PSYCHOMETRIKA—VOL. 68, NO. 4, 493–517
DECEMBER 2003
PRESIDENTIAL ADDRESS

PREDICTION AND CLASSIFICATION IN NONLINEAR DATA ANALYSIS: SOMETHING OLD, SOMETHING NEW, SOMETHING BORROWED, SOMETHING BLUE

JACQUELINE J. MEULMAN

LEIDEN UNIVERSITY

Prediction and classification are two very active areas in modern data analysis. In this paper, prediction with nonlinear optimal scaling transformations of the variables is reviewed, and extended to the use of multiple additive components, much in the spirit of statistical learning techniques that are currently popular, among other areas, in data mining. Also, a classification/clustering method is described that is particularly suitable for analyzing attribute-value data from systems biology (genomics, proteomics, and metabolomics), and which is able to detect groups of objects that have similar values on small subsets of the attributes.

Key words: multiple regression, optimal scaling, optimal scoring, statistical learning, data mining, boosting, forward stagewise additive modeling, additive prediction components, monotonic regression, regression splines, distance based clustering, clustering on variable subsets, COSA, genomics, proteomics, systems biology, categorical data, ordinal data, ApoE3 data, cervix cancer data, Boston housing data.

1. Preface

When I delivered this Presidential Address at the 68th Annual Psychometric Society Meeting in Sardinia, I intentionally included some very personal remarks. I have been deliberating for quite some time whether to include these in this written version or not. Only at the last moment I decided that I would. Among the many reasons to omit such personal notes, is, of course, that doing so creates a mixture of the personal and professional that readers may find distracting. I was motivated to include them by the fact that the contents of this Presidential Address to a large extent came from activities that kept me alive in a very difficult period that started just after I was made President-Elect of the Society.

In June 2001, I was diagnosed with a very rare and extremely aggressive form of cancer. Early detection (thanks to the "Powerplate" in my fitness center, a muscle strengthening, vibrating machine that made the tumor bleed "prematurely") , extremely quick actions by the Leiden University Medical Center (especially urologist Jaap Zwartendijk), and heavy chemo and radiation therapy afterwards, saved my life. Because of the cancer and recovery from the treatment, my Presidential term and this Presidential Address have certainly turned out different from what they probably would have been otherwise.

I started my Presidential Address in Sardinia by quoting a fortune cookie that I received just before I became President-Elect. It said: "Something you've been thinking about could work out well. You're following the blueprints, but don't let them limit you. A combination of planning, foresight and flying by the seat of your pants will get you there this time". When you become President-Elect of this Society, you certainly think about planning and foresight. Fate decided that in my case there was no planning and no foresight. What I was left with, was flying by the seat of my pants. I got there, but only with a lot of help from my friends. (I love the Joe Cocker version of the Beatles song, but I really couldn't write down that the help was little.) I use the term "friends" as a generic one, to refer to persons, but also to projects, such as the "COSA"

This article is based on the Presidential Address Jacqueline Meulman gave on July 9, 2003 at the 68th Annual Meeting of the Psychometric Society held near Cagliari, Italy on the island of Sardinia.—Editor

Special thanks are due to Brian Junker who gave me very helpful comments, and to Tim Null who made the printed version look as good as it does. Both waited patiently for me to finish, for which I'm forever grateful.

Requests for reprints should be sent to Jacqueline J. Meulman, Data Theory Group, Department of Education, Leiden University, P.O. Box 9555, 2300 RB Leiden, THE NETHERLANDS. E-Mail: meulman@fsw.leidenuniv.nl

0033-3123/2003-4/2003-1146-PA $00.75/0

project with Jerome Friedman, and the "ApoE3" project with Jan van der Greef, that helped me to "stay alive" since June 2001. (More about these projects later.)

First of all, I wish to express my great gratitude towards my friends in the organizing committee of the Sardinia IMPS2003 meeting, in particular Francesco Mola, Roberta Siciliano and Claudio Conversano (and AJ van der Meer). I have omitted Willem Heiser from this list, and not only because he is not from Naples. Willem deserves a special place, as then President-Elect of the Psychometric Society, and Chair of the Program Committee. At the time in July 2001, I took responsibility for the Scientific Program, but only to find out later that I simply couldn't do it. Willem relieved me from any possible duty, so that I could concentrate on my Presidential Address. Without him, there wouldn't have been the wonderful Scientific Program at the Sardinia Meeting. My gratitude to him is beyond words.

On the occasion of my Presidential Address, I also would like to thank my friends in the 2001 Organizing Committee of the International Meeting of the Psychometric Society in Osaka, Japan. Being a strong advocate of bringing the Society outside North America, it was time to organize a meeting in Japan. I had the pleasure to be very involved with the early preparations; I went to Japan to look at possible conference sites, and we choose Osaka University, close to beautiful Kyoto. Sometimes a race horse stumbles just before he reaches the finish; in my case, I had to start my chemo treatment just before the start of the meeting in Japan. The fact that I missed the very successful Osaka meeting still saddens me deeply. I wish to thank Haruo Yanai, Akinori Okada, Kazuo Shigemasu, and Yutaka Kano for the great meeting in Japan that they organized, and for producing the memorable proceedings volume (containing state-of-the art of psychometrics) that was the result of this meeting (Yanai, Okada, Shigemasu, Kano, & Meulman, 2003).

Before launching into the technical part of this paper, I wish to thank Larry Hubert. In the middle of my difficult recovery, he got me back to computing again, by supplying me with the wonderful Mac G4 powerbook (the Mac being my all-time favorite computer since 1985). He also put a lot of software on it, so I finally had no excuse anymore for not starting to work on my Presidential Address. It turned out to be the definite push that I needed. Larry also gave me the Florence Nightingale (one of my major heroes) scarf that I wore during my Presidential Address for good luck; Flo's inspiration never fails.

This Presidential Address has two main areas in its title: prediction and classification. It must be clear that from these two very large domains, we can only look at very modest subgroups. With respect to prediction, I've chosen to look at prediction in multiple regression with optimal transformations, and this will be done in the context of how prediction is used in statistical learning, and applied, a.o., in data mining (Hastie, Tibshirani, & Friedman, 2001). Classification will be used in the sense of clustering, and we will look at it in the context of analyzing a large number of variables while having only a small number of samples. Typical examples for this data analytical situation are to be found in the life sciences, with applications in genomics, proteomics, and metabolomics, recently subsumed under the name *systems biology* (e.g., see van der Greef et al., 2003). Both statistical learning, data mining, and systems biology are very exiting data analytic areas that in my opinion have not been given appropriate attention by psychometricians thus far.

2. Something Old

The term *nonlinear* in the title of this paper refers to nonlinear transformation of the variables in multivariate data. In the nonlinear transformation process, an appropriate quantification level has to be chosen for each of the variables. The most restricted transformation level is called *numeric*; it applies a linear transformation to the original scale values, so that the resulting variables will be standardized. Instead of a linear transformation, we have the choice between different nonlinear transformations, and these can either be *monotonic* with the original order of the objects in the data, or *nonmonotonic*. When the data are categorical, and the only fact we will take into account, is that a particular subset of the objects is in the same category (while others are in different ones), we talk about a nominal quantification (or a nonmonotonic transformation). The quantifications only maintain the class membership, and the categories obtain an optimal ordering. Nonmonotonic functions can also be used for continuous (numeric) and ordinal variables when nonlinear relationships among the variables are assumed. In these cases, we can collapse the data in a limited number of categories (sometimes called binning), and find an optimal quantification for the categories, but we can also fit a nonmonotonic spline transformation with a limited number of parameters.

Within the domain of either monotonic or nonmonotonic transformations, at least two approaches are available; we will consider fitting *optimal step functions* and *optimal spline functions*. The use of monotonic regression to transform the data was proposed originally in nonmetric multidimensional scaling analysis in Kruskal (1964a, 1964b). In the equivalent approach in nonlinear data analysis, the number of parameters that is fitted can be as large as the number of categories. Since this could lead to overfitting, more restricted classes of transformations were introduced into the psychometric literature. The most important ones form the class of regression splines. For splines, the number of parameters is determined by the degree of the spline that is chosen, and the number of interior knots. Because splines use less parameters, they usually will be smoother and more robust, albeit this is at the cost of less goodness of fit with respect to the overall loss function that is minimized, and at losing invariance of the results under one-to-one transformations of the data.

In this section we will review multiple regression with optimal scaling, a technique that originates with Kruskal's 1965 nonmetric version of ANOVA. This approach was followed upon by ADDALS (de Leeuw, Young, & Takane, 1976) and MORALS (Young, de Leeuw, & Takane, 1976). The collective work by the Leiden group at the department of Data Theory resulted in Gifi (1990). Winsberg and Ramsay (1980) replaced the original monotonic regression approach (that produced step functions) by monotonic regression splines (that produce smooth piecewise polynomial functions); a nice review is given in Ramsay (1988). In the meantime, optimal scaling had entered the mainstream statistical literature in the Breiman and Friedman (1985) paper on alternating conditional expectations (ACE), which became the *JASA* paper of the year. Very closely related is work on generalized additive models (GAM), extensively described in Hastie and Tibshirani (1990). Finally, regression with optimal scaling became widely available in statistical packages such as SAS (in a procedure called TRANSREG) and in SPSS Categories (van der Kooij & Meulman, 1999).

The transition from linear regression to regression with nonlinear optimal scaling transformations can be described as follows. In linear regression, an outcome (response) variable y is to be predicted from a set of J predictor variables in X, $X = \{x_j, j = 1, \ldots, J\}$, both y and $\{x_j\}$ having measurements on N objects. The optimization task is to find a linear combination Xw that correlates maximally with y, where the vector w contains J regression weights. It is convenient to write the optimization task in the form of a least squares loss function

$$L(w) = ||y - Xw||^2 = ||y - \sum_{j=1}^{J} w_j x_j||^2, \tag{1}$$

where $|| \cdot ||^2$ denotes the squared Euclidean norm. Loss function (1) has to be minimized over the regression weights $w = \{w_j, j = 1, \ldots J\}$ (we will assume that the predictor variables in X are standardized to have a mean of zero and variance equal to one, so we do not need to fit an intercept). The well-known analytic solution is given by

$$w = (X'X)^{-1}X'y, \tag{2}$$

where $(X'X)^{-1}$ denotes the inverse of the correlation matrix between the predictor variables.

If we include optimal scaling of the variables, we write a predictor as $\varphi_j(x_j)$, where $\varphi_j(x_j)$ denotes a one-to-one nonlinear transformation of x_j. The corresponding loss function becomes

$$L(w, x) = ||y - \sum_{j=1}^{J} w_j \varphi_j(x_j)||^2, \tag{3}$$

where $L(w, x)$ indicates that the arguments over which the function is to be minimized are the weights $w = \{w_j, j = 1, \dots J\}$ and x, which stands for the multivariate argument of functions of x_j, that is, the set of nonlinear transformations $x = \{\varphi_j(x_j), j = 1, \dots, J\}$. The nonlinear transformation process has been denoted by various names in the literature: in psychometrics it was called *optimal scaling* (a term originally coined by Bock, 1960), Nishisato (1980; 1994) called it *dual scaling*, Buja (1990) reintroduced the older term *optimal scoring*, and when the predictor variable is categorical, we usually use the term "quantification". (Quantification is one of the key terms in the "optimal scaling" system that was developed by Hayashi (1952).) To show how categorical, discrete variables fit into the optimal scaling framework, we introduce an $N \times C_j$ indicator matrix $G_j(x_j)$ that replaces a categorical variable x_j. The number of different categories in the variable x_j is indicated by C_j, and each column of $G_j(x_j)$ shows by 1-0 coding whether an object i scores in category c_j of variable $x_j, c_j = 1, \dots C_j$. For each variable x_j, we search for quantifications v_j that minimize the loss function, in the case of indicator matrices written as

$$L(w, v) = ||y - \sum_{j=1}^{J} w_j G_j(x_j) v_j||^2, \tag{4}$$

where v collects all quantifications $\{v_j\}$ for the J variables in a $\sum_j C_j$-vector. Because $G_j(x_j)v_j$ can be written as $G_j(x_j)v_j = \varphi_j(x_j)$, a function of the original predictor x_j, it is easy to see how numeric, continuous variables, and categorical, discrete variables, can be dealt with in the same framework.

