Men in Science/Nonscience Careers", *Psychology of Women Quarterly*, 23, 763-780.

GOODMAN, L. A. (1974), "Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models", *Biometrika*, 61, 215-231.

HABERMAN, S. J. (1974), "Log-Linear Models For Frequency Tables Derived By Indirect Observation: Maximum Likelihood Equations", *The Annals of Statistics*, 2, 911-924.

HABERMAN, S.J. (1979), *Analysis of Qualitative Data*: (Vol. 2), New Developments, Academic Press.

HAGENAARS, J.A., and MCCUTCHEON, A.L. (2002). *Applied Latent Class Analysis*, Cambridge University Press: Cambridge.

HUBERT, L,, and ARABIE, P. (1985), "Comparing Partitions", *Journal of Classification*, 2, 193-218.

LARSEN, M. D. (2002). "Latent Class Analysis of 1997 NSF Survey Data on Science PhDs", submitted to Proceedings of *Retention of Women Graduate Students and Early Career Academics in Science, Mathematics, Engineering, and Technology*, Iowa State University, October 17-20, 2002.

LARSEN, M. D., and RUBIN, D. B. (2001). "Iterative Automated Record Linkage Using Mixture Models", *Journal of the American Statistical Association*, 96, 32-41.

MCCUTCHEON, A.L. (1987). *Latent Class Analysis*. Sage Publications.

MCLACHLAN, G. J., and PEEL, D. (2000). *Finite Mixture Models*, John Wiley & Sons, Inc.: New York.

MILLIGAN, G. W., and COOPER, M. C. (1986), "A Study of the Comparability of External Criteria for Hierarchical Cluster Analysis", *Multivariate Behavioral Research*, 21, 441-458.

NATIONAL SCIENCE FOUNDATION. (1999), *SESTAT: A Tool for Studying Scientists and Engineers in the United States* (NSF 99-337). Authors, N. Kannankutty and R. K. Wilkinson, Arlington, VA.

RAND, W.M. (1971), "Objective Criteria for the Evaluation of Clustering Methods", *Journal of the American Statistical Association*, 66, 846-850.

TOP, T.J. (1991), "Sex Bias in the Evaluation of Performance in the Scientific, Artistic, and Literary Professions: A Review", *Sex Roles*, 24, 73-106.

STATISTICS IN TRANSITION, October 2002
Vol. 5, No. 5, pp. 857–872

AN APPROACH TO THE PROBLEM OF SPATIAL DIFFERENTIATION OF MULTI–FEATURE OBJECTS USING METHODS OF GAME THEORY

Andrzej Młodak[1]

ABSTRACT

In this paper we try to apply some specific theoretical methods of game theory to research of spatial differentiation of multi–feature objects. The aim of presented research is to determine a participation of each object in general level of development of given domain in whole region or subregion to which this object belongs. In order to realize this intention we apply a model of cooperative game and its particular cases, i.e. games with a priori unions and so–called "airport" game. As a numerical example was assumed data on state of technical infrastructure in the Wielkopolskie Voivodship in 1999 presented by poviats.

Key words*: spatial differentiation, feature, co–operative game, Shapley value, airport game, infrastructure.*

1. Introduction

Classical analysis of spatial differentiation of multi – feature objects is conducted for the purpose of construction of synthetic measure of their development in terms of researched area. As these objects are usually understood units of administrative or geographical division of considered territory. The above – mentioned synthetic measure is usually called a "meta – feature" and it describes an aggregated level of deviation of state of analyzed area within given object from assumed general development pattern. Moreover, one can formulate other important question connected with this issue, namely, which is the participation of given object in total state of studied domain in a whole-analyzed territory containing this object?

One attempt at solution of this problem was presented in article by A. Młodak (2002) in consideration to the situation of labour market in the

[1] Statistical Office in Poznań, Branch in Kalisz, pl. J. Kilińskiego 13, 62–800 Kalisz, POLAND; e–mail: amlodak@stat.gov.pl

Wielkopolskie Voivodship. For this purpose a new type of statistical feature, i.e. so called complex feature was there constructed. Main difference which distinguishes this model from the "meta – feature" approach is the fact, that values of the complex feature can be practically interpreted as an aggregated measure of state of researched field. It can be easily converted into its non – negative counterpart (almost without loss of variation), what is necessary for further calculations (requirement for the model). On the basis of this construction and by application of special type of cooperative game[1], i.e. so–called "airport" game (which belongs also to the class of allocation cost games), a complex feature was constructed. Next, the Shapley values (as a one type of solution of this game) of poviats (understood as players of the game) were calculated. By these methods we have determined quantities participation of particular poviats in total state of the labour market in Wielkopolska.

The present paper is in theoretical sense a continuation and an extension of the above – mentioned article. Since introduction of new territorial division in Poland in 1999 poviats a partition of collection of Wielkopolska into five subregions is used for the statistical purposes. In present consideration we would like to determine a share of every poviat in total state of a given domain against the subregion, to which it belongs. To realize this purpose, we will apply a specific model of co–operative game, i.e. a game with a priori unions. That is, a player set (which members in our case are the poviats) is divided into disjoint, nonempty subsets called a priori unions or percolation. We assume that the percolations are here understood as subregions of Wielkopolska. On the basis of values of complex feature we will construct one of most interesting solutions of these games (in the "airport" version), i.e. Shapley value with appropriate modification for a priori unions. Numerical data, which will be used in this article, concern state of technical infrastructure (water and sewage systems, dwelling stocks, public roads, etc.) in the Wielkopolskie Voivodship. As a source of this information was assumed the publication by the Statistical Office in Poznań (2001).

The paper is organized as follows. First, we present main consideration concerning the co–operative game theory and a mathematical model of game with a priori unions, its "airport" case and the Shapley value. In the second chapter a method of construction of a complex feature for multi – feature objects applying the Weber median is described. Last, we analyze the results of calculation of our numerical models adapted to the problem of spatial differentiation of technical infrastructure in Wielkopolska.

---

[1] A model of co–operative game was yet mainly applied to the analysis of decisions making processes in collective organs, for example in: Parliament of Catalonia, Spain (Carreras and Owen (1988)), Italian Chamber of Deputies (Gambarelli (1997)), Sejm of the Republic of Poland (Sosnowska (1993, 1995), Młodak (2000)) and Town Council in Kalisz (Młodak (1998)) as well as a tool used to determine an allocation of airport costs (Littlechild (1974), Malawski *et al.* (1997), Vázquez – Brage *et al.* (1996).

2. A model of co–operative game and its "airport" case

First we will introduce necessary notions and facts connected with considered theory, i.e. a description of general model of co – operative game and its solutions.

Let n be a natural number; An *n – person co–operative game* is uniquely defined by the pair (N, v), where N = {1, 2, … , n} is a set of players taken part in the game and v: $2^N \rightarrow \mathbf{R}$ is a function called *characteristic function* of this game. This function assigns to each subset of the set N (called *a coalition* of players they belong to) some real number. It is assumed that a value of the characteristic function for empty coalition, i.e. the one to which nobody belongs, amounts to 0. Thus $v(\varnothing) = 0$. Because of fact, that the characteristic function determines uniquely the game, in further part of this article we will use a short description "n –person game v".

As a *solution* of co–operative game is understood a function φ, which assigns to each n – person game v a vector from n–dimensional $\mathbf{R}^n$ space. In practice, the most important are these solutions, which constitute a division of value of the characteristic function for the full coalition (i.e. the one containing all players) between participants of the game. In other words, a vector $(x_1, x_2, \ldots, x_n)$ being a solution of game v is a *division* (or *preimputation*) of this game, if

$$\sum_{i=1}^{n} x_i = v(N)$$

The oldest and most frequently used solution of co–operative game is the *Shapley value,* introduced by Shapley (1953) and given by the formula:

$$Sh_i(v) = \sum_{K \subseteq N} \frac{k!(n-k-1)!}{n!}\left[v(K \cup \{i\}) - v(K)\right] \qquad (1)$$

where[1] $k = card(K)$ for any $K \subseteq N$.

This value is interpreted as an expected value of increase of value of characteristic function for a coalition after co–opting an i – th player, i = 1, 2, …, n. Note that the solution $Sh(v) = (Sh_1(v), Sh_2(v), \ldots, Sh_n(v))$ is a preimputation for any n – person game v.

Now we will present a generalized version of construction of solution with a priori unions proposed by G. Owen (1977) and applied to derivation of the Shapley value with a priori unions. But first we define the a priori unions' structure.

Let m be a natural number not greater than n. A system of m subsets of the set N, $T = (T_1, T_2, \ldots, T_m)$ is called *an a priori unions structure* in the game v, if it

[1] The symbol card (K) denotes a number of elements of a set K.

is a partition of N, i.e. if $\emptyset \neq T_i \subseteq N$ for any $i \in M = \{1,2, \ldots, m\}$, $T_i \cap T_j = \emptyset$ whenever $i \neq j$, $i, j \in M$ and

$$\bigcup_{j=1}^{m} T_j = N .$$

The sets T_i are regarded as *a priori unions*.

A *division game* based on the cooperative game v is such a game v*, whose players are the a priori unions from structure T, and its characteristic function is given as

$$v^*(S) = v(\bigcup_{c \in S} T_c)$$

for any $S \subseteq M$.

Construction. Let $\varphi(v)$ be a solution of the defined n – person cooperative game v. Suppose that for this game a priori union's structure T is defined. A solution $\varphi(v,T)$ with a priori unions can be constructed in two steps:

Step1. Let $j \in M$ and K be a subset of T_j, $K'=T_j \setminus K$ (i.e. a complement of the set K in T_j). Consider a game $v_{T,K}$, whose players are the a priori unions of T and

$$v_{T,K}(S) = v(\bigcup_{c \in S} T_c \setminus K')$$

for any subset S of the set M.

Let vj be a game on the set Tj such that vj(K)□□j(vT,K) for any K □Tj.

Step 2. We find a solution □(vj) of a game vj.

The solution □(v,T) = (□1(v,T), □2(v,T), ... , □n(v,T)), where□i(v,T) = □i(vj) for any i □ Tj and j □ M is called a solution □ with a priori unions.

Note, that if □(v) is a preimputation of any n – person game v, then □(v,T) is also a preimputation of this game. Indeed, in this case, $\varphi(v_j)$ is a preimputation of v_j, for all $j \in M$ and hence

$$\sum_{i \in T_j} \varphi_i(v,T) = \sum_{i \in T_j} \varphi_i(v_j) = v_j(T_j) = \varphi_j(v_{T,T_j}) = \varphi_j(v^*) .$$

Thus

$$\sum_{i=1}^{n} \varphi_i(v,T) = \sum_{j \in M} \sum_{i \in T_j} \varphi_i(v,T) = \sum_{j \in M} \varphi_j(v^*) = v^*(M) = v(N).$$

A Shapley value with a priori unions constructed in the cited paper (i.e. in article by Owen (1977)) by above – mentioned method can be expressed as

$$Sh_i(v,T) = \sum_{S\subseteq M}\sum_{K\subseteq T_j} \frac{s!(m-s-1)!}{m!}\frac{k!(t_j-k-1)!}{t_j!}\left[v(Q_S\cup K\cup\{i\}) - v(Q_S\cup K)\right] \quad (2)$$

for any $i\in T_j$ and $j\in M$, where $t_j = \text{card}(T_j)$.

