Statistical Science
2019, Vol. 34, No. 3, 361–390
https://doi.org/10.1214/19-STS697

ROS Regression: Integrating Regularization with Optimal Scaling Regression

Jacqueline J. Meulman, Anita J. van der Kooij and Kevin L. W. Duisters

Abstract. We present a methodology for multiple regression analysis that deals with categorical variables (possibly mixed with continuous ones), in combination with regularization, variable selection and high-dimensional data ($P \gg N$). Regularization and optimal scaling (OS) are two important extensions of ordinary least squares regression (OLS) that will be combined in this paper. There are two data analytic situations for which optimal scaling was developed. One is the analysis of categorical data, and the other the need for transformations because of nonlinear relationships between predictors and outcome. Optimal scaling of categorical data finds quantifications for the categories, both for the predictors and for the outcome variables, that are optimal for the regression model in the sense that they maximize the multiple correlation. When nonlinear relationships exist, nonlinear transformation of predictors and outcome maximize the multiple correlation in the same way. We will consider a variety of transformation types; typically we use step functions for categorical variables, and smooth (spline) functions for continuous variables. Both types of functions can be restricted to be monotonic, preserving the ordinal information in the data. In combination with optimal scaling, three popular regularization methods will be considered: Ridge regression, the Lasso and the Elastic Net. The resulting method will be called ROS Regression (Regularized Optimal Scaling Regression). The OS algorithm provides straightforward and efficient estimation of the regularized regression coefficients, automatically gives the Group Lasso and Blockwise Sparse Regression, and extends them by the possibility to maintain ordinal properties in the data. Extended examples are provided.

Key words and phrases: Lasso and Elastic Net regularization for nominal and ordinal data, monotonic group Lasso, regularization for categorical high-dimensional data, optimal scaling, linearization of nonlinear relationships, monotonic step functions and splines.

1. INTRODUCTION

Multiple regression investigates the relationship between an outcome (response) variable and a set of predictor variables, and can be used to estimate a model for predicting future outcomes. Ordinary least squares (OLS) regression is known for not performing well with respect to both model complexity and prediction accuracy, and breaks down under multicollinearity, for

Jacqueline J. Meulman is Professor, Mathematical Institute, Leiden University, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands (e-mail: jmeulman@math.leidenuniv.nl) and Adjunct Professor, Department of Statistics, Stanford University, Stanford, Californoa 94305, USA (e-mail: jmeulman@stanford.edu). Anita J. van der Kooij is Senior Researcher, Institute of Psychology, Division of Methodology and Statistics, Wassenaarseweg 52, 2323 CA Leiden, The Netherlands (e-mail: kooij@leidenuniv.nl). Kevin L.W. Duisters is Ph.D. candidate at the Mathematical Institute, Leiden University, Leiden, The Netherlands (e-mail: k.l.w.duisters@math.leidenuniv.nl).

example, when the number of predictors is larger than the number of observations. Regularization improves prediction accuracy, and using L_1-norm regularization (through the Lasso or the Elastic Net), also decreases the model complexity. When variables are categorical and/or when relations among variables are not linear, standard OLS methods have to be further adjusted. Usually, a categorical predictor is handled by replacing it with a set of dummy variables. Nonlinear relations for continuous variables are usually dealt with by replacing predictors by basis functions such as polynomials. (For a state-of-the-art overview, see Hastie, Tibshirani and Friedman, 2009.)

There are two separate yet equally important reasons to replace OLS regression with optimal scaling regression. In optimal scaling, each categorical predictor variable is replaced by a set of quantifications. Instead of creating dummy variables, optimal quantifications are assigned directly to the categories of the predictor. In case of continuous predictor variables, optimal scaling deals with nonlinear relationships by transforming the predictors, typically by smooth spline functions. The basis of OS regression is the "one-variable-at-a-time" approach, also known as "coordinate descent", originally used to find transformations of the data (De Leeuw, Young and Takane, 1976; Friedman and Stuetzle, 1981; Gifi, 1990; Breiman and Friedman, 1985; Buja, Hastie and Tibshirani, 1989; Hastie and Tibshirani, 1990). A special property of optimal scaling is that these quantifications and transformations can be either monotonic or nonmonotonic with the originally given coding of the categories. In addition, the same set of quantifications and transformations can be applied to the (possibly categorical) outcome.