Within the class of nonlinear transformations, we make the following distinctions. We call a transformation *nominal* if we merely maintain the class membership information—coded in $G_j(x_j)$ - in the quantified variable $G_j(x_j)v_j$, or equivalently in the transformation $\varphi_j(x_j)$: if $x_{ij} = x_{i'j} \implies \varphi_j(x_{ij}) = \varphi_j(x_{i'j})$. If the (categorical) predictor variable contains *order information* on the objects, this information can be preserved in the quantifications $v_j, v_1^j \leq \dots \leq v_C^j$, and thus in the *ordinal* transformation $\varphi_j(x_j)$. In the latter case, x_j and $\varphi_j(x_j)$ are related by a monotonic function. Numeric, continuous variables, finally, are integrated in the framework by a linear transformation that renders them standardized. Within monotonic and nonmonotonic transformations, we distinguish two classes, step functions and spline functions. Step functions are usually associated with categorical data, with a limited number of categories; splines are usually associated with continuous data. Because a continuous variable can be considered as a discrete variable with N (the number of objects) categories, we need to limit the number of parameters that are fitted. By using so-called regression splines, we limit the number of parameters by restricting the degree of the spline and the number of interior knots. Alternatively, we could discretize a continuous variable to have a fixed number of categories (a process called binning in computer science). The relation between regression splines and step functions is given by the fact that they are equivalent when the number of parameters fitted in the spline function is equal to the number of categories that is involved in the step function. So-called smoothing splines use a maximum of number of parameters, similar to step functions, but with smoothing splines some form of regularization is performed, usually on the spline coefficients.

2.1. *Computation of Regression Weights and Transformation Parameters*

In the regression problem (3), when the predictor variables are correlated, the optimal transformations $\varphi_j(x_j)$ are also interdependent. To solve for each $\varphi_j(x_j)$ separately, we use an elegant approach that separates each transformed variable and its weight from the remainder of the weighted predictors, isolating the current target part $w_j\varphi_j(x_j)$ from the remainder, denoted as

$\sum_{k \neq j} w_k \varphi_k(x_k)$. This approach to separate each variable from the rest has been called *block modeling*, the Gauss-Seidel algorithm, and *backfitting*; its application in the optimal scaling framework can be found in Kruskal (1965), de Leeuw, Young, and Takane (1976), Friedman and Stutzle (1981), Gifi (1990), and Hastie and Tibshirani (1990). The trick is as follows. We rewrite the loss function as

$$L(w, x) = ||y - \sum_{k \neq j} w_k \varphi_k(x_k) - w_j \varphi_j(x_j)||^2, \; j = 1, ..J, \tag{5}$$

where w again denotes $w = \{w_j\}$ and $x = \{\varphi_j(x_j), j = 1, \ldots, J\}$. Then we turn the original multivariate problem into a univariate one by creating an auxiliary variable u_j, defined as

$$u_j \equiv y - \sum_{k \neq j} w_k \varphi_k(x_k), \tag{6}$$

and we minimize

$$L(w_j, \varphi_j(x_j)) = ||u_j - w_j \varphi_j(x_j)||^2, \tag{7}$$

over w_j and $\varphi_j(x_j)$. After fitting the weight w_j with respect to fixed u_j and $\varphi_j(x_j)$, we minimize (7) over $\varphi_j(x_j)$ with respect to fixed u_j and new w_j. Then, we turn to the next regression weight and variable to be transformed. Using alternating least squares, we minimize over w_j and $\varphi_j(x_j)$ separately. To ensure that we can compute the regression weight w_j separately from the transformation, we standardize the transformed variable $\varphi_j(x_j)$ to have variance equal to one.

Keeping w_j fixed, we minimize (7) over all $\varphi_j(x_j) \in \mathbb{C}_j(x_j)$, where $\mathbb{C}_j(x_j)$ specifies the cone $\mathbb{C}_j$ that contains all admissible transformations of the variable x_j. In the case of a nominal transformation, the cone $\mathbb{C}_j(x_j)$ is defined by

$$\mathbb{C}_j(x_j) \equiv \{\varphi_j(x_j) | \varphi_j(x_j) = G_j(x_j) v_j\}, \tag{8}$$

and we define the metric projection $P_{\mathbb{C}_j(x_j)}$ as

$$P_{\mathbb{C}_j(x_j)} \equiv \min_{v_j} ||u_j - w_j G_j(x_j) v_j||^2. \tag{9}$$

This metric projection is equivalent to applying *equality restrictions* to $w_j^{-1} u_j$, so that objects in the same group according to variable j obtain the same quantification in the transformed variable $\varphi_j(x_j) = G_j(x_j) v_j$. (In fact, we only need the sign of w_j because the transformed variable $\varphi_j(x_j)$ will be standardized.)

For ordinal transformations, the cone $\mathbb{C}_j$ that contains all monotonic transformations of x_j is defined by

$$\mathbb{C}_j(x_j) \equiv \{\varphi_j(x_j) | \varphi_j(x_j) = \text{mon}_j(x_j)\}, \tag{10}$$

where $\text{mon}_j(x_j)$ denotes a monotonic transformation of x_j. The metric projection is written as

$$P_{\mathbb{C}_j(x_j)} \equiv \min_{\text{mon}_j(x_j)} ||u_j - w_j \text{mon}_j(x_j)||^2, \tag{11}$$

which amounts to applying *monotonic regression* of $w_j^{-1} u_j$ onto x_j. (The monotonic regression can either be increasing or decreasing, which ever gives the lesser loss value.)

In the case of spline transformations, one possibility is to construct an I-spline basis matrix $S_j(x_j)$ (see Ramsay, 1988, for details) and minimize

$$L(b_j) = ||u_j - w_j S_j(x_j) b_j||^2, \tag{12}$$

where $b_j = \{b_t^j, t = 1, \ldots, T\}$, the T-vector with spline coefficients that have to be estimated, and where T is dependent on the degree of the spline and the number of interior knots. If the

I-spline transformation does not have to follow the order of the values in x_j, we can compute the analytical solution for b_j directly, since (12) is a straightforward regression problem (with the columns of $S_j(x_j) = \{s_t^j, t = 1, \ldots, T\}$ as predictors). If, however, the I-spline transformation is required to be monotonic with the original order, we have to minimize (12) under the restriction that the vector b_j contains nonnegative elements. To ensure the nonnegativity of the entries in b_j, the problem is further partitioned by separating the t-th column of the spline basis matrix $S_j(x_j)$ (denoted by s_t^j) from the other columns $\{s_r^j, r \neq t\}$ and the t-th element (b_t^j) of the spline coefficient vector b_j from the remaining elements $\{b_r^j, r \neq t\}$. Next, we minimize iteratively

$$L(b_t^j) = ||(u_j - w_j \sum_{r \neq t} b_r^j s_r^j(x_j)) - w_j b_t^j s_t^j(x_j)||^2 \tag{13}$$

over $b_t^j \geq 0$, for $t = 1, \ldots, T$. There are some further complications if we take the normalization condition $b_j' S_j' S_j b_j = N$ into account, to ensure that the transformed variable is standardized. To go into the details of this problem at this point would lead us too far off the subject of the present paper, and we refer for the specifics to Groenen, van Os, and Meulman (2000).

Last, when the transformation level is chosen to be linear, the cone is defined by

$$\mathbb{C}(x_j) \equiv \{\varphi_j(x_j) | \varphi_j(x_j) = a_j x_j + \alpha_j\}, \tag{14}$$

which in the least squares framework comes to applying linear regression, which is equivalent to standardizing x_j.

After finding the optimal transformations by the projection on the cone $\mathbb{C}(x_j)$, the regression weight w_j is computed as

$$w_j = u_j' \varphi_j(x_j). \tag{15}$$

There is another important issue in multiple regression with optimal scaling that we cannot go into further at this point. This is the occurrence of possible local optima. How these may come about, and how they can be dealt with, is discussed in van der Kooij, Meulman and Heiser (2003).

3. Nonlinear Transformations Towards Independence

This section will contain some speculation, based on Meulman and van der Kooij (2000). They hypothesized that the particular transformations resulting from the regression problem would have a particular influence on the structure of the correlation matrix between the predictors after optimal scaling. The structure of the correlation matrix before transformation can be captured in a number of properties that indicate multicollinearity and dependence among predictor variables, a.o., the size of the smallest eigenvalues and the distribution of the eigenvalues. (If the predictors are independent, the correlation matrix is the identity matrix, with equal eigenvalues; increasing correlation among the predictors decreases the size of the smaller eigenvalues.) Regarding the independent contributions of a particular variable, a useful measure is given by the inverse of its associated diagonal element of the inverse of the correlation matrix. The inverse of the correlation matrix is sometimes called the decision matrix; its usefulness was already pointed out by Guttman (1950), and is also used in graphical modeling (Whittaker, 1990).

Meulman and van der Kooij (2000) performed a modest Monte Carlo study, in which they varied the values for the multiple correlation r^2, different forms of nonlinearity, and degrees of multicollinearity, as defined by the different sizes of the eigenvalues of the correlation matrix between the original predictors. The number of objects was 200, and the number of predictors was three. The data were sampled from a multivariate normal distribution, and variables were made discrete by grouping the samples optimally into seven categories (Max, 1960).

On the basis of their results, Meulman and van der Kooij (2000) speculated that when the multiple correlation is small due to a *nonlinear* relationship between predictors and the response, nominal transformation *linearizes* the relationship, and by doing so increases the dependence among the predictors (the transformation increases r^2, by definition). Ordinal transformation increases r^2, and decreases the dependence between the predictors. When multicollinearity exists, nonlinear transformations increases r^2 by increasing the independence among the predictors (i.e., the nonlinear transformation decreases the multicollinearity). The effect of the nonlinear transformation on the (in)dependence among the predictors depends on the size of r^2 and the smoothness of the transformation (the smoother the transformation, the smaller the effect). Unfortunately, these conclusions are still somewhat speculative, and the conjectured role of optimal scaling in the decrease of interdependence deserves additional study. Some empirical results will be given in the first application in section 6.

4. Some Connections

Until now, we described the situation with qualitative, or categorical predictor variables, but we have said nothing about the character of the outcome variable. The classical framework says to use regression when the outcome variable is continuous, and (canonical) discriminant analysis (e.g., McLachlan, 1992) when the outcome variable is categorical. A similar distinction is made in CART, classification and regression trees, originally proposed by Breiman, Friedman, Olshen and Stone (1984). Classification and regression trees (jointly called decision trees) have many attractive properties (that they share, in fact, with optimal scaling techniques). The data don't require ad-hoc transformations beforehand, missing data are elegantly treated in the overall analysis approach, mixtures of numeric and categorical data can be analyzed, and no distributional assumptions about the data or the error terms need to be made. In CART, a *numeric* outcome variable is predicted by a *regression tree*, and a *categorical* outcome variable by a *classification tree*. Both canonical discriminant analysis and classification trees do not take the order of the categories of the outcome variable into account.