Now we will present specific type of co–operative game, i.e. so called *airport game*. introduced by several scientists in the 1970's. As one of the original presentations of this approach can be regarded articles by. Littlechild and Owen (1973) and by Littlechild (1974)). In this model it is assumed, that there exist r types of players (where r is a natural number not greater than n). Let $R = \{1, 2, \ldots, r\}$ be a set of these types, r_i denotes a type of player i, $i\in N$, and R_t be a set of all players of type t, $t = 1, 2, \ldots, r$.

Let introduce some coherent, antisymmetric and transitive[1] relation „$\prec$" which orders the set of types R. If $t \prec u$ for $t, u\in R$, $t\neq u$, then type t is said to be weaker than u (or, type u is said to be stronger than t) according to the relation „$\prec$". For simplification, without loss of generality, we assume that $1 \prec 2 \prec \ldots \prec r$.

Let r(S) denotes the highest (according to the relation „$\prec$") type represented in coalition $S\subseteq N$, i.e. such a type, that coalition S contains at least one player of this type, and doesn't contain players of types greater than r(S). More formally[2], $t = r(S) \Leftrightarrow S\cap R_t\neq\varnothing$ and $r_i \preceq t$ for every $i\in S$. To any type $r\in R$ is assigned some real number c_r. Thus, a characteristic function of the game is given as follows:

$$v(S) = c_{r(S)}$$

for any $S\subseteq N$.

In a research practice, this model is usually used to analysis of possibilities of optimal allocation of costs of building of runway at the airport to arriving airplanes. Such approach was presented by Littlechild (1974) upon the example of the Birmigham airport (Great Britain) and Vázquez–Brage *et al.* (1996) exploiting the data concerning number and costs of aircraft movement at the airport Labacolla in Santiago de Compostela (Spain). A utility of the model results from the fact, that the greater measurements has the airplane, the higher are costs of building of runway, on which can he touch down. Therefore it is assumed, that participants of this game are particular landings on analyzed runway (each airplane can touch down more than once) of r types of airplanes. By c_r a cost of building of a runway which can be exploited by planes of i – th type. (i = l, 2, …, r) is denoted. Moreover, it is assumed that

---

[1] That is, relation satisfying three following conditions: 1) if $x \neq y$ then either $x \prec y$ or $y \prec x$ (coherency), 2) if $x \prec y$ then not $y \prec x$ (antisymmetry) and 3) if $x \prec y$ and $y \prec z$ then $x \prec z$ (transitivity).

[2] Symbol $x \preceq y$ denotes, that y is not weaker than x according to the relation "$\prec$", that is; either $x = y$ or $x \prec y$.

$$0 = c_0 < c_1 < c_2 < \ldots < c_r \tag{3}$$

what determines a relation ordering the types of players. As a value of characteristic function for lands coalition S is defined a cost of building a runway suitable for the greatest airplane belonging to S. In this way, v(S) is the (fixed) costs that would be incurred when the runway can be constructed that could accommodate all the movements in the set S. By application of solutions of co–operative games the allocation of cost c_r (i.e. total costs of building of this runway) to all the movements on the airport is obtained. Then all the landings of airplanes of one type are burden identically.

In our consideration we assume that there are n players and one player of each type, that is r = n and $R_i = \{i\}$, i = 1, 2, …, n. Introduce an ordering relation in the set of players N = {1, 2, …, n} by formula (3).

A Shapley value defined by (1) can be in this case written as (cf. for example Littlechild and Owen (1973) or Malawski *et al.* (1997))

$$Sh_i(v) = \sum_{k=1}^{i} \frac{c_k - c_{k-1}}{n-k+1} \tag{4}$$

In this model as members of a priori unions structure are regarded the air companies, which use the mentioned airplanes. Let $T = (T_1, T_2, \ldots, T_m)$ be this a priori unions structure. Denote by M_i a number of a priori unions, such that a maximal type of player of them is at least type i. That is, $M_i = \text{card}(\{j \in M: i \preceq r(T_j)\})$, where i = 1, 2, …, r. Let T_{ij} be a number of players of type not smaller than $i \in R$ according to (3), which belong to the set T_j. Then the Shapley value with a priori unions of any player $i \in T_j$ (cf. (2)) in its "airport" version is given by the formula (cf. Vázquez – Brage *et al.* (1996))

$$Sh_i(v,T) = \sum_{k=1}^{i} \frac{c_k - c_{k-1}}{M_k T_{kj}} \tag{5}$$

3. An application of games in taxonomic model.

Assume that there are n researched objects (where n is a natural number), which can be characterized by system of statistical features. We would like to obtain synthetic measure concerning aggregated level of development of these objects with respect to the field of interest.

Consider a collection of features characterizing of analyzed phenomenon and let make the introductory operations. First, we convert features – destimulants (i.e. these variables, which lower values imply higher level of intensity of analyzed phenomenon) into stimulants (with which an incline in value represents a higher development). Next, we conduct all necessary procedures of verification of this set, i.e. eliminate features with low variation (having minimal influence on

the intensity of the spatial differentiation)[1] and high correlated with others (that is, the one being carrier of similar information to other) were eliminated. As a result of these processes we obtain a set of m *diagnostic features* $X_1, X_2, \ldots, X_m$,. Each of them is represented by n – dimensional vector $X_j = (x_{1j}, x_{2j}, \ldots, x_{nj})$, j = 1,2,...,m.

Now there a standardization of the diagnostic features will be made. We apply the approach based on *the Weber median* vector presented by Wagner *et al.* (2000). It is defined to be such a vector from m – dimensional space $\mathbf{R}^m$, that a sum of its Euclidean distances from m – dimensional vectors describing values of particular features for all objects is the smallest. Formally, a vector $\theta_0 = (\theta_{01}, \theta_{02}, \ldots, \theta_{0m})$ is defined to be the Weber median if it satisfies following optimization equality:

$$\sum_{i=1}^{n}\left[\sum_{j=1}^{m}\left(x_{ij}-\theta_{0j}\right)^2\right]^{\frac{1}{2}} = \min_{\theta\in R^m}\left(\sum_{i=1}^{n}\left[\sum_{j=1}^{m}\left(x_{ij}-\theta_{j}\right)^2\right]^{\frac{1}{2}}\right).$$

We create features $\widetilde{Z}_j = (\widetilde{z}_{1j}, \widetilde{z}_{2j}, \ldots, \widetilde{z}_{nj})$, such that,

$$\widetilde{z}_{ij} := \frac{x_{ij} - \theta_{0j}}{1{,}4826 \cdot \tilde{mad}(X_j)} \tag{6}$$

where $\tilde{mad}(X_j)$ denotes *absolute median deviation*, i.e. $\tilde{mad}(X_j) = med(\widetilde{Y}_j)$, and $\widetilde{Y}_j = (\widetilde{y}_{1j}, \widetilde{y}_{2j}, \ldots, \widetilde{y}_{nj})$ $\widetilde{y}_{ij} = |x_{ij} - \theta_{0j}|$ i = 1,2, …, n, j = 1,2,...,m.

The name of this notion comes from the name of the famous German sociologist and economist, Weber, who proposed it (Weber (1909, reprint 1971)) in 1909 as a solution to a transportation cost minimization problem.

The main purpose of standardization of the diagnostic features is a transformation of them into new forms, which ensure their comparability and normalization of their basic statistical measures. In our positional approach (given by formula (6)) it should be $med(\widetilde{Z}_j) = 0$ and $\tilde{mad}(\widetilde{Z}_j) = 1$, j = 1, 2, …, m. In practice, there exists some minute deviations from these quantities being a penalty for exploiting mutual connections between the features by transformation (8).

In this type of analysis a taxonomic development standard, i.e. a m – dimensional vector representing optimal levels of values of particular features, consists of maximal values of them. In other words, it is a vector $\xi = (\xi_1, \xi_2, \ldots, \xi_m)$, such that

[1] Methods of conversion of destimulants into stimulants were described by Malina and Zeliaś (1998) or Wagner *et. al.* (2000).

$$\xi_j = \max_{i=1,2,..,n} \widetilde{z}_{ij}$$

j = 1, 2, …, m. (cf. Śmiłowska (1997), Wagner *et. al.* (2000)).

In our approach we introduce some modification of the above – mentioned construction. Let define a *complex feature* C = (c_1, c_2, …, c_n), as a linear combination of normalized features with weights assumed as importance coefficients of diagnostic features. That is

$$c_i = \sum_{j=1}^{m} w_j \widetilde{z}_{ij} \tag{7}$$

i = 1,2,...,n, where w_j is *an importance coefficient* of j – th feature X_j, given by the formula

$$w_j = \frac{V_j}{\sum_{k=1}^{m} V_k}$$

and V_j denotes a coefficient of variation of feature X_j, i.e.

$$V_j = \frac{S_j}{\overline{x}_j}$$

where $\overline{x}_j$ is an arithmetic mean of values of X_j and S_j denotes its standard deviation, j = 1,2,...,m. This combination reflects the worth of particular features in the model as well as connections between them.

From the economical point of view, the quantities c_i can be interpreted as aggregated measure of intensity of analyzed phenomenon taking into account levels of influence of particular diagnostic features on the situation in that field and involving also a classical postulate of comparability of the features.

If in some cases the feature C possess negative values, then this feature can be corrected by adding to each of its value an absolute value of integral part[1] of its minimum. That is, we assume:

$$c_i^* = c_i + \left| \left[\min_{k=1,2,\ldots,n} c_k \right] \right| \tag{8}$$

for any i = 1, 2, …, n. The complex feature describes a level of total development of researched domain. As a development standard we assume a maximal value of the feature C.

[1] An integral part of real number is the greatest integral number not greater than this real number. Integral part of number x is usually denoted by [x].

The aim of our work is to determine a participation of each object in general level of intensity of analyzed phenomenon of total collection to which it belongs. The above – mentioned partition can be established on account of various between – object connections. For example, in the case of administrative units in Poland, as the objects can be regarded poviats of a given voivodship. As partition of their set is then assumed collection of subregions used since 1999 as an level of data aggregation for the statistical purposes. Moreover, we would like a development pattern of area consisting of subset of analyzed objects to be a maximal value of complex feature restricted to the objects of this area. This assumption is essential, because in context of conducted analysis we should take into account a development of subsets of analyzed collection of objects. Such subset can be e.g. a set of poviats within borders of the voivodship or the subregion created on account of neighborhood or other socio–economical premises.

All the values of complex feature (maybe after correction (8) if it is necessary) satisfy assumptions of the "airport model" (introduced in Section 1) in terms of "cost indicators", c_i. Therefore we can use a game of "airport" type, whose players are particular objects, set of types coincides with the set of players, so that each type is represented by exactly one player and c_i is a value of complex feature i = 1, 2, …, n. Moreover[1] $0 = c_0 < c_1 < c_2 < \ldots < c_n$. As a priori unions can here be regarded all the elements of the partition of the collection of researched objects. Thus on the basis of Shapley values (4) one can determine composition of participation of particular objects in aggregated development of the collection of interest. By (4) and (5) we conclude moreover, that application of these formulas reduces almost all loss of variation constituting a result of correction (8), if, of course, this correction is necessary.