Existing methods that apply regularization to categorical predictors by creating augmented data do not give regularized coefficients for each predictor, but regularized coefficients for each dummy variable. To remedy this, the Group Lasso (Yuan and Lin, 2006) and Blockwise Sparse Regression (Kim, Kim and Kim, 2006) were proposed, regularizing a group or block instead of the individual dummy respectively substitute variables, by applying a norm restriction to the coefficients in the group or block. However, the Group Lasso can only deal with nominal predictors, and does not apply to a categorical outcome variable. In contrast, Regularized Optimal Scaling (ROS) Regression, as proposed in this paper, does not use sets of dummies, but applies the above-mentioned quantification or optimal transformation of predictor variables to give regularized coefficients in a straightforward way. ROS regression can maintain the ordinal properties of ordinal predictors, and can deal with an ordered categorical outcome. In addition, we can easily generalize from categorical to continuous variables, also allowing for mixtures of categorical and continuous variables (again, without using dummy variables). It turns out that the "one-variable-at-a-time" approach makes the computation of regularized coefficients for the Lasso and subsequently for the Elastic Net trivially simple, even when the number of predictors is much larger than the number of observations.

1.1 Related Methods

Since optimal scaling regression is a particular nonlinear generalization of OLS, we mention several related methods.

- Alternating Conditional Expectation (ACE, Breiman and Friedman, 1985) allows for nonlinear transformations of both outcome and predictors, thereby being closely related to optimal scaling. ACE can handle nominal variables, but it does not allow restrictions for ordinal categorical variables, as in optimal scaling.
- Generalized Additive Models (GAM; Hastie and Tibshirani, 1990) extend linear regression by allowing nonlinear transformation of the predictors using scatterplot smoothers (GAM). When the predictors are all continuous, optimal scaling methods are equivalent to generalized additive models (GAMs).
- Multivariate Adaptive Regression Splines (MARS, Friedman, 1991) extends linear regression by replacing each predictor by a set of basis splines.
- Copula based regression (Sklar, 1959; Kolev and Paiva, 2009; Trivedi and Zimmer, 2005) has become a popular way of describing (nonlinear) dependence between outcome and predictors in the financial and actuarial field. It can be considered as applying monotone transformations to both outcome and predictors coupled with distributional assumptions, and is known to work best for continuous data (Parsa and Klugman, 2011, Genest and Nešlehová, 2007). Since copula based regression is discussed less frequently in the statistical literature than GAM or GLM, an elaborate comparison with (regularized) optimal scaling is included in Appendix A.
- Generalized linear models (GLM/GLIM, Nelder and Wedderburn, 1972), such as logistic regression, involve a different type of nonlinearity. They handle nonnormal error terms in the linear regression model through a link function, giving rise

to a nonlinear relation between (expected) outcome and the linear combination of predictors. Thus, an important distinction with respect to optimal scaling is the distributional assumption underlying these methods and the fact that the nonlinearity is captured in the link function.

Thus, some are quite similar in spirit to optimal scaling, because they involve optimal transformations of the predictors while the relation between transformed predictors and outcome remains linear. An advantage of these methods is that they are relatively insensitive to misspecification given their lack of distributional assumptions and abundant (semiparametric) flexibility. In other words, instead of specifying functions beforehand, these methods (including OS) allow the analysis to reveal the appropriate functional form. Other nonlinear generalizations, however, are of a very different type because they transform the relation between the predictors and outcome, which relation (the link) becomes nonlinear.

1.2 Outline

In summary, we present a methodology for multiple regression analysis that deals with categorical variables (possibly mixed with continuous ones), in combination with regularization. ROS regression handles highly correlated predictors, provides variable selection and can be used with high-dimensional data ($P \gg N$). Transformations of predictors and outcome can be both nonmonotonic as well as monotonic.

The remainder of this paper is organized as follows. Section 1 is concluded by a short example to illustrate some of the introduced benefits of ROS regression. We will use the well-known Marketing Data from Hastie, Tibshirani and Friedman (2009), abbreviated HTF, and compare our analysis with an approach analogue to the analysis presented there. Section 2 gives a brief history and description of the basic OS regression approach including computational details. Section 3 presents three illustrations: (a) a small example with simulated data showing nonlinear relationships between predictor and outcome, (b) the full analysis of the Marketing Data with mixed nominal and ordinal predictors, and an ordinal outcome, and (c) an analysis of cervix cancer data, with a mixture of ordinal and continuous predictors and with an ordinal outcome. A variety of different models is fitted and the results are compared through the use of diagnostics and cross-validation. Section 4 describes the ROS Regression methodology proposed in this paper, including a short literature review that led to its development. This section contains details on how regularization with Ridge, Lasso and Elastic Net penalties is incorporated in optimal scaling, and discusses selection of the Ridge and Lasso penalties. In this section, it is also shown that ROS regression in specific situations is equivalent to the above-mentioned Group Lasso (Yuan and Lin, 2006) and Blockwise Sparse Regression (Kim, Kim and Kim, 2006). Section 5 presents three different applications. The first revisits the simulated data with nonlinear relationships from Section 3, the second shows an extended analysis of data concerning the 50 states, and the third shows an application in a high-dimensional data setting ($P \gg N$), with metabolomic data from LC-MS (Liquid Chromatography Mass Spectrometry) measurements of plasma lipids. The paper concludes with a discussion and suggestions for further research.