In the optimal scaling framework, we deal somewhat differently with a categorical outcome variable. Obviously, we can apply a nonlinear variant of canonical discriminant analysis (see, e.g., Gifi, 1990; Hastie, Tibshirani, & Buja, 1998; Meulman, 2000), where nonlinear applies to the treatment of the predictor variables. The categories of the outcome variable are replaced by columns of an indicator matrix $G(y)$ (similar as described in sec. 2), and multiple y_j's (collected in the multidimensional Y_j) are fitted simultaneously if we wish discrimination in more than one dimension. If we have done so, we are in the nominal transformation framework for the outcome variable: the original categories will obtain an optimal ordering in the quantification (one for each dimension) that maximizes the between-to-total ratio when the variances for the different groups are considered. When the outcome variable is categorical, but the categories have an ordering that we would like to maintain, we could use the optimal scaling tool, and choose the *regression* approach, and transform the categories of the outcome variable to be monotonic with the original ones. But even if the categorical outcome variable is *nominal*, using optimal scaling we can also opt for a regression instead of a discriminant analysis. Thus, it is not appropriate to use the character of the outcome variable (continuous of categorical) to define the analysis. Therefore, we will use the term *prediction* for analyses that involve an outcome variable, whether the outcome is numeric, ordinal or nominal (in pattern recognition and statistical learning, this is called "supervised learning"; see Duda, Hart, & Stork, 2000). We will use the term *classification* for the task of assigning objects into groups, which is usually done *without* the help of an outcome variable. In statistical learning, this is called "unsupervised learning".

Without any attempt to make this section comprehensive, it is worthwhile to mention the relation between optimal scaling and so-called "support vector machines" (Vapnik, 1996; also, see Hastie, Tibshirani, & Friedman, 2001, pp. 371-389). Support vector machines perform prediction in an artificial high-dimensional space that is made up by an extremely large number of transformations of the predictors (e.g., the variables taken to their 2nd, 3rd, 4th, and 5th power, and all the cross products of the resulting new variables, and higher order interaction terms). The term "support vectors" refers to a limited number of point in this high-dimensional space that define boundaries on the basis of which discrimination between groups can be made. The high-dimensional space is called artificial because through a very clever "change of variables" of the inner products in the space, all computations can be performed in terms of other inner products in a low-dimensional space. Support vector machines are an important statistical learning tool and they are becoming popular in data mining. They are mentioned here in this section because a set of transformations of a predictor variable in support vector machines is similar to the cone of admissible transformations for a predictor in optimal scaling.

Finally, the relation between optimal scaling and neural networks (e.g., Ripley, 1996) should be mentioned. Since space is limited, we refer to the series of papers by Takane and Oshima-Takane, for example, Takane (1998) and Takane and Oshima-Takane (2002). A particular form of neural networks is equivalent to projection pursuit regression (Friedman & Stuetzle, 1981). We will return to neural networks and projection pursuit regression very briefly in section 9.

5. Something New

Suppose we have a categorical outcome variable; we wish to maintain the order of the categories in the quantifications, but we also would like to have multiple quantifications, solutions, dimensions, as in canonical discriminant analysis. This section proposes the following approach. We will call a linear combination of optimally transformed predictors a *prediction component* (in analogy with a principal component), and propose to compute multiple, additive prediction components sequentially over different nonlinear transformations of the predictors.

This can be formalized as follows. We write a forward stagewise additive model as

$$L(x) = ||y - f_m(x)||^2, \tag{16}$$

where

$$f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J} w_{jm}\varphi_{jm}(x_j), m = 1, \ldots, M \tag{17}$$

(cf. Hastie, Tibshirani & Friedman, 2001, pp. 304–305). A weight w_{jm} and a transformation $\varphi_{jm}(x_j)$ are confounded, but as was mentioned before, w_{jm} can be identified when $\varphi_{jm}(x_j)$ is standardized. As an example of an additive prediction components model, consider the case with three prediction components (so $M = 3$), where the three components

$$\sum_{j=1}^{J} w_{j1}\varphi_{j1}(x_j), \quad \sum_{j=1}^{J} w_{j2}\varphi_{j2}(x_j), \quad \text{and} \quad \sum_{j=1}^{J} w_{j3}\varphi_{j3}(x_j)$$

are computed sequentially to predict the outcome variable y. In M consecutive steps, we minimize for $m = 1, \ldots, M$,

$$L(w, x)_m = ||y - \sum_{j=1}^{J} w_{jm}\varphi_{jm}(x_j)||^2. \tag{18}$$

For example, if we first fit a *linear prediction component*, defined by a weighted sum of *linear* transformations of the predictor variables, and denoted as $\sum_j w_j^l \varphi_j^l(x_j)$, the loss can be written as

$$L(w, x)_{\text{lin}} = ||y - \sum_{j=1}^{J} w_j^l \varphi_j^l(x_j)||^2, \tag{19}$$

and the residual r_{lin} from the linear component is defined as

$$r_{\text{lin}} \equiv y - \sum_{j=1}^{J} w_j^l \varphi_j^l(x_j). \tag{20}$$

When we next minimize

$$L(w, x)_{\text{lin+ord}} = ||r_{\text{lin}} - \sum_{j=1}^{J} w_j^o \varphi_j^o(x_j)||^2, \tag{21}$$

with respect to the residual vector r_{lin} as defined in (20), and where $\varphi_j^o(x_j)$ denotes an ordinal transformation, then

$$r_{\text{ord}} \equiv r_{\text{lin}} - \sum_{j=1}^{J} w_j^o \varphi_j^o(x_j), \tag{22}$$

is the residual. When we consider the residual vector r_{ord}, (22), and finally fit a nominal prediction component, written as the minimization of

$$L(w, x)_{\text{lin+ord+nom}} = ||r_{\text{ord}} - \sum_{j=1}^{J} w_j^n \varphi_j^n(x_j)||^2, \tag{23}$$

we obtain the minimum loss value $L_{\text{lin+ord+nom}}$. This value $L(w, x)_{\text{lin+ord+nom}}$ is equal to $L(w, x)_{\text{nom}}$, where $L(w, x)_{\text{nom}}$ is defined by

$$L(w, x)_{\text{nom}} = ||y - \sum_{j=1}^{J} w_j^n \varphi_j^n(x_j)||^2. \tag{24}$$

The conjecture here is that fitting a sequential series of prediction components with transformations (cf. (19), (21), (23)) is equivalent to fitting a single prediction component (cf. (24)) iff the transformations are less restricted in each sequential step, and if the optimal solution has been obtained for both the weights and the transformations in the last (the M-th) step. Because the optimizing transformations $\varphi_j^n(x_j)$ are defined with respect to different targets: r_{ord} in (23) and y in (24), the transformation results will typically not be the same. In the example constructed above, we used linear, ordinal and nominal transformations, sequentially, but we could also have used a sequence of spline transformations with an increasing degree, an increasing number of knots, or combinations thereof, as long as the number of parameters fitted in the m-th prediction component is larger than the number in the $(m - 1)$-th one.

Although we might not have gained anything in terms of the loss value by computing the standard weighted linear combination of predictors (the "optimal" prediction component) as the sum of multiple prediction components, there are other considerations, similar as in prediction methods used for data mining. First, we have to make a distinction between "prediction" using the observed predictors to approximate the observed outcome, and "prediction" using the observed predictors and the observed outcome to obtain parameter values that are subsequently applied to new observations to predict future, unknown outcomes. To the best of my knowledge, prediction of future observations has never been an issue in the psychometric optimal scaling literature: the objective was always to minimize (3) to find a global minimum. In what is called *statistical learning*, prediction is only used for the second situation, and we will use it here in the same vein. Concluding: it might be true that the loss values $L(w, x)_{\text{lin+ord+nom}}$ in (23) and $L(w, x)_{\text{nom}}$ in (24) are equal with respect to the *observed* outcomes, this is not necessarily the case for *future* outcomes. At this point, however, we will also take another consideration into account.

Especially when the data contain a lot of noise, it turns out that optimal fitting on the data can lead to poor prediction for future observations (this is sometimes called the trade-off between bias and variance; Hastie, Tibshirani, & Friedman, 2001, pp. 39, 194). Instead of searching for an optimized fit (the global, or possibly a local, optimum), it may be beneficial to obtain a suboptimal solution that may have better prediction results for future observations. Prediction results for future observations can be obtained by simulation, for example, by subsampling methods such as cross-validation. (For completeness: we set a particular percentage of the data, say 10 percent of the objects, apart and fit the model (16), (17) on the remaining 90 percent of the objects. In the second step, we apply the results—the regression weights and the optimal transformations—to the left-out 10 percent of the observations, and inspect how well the sum of the prediction components approximates the 10 percent of the left-out outcome values. This process is repeated a large number of times, and the error rates, both for the "training set"—the 90% sample—and the test set—the 10% sample—are averaged to establish the overall prediction success.)

Since I learned from the statistical learning literature that to iterate to eternity is often not beneficial for prediction, I had to give up my old adage:

"To Iterate is Heaven
To Converge Divine"

The following adaptation of (16), (17) was inspired by so-called *boosting* of regression and classification trees (Friedman, 2001). Boosting (Freund & Shapire, 1996; Friedman, Hastie, & Tibshirani, 2000) is a revolutionary idea. With respect to decision trees, boosting deals with the fact that trees are rather inaccurate in terms of predicting future observations, which is due to the "greediness" that is inherent to the sequential nature of the partitioning algorithm. The idea of boosting is as follows. A *small* tree is fitted to the data, so that there is still quite a lot of residual variance to be accounted for. Next, a second small tree is fitted to the residual. This process continues, where each small tree gets only a *small* weight. It was shown in Friedman (2001) and in Friedman and Meulman (2003a) that the weighted sum of many small trees (where the weights are small) gives a dramatic improvement of prediction accuracy compared to a single large, optimized tree fitted to the outcome.

A similar approach can be applied in additive prediction components modeling. In the first step, we fit a particular model with nonlinear transformations, but we do not iterate until convergence. We stop *prematurely* (this is sometimes called regularization, or shrinking), and we continue to fit on the residual from the first step with a new set of nonlinear transformations, and stop early *again*. And we continue. It may look as if this approach has a great risk of overfitting since the number of parameters we are using is increasing with every new prediction component. But overfitting is controlled by checking against the error rate that is computed for the left-out sample from the data. So we use forward stagewise additive regression as in (16), (17), but as in boosting, the M consecutive steps result in *suboptimal* solutions to the current, m-th, problem, giving suboptimal prediction components

$$\sum_j w_{j1}\varphi_{j1}(x_j), \ldots, \sum_j w_{jm}\varphi_{jm}(x_j), \ldots, \sum_j w_{jM}\varphi_{jM}(x_j).$$

When to stop adding components can again be determined by cross-validation. Summarizing: additive prediction components modeling constructs the approximation of y, $\hat{y}$, by adding two or more components, sequentially, thus

$$\hat{y} = \sum_{m=1}^{M} \hat{y}_m = \sum_{m=1}^{M} \sum_{j=1}^{J} w_{jm}\varphi_{jm}(x_j). \tag{25}$$

Experience up until now has shown that if we use prediction components additively, prediction is improved. If sometimes not really dramatically, this is compensated by the fact that the number of iterations that is needed is greatly reduced (see the next section for examples).

As we mentioned above, the sequential nature has influence on the transformations of the predictors (as in (23) and (24)). We have not said anything yet about possible transformation of the outcome variable. If we allow the outcome to be transformed as well, we would minimize

$$L(w, x, \psi(y)) = ||\psi(y) - f_{m-1}(x) - \sum_{j=1}^{J} w_{jm}\varphi_{jm}(x_j)||^2, m = 1, \ldots, M. \tag{26}$$

Here we would transform the outcome variable only *once*, that is, only in the first step when we form the first prediction component. In the second step, we form the second prediction component with respect to the residual from the first step,

$$r_1 = \psi(y) - f_{m-1}(x), \tag{27}$$

and so on (cf. (20), (22)). Transforming the outcome directly with respect to nominal transformations of the predictors may give very different results compared to transforming it with respect to ordinal transformations of the predictors (where the nominal transformations of the predictors are fitted in a second step). We cannot say anything conclusive at this point since more simulation studies are needed, many more than could be performed while this Presidential Address was written.