For the purpose of obtaining percentages we divide traditionally the value of computed solution by sum of all the values of this solution for all players

$$\frac{Sh_i(v)}{\sum_{d=1}^{n} Sh_d(v)} = \frac{Sh_i(v)}{c_n} \tag{9}$$

for any i = 1,2…,n; or by sum of all players within given a priori union (in the case of game with a priori unions)

$$\frac{\mathrm{Sh}_i(\mathrm{v},\mathrm{T})}{\sum_{\mathrm{d}\in\mathrm{T}_j}^{\mathrm{n}} \mathrm{Sh}_\mathrm{d}(\mathrm{v},\mathrm{T})} = \frac{\mathrm{Sh}_i(\mathrm{v})}{\mathrm{Sh}_j(\mathrm{v}^*)} \tag{10}$$

for any $i \in T_j$, and any j = 1, 2, …, m and obtained result multiply by 100.

[1] By simple ordering of objects by values c_i (if they are different within pairs) i = 1,2,...,n and respective renumbering of them we can always easily obtain this requirement.

4. Technical infrastructure in Wielkopolska

A start point of calculations was a set of diagnostic features describing spatial differentiation of technical infrastructure of the Wielkopolskie Voivodship in 1999 defined as follows:

1. Length of hard surface public roads in poviats and gminas (communities) in km per 100 km^2 (analytical symbol – X_1),
2. Length of water – line systems in km per 10 thous. population (X_2),
3. Length of sewerage systems in km per 10 thous. population (X_3),
4. Inhabited dwellings per 1000 population (X_4),
5. Average usable floor space of dwellings per one living person in m^2 (X_5),
6. Average number of persons per one room in dwelling (X_6),
7. Dwellings completed per 1000 population (X_7)

Data presented in table 1. occur from publication by the Statistical Office in Poznań (2001). The feature X_6 is destimulant, others are stimulants. The highest variation possess X_7 and X_1 (76,3% and 69,3%, respectively). Values of coefficient of variation for other features amount to between 4,7% (X_5) and 44,4% (X_2).

We will apply to the collection of features all operations described in the previous section. First, we convert the destimulant (X_6) into the stimulant (taking its opposite values) and make the standardization of all the variables according to the formula (6) and construct the complex feature by (7). Its values are presented in the third column of the table 2.

Table 1. Values of diagnostic features

Specification	X_1	X_2	X_3	X_4	X_5	X_6	X_7
WIELKOPOLSKA	**64,3**	**73,4**	**11,5**	**285**	**19,3**	**0,93**	**19**
poviats							
Chodzieski	62,2	97,7	15,1	281	19,4	0,94	13
Czarnkowsko – trzcianecki	25,3	55,9	10,0	272	19,0	0,94	8
Gnieźnieński	84,7	81,1	14,3	283	17,6	0,99	11
Gostyński	96,5	82,1	10,8	258	19,8	0,97	2
Grodziski	59,0	79,6	18,7	253	19,8	0,98	10
Jarociński	83,0	87,3	17,7	263	19,4	0,95	8
Kaliski	73,3	166,5	4,4	236	18,5	1,11	6
Kępiński	65,4	94,2	14,4	256	20,4	0,95	14
Kolski	80,4	148,0	9,3	292	18,5	0,95	21
Koniński	74,9	167,6	10,2	247	17,6	1,06	3

Specification	X_1	X_2	X_3	X_4	X_5	X_6	X_7
Kościański	57,9	73,5	11,6	264	18,9	0,99	11
Krotoszyński	98,8	93,3	12,2	265	20,6	0,95	7
Leszczyński	60,6	88,0	5,3	236	19,7	1,02	6
Międzychodzki	26,2	71,4	13,3	284	19,6	0,93	17
Nowotomyski	36,8	61,6	9,1	272	20,0	0,93	4
Obornicki	42,1	69,9	5,4	275	18,5	0,99	15
Ostrowski	79,7	74,7	10,2	267	19,3	0,98	11
Ostrzeszowski	70,6	140,4	11,8	252	20,3	0,96	9
Pilski	54,3	46,9	13,1	282	18,0	0,95	30
Pleszewski	86,9	114,3	5,8	252	19,5	0,98	7
Poznański	60,0	73,2	9,1	269	20,1	0,94	41
Rawicki	81,7	71,3	9,6	280	20,5	0,93	10
Słupecki	72,2	121,5	11,2	274	19,4	0,94	5
Szamotulski	53,8	102,7	16,4	278	19,4	0,94	6
Średzki	68,5	105,6	11,7	276	19,2	0,96	7
Śremski	70,8	75,6	16,1	275	18,2	0,96	11
Turecki	57,3	117,4	11,9	269	17,6	1,00	12
Wągrowiecki	57,8	127,8	13,1	272	18,6	0,98	11
Wolsztyński	40,7	55,3	14,5	240	18,8	1,00	12
Wrzesiński	70,2	92,7	8,4	275	19,2	0,94	6
Złotowski	34,1	75,2	14,7	270	18,4	0,95	9
Cities with the powiat status							
Kalisz	271,4	18,9	11,0	340	18,7	0,91	41
Konin	122,4	19,4	14,3	310	17,4	0,90	22
Leszno	247,6	20,9	19,4	310	19,7	0,85	32
Poznań	241,5	13,1	10,8	343	20,9	0,81	39
Importance coefficient in %	***29,0***	***18,6***	***12,7***	***3,6***	***2,0***	***2,2***	***31,9***

Next, the model of "airport game" is adapted to our problem. We assume that players of this game are the poviats, the coefficients c_i coincide with values of the complex feature ordered increasingly and as the a priori unions are regarded the five subregions of the Wielkopolskie Voivodship: Kaliski, Koniński, Pilski, Poznański and the Poznań City.

In the fourth and sixth column of the table (2) are presented the "normal" Shapley values and the Shapley values with a priori unions calculated using the formulas (4) and (5), respectively.

The fifth and seventh columns of the table contain percentages of participation of particular poviats in development of infrastructure of the voivodship and subregion computed on the basis of the Shapley values ("normal" and with a priori unions) as (9) and (10).

The poviats are ordered increasingly by values of complex feature within each subregion.

The first conclusion which can be formulated from the above – mentioned calculations shows that although poviats of the Koniński Subregion have small share in total state of the voivodships infrastructure, then their differentiation within this subregion is rather not great. In this case a large city with the poviat status (i.e. Konin city) has smaller Shapley value (and it's percentage recalculation) than the Kolski poviat.

Similarly, one can consider, that in the Pilski Subregion beside the Pilski powiat, the great importance have also the Chodzieski and Wągrowiecki poviats. A different situation is in the case of the Kaliski and Poznański subregions. In each of them, there is a city which concentrates a subregional development (Kalisz and Leszno, respectively) and the worth of other poviats which importance is rather small. This is also probably a reason of fact, that for some poviats, particularly of two above – mentioned subregions the Shapley values (without a priori unions) are greater than the Shapley values with a priori unions.

As one can consider from the construction 1., the sum of the Shapley values of all poviats belonging to given subregion is equal to the Shapley value of this subregion. Therefore this sum expresses an importance of infrastructure of this subregion in the voivodship.

At the end, it is worth to note, that the computing model concerning the results presented in table 2. in percentages (columns 5 and 7) could be among others regarded as suggestion of allocation of subject EU regional subventions to the candidate countries to particular regions according to level of development of the field of interest observed across them.

Table 2. Results of calculations of complex feature and Shapley values without a priori unions and with a priori unions

Powiats	Subregions	Values of complex feature	Shapley values (without a priori unions)	Participation in regional development calculated on the basis of Shapley values (without a priori unions) in %	Shapley values with a priori unions	Participation in subregional development calculated on the basis of Shapley values with a priori un-ions in %
1	2	3	4	5	6	7

Powiats	Subregions	Values of complex feature	Shapley values (without a priori unions)	Participation in regional development calculated on the basis of Shapley values (without a priori unions) in %	Shapley values with a priori unions	Participation in subregional development calculated on the basis of Shapley values with a priori un-ions in %
Gostyński	Kaliski	106,529	3,564	0,648	2,131	1,11
Rawicki		115,792	3,965	0,721	2,336	1,21
Ostrowski		117,149	4,032	1,230	2,370	1,23
Pleszewski		118,355	4,096	0,745	2,405	1,25
Kaliski		119,068	4,135	0,752	2,429	1,26
Kępiński		139,861	5,576	1,014	3,260	1,69
Jarociński		145,697	6,107	1,111	3,552	1,84
Ostrzeszowski		146,555	6,200	1,128	3,609	1,87
Krotoszyński		151,828	6,859	1,248	4,137	2,15
Kalisz[1]		549,757	163,014	29,652	166,552	86,39
Słupecki	Koniński	115,976	3,974	0,723	4,639	9,76
Turecki		121,528	4,280	0,779	4,917	10,35
Koniński		130,888	4,885	0,889	5,541	11,66
Konin[1]		220,664	18,093	3,291	15,989	33,65
Kolski		222,405	18,441	3,354	16,424	34,57
Czarnkowsko - trzcianecki	Pilski	0,780	0,022	0,004	0,031	0,10
Złotowski		50,666	1,536	0,279	2,525	7,80
Wągrowiecki		132,354	4,990	0,908	7,971	24,63
Chodzieski		138,455	5,459	0,993	8,581	26,51
Pilski		161,849	8,290	1,508	13,260	40,96
Nowotomyski	Poznański	2,333	0,068	0,012	0,033	0,02
Leszczyński		46,552	1,408	0,256	0,714	0,51
Obornicki		54,717	1,667	0,303	0,850	0,61
Wolsztyński		54,840	1,671	0,304	0,852	0,61
Międzychodzki		75,106	2,370	0,431	1,257	0,90
Kościański		85,229	2,732	0,497	1,482	1,06
Wrzesiński		85,242	2,732	0,497	1,483	1,06
Szamotulski		98,252	3,232	0,588	1,854	1,33
Średzki		109,875	3,703	0,674	2,242	1,60
Grodziski		113,493	3,860	0,702	2,386	1,71
Śremski		126,079	4,564	0,830	3,016	2,16

[1] City with the poviat status.

Powiats	Subregions	Values of complex feature	Shapley values (without a priori unions)	Participation in regional development calculated on the basis of Shapley values (without a priori unions) in %	Shapley values with a priori unions	Participation in subregional development calculated on the basis of Shapley values with a priori un-ions in %
Gnieźnieński		145,928	6,130	1,115	4,339	3,11
Poznański		230,803	20,541	3,736	14,900	10,66
Leszno[1]		496,714	109,970	20,003	104,330	74,66
Poznań[1]	Poznań city	491,955	107,591	19,571	137,359	100,00

Source: Own calculations conducted by application of results of calculations of the Weber median made by Wysocki and Lira (Agricultural University in Poznań) for the purpose of a publication by the Statistical Office in Poznań (2001).