1.3 An Introductory, Abbreviated Example

To illustrate the optimal scaling approach to categorical data, we use the Marketing Data, as described in Hastie, Tibshirani and Friedman (2009), pages 492–494. The data consist of an ordinal (ordered) outcome variable (Annual Income) ($\mathbf{y}$), having 9 levels and 13 predictor variables ($\mathbf{x}_k$) for 8993 customers in a San Francisco shopping mall. The predictor set consists of a mixture of ordinal (ordered) and nominal (unordered) categorical variables. For a standard regression analysis (a), we follow a procedure similar as was described in HTF: observations with missing data were removed, leaving 6876 objects. Then each ordinal predictor was cut at its median and coded by one dummy variable; each categorical predictor with C_k categories was coded by $C_k - 1$ dummy variables. This resulted in a 6876×35 matrix of 6876 observations on 35 dummy variables (predictors). The ordinal outcome variable (annual income) was treated as a numerical variable with nine different values. The result of this analysis can be found in the first row of Table 1. We give the apparent prediction error (APE), the mean squared error obtained for the total sample, and the estimated expected prediction error ($\widehat{\text{EPE}}$), the mean squared error obtained by some hold-out method, in this example 10-fold cross-validation.

For the optimal scaling regression analysis (b), we analyze the 6876 observations directly, choosing a nominal scaling level for the unordered categorical predictors, and an ordinal scaling level for the ordered predictors. The outcome variable (Annual Income) with nine categories was given an ordinal scaling level as well. The ordinal scaling level results in optimal quantifications, and these are used in the computation of the

TABLE 1

Results for two different regression analyses of the HTF (2009) Marketing Data. APE *gives the apparent prediction error (the mean of the squared differences between the linear combination of (transformed) predictors and the (transformed outcome)), and* $\widehat{\text{EPE}}$ *the estimate of the expected prediction error with associated standard error (s.e.), obtained by 10-fold cross-validation*

Observed Data	Outcome Ordinal	Predictors Ordinal	Predictors Nominal	APE	$\widehat{\text{EPE}}$ (s.e.)
Standard Treatment (a)	numeric ($\mathbf{y}$)	binary ($\mathbf{x}_k$)	dummies ($\mathbf{x}_k$)	0.534	0.539 (0.010)
Optimal Scaling (b)	ordinal ($\mathbf{y}$)	ordinal ($\mathbf{x}_k$)	nominal ($\mathbf{x}_k$)	0.483	0.492 (0.011)

MSE. (Details are given in Section 2.) The results for the two analyses are compared in Table 1.

It is clear that incorporating optimal transformations for both the ordered and the unordered categorical variables is beneficial in terms of prediction accuracy (as estimated in 10-fold cross-validation); the decrease in the estimate of the expected prediction error in analysis (b) compared to analysis (a) is more than 9% (0.539 versus 0.492). The associated standard errors are comparable (0.010 and 0.011), even though 40% more parameters were fitted in the transformation of the predictors (49 versus 35). We will show how the optimal transformations are obtained in Section 2, and will present them in Figures 2 and 3 in Section 3.

2. OPTIMAL SCALING REGRESSION

In this section, the optimal scaling methodology is treated in detail. After some background, the OS loss function is defined for general transformations. Then, focusing on transformations of categorical variables as running example, computational details are provided. Explanations of other transformation possibilities and an algorithmic overview are included in Appendix B.

2.1 Background

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* (also used in Hastie, Tibshirani and Buja, 1994), and when the (predictor and outcome) variables are all categorical, the term quantification is used (Gifi, 1990). Quantification is also one of the key terms in the data analysis framework developed by Hayashi (1952). In the psychometric literature, nonlinear regression with optimal scaling has been extensively explored, starting with Kruskal's (1965) nonlinear, monotonic, version of ANOVA. This approach was followed upon in additive modeling (ADDALS; De Leeuw, Young and Takane, 1976) and multiple regression (MORALS; Young, De Leeuw and Takane, 1976); also, see the review paper Young (1981). The collective work by the Leiden group at the department of Data Theory resulted first in Gifi (1981), later officially published as Gifi (1990). Winsberg and Ramsay (1980) replaced Kruskal's original monotonic regression approach (that produces step functions) by monotonic regression splines (that produce smooth piecewise polynomial functions); a nice review is given in Ramsay (1988). In the meantime, optimal transformations in regression had entered the mainstream statistical literature in the Breiman and Friedman (1985) paper on Alternating Conditional Expectations (ACE) and the Tibshirani (1988) paper on Additivity Variance Stability (AVAS). Finally, regression with optimal scaling became widely available in statistical packages such as SAS/STAT (in a procedure called TRANSREG) (SAS/STAT, 1990) and in the CATREG procedure in SPSS Categories 8.0 (Meulman, Heiser and SPSS, 1998).