6. Applications

The first data set used in this application section was collected at the Leiden Cytology and Pathology Laboratory, and concerns characteristics of cells obtained from patients with various grades of cervical preneoplasia and neoplasia (cancer). To obtain the samples, which were taken from the ectocervix as well as the endocervix, special sampling and preparation techniques were used (Boon, Zeppa, Ouwerkerk-Noordam, & Kok, 1990). The "correct" histological diagnosis was known by a subsequently taken biopsy. Previous analyses (Meulman, Zeppa, Boon, & Rietveld, 1992) used a subset of the data, containing 50 cases with mild dysplasia (Histological Group 1), 50 cases with moderate dysplasia (Histological Group 2), 50 cases with severe dysplasia (Histological Group 3), 50 cases with carcinoma in situ (Histological Group 4), and 42 cases with invasive squamous cell carcinoma (Histological Group 5). For each of the 242 cases, 11 attributes were determined. Four of these were quantitative (counts) with many distinct values, being the number of abnormal cells per fragment, the total number of abnormal cells, the number of mitoses, and the number of nucleoli. In addition, seven qualitative variables were available, being abnormality ratings from 1 (normal) to 4 (very abnormal) of various aspects of each cell. These variables are nuclear shape, nuclear irregularity, chromatin pattern, chromatin distribution, nucleolar irregularity, nucleus/nucleolus ratio and nucleus/cytoplasm ratio. Most of the qualitative variables had only three categories, because the value 1 was extremely rare. Summarizing the predictors: we have $J = 11$ manually observed features of cells taken from pap smears of $N = 242$ patients with cervical cancer or precursor lesions (dysplasias).

A (limited) number of combinations (models) with additive prediction components were tried. If we refer to fitting successive prediction components as major, *outer* iterations, $m = 1, \ldots, M$, then steps that are taken within prediction component fitting could be called *inner* iterations. The number of inner iterations was chosen to be either 1 or 50 (where 50 implies convergence), and either one or two prediction components were fitted. The different transformations applied were either linear, ordinal, or nominal step functions for the qualitative variables, and either linear, monotonic or nonmonotonic regression spline functions for the quantitative variables. The splines were of order two (quadratic) and had one interior knot (so the number of free parameters was three, similar to those in most of the step functions). In the cross validation, 10%

TABLE 1.

Training Set		Test Set		Prediction
RMS(e)	Stdev	RMS(e)	Stdev	Components
.034077	.000522	.035982	.004924	lin
.025992	.000484	.030070	.004090	nom(50)
.026498	.000493	.029850	.004089	ord(5)
.026483	.000495	.029838	.004090	ord(50)
.026040	.000482	.029821	.003956	nom(1), nom(4)
.026046	.000483	.029812	.003942	ord(1), nom(5)
.026051	.000481	.029790	.003936	nom(5)
.026063	.000482	.029787	.003930	ord(1), nom(4)

Notes: Prediction for the Cervix data by cross validation, root mean squared errors—RMS(e)—and standard deviations (Stdev) in 1000 cross validation samples. Rows are sorted by RMS(e) for the test set. In the last column, "lin" indicates linear transformation, "ord" either ordinal or monotonic spline transformation, "nom" either nominal or monotonic spline transformation. All splines were of order two and had one interior knot. The numbers within parentheses indicate the number of inner iterations.

(24) of the patients was left out, and the number of cross validation samples was set to 1000. To minimize uninteresting sampling effects, the same 1000 sets of cases were left out in the cross validation for each of the different models. The results are given in Table 1.

It is clear from Table 1 that linear transformation ("lin") does not predict very well, and neither do the single ordinal and nominal transformations, except for "nom(5)". The best result is obtained for model "ord(1), nom(4)", which indicates a first prediction component with ordinal and a second component with nominal transformations, fitted with 1 and 4 inner iterations, respectively. If the number of inner iterations for nominal is increased—model "ord(1), nom(5)"—the root mean squared error increases. One might argue that the results for the models "nom(5)" and "ord(1), nom(4)" are very much alike. We do chose to look at the results for model "ord(1), nom(4)", however, because there is another issue at stake here, which is the tradeoff between "interpretation" and "prediction".

If prediction is the one and only objective, we do not necessarily need to interpret the different contributions of the variables to the prediction. This is typical of neural networks: they can be trained to predict extremely well, but the prediction success in the end cannot be ascribed to any of the original predictors, thus interpretation is almost impossible. However, even in noisy data situations as is usually the case in data mining, the analysis should give a result that can be interpreted in terms of its contributors to the goodness-of-prediction. That is why neural nets are not always the method of choice in data mining.

So in the present case, the ordinal transformations in the first prediction component are primarily used for interpretation (ordinal transformations are generally easier to interpret than nominal transformations), while the nominal transformations in the second prediction component are used to further improve upon the prediction.

Figure 1 shows the two-dimensional results, with the first prediction component on the horizontal axis, and the second prediction component on the vertical axis; the components were normalized to have the same variance; the transformed variables were projected into the space and displayed by arrows (cf. the biplot representation in principal components analysis). The directions of the red arrow heads (representing the variables from the first, ordinal, prediction component) highly correlate with the horizontal dimension that is very close to the outcome variable (diag). In the vertical dimension, we notice that four variables stand out: these are transformed variables from the second (nominal) prediction component (with blue arrow heads), 9 = total number of abnormal cells, 11 = number of nucleoli, 6 = nucleolar irregularity and 8 =

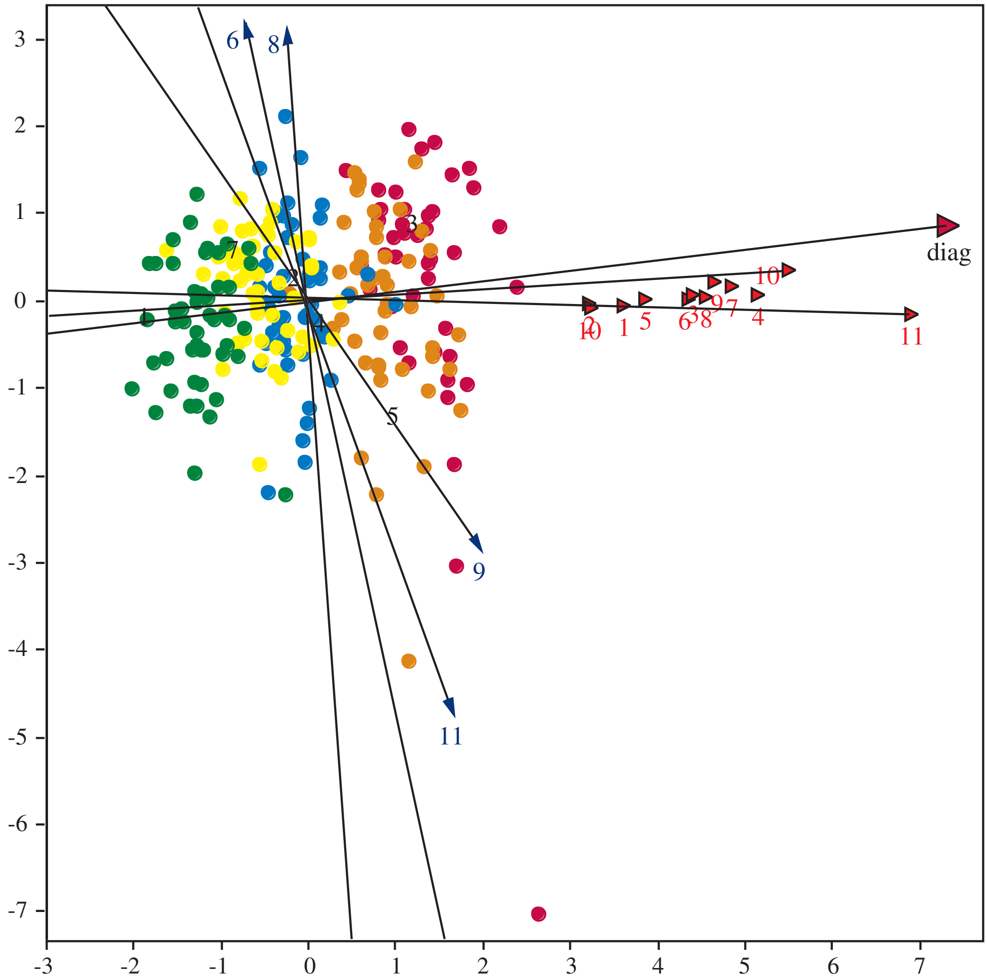

FIGURE 1.
Second prediction component (vertical axis) versus first prediction component (horizontal axis). The 242 patients are displayed with five different colors, indicating their diagnosis. Green = mild dysplasia, yellow = moderate dysplasia, blue = severe dysplasia, orange = carcinoma in situ, and red = invasive carcinoma. The arrow heads indicate the directions of the transformed variables that are projected into the two-dimensional space (not all arrows were drawn). The labels are: 1 = nuclear shape, 2 = nuclear irregularity, 3 = chromatin pattern, 4 = chromatin distribution, 5 = number of abnormal cells per fragment, 6 = nucleolar irregularity, 7 = nucleus/nucleolus ratio, 8 = nucleus/cytoplasm ratio, 9 = total number of abnormal cells, 10 = number of mitoses, and 11 = number of nucleoli.

nucleus/cytoplasm ratio. For the interpretation of the direction, we need to look at the transformations in Figure 2.

Here each separate graph contains two transformations: one ordinal (in red), and one nominal (nonmonotonic, in blue). Considering the blue curves, we notice that most of the qualitative variables (with pink background) are transformed so that the middle values are distinguished from the low and higher ones, which obviously are closer than in the ordinal transformation. In the nominal transformation of variable 6 = nucleolar irregularity (second row, middle graph) the value 1 is mostly contrasted to the value 2, which is close to 4; for variable 8 = nucleus/cytoplasm ratio (third row, first graph), the values 4 and 2 are distinguished from the value 3. The fitted nonmonotonic transformations for 9 = total number of abnormal cells, and 11 = number of nucleoli, came out as monotonic, clearly distinguishing some special cases in the very high regions of the original variables (500–600, and 15–25, respectively).