Acknowledgements

I am grateful to the anonymous referees for careful reading of the article and useful remarks, comments and suggestions.

REFERENCES

CARRERAS,F.and OWEN, G. (1988) *Evaluation of the Catalonian Parliament, 1980 – 1984, Mathematical Social Sciences,* vol. 15, pp. 87-92.

GAMBARELLI, G. (1997) *Takeover,* Department of Mathematics, Statistics, Intormatics and Applications University of Bergamo (mineo).

LITTLECHILD, S.C. and. OWEN, G. (1973), *A simple expression for the Shapley value in the special case*, Management Sciences vol. 20., pages 370-372.

S. C. LITTLECHILD (1974) *A simple expression for Nucleolus in the special case.* International Journal of Game Theory vol. 3., pages 21-29.

MALAWSKI, M., WIECZOREK, A.and SOSNOWSKA, H. (1997) *Konkurencja i kooperacja. Teoria gier w ekonomii i naukach społecznych*. (*Competition and Co–operation. Game theory in economics and social sciences),*

Wydawnictwo Naukowe PWN [PWN Scientific Publishing House], Warsaw.

MALINA, A.and. ZELIAŚ, A. (1998) *On building taxonometric measures on living conditions*. Statistics in Transition, vol. 3, No. 3, pages 523-544.

MŁODAK, A. (1998) *Zastosowanie niektórych rozwiązań gier kooperacyjnych z prekoalicjami w badaniu ważonych gier większości na przykładzie wyników wyborów do Rady Miejskiej Kalisza. (An application of some solutions of co–operative games in a research of weighted majority games upon the example of results of election to the Town Council in Kalisz)*. Instytut Ekonometrii Szkoły Głównej Handlowej w Warszawie, Prace z ekonomii matematycznej [Institute of Econometrics, Warsaw School of Economics, Series: Mathematical Economics] No. 1/EM/98.

MŁODAK, A. (2000) *Zastosowanie rozwiązań gier kooperacyjnych w analizie tworzenia się koalicji większościowych w ciałach decyzyjnych na przykładzie Sejmu RP (An application of the solutions of co–operative games in an analysis of creating majority coalitions in decisions – making organs upon the example of the Sejm of the Republic of Poland),* Przegląd Statystyczny [Statistical Review], R. XLVII, pages 145-160.

MŁODAK, A. (2002) *Gry kooperacyjne w analizie taksonomicznej na przykładzie rynku pracy* (*Co–operative games in a taxonomic analysis for instance of labour market)*, Wiadomości Statystyczne [Statistical News] No. 3/2002, pp.6-17.

OWEN, G. (1977) *Values of games with a priori unions*. Mathematical Economics and Game Theory, Ed. by R. Hennan and O. Moeschlin, Berlin, pages 77-88.

SHAPLEY, L.S. (1953) *A value for n – person game,* Annals of Mathematical Studies, vol. 28, pages 307-317.

SOSNOWSKA, H. (1993) *O pewnej metodzie analizy tworzenia się koalicji parlamentarnych. (About some method of analysis of creation of parliamentary coalitions),* Przegląd Statystyczny [Statistical Review], R. XL, pages 229-231.

SOSNOWSKA, H. (1995) *Analiza programów wyborczych i wyników wyborów za pomocą wartości Shapleya z prekoalicjami na przykładzie wyborów do Sejmu z 19.09.1993 r.* (*Analysis of electoral programmes and results of election made with the help of a Shapley value with a priori unions upon the example of the general election to Sejm of the Republic of Poland conducted 19th September 1993)*. Roczniki Kolegium Analiz Ekonomicznych [Yearbooks of Economical Analysis Department of the Warsaw School of Economy], issue 2/95, pages 181-188.

ŚMIŁOWSKA, T. (1997) *Statystyczna analiza poziomu życia ludności w ujęciu przestrzenny*m (*Statistical analysis of living conditions of Polish population in territorial section),* Studia i Prace. Z prac Zakładu Badań Statystyczno – Ekonomicznych GUS i PAN. [Studies and Papers. Working Papers of the Research Centre for Economic and Statistical Studies]. No. 247.

STATISTICAL OFFICE IN POZNAŃ (2001) *Przestrzenne zróżnicowanie gospodarki województwa wielkopolskiego w 1999 r.* (*Spatial differentiation of the economy of the Wielkopolskie Voivodship in 1999).*

VÁZQUEZ – BRAGE, M., van DEN NOUWELAND, A.I. GARCÍA–JURADO, I. (1996) *Owen's coalitional value and aircraft landing fees* (typescript).

WAGNER,W., WYSOCKI, F. and LIRA, J (2000) *Mediana w zagadnieniach porządkowania obiektów wielocechowych (Median in ordering problem of multi–feature objects),* 4th Międzynarodowa Konferencja Naukowa „Statystyka regionalna w służbie samorządu lokalnego i biznesu" Poznań – Kiekrz (typescript).

WEBER, A. (1909, reprint 1971). *Theory of location of industries*, Translated with an introduction and notes by Carl J. Friedrich, Ed. by Russel & Russel, New York, 1971.

STATISTICS IN TRANSITION, October 2002
Vol. 5, No. 5, pp. 873–886

AN ALTERNATIVE TO AN IMPROVED RANDOMIZED RESPONSE STRATEGY

Housila P. Singh and Nidhi Mathur[1]

ABSTRACT

In this paper, we have suggested a class of estimators $\hat{\pi}_M$, say, for estimating π, the proportion of a population having a sensitive attribute. It is shown that the suggested class of estimators of $\hat{\pi}_M$ of π is more efficient than by Mangat's (1994) estimator $\hat{\pi}_m$. Since $\hat{\pi}_M$ involves an unknown population parameter M, it has , therefore little practical utility. Replacing M in $\hat{\pi}_M$ by its different estimated values, various estimators $\hat{\pi}_M^{(i)}$; i = 1,2, h; have been suggested for use in practice. Exact efficiency of $\hat{\pi}_M^{(i)}$; i = 1,2,h; have been worked out theoretically and numerically. Approximate variance expression for the proposed estimator is also given.

Key words: Randomized response technique, Sensitive attribute, Class of estimators, Equal probabilities with replacement, Variance expression.

1. Introduction

This rising concern about "invasion of privacy" demonstrates an important challenge to the applied statistician to formulated new theory and procedures for the collection of sensitive data. Warner (1965) introduced an skilful interviewing procedure known as randomized response technique (RRT) for estimation of proportion π of population belonging to sensitive characteristic A. Subsequently, many others have reported different RRT, for instance, see Hedayat and Sinha (1991). Mangat and Singh (1990) suggested a two stage RRT which requires the use of two randomized devices, this makes the interview procedure a little cumbersome. Owing to this Mangat (1994) suggested a relatively simple RRT described as below:

[1] School of Studies in Statistics, Vikram University, Ujjain 456010, (M.P.), India.

Each of n respondents assumed to be selected by simple random sampling with replacement (SRSWR) scheme, is instructed to say 'yes' if he or she has the attribute A. If he or she does not have attribute A, the respondent is required to use the Warner randomization device consisting of two statements:

i. 'I belong to attribute A' and
ii. 'I do not have attribute A'

represented with probability p and (1-p) respectively. Then he or she is to report 'yes' or 'no' according to the outcome of this randomization device and the actual status that he or she with respect to attribute A.

The whole procedure is completed by the respondent unobserved by the interviewer.

The probability of a 'yes' answer for this method is given by

$$\alpha = \pi + (1-\pi)(1-p) \tag{1.1}$$

For estimating π, Mangat (1994) suggested an unbiased and maximum likelihood estimator of π as

$$\hat{\pi}_m = \frac{(\hat{\alpha} - 1 + p)}{p} \tag{1.2}$$

where $\hat{\alpha} = \frac{n_1}{n}$, is the proportion of 'yes' answers in the sample. n is the number of respondents selected by SRSWR and n_1 is the number of 'yes' answers out of n responses.

The variance of $\hat{\pi}_m$ is given by

$$V(\hat{\pi}_m) = \frac{\pi(1-\pi)}{n} + \frac{(1-\pi)(1-p)}{np} = \frac{\alpha(1-\alpha)}{np^2} \tag{1.3}$$

In this paper, we have suggested a class of estimators of π and their properties are studied.

2. The Class of Estimators

Motivated by Searls (1964), we suggest the following class of estimators of π as

$$\hat{\pi}_M = M\hat{\pi}_m \tag{2.1}$$

where M is constant to be chosen suitably.

The mean square error (MSE) of $\hat{\pi}_M$ is given by

$$MSE(\hat{\pi}_M) = \frac{1}{p^2}\left[M^2 \frac{\alpha(1-\alpha)}{n} + (M-1)^2(\alpha - \bar{p})^2\right] \tag{2.2}$$

where $\bar{p} = (1-p)$.

The $MSE(\hat{\pi}_M)$ at (2.2) is minimized for

$$M = \frac{(\alpha - \bar{p})^2}{(\alpha - \bar{p})^2 + \frac{\alpha(1-\alpha)}{n}} = M_{opt} \text{ (say)} \tag{2.3}$$

which essentially lies between 0 and 1.

Thus the minimum MSE of $\hat{\pi}_M$ is given by

$$\text{min. MSE}(\hat{\pi}_M) = \frac{\alpha(1-\alpha)(\alpha - \bar{p})^2}{p^2\left[n(\alpha - \bar{p})^2 + \alpha(1-\alpha)\right]} \tag{2.4}$$

From (1.3) and (2.4) we have

$$V(\hat{\pi}_m) - \text{min.MSE}(\hat{\pi}_M) = \frac{\alpha(1-\alpha)}{np^2} - \frac{M_{opt}^2\,\alpha(1-\alpha)}{np^2} - \frac{(M_{opt}-1)^2(\alpha - \bar{p})^2}{p^2}$$

$$= \frac{\alpha(1-\alpha)(1-M_{opt})}{np^2}\left[\frac{\{\alpha(1+\alpha) + n(\alpha - \bar{p})^2\}}{\{\alpha(1-\alpha) + n(\alpha - \bar{p})^2\}}\right] \tag{2.5}$$

Clearly the right hand side of the expression (2.5) is now positive because $0 < M_{opt} < 1$. Hence we conclude that the minimum MSE of $\hat{\pi}_M$ in (2.4) is less than that of Mangat's estimator $\hat{\pi}_m$. For other relevant references in this context reader is referred to Mangat et. al. (1991) and Sampath et.al.(1995).

3. Proporties of the Estimator $\hat{\pi}_M$ when $0 < M \leq 1$

The MSE of $\hat{\pi}_M$ is given by

$$\text{MSE}(\hat{\pi}_M) = \frac{1}{p^2}\left[M^2\left\{\frac{\alpha(1-\alpha)}{n} + (\alpha - \bar{p})^2\right\} - 2M(\alpha - \bar{p})^2 + (\alpha - \bar{p})^2\right]$$

which is less than the variance of $\hat{\pi}_m$ if

$$\frac{\left\{(\alpha - \bar{p})^2 - \frac{\alpha(1-\alpha)}{n}\right\}}{\left\{(\alpha - \bar{p})^2 + \frac{\alpha(1-\alpha)}{n}\right\}} < M \leq 1.$$

The regions of M have been computed for different values of n, p and π for which the suggested estimator $\hat{\pi}_M$ is more efficient than Mangat (1994) estimator $\hat{\pi}_m$ presented in Table 3.1.