2.2 The OS Regression Loss Function

In linear regression problems, we have a system consisting of a random "outcome", "response" or "dependent" variable Y and a set of random "explanatory", "predictor", or "independent" variables $X = \{X_k\}_{k=1}^{P}$, where P denotes the number of predictors. The problem defines a "training" sample, $\{y_i, \mathbf{x}_i\}_1^N$ of known values for Y and X, where $(y_i, \mathbf{x}_i)$ links the predictor variables of the ith object with the ith value of the outcome variable, and where $i = 1, \ldots, N$. Using the training data, the model can be written as

$$y_i = \boldsymbol{\beta}'\mathbf{x}_i + \epsilon_i = \sum_{k=1}^{P} \beta_k x_{ik} + \epsilon_i,$$

or (in vector notation) as

$$\mathbf{y} = \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \cdots + \beta_P\mathbf{x}_P + \boldsymbol{\epsilon}, \tag{2.1}$$

where $\boldsymbol{\epsilon} = \{\epsilon_i\}_{i=1}^{N}$ are the residuals, and the linear combination of predictor variables is formed through

a set of regression coefficients in $\boldsymbol{\beta} = \{\beta_k\}_{k=1}^{P}$. We assume from the start, without loss of generality, that the outcome and predictor variables are standardized (i.e., centered and normalized to have a standard deviation of one), thus there is no need to fit an intercept. The optimal coefficients are estimated by minimizing the (mean) squared error, thus we write the optimization task in the form of a least squares loss function:

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{N} \| y_i - \boldsymbol{\beta}' \mathbf{x}_i \|^2 = \left\| \mathbf{y} - \sum_{k=1}^{P} \beta_k \mathbf{x}_k \right\|^2, \tag{2.2}$$

where $\|\cdot\|^2$ denotes the squared Euclidean norm. Loss function (2.2) has to be minimized over the vector of coefficients $\boldsymbol{\beta}$, and solving for the optimal $\boldsymbol{\beta}$ will give a maximum correlation between the linear combination of the predictor variables and the outcome variable.

For the optimal scaling algorithm, the loss function is written as

$$L(\boldsymbol{\beta}, \varphi, \vartheta) = \left\| \vartheta(\mathbf{y}) - \sum_{k=1}^{P} \beta_k \varphi_k(\mathbf{x}_k) \right\|^2. \tag{2.3}$$

The arguments over which the function has to be minimized are the weights $\boldsymbol{\beta} = \{\beta_k\}_{k=1}^{P}$, the transformation $\vartheta(\mathbf{y})$ of Y, and φ that stands for functions $\varphi_k(\mathbf{x}_k)$, that is, the set of nonlinear transformations $\varphi = \{\varphi_k(\mathbf{x}_k)\}_{k=1}^{P}$.

2.3 Transformation in Optimal Scaling

In the optimal scaling approach, there is a large emphasis on the analysis of categorical data; we therefore at the outset introduce an $N \times C_k$ indicator matrix $\mathbf{G}_k$ for each categorical predictor X_k. The number of different categories in X_k is indicated by C_k, and each column of $\mathbf{G}_k = G_k(\mathbf{x}_k)$ shows by 1–0 coding whether or not an object i scores in category c_k of X_k, $c_k = 1, \ldots, C_k$. For each variable, we search for a size C_k vector of quantifications $\mathbf{v}_k$ that minimizes the overall value of the associated loss function, now written as

$$L(\boldsymbol{\beta}, V, \vartheta) = \left\| \vartheta(\mathbf{y}) - \sum_{k=1}^{K} \beta_k \mathbf{G}_k \mathbf{v}_k \right\|^2, \tag{2.4}$$

where V represents the super vector of concatenated quantifications $\{\mathbf{v}_k\}_{k=1}^{P}$. Thus, the optimal scaling mechanism first involves the expansion of $\mathbf{x}_k$ in $\mathbf{G}_k$, followed by a contraction in $\mathbf{G}_k \mathbf{v}_k$, and the result is the transformation $\varphi_k(\mathbf{x}_k)$. Since a continuous variable can be viewed as a variable with N (number of objects) categories, numeric, continuous variables and categorical, discrete variables, can be dealt with in the same framework. It should be noted from the start that we only use indicator matrix notation in the equations to show how to obtain optimal quantifications. We do not use indicator matrices (that are extremely sparse) in the computations. In an efficient algorithm, matrix multiplications that involve $\mathbf{G}_k$ are replaced by simple additions.