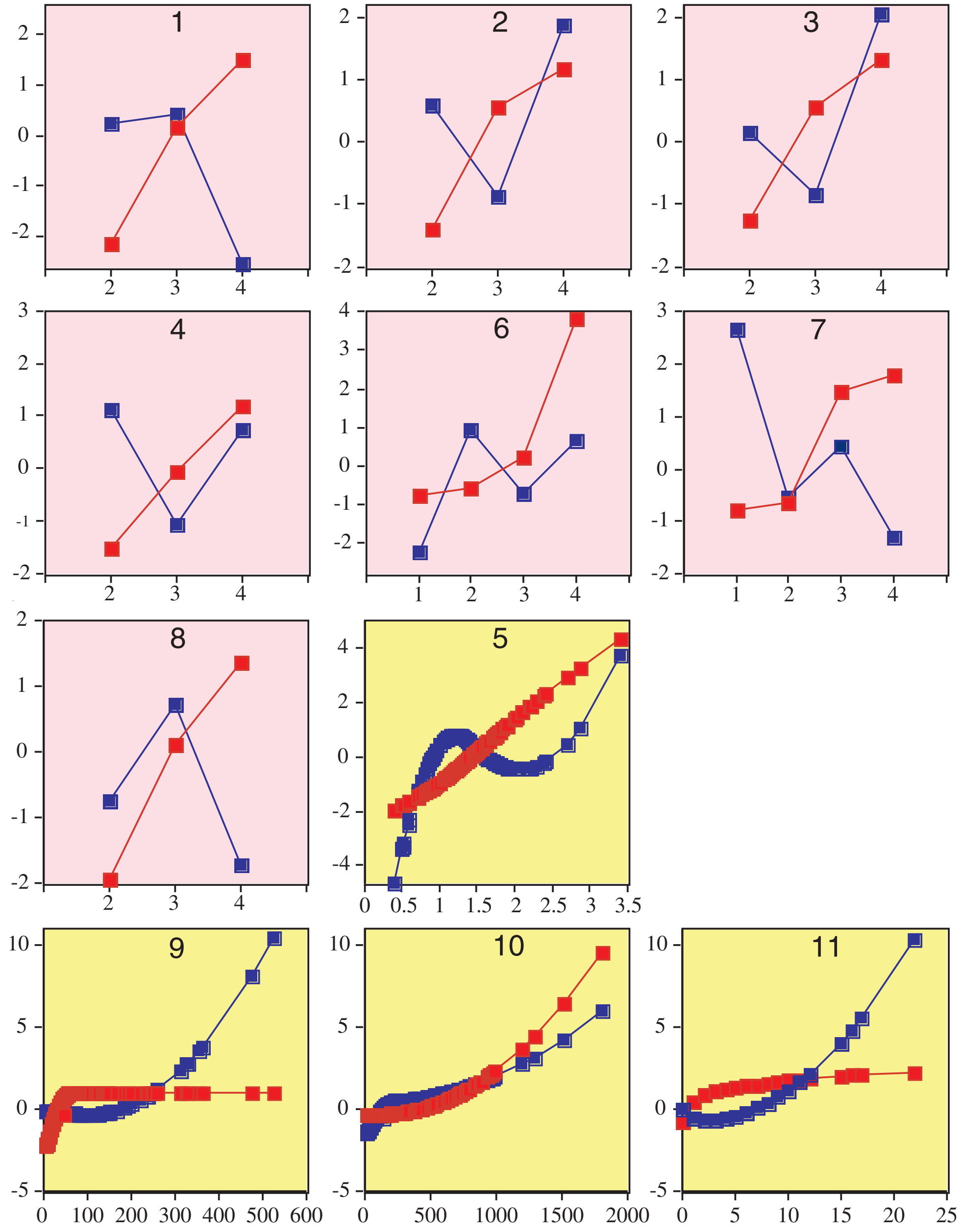

FIGURE 2.
Transformations first prediction component (ordinal, in red) and second prediction component (nominal, in blue). First row: 1 = nuclear shape, 2 = nuclear irregularity, 3 = chromatin pattern. Second row: 4 = chromatin distribution, 6 = nucleolar irregularity, 7 = nucleus/nucleolus ratio. Third row: 8 = nucleus/cytoplasm ratio, 5 = number of abnormal cells per fragment. Fourth row: 9 = total number of abnormal cells, 10 = number of mitoses, 11 = number of nucleoli. Qualitative variables with pink, and quantitative variables with yellow background.

Figure 3 displays three separate graphs. Instead of individual patients as in Figure 1, here the means for each diagnostic group are displayed. The graph at the left shows the averages per diagnostic group for the first (in red) and the second component (in blue) on the vertical axis versus the values of the outcome variable on the horizontal axis. The graph in the middle gives the averages for the *sum* of the first and second prediction component versus the outcome, and

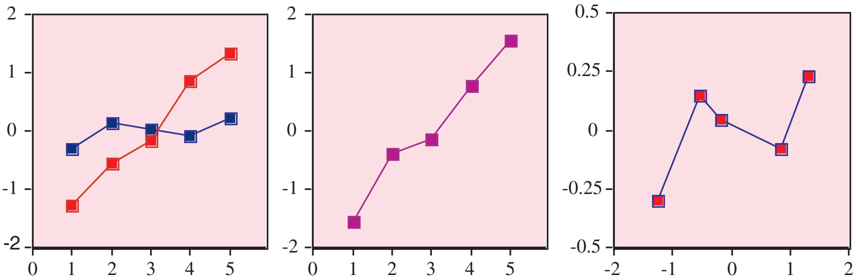

FIGURE 3.

Left panel: averages per diagnostic group versus diagnosis category (outcome). In red: averages for the first prediction component; in blue: averages for the second prediction component. Middle panel: sum of the averages over two prediction components versus values of the outcome variable. Right panel: averages second predictor component versus averages first predictor component.

the graph at the right shows the averages of the second prediction component versus the first. The latter graph can be directly compared to Figure 1 since it contains the group points (the two-dimensional averages) of the individual patients in the prediction components space. The average values of the first prediction component are monotonically increasing with the values of the outcome variable; although the effect is much smaller, the second prediction component emphasizes the difference between the second, third and fifth diagnosis group, and the first and fourth. The effect can be seen in the middle graph, where the mild dysplasia group (1) has been separated from the moderate and severe dysplasia groups (2 and 3), and where the severe dysplasia (3) and the carcinoma groups (4 and 5) are on a straight line. The graph at the right shows again that the middle groups 2 and 3 are closer to Group 5 in some respects than Group 4 is (compare the transformation for variables 6 and 8 in the second prediction component).

To finish this application, we look at the diagnostics that were mentioned in section 3, which are the eigenvalues of the correlation matrices, both before and after transformation, and the associated independence values for the predictor variables (see Table 2). The eigenvalues are more evenly distributed in “ord(1)” than in “lin” (as is summarized by the variances), but the

TABLE 2.

	Eigenvalues correlation matrix			Independence separate variables		
nr	lin	ord(1)	nom(4)	lin	ord(1)	nom(4)
1	4.244	3.832	1.848	.776	.785	.882
2	1.405	1.423	1.615	.799	.826	.933
3	1.136	1.039	1.327	.683	.670	.943
4	.985	.843	1.176	.673	.675	.956
5	.691	.766	.961	.654	.686	.847
6	.629	.651	.934	.521	.725	.836
7	.578	.621	.715	.556	.643	.831
8	.482	.572	.662	.799	.770	.900
9	.369	.497	.647	.397	.589	.759
10	.283	.475	.579	.287	.712	.797
11	.198	.281	.536	.335	.365	.743
Mean	1.000	1.000	1.000	.589	.677	.857
Variance	1.176	.889	.178	.031	.014	.005

Notes: Eigenvalues of the correlation matrix for model “lin”, and for model “ord(1), nom(4)”. Next, the independence values (the inverse of the diagonal elements of the inverse of the correlation matrix) for the 11 predictor variables.

TABLE 3.
Attributes for the Boston Housing Data.

Label	Description
CRIME	per capita crime rate by town
LLOT	land zoned for lots over 25,000 sq.ft.
INDU	nonretail business acres per town
CHAS	Charles River adjacency
NOX	nitric oxides concentration
ROOM	number of rooms per dwelling
AGE	proportion of units built prior to 1940
DIST	distances to 5 Boston employment centres
HIGH	accessibility to radial highways
TAX	property-tax rate
PTRA	pupil-teacher ratio
FBL	$(\text{Bk-}0.63)^2$, Bk = fraction of blacks
LSTA	% lower status
MEDV	value of owner-occupied homes

Notes: MEDV is the outcome variable.

variance is dramatically reduced for "nom(4)", from which we conclude that the predictors were (much) less correlated in the "ord(1)" and "nom(4)" transformation than in the linear data. From the independence values we can derive information on the separate variables. It is clear that the overall independence increases, and that this is especially so for the variables 9, 10, and 11 in the "nom(4)" transformation.

The second data set analyzed is usually called the Boston Housing data. The data were collected by Harrison and Rubingeld (1978), and have been used extensively to demonstrate the results of new regression methods. The data comprise 506 observations for each census district of the Boston metropolitan area, the variables are given in Table 3. The last attribute, labeled MEDV, indicates the value of the owner-occupied houses, and is the outcome variable to be approximated by predictor components, the predictors being described in the first 13 rows of Table 3. The 506 census districts form 92 neighborhoods, originally labeled from 0-91 (we will follow this labeling in the plots). The 92 neighborhoods and 506 districts were grouped (rather arbitrarily in groups of about 50 observations) and labeled as is shown in Table 5.

The results from the cross validation study are given in Table 4 that shows again that the one-component models do not perform very well, except, surprisingly, the model with monotone second order splines with one interior knot, fitted with one inner iteration—denoted as "mon[2, 1](1)". The best result is obtained for the model that uses four prediction components, all consisting of monotonic splines, and each fitted with one inner iteration only—model "mon[2, 1](1)(1)(1)(1)'. The fit for the training set, and the standard deviations are also best for this model.

The results are depicted in Figure 4; here the individual 506 districts are displayed in a two-dimensional graph. The points in Figure 4 are colored according to the coding in Table 5.

The first dimension in Figure 4 is the first prediction component; in the second dimension, the *sum* of the second, third, and fourth prediction components is displayed. The transformed predictor variables are again shown as vectors in the two-dimensional space. The outcome MEDV vector is close to the first dimension, and is highly positively associated with ROOM, and negatively with LSTA and a cluster containing, o.a., NOX, PTRA, TAX, and CRIM (arrow heads in red). In the second dimension, we displayed "room4" (arrow head in blue), an important transformed variable from the *fourth* prediction component.

The dotted lines in Figure 4 were drawn perpendicular to the directions of ROOM and room4. In this way, the objects in between the dotted lines at the top indicate houses with rela-

TABLE 4.

Training Set		Test Set		Prediction
RMS(e)	Stdev	RMS(e)	Stdev	Components
.01124	.000336	.01943	.004069	mon[2, 1](50)
.01723	.000496	.01940	.004596	mon[1, 1](25), [2, 2](25)
.01718	.000494	.01938	.004611	nom[2, 2](50)
.01723	.000491	.01935	.004613	lin, nom[2, 2](25)
.01722	.000495	.01925	.004631	nom(3)(3)(3)[2, 2]
.01304	.000338	.01920	.004128	mon[2, 1](1)
.00921	.000218	.01890	.004367	mon[2, 1](1)(1)(1)(1)

Notes: Prediction for the Boston housing data determined by cross validation, root mean squared errors—RMS(e)—and standard deviations (Stdev) in 150 cross validation samples. Rows are sorted by RMS(e) for the test set. In the last column, "lin" indicates linear transformation, "mon" a monotonic spline transformation, "nom" a nonmonotic spline transformation. The two numbers within each bracket indicate the order of the spline and the number of interior knots, respectively. The numbers within parentheses indicate the number of inner iterations used to fit the model.