Table 3.1. The range of M for different values of n, p, π.

π=0.05				
p↓ n→	5	10	20	50
0.6	0~1	0~1	0~1	0~1
0.7	0~1	0~1	0~1	0~1
0.8	0~1	0~1	0~1	0~1
0.9	0~1	0~1	0~1	0~1
π=0.1				
0.6	0~1	0~1	0~1	0~1
0.7	0~1	0~1	0~1	0.02~1
0.8	0~1	0~1	0~1	0.23~1
0.9	0~1	0~1	0.03~1	0.45~1
π=0.2				
0.6	0~1	0~1	0.07~1	0.49~1
0.7	0~1	0~1	0.23~1	0.60~1
0.8	0~1	0.05~1	0.38~1	0.69~1
0.9	0~1	0.23~1	0.53~1	0.78~1

To have the tangible idea about the performance of $\hat{\pi}_M$, we have computed the percent relative efficiency (PRE) of $\hat{\pi}_M$ w.r.t. to $\hat{\pi}_m$ for different values of n, p,π and M using the formula :

$$\text{PRE}(\hat{\pi}_M, \hat{\pi}_m) = \left[M^2 + \frac{n(\alpha - \bar{p})^2}{\alpha(1-\alpha)}(M-1)^2 \right]^{-1} \times 100$$

The results have been presented in Table 3.2.

Table 3.2. Percent Relative Efficiency of $\hat{\pi}_M$ w.r.t. $\hat{\pi}_m$ for different values of n,p,π and M.

π=0.05					
M	**p↓ n→**	**5**	**10**	**20**	**50**
0.25	0.6	1373.11	1202.58	963.30	603.23
	0.7	1282.62	1070.31	804.11	460.50
	0.8	1147.17	894.12	620.41	323.40

π=0.05					
M	**p↓ n→**	**5**	**10**	**20**	**50**
	0.9	922.18	647.76	406.08	191.61
0.50	0.6	392.79	385.83	372.63	337.95
	0.7	289.30	379.15	360.37	313.74
	0.8	383.19	367.74	340.30	278.05
	0.9	369.80	343.84	301.50	220.18
0.75	0.6	177.42	177.06	176.34	174.22
	0.7	177.24	176.70	175.63	172.51
	0.8	176.92	176.06	174.38	169.52
	0.9	176.18	174.61	171.55	162.99
π=0.1					
0.25	0.6	968.42	694.34	443.37	212.72
	0.7	822.22	553.27	334.46	152.97
	0.8	658.82	414.81	238.30	104.67
	0.9	475.00	278.90	152.76	064.82
0.50	0.6	372.97	349.37	310.11	231.93
	0.7	361.96	330.52	281.61	195.02
	0.8	345.21	303.61	244.66	154.60
	0.9	316.67	262.07	194.87	110.14
0.75	0.6	176.36	174.96	172.23	164.53
	0.7	175.73	173.72	169.84	159.19
	0.8	174.70	171.72	166.06	151.12
	0.9	172.73	167.96	159.16	137.56
π=0.2					
0.25	0.6	444.92	258.39	140.54	059.34
	0.7	349.38	196.10	104.45	043.48
	0.8	266.67	145.45	076.19	031.37
	0.9	194.36	103.46	053.46	021.82
0.50	0.6	310.45	253.66	185.71	102.97
	0.7	286.18	222.78	154.39	080.37
	0.8	257.14	189.47	124.14	061.02
	0.9	221.78	153.42	094.92	044.27
0.75	0.6	172.26	167.07	157.58	134.63
	0.7	170.25	163.34	151.07	123.29
	0.8	167.44	158.24	142.57	109.92
	0.9	163.21	150.84	130.99	093.92

Table 3.1, shows that the proposed estimator $\hat{\pi}_M$ is more efficient than Mangat's estimator $\hat{\pi}_m$ for full range of M (i.e. 0<M<1) and all values of n when the value of π(=0.05) is small i.e. the proportion of persons possessing attribute is

small, which happened in many practical situations. The range of M for $\hat{\pi}_M$ to be more efficient than $\hat{\pi}_m$ decreases as the value of π and n increase.

From Table 3.2, it is observed that the saving in MSE due to $\hat{\pi}_M$ compared to $\hat{\pi}_m$ is large for smaller values of n, π and M. It may be noted that when n=5, π=0.05 and M=0.25, the MSEs of $\hat{\pi}_M$ are very small if p=0.6. Therefore in these cases the percentage of efficiencies are quit large.

Finally, with these numerical illustrations we conclude that the proposed estimator is to be recommended for its use in practice for small values of π, n, M and p. In practice, small sample sizes are desirable when the survey procedure, like RRT is lavish.

4. Estimators Based on Estimated Optimum Constant

The optimum value of M in (2.3) can be rewritten as

$$M_{opt}=\frac{\pi^2}{\pi^2+V(\hat{\pi}_m)} \tag{4.1}$$

A consistent estimate of M_{opt} is given by

$$\hat{M}_{opt}^{(1)}=\frac{\hat{\pi}_m^2}{\left\{\hat{\pi}_m^2+\frac{\hat{\alpha}(1-\hat{\alpha})}{np^2}\right\}} \tag{4.2}$$

Substitution of $\hat{M}_{opt}^{(1)}$ in place of M in (2.1) yields an estimator of π as

$$\hat{\pi}_M^{(1)}=\frac{\hat{\pi}_m^3}{\left\{\hat{\pi}_m^2+\frac{\hat{\alpha}(1-\hat{\alpha})}{np^2}\right\}} \tag{4.3}$$

Replacing π by $\hat{\pi}_m$ and $V(\hat{\pi}_m)$ by its unbiased estimator $\hat{V}(\hat{\pi}_m)=\frac{\hat{\alpha}(1-\hat{\alpha})}{(n-1)p^2}$ in (4.1) we get another consistent of M_{opt} as

$$\hat{M}_{opt}^{(2)}=\frac{\hat{\pi}_m^2}{\left\{\hat{\pi}_m^2+\frac{\hat{\alpha}(1-\hat{\alpha})}{(n-1)p^2}\right\}} \tag{4.4}$$

and thus the resulting estimator of π is given by

$$\hat{\pi}_{M}^{(2)} = \frac{\hat{\pi}_{m}^{3}}{\left\{\hat{\pi}_{m}^{2} + \frac{\hat{\alpha}(1-\hat{\alpha})}{(n-1)p^{2}}\right\}} \tag{4.5}$$

A more flexible estimator of π is given by

$$\hat{\pi}_{M}^{(h)} = \frac{\hat{\pi}_{m}^{3}}{\left\{\hat{\pi}_{m}^{2} + \frac{h\hat{\alpha}(1-\hat{\alpha})}{np^{2}}\right\}}, \tag{4.6}$$

where h(>0) is constant to be chosen suitably. For h=1, $\hat{\pi}_{M}^{(h)}$ reduces to $\hat{\pi}_{M}^{(1)}$ in (4.3), while for h=$\left(\frac{(n-1)}{n}\right)^{-1}$ it boils down to $\hat{\pi}_{M}^{(2)}$ in (4.5).

To obtain the approximate mean square error of $\hat{\pi}_{M}^{(h)}$, we write

$$\hat{\alpha} = \alpha(1+e) \text{ such that E(e) =0 and } E(e^{2}) = (1-\alpha)/n\alpha$$

Expressing $\hat{\pi}_{M}^{(h)}$ at (4.6) in term of e's, we have

$$\hat{\pi}_{M}^{(h)} = \pi\left(1+\frac{\alpha e}{p\pi}\right)\left[1+\frac{h\alpha(1-\alpha)}{np^{2}\pi^{2}}(1+e)\left\{1-\frac{\alpha e}{(1-\alpha)}\right\}\left\{1+\frac{\alpha e}{p\pi}\right\}\right]^{-1}$$

or
$$\left(\hat{\pi}_{M}^{(h)} - \pi\right) \cong \frac{\alpha e}{p} - \frac{h\alpha(1-\alpha)}{np^{2}\pi}\left[1+\frac{\alpha e}{p\pi}\left\{\frac{(1-2\alpha)p\pi}{(1-\alpha)}-1\right\}\right]$$

Squaring both sides of above expression and then taking expectations, we get the MSE of $\hat{\pi}_{M}^{(h)}$ to terms of order n^{-2} as

$$MSE\left(\hat{\pi}_{M}^{(h)}\right) = V(\hat{\pi}_{m}) + \frac{h\{V(\hat{\pi}_{m})\}^{2}}{\pi^{2}}\left[h - 2\left\{\frac{(1-2\alpha)p\pi}{\alpha(1-\alpha)}-1\right\}\right]$$

which is less than the variance of $\hat{\pi}_{m}$ if either

$$0 < h < 2\left\{\frac{(1-2\alpha)p\pi}{\alpha(1-\alpha)} - 1\right\}$$

or

$$2\left\{\frac{(1-2\alpha)p\pi}{\alpha(1-\alpha)} - 1\right\} < h < 0$$

Further, the exact MSE of an estimator $d = \hat{\pi}_M^{(1)}, \hat{\pi}_M^{(2)}, \hat{\pi}_M^{(h)}$ of π is given by

$$MSE(d) = \sum_{n_1=0}^{n} (d-\pi)^2 \; {}^nC_{n_1} \alpha^{n_1} (1-\alpha)^{n-n_1} \tag{4.7}$$

as n_1 follows the Binomial distribution B(n,α), where ${}^nC_{n_1} = \frac{n!}{n_1!\,(n-n_1)!}$.

Thus the percent relative efficiency of an estimator d with respect to $\hat{\pi}_m$ is computed from:

$$PRE(d, \hat{\pi}_m) = \frac{V(\hat{\pi}_m)}{MSE(d)} \times 100$$

$$= \frac{\alpha(1-\alpha)}{np^2}\left[\sum_{n_1=0}^{n} (d-\pi)^2 \; {}^nC_{n_1} \alpha^{n_1} (1-\alpha)^{n-n_1}\right]^{-1} \times 100 \tag{4.8}$$

To have tangible idea about the performance of the various estimators of π. We have computed the PRE's of d for different values of n, p,π and α in the Table 4.1(a),4.1(b) and 4.1(c).