Within the class of nonlinear transformations, we make the following distinctions. We call a quantification *nominal* if we merely maintain the class membership information in the quantified variable $\mathbf{G}_k \mathbf{v}_k$, or equivalently, in the nominal transformation $\varphi_k(\mathbf{x}_k)$; if two objects i and i' belong to the same category of variable k, then

$$x_{ik} = x_{i'k} \quad \Longrightarrow \quad (\mathbf{G}_k)_i \mathbf{v}_k = (\mathbf{G}_k)_{i'} \mathbf{v}_k, \tag{2.5}$$

where $(\mathbf{G}_k)_i$ denotes the ith row of $\mathbf{G}_k$. If a categorical predictor variable contains *order information* on the objects, this information can be preserved in the transformation:

$$x_{ik} < x_{i'k} \quad \Longrightarrow \quad (\mathbf{G}_k)_i \mathbf{v}_k \leq (\mathbf{G}_k)_{i'} \mathbf{v}_k, \tag{2.6}$$

restricting the ordinal quantifications in $\mathbf{v}_k$ so that they are nondecreasing, and we call the transformation $\varphi_k(\mathbf{x}_k)$ *ordinal*. In the latter case, $\mathbf{x}_k$ and $\varphi_k(\mathbf{x}_k)$ are related by a monotonic step function. A linear transformation is a further restriction by preserving interval information as well, and amounts to standardizing the original variable. If the original variable is continuous, and we wish to apply less restrictive transformations than linear ones, we need to limit the number of parameters that are fitted in the nonlinear transformation. For instance, we can use regression splines in which the number of parameters is limited by restricting the degree of the spline and the number of interior knots. Alternatively, we could first make a continuous variable discrete with a fixed number of categories (binning), and subsequently apply optimal category quantification, resulting in a step function. The relation between regression spline functions 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 $C_k - 1$, where C_k is the number of categories that is quantified.

Since the predictor variables are usually correlated in the regression problem (2.2), the optimal transformations $\varphi_k(\mathbf{x}_k)$ in (2.3) (e.g., the quantifications $\mathbf{v}_k$ in (2.4)) are also interdependent. For the moment, we assume the transformation of the outcome $\vartheta(\mathbf{y})$ to be fixed to ϑ^*.

2.4 Computation and Convergence

To solve for each $\varphi_k(\mathbf{x}_k)$, we rewrite the loss function in (2.4) as

$$L(\boldsymbol{\beta}, V, \vartheta^*) = \Big\| \vartheta^* - \sum_{l \neq k} \beta_l \mathbf{G}_l \mathbf{v}_l - \beta_k \mathbf{G}_k \mathbf{v}_k \Big\|^2, \tag{2.7}$$

where $\boldsymbol{\beta}$ again denotes $\{\beta_k\}_{k=1}^P$ and V the super vector of quantifications $\{\mathbf{v}_k\}_{k=1}^P$. A superscript asterix in the argument list indicates a parameter held fixed in minimizing the loss function. Thus, we separate a variable and its weight from the linear combination of predictors, isolating the target part $\beta_k \mathbf{G}_k \mathbf{v}_k$ from the remainder, denoted as $\sum_{l \neq k} \beta_l \mathbf{G}_l \mathbf{v}_l$. In short, OS not only alternates between optimizing $\boldsymbol{\beta}$ and φ, it also turns the original multivariate problem into a series of univariate ones: we update the weight and transformation for one predictor at a time, holding the weights and transformations for the other predictors fixed, and iterate across predictors. This estimation strategy has been called "alternating least squares" or "conditional least squares"; in statistics it was labeled "backfitting", following Friedman and Stuetzle (1981). Other terms found in the literature are "the Gauss-Seidel algorithm", "Newton–Raphson", "Componentwise update", "Block Relaxation", "one-variable-at-a-time" and "Coordinate Descent".

De Leeuw, Young and Takane (1976) have established the convergence of the OS algorithm by showing that it can be viewed as involving a cyclically repeated series of optimal conic projections. Convergence of such series can be proven by theorems available in literature (Gurin, Poljak and Raĭk, 1967; Céa and Glowinski, 1973; Oberhofer and Kmenta, 1974; and Zangwill, 1969/70). A formal proof of convergence for the closely related ACE procedure was given in Breiman and Friedman (1985). The full optimal scaling algorithm is described in the SPSS Algorithms documentation (IBM Corp., 2010), of which a concise version is included in Appendix B. Here we describe the update of the coefficients and the quantifications/transformations.