TABLE 5.
Coding of the Boston Housing Data

Group	Neighborhoods/Suburbs	Census Districts	Color	
1	0–7	1–50	light pink	■
2	8–23	51–100	pink	■
3	24–28	101–172	light blue	■
4	29–34	173–199	orange	■
5	35–42	200–254	red	■
6	43–55	255–298	salmon	■
7	56–69	299–349	yellow	■
8	70–80	350–406	magenta	■
9	81–83	407–456	blue	■
10	84–91	457–506	purple	■

Notes: 0 = Nahant, 1 = Swampscott, 2 = Marblehead, 3 = Salem, 4 = Lynn, 5 = Sargus, 6 = Lynnfield, 7 = Peabody, 8 = Danvers, 9 = Middleton, 10 = Topsfield, 11 = Hamilton, 12 = Wenham, 13 = Beverly, 14 = Manchester, 15 = North Reading, 16 = Wilmington, 17 = Burlington, 18 = Woburn, 19 = Reading, 20 = Wakefield, 21 = Melrose, 22 = Stoneham, 23 = Winchester, 24 = Medford, 25 = Malden, 26 = Everett, 27 = Somerville, 28 = Cambridge, 29 = Arlington, 30 = Belmont, 31 = Lexington, 32 = Bedford, 33 = Lincoln, 34 = Concord, 35 = Sudbury, 36 = Wayland, 37 = Weston, 38 = Waltham, 39 = Watertown, 40 = Newton, 41 = Natick, 42 = Framingham, 43 = Ashland, 44 = Sherborn, 45 = Brookline, 46 = Dedham, 47 = Needham, 48 = Wellesley, 49 = Dover, 50 = Medfield, 51 = Millis, 52 = Norfolk, 53 = Walpole, 54 = Westwood, 55 = Norwood, 56 = Sharon, 57 = Canton, 58 = Milton, 59 = Quincy, 60 = Braintree, 61 = Randolph, 62 = Holbrook, 63 = Weymouth, 64 = Cohasset, 65 = Hull, 66 = Hingham, 67 = Rockland, 68 = Hanover, 69 = Norwell, 70 = Scituate, 71 = Marshfield, 72 = Duxbury, 73 = Pembroke, 74 = Boston Allston-Brighton, 75 = Boston Back Bay, 76 = Boston Beacon Hill, 77 = Boston North End, 78 = Boston Charlestown, 79 = Boston East Boston, 80 = Boston South Boston, 81 = Boston Downtown, 82 = Boston Roxbury, 83 = Boston Savin Hill, 84 = Boston Dorchester, 85 = Boston Mattapan, 86 = Boston Forest Hills, 87 = Boston West Roxbury, 88 = Boston Hyde Park, 89 = Chelsea, 90 = Revere, 91 = Winthrop.

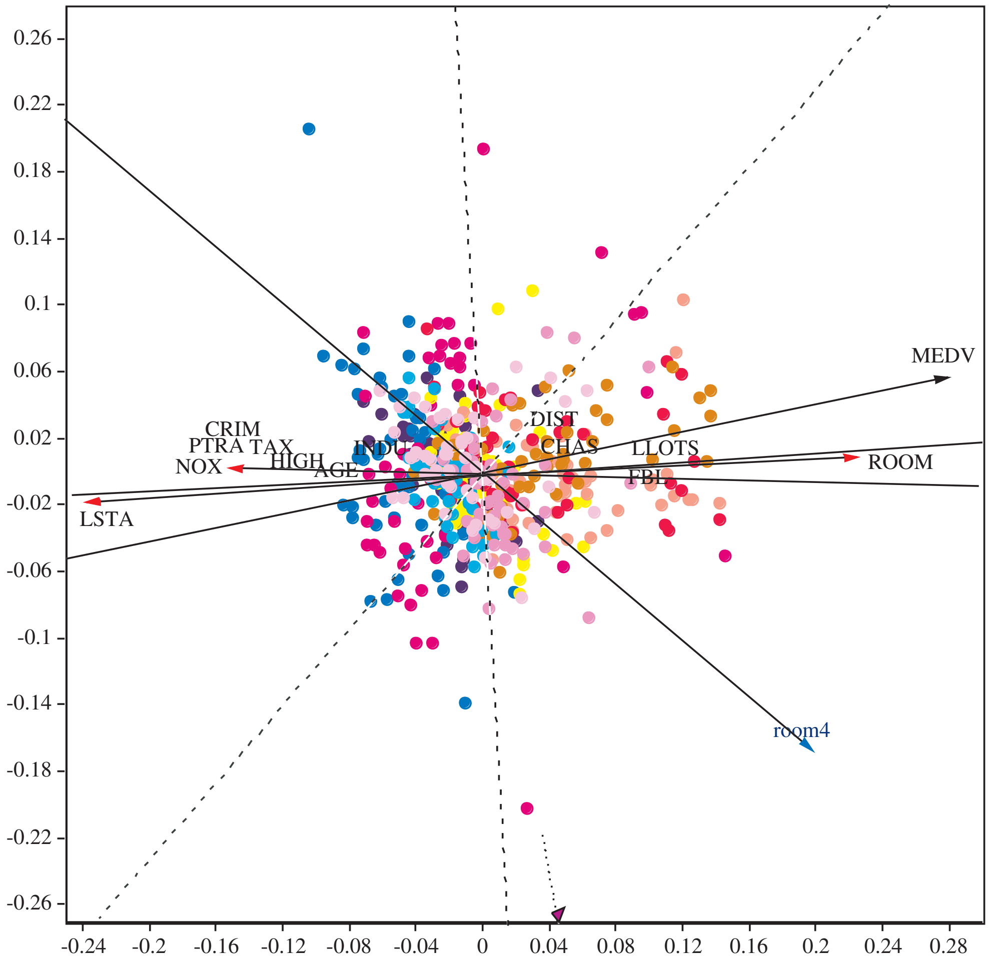

FIGURE 4.
Two-dimensional space based on four prediction components for the Boston Housing data. Sum of the second, third and fourth components on the vertical axis versus the first component on the horizontal axis. Arrows represent transformed variables, projected into the space. Dotted arrow with magenta arrowhead denotes the position of an outlier (not plotted). Description of the colors is given in Table 5.

tively many rooms according to the first (monotonic) transformation (in ROOM) and relatively few rooms according to the fourth (monotonic) transformation (in room4). For the houses in between the dotted lines at the bottom, the reverse applies: relatively few rooms according to the first and relatively many rooms according to the fourth transformation. In the first dimension, many rooms is associated with high values for MEDV, and low values for LSTA, NOX, PTRA, TAX and CRIM; the second dimension shows that neighborhoods with low status, high pupil-teacher ratio, and high crime rate may have many rooms as well.

Figure 5 displays the points for the 92 neighborhoods (the averages of the 506 individual points displayed in Figure 4), labeled with the original labels (from 0–91), again with colors according to Table 5. Points with the same color refer to a number of neighborhoods that were arbitrarily grouped according to their order in the original 506 rows table. Although there are some remarkable exceptions (e.g., point 76—Boston Beacon Hill—is far away from the other magenta points in the first dimension, being at the expensive side), there is quite some structure displayed. In general, the cheaper houses are represented by points in magenta, blue and light blue, and the expensive houses by red and orange points.

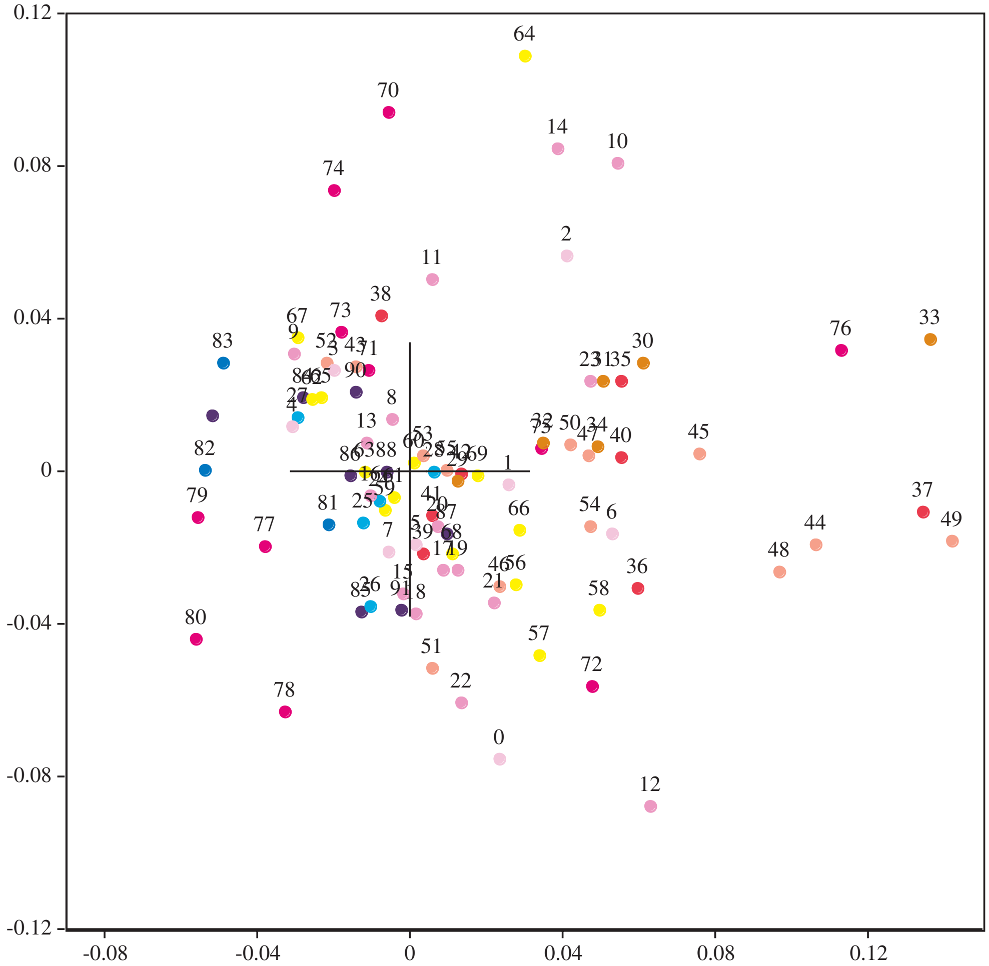

FIGURE 5.
Two-dimensional space based on four prediction components for the Boston Housing data. Sum of the second, third and fourth components on the vertical axis versus the first component on the horizontal axis. Here each point represents the average over districts that belong to the same neighborhood. Description of the colors is given in Table 5.

7. Something Borrowed

Until now we did NOSTRA, an acronym for Nonlinear Optimal Scaling TRAnsformations, which we applied to prediction. So it's now time for classification under the name COSA, which stands for *clustering objects on subsets of attributes*. This part is the borrowed part of my Presidential Address, particularly from joint work with Jerome Friedman. This research took part from April 2001 until the present day (and it kept me going during the more difficult parts of dealing with my disease and its treatments). COSA's fascinating technical details, along with many examples from different areas, will be published separately in a discussion paper in *JRSS-Series B* (Friedman & Meulman, in press). Here we will only allude to its concepts, in a nontechnical way, and apply COSA to real data from the life sciences.

The objective of COSA is to cluster the objects in attribute-value data, and its main motivation was given by considering the analysis of a particular kind of data, such as in genomics, proteomics, and metabolomics, areas recently subsumed under the name *systems biology*. The target data for the particular clustering approach are large data sets with a *very small* number of objects compared to a *very large* number of attributes (variables). In genomics, for example, we deal with gene expression data in micro arrays, where we have very many genes (say, 1,500–40,000), and very few objects (say, 20–250). Usually we know some characteristics from the objects: their gender (female or male), their condition (healthy or ill, untreated or experimental). It is clear that a standard statistical analysis, depending on a large sample-to-attribute ratio, cannot do much here, especially not in regard to using the gene expressions (from low to high) to predict the conditions of the objects. Therefore, common approaches are to cluster the attributes (genes) first, and only after having reduced the original many-attribute data set to a much, much smaller one, one may try to classify, or cluster, the objects. The problem here, of course, is that we would like to select those attributes that discriminate most among the objects (so we have to do this while regarding all attributes multivariately), and it is usually not good enough to inspect each attribute univariately, because important clustering attributes usually don't have their influence as single agents, but in a (small) group.

Therefore, we have a double task: to select clusters of variables (subsets of attributes) and group the objects into homogeneous clusters. In the present case, we don't use the background information of the cases in the analysis task (unsupervised learning), but we will use it afterwards to validate the solution found by the procedure. An additional problem is that with a large numbers of variables, objects are very unlikely to cluster on all, or even a large number of them. Indeed, objects might cluster (be close) on some, and be far apart on all others, and our task is to find *that* (unique) set of attributes that a particular group of objects is clustering on. The objective of COSA is to find *multiple* groups of objects that cluster *not on all* variables but on *subsets* of attributes, and these subsets of variables may be different for each separate group of objects. Consequently, the subsets of attributes that we are looking for may be nonoverlapping, or partially, or completely overlapping. This objective creates computational problems of various kinds.