Table 4.1 (a). Percent Relative Efficiency $\hat{\pi}_M^{(1)}, \hat{\pi}_M^{(2)}$ and $\hat{\pi}_M^{(h)}$ w.r.t. $\hat{\pi}_m$

π = **0.05**					
p↓	**n→ Estimator↓**	**5**	**10**	**20**	**50**
0.6	$\hat{\pi}_M^{(1)} = \hat{\pi}_M^{(h=1)}$	159.92	182.12	189.77	182.55
	$\hat{\pi}_M^{(2)}$	169.70	189.68	193.87	183.96
	$\hat{\pi}_M^{(h=6)}$	252.58	472.74	534.66	422.95
	$\hat{\pi}_M^{(h=16)}$	285.36	877.15	970.19	564.55
0.7	$\hat{\pi}_M^{(1)} = \hat{\pi}_M^{(h=1)}$	153.81	178.79	184.85	172.63
	$\hat{\pi}_M^{(2)}$	161.10	185.84	188.66	173.82
	$\hat{\pi}_M^{(h=6)}$	221.60	411.23	468.42	342.07

π = **0.05**					
p↓	**n→ Estimator↓**	**5**	**10**	**20**	**50**
	$\hat{\pi}_M^{(h=16)}$	247.78	603.53	731.45	402.78
0.8	$\hat{\pi}_M^{(1)} = \hat{\pi}_M^{(h=1)}$	143.77	171.92	176.62	157.96
	$\hat{\pi}_M^{(2)}$	150.20	177.90	179.62	158.85
	$\hat{\pi}_M^{(h=6)}$	200.80	305.12	375.25	255.62
	$\hat{\pi}_M^{(h=16)}$	220.49	351.04	485.26	266.61
0.9	$\hat{\pi}_M^{(1)} = \hat{\pi}_M^{(h=1)}$	148.98	150.56	158.97	134.50
	$\hat{\pi}_M^{(2)}$	156.57	154.10	161.25	134.97
	$\hat{\pi}_M^{(h=6)}$	214.77	209.53	240.56	164.12
	$\hat{\pi}_M^{(h=16)}$	231.76	217.26	243.80	150.92

Table 4.1(b). Percent Relative Efficiency $\hat{\pi}_M^{(1)}$, $\hat{\pi}_M^{(2)}$ and $\hat{\pi}_M^{(h)}$ w. r. t. $\hat{\pi}_m$.

	n→	**π=0.1**			
p↓	**Estimator↓**	**5**	**10**	**20**	**50**
0.6	$\hat{\pi}_M^{(1)} = \hat{\pi}_M^{(h=1)}$	155.49	170.23	165.71	138.78
	$\hat{\pi}_M^{(2)}$	164.43	176.35	168.40	139.31
	$\hat{\pi}_M^{(h=6)}$	238.60	363.31	313.74	182.44
	$\hat{\pi}_M^{(h=16)}$	263.29	524.10	376.67	174.89
0.7	$\hat{\pi}_M^{(1)} = \hat{\pi}_M^{(h=1)}$	149.16	163.52	154.89	124.29
	$\hat{\pi}_M^{(2)}$	155.97	168.87	157.03	124.60
	$\hat{\pi}_M^{(h=6)}$	210.28	307.84	252.45	139.43
	$\hat{\pi}_M^{(h=16)}$	229.85	380.98	273.06	124.92
0.8	$\hat{\pi}_M^{(1)} = \hat{\pi}_M^{(h=1)}$	137.55	152.98	140.68	109.15
	$\hat{\pi}_M^{(2)}$	142.90	157.18	142.17	109.25
	$\hat{\pi}_M^{(h=6)}$	183.53	233.61	191.16	103.28
	$\hat{\pi}_M^{(h=16)}$	196.26	246.50	187.24	086.41
0.9	$\hat{\pi}_M^{(1)} = \hat{\pi}_M^{(h=1)}$	132.30	132.36	121.74	094.94
	$\hat{\pi}_M^{(2)}$	137.65	134.49	122.46	094.86
	$\hat{\pi}_M^{(h=6)}$	175.03	159.79	130.29	073.30

p↓	n→ Estimator↓	π=0.1			
		5	10	20	50
	$\hat{\pi}_M^{(h=16)}$	181.44	153.77	114.74	056.08

Table 4.1(c). Percent Relative Efficiency $\hat{\pi}_M^{(1)}$, $\hat{\pi}_M^{(2)}$ and $\hat{\pi}_M^{(h)}$ w. r. t. $\hat{\pi}_m$.

p↓	n→ Estimator↓	π=0.2			
		5	10	20	50
0.6	$\hat{\pi}_M^{(1)} = \hat{\pi}_M^{(h=1)}$	140.27	137.70	118.12	090.49
	$\hat{\pi}_M^{(2)}$	146.07	104.43	118.68	090.38
	$\hat{\pi}_M^{(h=6)}$	252.58	472.74	534.66	422.95
	$\hat{\pi}_M^{(h=16)}$	188.47	199.87	115.88	053.93
0.7	$\hat{\pi}_M^{(1)} = \hat{\pi}_M^{(h=1)}$	132.32	127.26	107.03	084.81
	$\hat{\pi}_M^{(2)}$	136.65	129.17	107.23	084.62
	$\hat{\pi}_M^{(h=6)}$	221.60	411.23	468.42	342.07
	$\hat{\pi}_M^{(h=16)}$	164.59	151.94	087.17	042.10
0.8	$\hat{\pi}_M^{(1)} = \hat{\pi}_M^{(h=1)}$	120.85	115.83	097.25	082.49
	$\hat{\pi}_M^{(2)}$	123.43	116.89	097.14	082.25
	$\hat{\pi}_M^{(h=6)}$	200.80	305.12	375.25	255.62
	$\hat{\pi}_M^{(h=16)}$	136.43	111.27	065.23	033.46
0.9	$\hat{\pi}_M^{(1)} = \hat{\pi}_M^{(h=1)}$	109.35	103.35	090.13	084.60
	$\hat{\pi}_M^{(2)}$	111.40	103.45	089.76	084.32
	$\hat{\pi}_M^{(h=6)}$	214.77	209.53	240.56	164.12
	$\hat{\pi}_M^{(h=16)}$	113.88	079.04	048.12	027.24

From Table 4.1(a), 4.1(b) and 4.1(c) show that the suggested estimators $\hat{\pi}_M^{(1)}, \hat{\pi}_M^{(2)}$ and $\hat{\pi}_M^{(h)}$ are preferable over $\hat{\pi}_m$ for small values of n, p and π. The estimator $\hat{\pi}_M^{(2)}$ appears to be more efficient than $\hat{\pi}_M^{(1)}$. In practice, small sample sizes are desirable when the survey procedure like RRT, is costly.

In Warner's(1965) strategy, the probability of 'yes' answers is defined by

$$\delta = \pi p + (1-\pi)(1-p) \tag{4.9}$$

For estimating π, Warner (1965) suggested an unbiased estimator

$$\hat{\pi}_W = \frac{\hat{\delta} - \bar{p}}{2p-1}; \tag{4.10}$$

where $\hat{\delta} = \frac{n_1}{n}$ is the proportion of 'yes' answers in the sample.

The variance of $\hat{\pi}_W$ is given by

$$V(\hat{\pi}_W) = \frac{\delta(1-\delta)}{n(2p-1)^2} \quad \text{for} \quad p \neq \frac{1}{2}. \tag{4.11}$$

Thus the percent relative efficiency of an estimator d with respect to Warner's (1965) estimator $\hat{\pi}_W$ is given by

$$\text{PRE}(d, \hat{\pi}_W) = \frac{\delta(1-\delta)}{n(2p-1)^2}\left[\sum_{n_1=0}^{n}(d-\pi)^2 \; {}^{n}C_{n_1}\alpha^{n_1}(1-\alpha)^{n-n_1}\right]^{-1} \times 100. \tag{4.12}$$

To have tangible idea about the performance the estimator $d = \hat{\pi}_M^{(1)}, \hat{\pi}_M^{(2)}, \hat{\pi}_M^{(h)}$ with respect to Warner's (1965) estimator $\hat{\pi}_W$. We have computed the PRE's of d for different values of n, p, π and h in the table 4.2(a), 4.2(b) and 4.2(c).

Table 4.2(a). Percent Relative Efficiency $\hat{\pi}_M^{(1)}, \hat{\pi}_M^{(2)}$ and $\hat{\pi}_M^{(h)}$ w. r. t. $\hat{\pi}_W$.

p↓	**n→ Estimator↓**	**π=0.05**			
		5	**10**	**20**	**50**
0.6	$\hat{\pi}_M^{(1)} = \hat{\pi}_M^{(h=1)}$	1420.50	1617.65	1685.59	1621.47
	$\hat{\pi}_M^{(2)}$	1570.40	1684.79	1722.12	1634.05
	$\hat{\pi}_M^{(h=6)}$	2243.58	4199.13	4749.11	3756.81
	$\hat{\pi}_M^{(h=16)}$	2534.74	7791.31	8617.74	5014.58
0.7	$\hat{\pi}_M^{(1)} = \hat{\pi}_M^{(h=1)}$	460.09	534.83	552.95	516.41
	$\hat{\pi}_M^{(2)}$	481.91	555.93	564.34	519.98
	$\hat{\pi}_M^{(h=6)}$	662.88	1230.14	1401.20	1023.26
	$\hat{\pi}_M^{(h=16)}$	741.20	1805.39	2188.02	1204.86
0.8	$\hat{\pi}_M^{(1)} = \hat{\pi}_M^{(h=1)}$	248.16	296.76	304.38	272.65
	$\hat{\pi}_M^{(2)}$	259.26	307.07	310.04	274.20
	$\hat{\pi}_M^{(h=6)}$	346.60	526.68	647.72	441.24
	$\hat{\pi}_M^{(h=16)}$	380.59	605.94	837.62	460.21

p↓	n→ Estimator↓	π=0.05			
		5	10	20	50
0.9	$\hat{\pi}_M^{(1)} = \hat{\pi}_M^{(h=1)}$	183.12	185.06	195.40	165.32
	$\hat{\pi}_M^{(2)}$	192.45	189.41	198.20	165.91
	$\hat{\pi}_M^{(h=6)}$	263.98	257.54	295.68	201.73
	$\hat{\pi}_M^{(h=16)}$	284.86	266.98	299.67	185.50

Table 4.2(b). Percent Relative Efficiency $\hat{\pi}_M^{(1)}, \hat{\pi}_M^{(2)}$ and $\hat{\pi}_M^{(h)}$ w. r. t. $\hat{\pi}_W$.

p↓	n→ Estimator↓	π=0.1			
		5	10	20	50
0.6	$\hat{\pi}_M^{(1)} = \hat{\pi}_M^{(h=1)}$	1372.37	1502.49	1462.55	1224.89
	$\hat{\pi}_M^{(2)}$	1451.26	1556.48	1486.27	1229.60
	$\hat{\pi}_M^{(h=6)}$	2105.94	3206.57	2769.13	1610.22
	$\hat{\pi}_M^{(h=16)}$	2323.80	4625.74	3324.49	1543.61
0.7	$\hat{\pi}_M^{(1)} = \hat{\pi}_M^{(h=1)}$	439.75	482.10	456.64	366.43
	$\hat{\pi}_M^{(2)}$	459.84	497.91	462.97	367.34
	$\hat{\pi}_M^{(h=6)}$	619.96	907.59	744.28	411.06
	$\hat{\pi}_M^{(h=16)}$	677.64	1123.21	805.03	368.29
0.8	$\hat{\pi}_M^{(1)} = \hat{\pi}_M^{(h=1)}$	233.37	259.55	238.68	185.19
	$\hat{\pi}_M^{(2)}$	242.46	266.68	241.21	185.36
	$\hat{\pi}_M^{(h=6)}$	311.39	396.35	324.34	175.24
	$\hat{\pi}_M^{(h=16)}$	332.99	418.22	317.68	146.61
0.9	$\hat{\pi}_M^{(1)} = \hat{\pi}_M^{(h=1)}$	160.59	160.65	147.77	115.24
	$\hat{\pi}_M^{(2)}$	167.08	163.25	148.64	115.14
	$\hat{\pi}_M^{(h=6)}$	212.45	193.95	158.14	088.98
	$\hat{\pi}_M^{(h=16)}$	220.23	186.65	139.27	068.07