For updating the target part $\beta_k \mathbf{G}_k \mathbf{v}_k$ in (2.7), we define an auxiliary variable $\mathbf{u}_k$

$$\mathbf{u}_k = \vartheta^*(\mathbf{y}) - \sum_{l \neq k} \beta_l \mathbf{G}_l \mathbf{v}_l, \tag{2.8}$$

thus $\mathbf{u}_k$ is the *partial residual*. Inserting $\mathbf{u}_k$ in (2.7), we find that

$$L(\beta_k, \mathbf{v}_k) = \|\mathbf{u}_k - \beta_k \mathbf{G}_k \mathbf{v}_k\|^2, \tag{2.9}$$

which is a function of β_k and $\mathbf{v}_k$ only. We minimize (2.9) over all $\mathbf{G}_k \mathbf{v}_k \in \mathbb{C}_k(\mathbf{x}_k)$, where $\mathbb{C}_k(\mathbf{x}_k)$ specifies the *cone* that contains all admissible transformations of the variable $\mathbf{x}_k$. In the case of a nominal transformation, the cone $\mathbb{C}_k(\mathbf{x}_k)$ is defined by

$$\mathbb{C}_k(\mathbf{x}_k) \equiv \big\{ \varphi_k(\mathbf{x}_k) | \varphi_k(\mathbf{x}_k) = \mathbf{G}_k \mathbf{v}_k \big\}, \tag{2.10}$$

and we define the metric projection $P_{\mathbb{C}_k(\mathbf{x}_k)}$ as

$$P_{\mathbb{C}_k(\mathbf{x}_k)} \equiv \min_{\mathbf{v}_k} \|\mathbf{u}_k - \beta_k \mathbf{G}_k \mathbf{v}_k\|^2. \tag{2.11}$$

This metric projection ensures that objects in the same category according to variable k obtain the same quantification in the transformed variable $\varphi_k(\mathbf{x}_k) = \mathbf{G}_k \mathbf{v}_k$. Minimizing over $\mathbf{v}_k$ results in the conditional estimate $\tilde{\mathbf{v}}_k$, defined as

$$\tilde{\mathbf{v}}_k = \beta_k^{-1} \mathbf{D}_k^{-1} \mathbf{G}_k' \mathbf{u}_k, \tag{2.12}$$

where $\mathbf{D}_k = \mathbf{G}_k' \mathbf{G}_k$, a diagonal matrix with the marginal frequencies of the categories in $\mathbf{x}_k$ on the main diagonal. Actually, only the sign of β_k is needed because the transformed variable $\varphi_k(\mathbf{x}_k)$ is standardized. The latter is ensured by setting

$$\mathbf{v}_k^* = N^{1/2} \tilde{\mathbf{v}}_k (\tilde{\mathbf{v}}_k' \mathbf{D}_k \tilde{\mathbf{v}}_k)^{-1/2}, \tag{2.13}$$

so that $(\mathbf{G}_k \mathbf{v}_k)'(\mathbf{G}_k \mathbf{v}_k) = N$. The standardization of the transformed variable, in addition to preventing the degenerate solution with all quantified values equal to zero, allows us to compute the regression weight β_k separately from the transformation. The current value for the regression weight β_k is obtained by minimizing $L(\beta_k, \mathbf{v}_k^*)$, resulting in

$$\beta_k^* = N^{-1} \mathbf{u}_k' \mathbf{G}_k \mathbf{v}_k^*. \tag{2.14}$$

For the other scaling levels the transformation amounts to restrictions of the nominal quantification in equation (2.12). For this purpose, we split the loss function (2.9) into loss of nominal transformation and loss due to the restriction:

$$\begin{aligned} L(\beta_k, \mathbf{v}_k) = {} & \|\mathbf{u}_k - \beta_k \mathbf{G}_k \mathbf{v}_k^{\text{nom}}\|^2 \\ & + \|\mathbf{G}_k \mathbf{v}_k^{\text{nom}} - \mathbf{G}_k \mathbf{v}_k^{\text{restr}}\|^2, \end{aligned} \tag{2.15}$$

where $\mathbf{v}_k^{\text{nom}}$ is the nominal quantification $\tilde{\mathbf{v}}_k$ given in equation (2.12), and $\mathbf{v}_k^{\text{restr}}$ is $\tilde{\mathbf{v}}_k$ restricted according to the chosen optimal scaling level. In Table 2, we describe three types of restrictions schematically, for ordinal, splines and numeric transformations, respectively. Details are fully described in Appendix B.

When all coefficients and predictor transformations have been updated in this way, one may transform the

TABLE 2
Restricting quantifications

Quantification		Ingredients
1. $\mathbf{v}_k^{\text{ord}}$:	weighted monotonic regression of $\mathbf{v}_k^{\text{nom}}$ on $\mathbf{x}_k^{\text{cat}}$ using $\mathbf{D}_k$	$\mathbf{x}_k^{\text{cat}} = C_k$-vector (*) with different categories or values in $\mathbf{x}_k$. $\mathbf{D}_k$ = diagonal matrix with marginal frequencies.
2. $\mathbf{v}_k^{\text{splin}}$:	weighted linear regression of $\mathbf{v}_k^{\text{nom}}$ on $\mathbf{S}_k\mathbf{b}_k$	$\mathbf{S}_k$ = matrix (*) with C_k different rows contain ing a spline basis, for example, for I-splines. $\mathbf{b}_k$ is a vector with spline coefficients to be estimated.
3. $\mathbf{v}_k^{\text{num}}$:	weighted linear regression of $\mathbf{v}_k^{\text{nom}}$ on $\mathbf{x}_k^{\text{cat}}$ using $\mathbf{D}_k$	$\mathbf{x}_k^{\text{cat}} = C_k$-vector with different categories or values in $\mathbf{x}_k$. $\mathbf{G}_k\mathbf{v}_k^{\text{num}}$ is standardized $\mathbf{x}_k$.