First of all, subset selection of variables is a combinatorial problem, which cannot be solved optimally for a large number of variables. A nice idea here is to replace subset selection (which is discrete) by differential weighting (which is continuous). A subset of attributes that most exhibit clustering obtains a weight set with high values. A crucial point is that these weight sets are allowed to be different for each different cluster of objects. An additional problem is that it is easy to find a group of objects that clusters on a *single* variable, and accordingly giving that particular variable a very large weight, and all other variables a zero weight. This result is trivial, however. As was mentioned above, important clustering variables usually act as a (small) group, so this solution needs to be avoided. We can do this by applying a negative penalty, or incentive that encourages an equal weight distribution among variables.

In the present paper, we will have to make a long story very short. It can be shown that differential optimal attribute weighting for each cluster of objects, boils down to the use of the inverse exponential mean (rather than the arithmetic mean) of the separate attribute distances, as the multivariate distance between objects. Some characteristics of the inverse exponential distance are: that it is very closely related to the harmonic mean, that it emphasizes small distances heavily, and that large distances are basically ignored. To attain its objectives, the COSA algorithm was designed to optimize a surrogate criterion with the same solution as the one we wish to obtain (an approach that can be compared to majorization; a.o., see de Leeuw & Heiser, 1980; Heiser, 1995), and during its iterations, COSA uses K-nearest neighbors (e.g., see Duda, Hart, & Stork, 2000) to define clusters among objects. COSA outputs a distance matrix that can be used as input for any distance based clustering method, as well as for multidimensional scaling. For the COSA software, see Friedman & Meulman (2003b).

7.1. A COSA Example with the "Leiden ApoE3 data"

Also borrowed are the exceptional data used for the application of COSA. Many thanks are due here to Jan van der Greef (Leiden University, TNO Zeist, and Beyond Genomics, Boston). Jan and his group of collaborators also kept me going during the past two years. During our collaboration, I analyzed a particular data set, which is called the “Leiden ApoE3 data”, and I’m grateful to Jan that I’m allowed to show some of the analysis results here.

The biochemical background is very briefly summarized as follows (with the help of Elwin Verheij). ApoE3 stands for Apolipoprotein E3 (one of many apolipoproteins) that, together with lipids, form lipoproteins (cholesterol particles), for example, LDL, VLDL, and HDL. The E3 “Leiden” is a human variant of ApoE3. When the lipoprotein is no longer recognized by special receptors in the liver, it prevents uptake of LDL cholesterol by the liver, and this results in strongly increased lipoprotein levels in the plasma. Eventually the latter condition results in atherosclerosis, which is hardening of the arteries, and if this blocks a blood vessel, it may lead to a stroke or a heart attack.

The objects in the experiment that gave the data for the COSA analysis are mice: we have a number of normal mice (called wildtype), and a set of experimental, transgenic mice, which contain the Human Leiden ApoE3 variety. On a high fat diet the mice develop severe atherosclerosis. In this study the mice were on a low fat diet. Other important features of the study are that the samples were collected at the age of 9 weeks, while atherosclerosis is manifest after

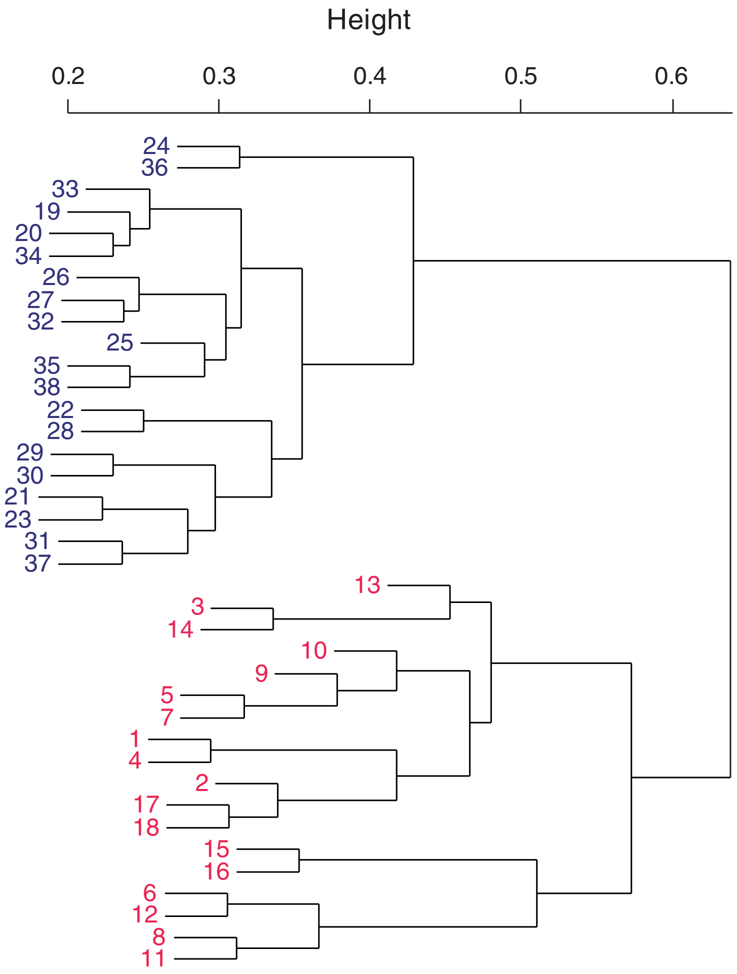

FIGURE 6.
Dendrogram for the ApoE3 data; hierarchical cluster analysis of the COSA distance matrix obtained for the ApoE3 Leiden data. Transgenic mice are in red and labeled from 1–18, and wildtype mice are in blue and labeled from 19–38. Complete linkage was used in the clustering.

20 weeks, and that big changes in metabolite profiles were not expected. The data analyzed are from LC-MS (mass spectrometry) analysis of plasma lipids, and consist of 1550 measurements (attributes) on 38 cases (consisting of two observations for each mouse). The original experiment was performed with 10 wildtype and 10 transgenic mice, but only 9 transgenic mice survived the experiment.

The COSA analysis resulted in a proximity matrix (a matrix with "distances") between the mice, and this matrix was subjected to a hierarchical cluster analysis in *R*, resulting in the dendrogram shown in Figure 6. (We choose to use the complete link option, but this choice was not essential for the clustering.)

There is a perfect separation between the transgenic mice (with labels 1–18) and the wildtype mice (with labels 19–38). The COSA procedure also provides so-called "importance values" that indicate the relevance of the variables to the clustering that has been found. For the ApoE3 data, only a small number (40–60) of the original 1550 attributes turned out to be important. To test whether the clustering found is not random, the following procedure can be performed. Ten random groups of the size of group of the wildtype cluster (which is 20) are sampled from the data, and for each of these random samples the importance values are computed. Then the actual importance values found are compared to those from the test, and in this way we can determine which variables are more important than can be attributed to chance. We also performed this test for the transgenic cluster; the results for both groups are depicted in Figure 7.

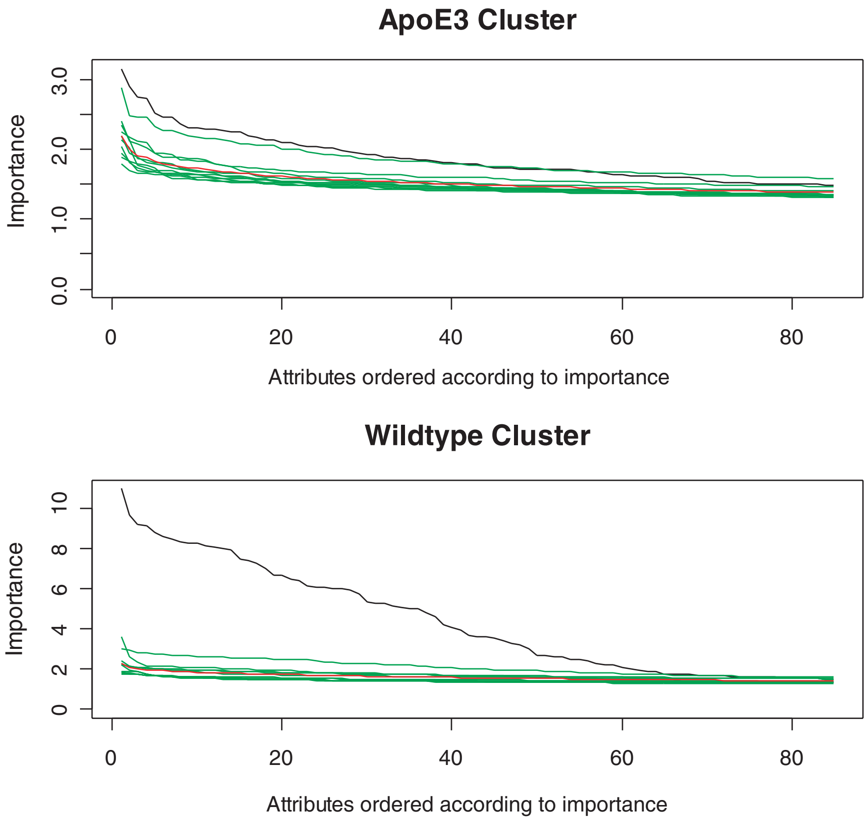

FIGURE 7.
The two black curves in the upper and lower graph indicate the 85 largest (out of 1550) importance values for the group of transgenic mice 1–18 in the ApoE3 Leiden data (at the top) and for the wildtype mice 19–38 (at the bottom). In each graph, the ten green curves indicate the 85 largest importance values for ten random groups of size 18 and 20, respectively. The two reds curves are the averages of each set of ten green curves.

Here the values for the 85 most important variables (out of 1550) are displayed. The black curve gives the observed importance values, the ten green curves are for the randomly generated samples, and the red curve is the average of the ten green curves. The difference between the importance values for the wildtype cluster and those for the 10 random groups is big; about 60 attributes appear to be important for the clustering of the wildtype group. The importance values for the transgenic cluster are somewhat less distinct. It is clear, however, that only 40–60 variables are truly important for the clustering of the transgenic mice.

8. Something Blue

Many alternatives have crossed my mind in choosing the subject under the label "blue". I said something about blueprints in the beginning, and there are a surprisingly large either good or bad connotations with "blue". Of course, the BLUE estimator comes immediately to mind in a classical statistical context, in which this paper definitely has not been written. Since this part is at the end of the paper, and I have some leeway since this is my Presidential Address, I choose to opt for other "friends" during the past two years, being the enormous piles of detective, murder, mystery, and police novels that I read while recovering. Also, two of my favorite TV series are Law & Order and NYPD Blue (not to mention Hill Street Blues), and it's this kind of "Blue" I'm referring to in this final section. I'm paraphrasing parts from a novel by Robert K. Tanenbaum, with the optimistic title "Reversible Error" (New York: Penguin Group, Putnam Inc., NY, 1992).

The setting is the District Attorney office in New York, where Marlene Ciampi (an assistant district attorney) asks the chief of the D.A. Office, without preamble, "Will they do it?" He answers: "What, your rape case correlations?... it's got no priority; it comes after the bookkeeping, the trial schedules, the rosters, everything". Ciampi says "We're talking about finding criminal patterns, catching multiple rapists. And they're worried about bookkeeping? How much time could it take? I thought computers worked like in no time at all". The Chief answers: "It's not the machines, it's the people...these guys in Data Processing, they can keep the payroll going, but correlations, social-science packages, ANOVA, it's out of their league". And he adds "University might be your best bet ... it's an intrinsically interesting project; it might make a good dissertation—an analytical method of discovering serial rapists. Get a criminology professor up at John Jay or NYU involved".