Table 4.2(c). Percent Relative Efficiency $\hat{\pi}_M^{(1)}, \hat{\pi}_M^{(2)}$ and $\hat{\pi}_M^{(h)}$ w. r. t. $\hat{\pi}_W$.

p↓	n→ Estimator↓	π=0.2			
		5	10	20	50
0.6	$\hat{\pi}_M^{(1)} = \hat{\pi}_M^{(h=1)}$	1246.25	1223.41	1049.45	803.99
	$\hat{\pi}_M^{(2)}$	1297.84	1247.65	1054.43	802.95
	$\hat{\pi}_M^{(h=6)}$	1649.89	1692.59	1137.34	619.75
	$\hat{\pi}_M^{(h=16)}$	1674.48	1775.79	1029.51	479.13
0.7	$\hat{\pi}_M^{(1)} = \hat{\pi}_M^{(h=1)}$	387.48	372.65	313.41	248.33
	$\hat{\pi}_M^{(2)}$	400.15	378.23	314.00	247.78
	$\hat{\pi}_M^{(h=6)}$	479.03	457.23	299.10	170.26
	$\hat{\pi}_M^{(h=16)}$	481.97	444.93	255.26	123.29
0.8	$\hat{\pi}_M^{(1)} = \hat{\pi}_M^{(h=1)}$	202.90	194.49	163.28	138.50
	$\hat{\pi}_M^{(2)}$	207.24	196.27	163.10	138.09
	$\hat{\pi}_M^{(h=6)}$	232.32	206.94	136.47	084.27
	$\hat{\pi}_M^{(h=16)}$	229.07	186.82	109.52	056.19
0.9	$\hat{\pi}_M^{(1)} = \hat{\pi}_M^{(h=1)}$	132.07	124.83	108.87	102.18
	$\hat{\pi}_M^{(2)}$	134.56	124.95	108.43	101.85
	$\hat{\pi}_M^{(h=6)}$	144.69	112.89	078.27	055.40
	$\hat{\pi}_M^{(h=16)}$	137.55	095.47	058.12	032.40

It is observed from Tables 4.2(a), 4.2(b) and 4.2(c) that :

i. When π=0.05, the performance of the suggested estimators $\hat{\pi}_M^{(1)}, \hat{\pi}_M^{(2)}$ and $\hat{\pi}_M^{(h)}$ {h=1,6,16} is better than Warner's estimator $\hat{\pi}_W$. The efficiency of the estimator $\hat{\pi}_M^{(h)}$ increases as h increases. Thus, the scalar 'h' plays a good role in improving the precision of the estimator $\hat{\pi}_M^{(h)}$.

ii. When π=0.1, the suggested estimators $\hat{\pi}_M^{(1)}, \hat{\pi}_M^{(2)}$ and $\hat{\pi}_M^{(h)}$ {h=1,6,16} perform better than Warner's estimator $\hat{\pi}_W$ except for higher values of n. It is noted that for smaller values of p and n, the efficiency of the estimator $\hat{\pi}_M^{(h)}$ increases as h increases, without loss of generality.

iii. When π=0.2, the suggested estimators $\hat{\pi}_M^{(1)}, \hat{\pi}_M^{(2)}$ and $\hat{\pi}_M^{(h)}$ {h=6,16} are more efficient than Warner's estimator $\hat{\pi}_W$ for smaller values of n.

iv. The gain in efficiency decreases as the values of π and p increase.

Finally, we conclude that the constructed estimators $\hat{\pi}_M^{(1)}, \hat{\pi}_M^{(2)}$ and $\hat{\pi}_M^{(h)}$ are more precise than Warner's estimator $\hat{\pi}_W$ for smaller values of π. The estimator $\hat{\pi}_M^{(2)}$ is more efficient than $\hat{\pi}_M^{(1)}$. It is further noted that substantial gain in efficiency due to suggested estimators $\hat{\pi}_M^{(1)}, \hat{\pi}_M^{(2)}$ and $\hat{\pi}_M^{(h)}$ {h=6,16}over Warner's estimator $\hat{\pi}_W$ is observed when sample size n is small. In practice, such sample sizes are desirable when the survey procedure, like Randomized Response Technique (RRT) is expensive. Comparing the results of the tables 4.1(a), 4.1(b),4.1(c),4.2(a),4.2(b) and 4.2(c), it is clear that the gain in efficiency by using proposed estimators over Warner's estimator $\hat{\pi}_W$ is more in comparison to Mangat's estimator $\hat{\pi}_m$.

Acknowledgement

Authors are thankful to the learned referee and the editor Prof. J. Kordos for their valuable suggestions regarding the improvement of the paper.

REFERENCES

HEDAYAT, A. S. and SINHA, B.K. (1991): Design and Inference in finite population sampling. New York :Wiley.

MANGAT, N. S. and SINGH, R. (1990): An alternative randomized response procedure. Biometrika, 77,2,439-441.

MANGAT, N. S., SINGH, R. and SINGH, S. (1991): Alternative estimators in randomized response technique. The Aligarh Journal of Statistics, 11, 75-80.

MANGAT, N.S. (1994): An improved randomized response sampling strategy. Journal of Royal Statistical Society B, 56, (1), 93-95.

SAMPATH, S., UTHAYAKUMARAN, N. and TRACY, D.S. (1995): On the alternative estimator for randomized response technique. Journal of Indian Society of Agricultural Statistics, 47(3), 243-248.

SEARLS, D. T. (1964): The utilisation of a known coefficient of variation in the estimation procedure. Journal of American Statistical Assocation, 59, 1225-1226.

WARNER, S. L. (1965): Randomized response: A survey technique for eliminating evasive answer bias. Journal of American Statistical Assocation, 60, 63-69.

STATISTICS IN TRANSITION, October 2002
Vol. 5, No. 5, pp. 887–890

REPORTS

INTERNATIONAL FEDERATION OF CLASSIFICATION SOCIETIES CONFERENCE – IFCS-2002

Cracov, July 16-19, 2002

"Data Analysis, Classification and Related Methods"

International Federation of Classification Societies is the scientific organization with the aim of development of the theoretical and practical issues related to the classification and data analysis methods. IFCS was founded in Cambridge in 1985. Currently there are 12 member societies, including Section of Classification and Data Analysis of Polish Statistical Association (SKAD).

It is worth to present the list of former and present IFCS Presidents.

The IFCS Presidents

Years	President
1986 – 1987	Hans-Hermann Bock (Germany)
1988 – 1989	Robert Sokal (United States)
1990 – 1991	John Gower (United Kingdom)
1992 – 1993	William Day (Canada)
1994 – 1995	Allan Gordon (United Kingdom)
1996 – 1997	Douglas Carroll (United States)
1998 – 1999	Chikio Hayashi (Japan)
2000 – 2001	Jean-Paul Rasson (Belgium)
2002 – 2003	Carlo Lauro (Italy)

At present Krzysztof Jajuga (Wrocław University of Economics) is the member of the Executive Committee of IFCS. Andrzej Sokołowski (Cracow University of Economics) and Marek Walesiak (Wrocław University of Economics) are the members of IFCS Council.

International Federation of Classification Societies holds its biannual conferences. The eighth conference took place in Cracow, at the campus of Cracow University of Economics, on July 16-19, 2002. It was organized by the

team chaired by Andrzej Sokołowski (official chairman of Local Organizing Committee). Krzysztof Jajuga was the official chairman of Scientific Program Committee, consisting of about 35 professors from many countries, including Poland. The other members of Scientific Programme Committee from Poland were: Zdzisław Hellwig, Kazimierz Zając, Aleksander Zeliaś, Andrzej Sokołowski, Józef Pociecha, Tadeusz Grabiński, Andrzej Barczak, Czesław Domański, Marek Walesiak, Józef Dziechciarz, Eugeniusz Gatnar.

It is worth to present the sites of the previous and future IFCS Conferences.

The IFCS Conferences

Year	City – host of the conference
1987	Aachen (Germany)
1989	Charlottesville (United States)
1991	Edinburgh (United Kingdom)
1993	Paris (France)
1996	Kobe (Japan)
1998	Rome (Italy)
2000	Namur (Belgium)
2002	Cracow (Poland)
2004	Chicago (United States)

More than 200 persons coming from more than 30 countries participated in the conference in Cracow. During the conference almost 150 papers were presented. They can be classified to 4 groups.

1. Keynote Lectures

They were presented by:

- Hans-Hermann Bock – "Clustering Methods: from Classical Models to New Applications"
- Frank Hampel – "Some Thoughts about Classification"
- Wojtek J. Krzanowski – "Orthogonal Components for Grouped Data – Review and Applications".

2. Invited Lectures

They were presented by:

- Henk A.L. Kiers – "Should We Use Standard Errors or Cross-Validation in Component Analysis Techniques?"
- Edwin Diday – "From Data to Knowledge: Symbolic Data Analysis, Mixture Decomposition and Spatial Pyramidal Clustering"

- Jean-Paul Rasson – "Divisive Classification and Segmentation Trees with the Poisson Processes Hypothesis"
- Maurizio Vichi – "Clustering and Reduction of Three-way Data"
- Hamparsun Bozdogan – "A New Generation Multivariate Mixture-model Cluster Analysis of Normal and Nonnormal Data Using Information Measure of Complexity"
- Klaus Obermayer – "New Methods for the Clustering, Visualization, and Classification of Proximity Data"
- Yoshiharu Sato – "The Performance of an Autonomous Clustering Technique"

3. Invited Sessions

Here some professors were asked to organize the session of the specialized area. Here are the titles and the organizers of the sessions:

- "Optimization Heuristics in Data Analysis": Javier Trejos
- "Dissimilarities in Clustering and Data Analysis": Jean-Pierre Barthélemy
- "Probability Models for Clustering": Hans-Hermann Bock
- "Classification and Regression Trees": Eugeniusz Gatnar
- "Application of Classification and Data Analysis in Marketing": Reinhold Decker, Daniel Baier
- "Optimization Methods and Algorithms in Classification and Clustering": Patrick Groenen, Hamparsun Bozdogan
- "Bioinformatics and Classification": Berthold Lausen
- "The WEB Mining Challenge": Wolfgang Gaul

4. Contributed Papers

All other papers (more than 100) are contributions of participants. They were divided into 27 contributed sessions, namely:

- Multivariate Data Analysis (2 sessions)
- Classification and Clustering Methods (4 sessions)
- Applications of Classification and Data Analysis in Economics
- Applications of Classification and Data Analysis in Medicine
- Classification and Regression Trees
- Categorical Data Analysis
- Dissimilarities and Similarities
- Neural Networks and Related Topics
- Mixture Models
- Symbolic Data Analysis
- Classification and Data Analysis– General and Special Problems
- Correspondence Analysis
- Phylogenetic Methods
- Clustering – Evaluation and Validation

- Multiway Data Analysis (2 sessions)
- Applications of Classification and Data Analysis in Social and Behavioral Sciences (2 sessions)
- Multivariate Statistics (2 sessions)
- Graphs
- Applications of Classification and Data Analysis in Environmental and Biological Sciences

53 papers were published in the proceedings of the conference:
Jajuga K., Sokołowski A., Bock H.-H. (editors),
Classification, Clustering, and Data Analysis. Recent Advances and Applications,
Springer, 2002, p.492, ISBN 3-540-43691-X.