*$\mathbf{x}_k^{\text{cat}}$ and $\mathbf{S}_k^{\text{cat}}$ are shortened and reordered versions of $\mathbf{x}_k$ and $\mathbf{S}_k$, with only one row per category.

outcome variable as well by defining $\mathbf{G}_{\text{y}}$ as the $N \times C_y$ indicator matrix for the outcome and $\mathbf{v}_{\text{y}}$ as the vector of category quantifications for the outcome. Then we may minimize the loss function (2.4) over $\mathbf{v}_{\text{y}}$ holding the coefficients and predictor quantifications fixed at $\boldsymbol{\beta}^*$ and $\mathbf{v}^*$, respectively,

$$L(\boldsymbol{\beta}^*, \mathbf{v}^*, \mathbf{v}_{\text{y}}) = \left\| \mathbf{G}_{\text{y}}\mathbf{v}_{\text{y}} - \sum_k \beta_k^* \mathbf{G}_k \mathbf{v}_k^* \right\|^2. \tag{2.16}$$

Setting partial derivatives in (2.16) to zero, gives

$$\tilde{\mathbf{v}}_{\text{y}} = \mathbf{D}_{\text{y}}^{-1}\mathbf{G}_{\text{y}}' \sum_k \beta_k^* \mathbf{G}_k \mathbf{v}_k^* \tag{2.17}$$

for the conditionally optimal nominal quantification. For the outcome variable, we have the same set of transformation options available as for the predictor variables. In practice, however, for reasons of interpretability of the final prediction model, it is preferable for a continuous outcome to choose a linear transformation or a spline transformation with very few degrees of freedom. A possible nonlinear relation between (a linear combination of) predictor variables and the outcome is preferably taken care of by nonlinear transformation of the predictors. If the outcome is categorical, we choose a monotonic step function if the categories are ordered, and a nonmonotonic step function if they are unordered. In the latter case, when all predictors are linearly transformed, optimal scaling is equivalent (up to a scaling factor) to classical linear discriminant analysis (Gifi, 1990).

3. APPLICATIONS OF OPTIMAL SCALING, INCLUDING DIAGNOSTICS

Before going to richer applications on empirical data, we first propose a number of diagnostics that are useful in evaluating the OS results. Next, we demonstrate OS regression with a small example, to show its properties when there are nonlinear relationships between the predictor variables and the outcome.

3.1 Diagnostics

The overall criterion that is optimized by the optimal scaling transformations is the multiple correlation R^2 between the optimal linear combination of transformed predictor variables and the (transformed) outcome (Gifi, 1990). An important diagnostic for a single predictor is its "predictability" from the other predictors, and the values for the so-called conditional independence are given by the inverse of the diagonal elements of the inverse of the correlation matrix $\mathbf{R}$ for the (transformed) predictors. The elements of this P-vector will be called *tolerance values*, defined by

$$\text{TOL} = \frac{1}{\text{diag}(\mathbf{R}^{-1})}.$$

Optimal scaling transformations for multiple regression will usually increase the average value of TOL over the various predictors. A suitable candidate for a diagnostic for the condition of the correlation matrix for (transformed) predictors is the so-called Log Determinant Divergence. This measures the difference between matrices by the log determinants of those matrices. In OS regression, we measure the divergence of the correlation matrix $\mathbf{R}$ and the identity matrix $\mathbf{I}$, because $\mathbf{I}$ is the correlation matrix when all predictors are completely uncorrelated. The Log Determinant Divergence (DLD) is then written as (adapted from Dhillon, 2008)

$$\begin{aligned} \text{DLD} &= \mathbf{D}_{\ell d}(\mathbf{R}, \mathbf{I}) \\ &= \text{tr}(\mathbf{R}) - \log\det(\mathbf{R}) - P \\ &= -\sum_{k=1}^{P} \log(\lambda_k). \end{aligned} \tag{3.1}$$

Note that this is a "degenerate" version of Stein's loss:

$$\operatorname{tr}(\mathbf{\Sigma}\hat{\Sigma}^{-1}) - \log\det(\mathbf{\Sigma}\hat{\Sigma}^{-1}) - P$$