Not long after this conversation Marlene Ciampi meets with a new rape victim, JoAnne Caputo. Marlene Ciampi says to her: "Oh sorry. It just occurred to me that your rapist may have done that trick before. I had a woman in here a couple of weeks ago, but I'm ashamed to say I've forgotten her name. I have them filed by name, and there are over two thousand". JoAnne Caputo leaned forward: "You don't have them cross-indexed?" Marlene answers "No, see, this is strictly amateur hour. It's a shoebox with cards". "Let me see the cards", said JoAnne Caputo. "I see the problem. A lot of key information is in text fields: what he said, what he did. You'd have to input the whole file as text and then do a string search subroutine to pull matches out". And then she says: "SPSS could handle it".

"You know about this stuff?" Marlene asked hopefully. "It's what I do. I told you I worked at NYU. I'm in social stat". says JoAnne. "I'm afraid to ask", said Marlene. "Would it be possible... ?" JoAnne: "Would it help to find that bastard?" Marlene: "Girl, it's about the only way there is". JoAnne: "OK, give me the boxes. I'll start right away."

It would have been too much if JoAnne had said that she was going to use multiple correspondence analysis (HOMALS) in SPSS Categories. But that technique could be very well used to solve the classification problem involved. And so could COSA. In any case, seeing your techniques alluded to in popular fiction, is more than any psychometrician could ask for.

9. A Short Epilogue

There are many other developments in modern regression that are very worthwhile to be integrated into optimal scaling, and I will only mention a few of them. These are "Bagging", which stands for bootstrapping the objects and averaging the results (Breiman, 1996a), combined with sampling from the predictors in so-called "Random Forests" (Breiman, 1996b). Both ideas originate in regression trees fitting, and it is worthwhile to note that averaging optimal scaling transformations over bootstrap solutions was already done in Gifi (1990). Instead of sampling randomly from the predictors, we could do sampling based on the correlation structure among the predictors. Predictors usually form clusters/groups, and we could sample one variable from each cluster, do the analysis, and repeat the process.

Another good idea would be to integrate ridge regression and the Lasso (Tibshirani, 1996) in multiple regression with optimal scaling. Both ridge regression and the Lasso are shrinkage/regularization methods, that are aimed at penalizing large values of the regression weights. In ridge regression, the restriction is on the sum of squares of the regression weights $\sum w_j^2$; in the Lasso, the restriction is on the sum of the absolute values $\sum |w_j|$. In Hastie, Tibshirani and Friedman (2001, pp. 328–330), it is shown that the Lasso can be approximated by forward stagewise regression when the regression weights are incremented with a very small amount for each important predictor, one at the time. This fits perfectly in the optimal scaling framework. Finally, optimal scaling is very suited to be used in the analysis of very large data sets that are common in data mining. Scalable algorithms for the analysis of very large data sets come natural when the data are discrete, since we can work with the categories and not with the individual objects.

It would be nice to compare the additive prediction components approach to projection pursuit regression (Friedman & Stuetzle, 1981) and (particular forms of) neural networks. Prediction in the two latter mentioned techniques is based on *nonlinear* transformations of *linear* combinations of variables, while additive prediction components are *linear* combinations of *nonlinear* transformations of variables. In this sense, prediction components are closer to the data, less of a black box method, and therefore hopefully easier to interpret.

It will be clear that more research is needed to confirm the usefulness of additive prediction component modeling. "Flying by the seats of my pants", that's how the major part of this Presidential Address came about. However, since in some cases there are no blueprints, it seems good to follow Ray Bradbury's advice that "sometimes you have to jump off a cliff and develop your wings on the way down".

References

Bock, R.D. (1960). *Methods and applications of optimal scaling* (Tech. Rep. 25). Chapel Hill, NC: University of North Carolina, L.L. Thurstone Psychometric Laboratory.

Boon, M.E., Zeppa, P., Ouwerkerk-Noordam, E., & Kok, L.P. (1990). Exploiting the tooth-pick effect of the cytobrush by plastic embedding of cervical samples. *Acta Cytologica*, *35*, 57–63.

Breiman, L. (1996a). Bagging predictors. *Machine Learning*, *26*, 123–140.

Breiman, L. (1996b). Stacked regressions. *Machine Learning*, *24*, 51–64.

Breiman, L., & 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.H., Olshen, R.A., & Stone, C.J. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth.

Buja, A. (1990). Remarks on functional canonical variates, alternating least squares methods and ACE. *Annals of Statistics*, *18*, 1032–1069.

de Leeuw, J., & Heiser, W.J. (1980). Multidimensional scaling with restrictions on the configuration. In P.R. Krishnaiah (Ed.), *Multivariate analysis, Vol. V* (pp. 501–522). Amsterdam: North-Holland.

de Leeuw, J., Young, F.W., & Takane, Y. (1976). Additive structure in qualitative data. *Psychometrika*, *41*, 471–503.

Duda, R., Hart, P. and Stork, D. (2000). *Pattern classification* (2nd ed.). New York, NY: John Wiley & Sons.

Freund, Y., & Schapire, R.E. (1996). Experiments with a new boosting algorithm. *Machine Learning: Proceedings of the Thirteenth International Conference* (pp. 148–156). San Francisco, CA: Morgan Kauffman.

Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232.

Friedman, J.H., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting (with discussion). *Annals of Statistics*, *28*, 337–307.

Friedman, J.H., & Meulman, J.J. (in press). Clustering objects on subsets of attributes, (with discussion). *Journal of the Royal Statistical Society*, *Series B*. Available at http://www-stat.stanford.edu/˜jhf/ftp/cosa.pdf

Friedman, J.H., & Meulman, J.J. (2003a). Multiple additive regression trees with application in epidemiology. *Statistics in Medicine*. *22*(9), 1365–1381.

Friedman, J.H., & Meulman, J.J. (2003b). *COSA* [Software]. Available at http://www-stat.stanford.edu/˜jhf/COSA.html

Friedman, J., & Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association*, *76*, 817–823.

Gifi, A. (1990). *Nonlinear multivariate analysis*. Chichester, U.K.: John Wiley & Sons. (First edition, 1981, University of Leiden, Department of Data Theory)

Groenen, P.J.F., van Os, B.J., & Meulman, J.J. (2000). Optimal scaling by alternating length constrained nonnegative least squares: An application to distance based principal components analysis. *Psychometrika*, *65*, 511–524.

Guttman, L. (1950). The principal components of scale analysis. In S.A. Stouffer, L. Guttman, E.A. Suchman, P.F. Lazarsfield, S.A. Star, & J.A. Clausen (Eds.), *Measurement and prediction*. Princeton, NJ: Princeton University Press.

Harrison, D., & Rubinfeld, D.L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics Management*, *5*, 81–102.

Hastie, T., & Tibshirani, R. (1990). *Generalized additive models*. New York, NY: Chapman and Hall.

Hastie, T., Tibshirani, R., & Buja, A. (1998). Flexible discriminant analysis by optimal scoring. *Journal of the American Statistical Association*, *89*, 1255–1270.

Hastie, T., Tibshirani, R., & Friedman, J.H. (2001). *The elements of statistical learning*. New York, NY: Springer-Verlag.

Hayashi, C. (1952). On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view. *Annals of the Institute of Statitical Mathematics*, *2*, 93–96.

Heiser, W.J. (1995). Convergent computation by iterative majorization: Theory and applications in multidimensional data analysis. In W.J. Krzanowski (Ed.), *Recent advances in descriptive multivariate analysis* (pp. 157–189). Oxford, U.K.: Oxford University Press.

Kruskal, J.B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*, 1–28.

Kruskal, J.B. (1964b). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, *29*, 115–129.

Kruskal, J.B. (1965). Analysis of factorial experiments by estimating monotone transformations of the data. *Journal of the Royal Statistical Society*, *Series B*, *27*, 251–263.

Max, J. (1960). Quantizing for minimum distortion. *Proceedings IEEE (Information Theory)*, *6*, 7–12.

McLachlan, G.J. (1992). *Discriminant analysis and statistical pattern recognition*. New York, NY: John Wiley & Sons.

Meulman, J.J. (2000). Discriminant analysis with optimal scaling. In R. Decker & W. Gaul (Eds.), *Classification and information processing at the turn of the millenium* (pp. 32–39). Heidelberg-Berlin, Germany: Springer-Verlag.

Meulman, J.J., Zeppa, P., Boon, M.E., & Rietveld, W.J. (1992). Prediction of various grades of cervical preneoplasia and neoplasia on plastic embedded cytobrush samples: Discriminant analysis with qualitative and quantitative predictors. *Analytical and Quantitative Cytology and Histology*, *14*, 60–72.

Meulman J.J., & van der Kooij, A.J. (2000, May). *Transformations towards independence through optimal scaling*. Paper presented at the International Conference on Measurement and Multivariate Analysis (ICMMA), Banff, Canada.

Nishisato, S. (1980). *Analysis of categorical data: Dual scaling and its applications*. Toronto, Canada: University of Toronto Press.

Nishisato, S. (1994). *Elements of dual scaling: An introduction to practical data analysis*. Hillsdale, NJ: Lawrence Erlbaum.

Ramsay, J.O. (1988). Monotone regression splines in action. *Statistical Science*, *4*, 425–461.

Ripley, B.D. (1996). *Pattern recognition and neural networks*. Cambridge, U.K.: Cambridge University Press.

Takane, Y. (1998). Nonlinear multivariate analysis by neural network models. In C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H.H. Bock, & Y. Baba (Eds.), *Data science, classification, and related methods* (pp. 527–538). Tokyo: Springer.

Takane, Y., & Oshima-Takane, Y. (2002). Nonlinear generalized canonical correlation analysis by neural network models. In S. Nishisato, Y. Baba, H. Bozdogan, & K. Kanefuji (Eds.), *Measurement and multivariate analysis* (pp. 183–190). Tokyo: Springer-Verlag.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, *Series B*, *58*, 267–288.

van der Greef J., Davidov E., Verheij E., Vogels J., van der Heijden R., Adourian A.S., Oresic M., Marple E.W., & Naylor S. (2003). The role of metabolomics in drug discovery: A new vision for drug discovery and development. In G.G. Harrigan & R. Goodacre (Eds.), *Metabolic profiling: Its role in biomarker discovery and gene function analysis* (pp. 170–198). Boston, MA: Dordrecht; London: Kluwer Academic Publishers.

van der Kooij, A.J., & Meulman, J.J. (1999). Regression with optimal scaling. In J.J. Meulman, W.J. Heiser, & SPSS Inc. (Eds.), *SPPS Categories 10.0.* (pp. 1–8, 77–101). Chicago, IL: SPSS.

van der Kooij, A.J., Meulman, J.J., & Heiser W.J. (2003). Local minima in categorical multiple regression. Manuscript mubmitted for publication.

Vapnik, V. (1996). *The nature of statistical learning theory*. New York, NY: Springer-Verlag.

Whittaker, J.L. (1990). *Graphical models in applied multivariate statistics*. New York, NY: John Wiley & Sons.

Winsberg, S., & Ramsay, J.O. (1980). Monotonic transformations to additivity using splines. *Biometrika*, *67*, 669–674.

Yanai, H., Okada, A., Shigemasu, K,. Kano, T., & Meulman, J.J. (Eds.) (2003). *New developments in psychometrics*. Tokyo: Springer-Verlag

Young, F.W., de Leeuw, J., & Takane, Y. (1976). Regression with qualitative and quantitative variables: An alternating least squares method with optimal scaling features. *Psychometrika*, *41*, 505–528.