In addition, the book of abstracts was published:
Sokołowski A., Jajuga K. (editors), IFCS 2002, Data Analysis, Classification and Related Methods, Program and Abstracts.
Cracow University of Economics Publishers, Kraków 2002, p.208. ISBN 83-7252-134-4.

The conference was one of the most important events in the area of statistics, not only in Poland. It is worth to mention that this conference was accompanied by the one-day conference, held in Tomaszowice near Cracow, celebrating the 90th anniversary of Polish Statistical Association.

STATISTICS IN TRANSITION, October 2002
Vol. 5, No. 5, pp. 891–895

REPORTS

THE NINETIETH ANNIVERSARY OF THE FOUNDATION OF THE POLISH STATISTICAL ASSOCIATION – THE SCIENTIFIC CONFERENCE,

Cracow, Poland, 14-15 July 2002

The Polish Statistical Association was founded in Cracow in 1912. As of the moment of its founding, the Polish Statistical Association (PSA) oriented its activities towards the development and promotion of statistics – its theory, methodology, and research practices.

During the initial phase of its existence, the Association played an outstanding role in establishing the fundamental principles of Polish statistics, publishing *"Statystyka Polski"* (*Statistics of Poland*) in 1915 which may be treated as the first statistical yearbook of the Polish nation under partition[1]. The Association not only contributed greatly to the development of statistical methodology and promoting statistical data, but also to the awakening and enriching awareness of those issues in the general public. The Association played a crucial role in developing statistics after Poland had gained independence in November 1918, as the Polish Central Statistical Office was founded in July 1918 already. members of the Association of that period constituted the core organizational and methodological staff of the Central Statistical Office.

During the subsequent years, the Association operated together with other (mainly economic) scientific societies, or suspended its activities for certain periods of time. The Association became very active in the two years preceding World War Two, operating as an independent Polish Statistical Association, as well as during the immediate post-war period. The Association's own statistical journal had been published at the time (*"Przeglad Statystyczny" – The Statistical Review*); statistical lectures were also held, during which crucial problems concerning further development of the Polish statistics were discussed.

After World War Two, the Association was reactivated in 1947. The publishing of the *Statistical Review* journal was also renewed; the first issue of

[1] A. Krzyzanowski and K. Kumaniecki, STATYSTYKA POLSKI (Statistics of Poland). Published by the Polish Statistical Association, Cracow, 1915.

which was published in 1950. The Association also undertook the ambitious task of promoting statistics; its members delivered various types of lectures on statistics. Nevertheless, it ought to be remembered that the Polish Statistical Association had then in its activities depended upon pre-war statisticians, a distinct proportion of whom held well-established democratic points of view towards both statistics and social issues. This contributed greatly to the problems experienced by the Association in the coming years[1].

In March 1953, a decision was taken to liquidate the Association for political reasons. The Polish Statistical Association was officially dissolved in April 1955. The attitude displayed by the economist community, reducing the role of statistics to that of a minor economic instrument, was one of the deciding factors, which caused the limitation of activities, and subsequently the liquidation of the PSA. This decision set a barrier to the Association, banning it from official activity for a period of nearly 25 years. Some of the Association members were still active in Statistical Sections of the Polish Economic Association.

In April 1981, on the wave of a general social boom, the Polish Statistical Association was reactivated, and began to deal primarily with the creation of professional consciousness, and the integration of the statistical community. A programme of re-establishing the importance of statistics within the economic environment and social consciousness had been then launched, and still is under implementation. Despite all the difficulties, and the dismembering of the professional community, the Association flourished mainly among professional statisticians, providing also a forum for linking practical activities in the field of statistics with the activities of the scientific circles. Ever since the founding assembly, i.e. since April 1981, until mid-December 1985, the organizational outlines of the Association were being defined. The work was carried out under extremely difficult political conditions, without any support in terms of equipment or financial resources.

The Association began to develop more widespread activities in 1986. At the Association's general assembly in mid-December 1985, a programme of activities was adopted and is consequently carried out and developed until today. Taking into account the scope and scale of the activities, only the major ones are mentioned here[2] .

Initially, after the Association had been reactivated in 1981, there were less than 300 members, while towards the end of 2002, the Association had approximately 800 members grouped in 10 field branches.

---

[1] Polish Statistical Association (1992), Polskie Towarzystwo Statystyczne 1912-1992 (Polish Statistical Association 1912-1992), Warsaw.

[2] J. Kordos, Activities of the Polish Statistical Association, *Statistics in Transition,* vol. 1, Number 1, 1993.

The PSA Publications

Of the widespread programme of activities accomplished by the Polish Statistical Association (the Main Board, its Presidium, and relevant agencies, as well as the field branches), the most important ones are presented here.

Biuletyn Informacyjny and Kwartalnik Statystyczny (The Information Bulletin and Quarterly Statistics)

During its plenary session in October 1986, the PSA Main Board took a decision to publish the ***Biuletyn Informacyjny (The Information Bulletin)*** in order to inform the PSA members and supporters on a regular basis about the activities of the Association, as well as to provide information on various issues related to the national and foreign statistics. It has been assumed that the *Biuletyn* ought to integrate the PSA members, and to unite statisticians representing various academic centres, scientific-and-research institutes, state statistical organs, enterprises and establishments, as well as different organisational entities. From 1986 to 1998 44 issues were published. *The Information Bulleti*n was closed in 1998, and replaced by a new journal ***Kwartalnik Statystyczny (Quarterly Statistics)*** in 1999**.**

Wiadomosci Statystyczne (Statistical News)[1]

The PSA is a co-editor of the *Wiadomosci Statystyczne* monthly journal. Since August 1989, the Polish Statistical Association, in co-operation with the Central Statistical Office, has been the co-editor, of the *Wiadomosci Statystyczne (the Statistical News)* monthly journal; two representatives of the Association are members of the editorial board of the periodical. The journal publishes a section on the "Activities of the Polish Statistical Association" which provides information on major events taking place in the Association.

Sylwetki statystyków polskich (The Biographies of Polish statisticians) [2]

In 1987, the *Sylwetki statystyków polskich (The Biographies of Polish Statisticians)* monograph was published; the English version of the publication was issued in 1989. In 1993, a more complete and revised version - both in Polish and English – was prepared. The monograph presents biographies of outstanding Polish statisticians, from the medieval to modern times. Professional and scientific achievements are presented alongside with the life history of the statisticians - special attention paid to promoting activities and publications of the statisticians.

---

[1] T. Walczak, „Wiadomosci Statystyczne" (Statistical News) - the major journal of Official Statistics in Poland,*Statistics in Transition,* vol. 3, Number 4, 1998.

[2] Polish Statistical Association (1989), *Biographies of Polish Statisticians*. Published by the Central Statistical Office, Warsaw.

Publication of a Statistical Journal in English – Statistics in Transition

Following lengthy discussions, a decision was made to publish in English a journal entitled the *STATISTICS IN TRANSITION* which is edited also by foreign statisticians, mainly from countries undergoing transformation into market economy systems. The first issue of the journal was published in 1993. Till the end of 2002 thirty issues of the journal in 5 volumes were published.

The PSA Conferences and Scientific Seminars

In accordance with a resolution taken by the PSA Main Board in 1987, the Association has initiated the organisation of annual scientific conferences on selected statistical topics.

International Statistical Conferences

Apart from the local statistical events, the Association has - since 1991 - initiated and co-organized the following international conferences:

1) in 1991, a conference on the ***Poverty Measurement for Economies in Transition in Eastern European Countries***, which took place in Warsaw, October 7th-9th, 1991;
2) in 1992, a conference on ***Small Area Statistics and Survey Design*****s**, which took place in Warsaw, September 30th until October 3rd 1992.
3) in 1994, ***an International Conference in Memory of the Hundredth Anniversary of the Birth of Jerzy Neyman*** ,Warsaw, 25-26 November 1994;
4) in 1995 an ***International Scientific Conference on Methodological Issues of Time Use Surveys: Design and Analysis.*** Warsaw, May 31- June 2 1995

The papers submitted for those conferences were published in English as proceeding of the conferences and were distributed among the participants and interested statisticians from various countries.

The Jubilee Conference in 2002

On the 14 and 15th of July 2002 the jubilee conference on the occasion of Ninetieth Anniversary of the Foundation of the Polish Statistical Association took place in Cracov, Poland. The conference gathered more than 50 statisticians from different part of Poland. It was organised by the Department of Statistics of the Cracow University of Economics. Professor Czesław Domański, President of the Polish Statistical Association, was the chairman of the Conference while Professor Aleksander Zeliaś, Deputy Chairman of the Polish Statistical Association, was the chairman of the Organising Committee. Tadeusz Toczyński, the President of the Central Statistical Office, and Peter Mach, the President of the Statistical Office of the Slovak Republic, were the honourable guests of the Conference. Moreover Prof. W. Welfe, corresponding member of the Polish

Academy of Sciences, and Prof. R. Rudzkis, President of the Lithuanian Statistical Association, were sending us sincere congratulations and best wishes on the occasion of the 90th Anniversary of the Foundation of the Polish Statistical Association.

The conference was opened by Prof. Cz. Domański. He outlined in his introductory speech the main facts about scientific activities of the Polish Statistical Association in the years 1912–2002.

The meeting was conducted in form of two plenary sessions (the chairpersons of the sessions were: Prof. A. Zeliaś, Prof. T. Walczak). During the two conference days 5 lectures were presented and there was a panel discussion as well on: *Past for the Future* led by Professor Andrzej Barczak. Prof. A. Barczak gave the introduction to the discussion.

During the two plenary sessions the following lectures were presented:

1. *A Challenge of Statistics at the Beginning of 21st Century (*by Czesław Domański).
2. *The Ninetieth Anniversary of the Foundation of the Polish Statistical Association, History, Achievements and Outlooks (*by Kazimierz Zając).
3. *Ethical Problems in Statistical Investigations (*by Józef Oleński).
4. *Statistics in Development Process of Information Society (*by Tadeusz Toczyński).
5. *Some Quality Aspects in Small Area Statistics (*by Jan Kordos).

All lectures presented at the conference met with great interest. The knowledge gained can now be used in practical daily work to improve statistical investigations. At the end of the conference Prof. Cz. Domański, President of the Polish Statistical Association, summarised the main conclusions of the conference and emphasised the role of the scientists in the development of statistics in Poland.

Aleksander Zeliaś