(James and Stein, 1961, page 376). Equation (3.1) shows that our diagnostic $\mathbf{D}_{\ell d}$ (DLD) boils down to a simple function of the eigenvalues of the correlation matrix between transformed predictors. In our experience, optimal scaling transformations for regression will usually decrease the value of $\mathbf{D}_{\ell d}(\mathbf{R}, \mathbf{I})$. A third diagnostic that can be used to evaluate the condition of $\mathbf{R}$ is the value of its smallest eigenvalue (SMEV). If $\mathbf{R}$ is ill-conditioned, the smallest eigenvalue will be small. In our experience, optimal scaling transformations will in general increase the value of the smallest eigenvalue if the predictors are highly correlated. We remark that if X is high dimensional, with $P \gg N$, and/or $\mathbf{R}$ has eigenvalues equal to 0, the diagnostics have to be adapted.

3.2 A Simple Example with Two Predictors, Nonlinearly Related to the Outcome Variable

This simple example has two predictor variables only, and we sampled X_1 and X_2 with $N = 1000$ from a multivariate normal distribution, with $\rho = 0.707$ being the population correlation. The outcome variable was constructed as $\mathbf{y} = \exp(\mathbf{x}_1) + |\mathbf{x}_2| + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, 1)$. The correlation between the predictors in the sample is 0.706. The results for three different models are shown in Table 3, which gives the results for the regression coefficients β (with their standard errors), the fit r^2, the correlation $r(\mathbf{x}_1, \mathbf{x}_2)$ and the diagnostics defined in Section 3.1. The standard error of the regression coefficients has been estimated by a bootstrap with 1000 samples, and the expected prediction error ($\widehat{\text{EPE}}$) and its standard error have been estimated by 10-fold cross-validation. When the predictors were transformed, nonmonotonic spline transformations (using second degree polynomials, with three internal knots) were fitted.

Model 1 gives results for OLS regression. The predictors are highly correlated, regression coefficients β_1 and β_2 are very different, and both the fit (r^2) and the estimated prediction accuracy ($\widehat{\text{EPE}}$) are rather poor. Because we only have two predictors, the smallest eigenvalue equals $1 - |r(x_1, x_2)|$. Because β_2 is very small, we transform $\mathbf{x}_2$, keeping $\mathbf{x}_1$ fixed (model 2); we observe that compared to model 1, the dependence among the predictors becomes minimal (the correlation between the predictors is now -0.050) and the conditional independence (tolerance) is close to maximal (0.998). The r^2 increases, as well as the regression coefficient β_2; the estimate of the expected prediction error decreases.

If we allow both predictors to be transformed (model 3), both r^2 and β_1 increase compared to model 2, while the tolerance values and β_2 decrease somewhat. The estimated prediction error is smallest for model 3, and compared to model 1, the overall improvement is obvious. Because the predictors are uncorrelated in model 2, the transformations in model 3 increase the value of the smallest eigenvalue.

Figure 1 shows the partial residual plots, with the partial residual plotted versus predictor k. (For example, the plot in the upper left panel depicts $\mathbf{u}_1 = \mathbf{y} - \beta_2\mathbf{x}_2$ on the vertical axis versus $\mathbf{x}_1$ on the horizontal axis.) These partial residual plots are given for both the original predictors $\mathbf{x}_1$ and $\mathbf{x}_2$ in the left panels, as well as for the transformed predictors $\varphi_1(\mathbf{x}_1)$ and $\varphi_2(\mathbf{x}_2)$ in the right panels. We observe that the transformations $\varphi_1(\mathbf{x}_1)$ and $\varphi_2(\mathbf{x}_2)$, shown in the left middle panels, are a nonlinear fit to the scatter in the partial residual plots in the left panels. The regression between the transformed predictors and the partial residuals in the right middle panels has been linearized, as is seen from the independently fitted smoothing splines (right panels). These functions are fitted to inspect whether the choice of transformation has been appropriate. If not, the plots on the far right hand side would indicate this by showing a nonlinear curve, implying there is still nonlinearity remaining after transformation.

TABLE 3
Results for three different regression models with two predictors

Transformation	r^2	β_1 (s.e.)	β_2 (s.e.)	$\widehat{\text{EPE}}$ (s.e.)	$r(x_1, x_2)$	SMEV	TOL*	DLD
1. lin(x_1), lin(x_2)	0.379	0.634 (0.027)	−0.027 (0.034)	0.661 (0.181)	0.706	0.294	0.502	0.690
2. lin(x_1), spl(x_2)	0.570	0.637 (0.029)	0.438 (0.020)	0.467 (0.146)	−0.050	0.950	0.998	0.002
3. spl(x_1), spl(x_2)	0.855	0.851 (0.031)	0.224 (0.029)	0.148 (0.007)	0.214	0.786	0.954	0.047

*In regression with two predictors, both obviously have the same value for the conditional independence.

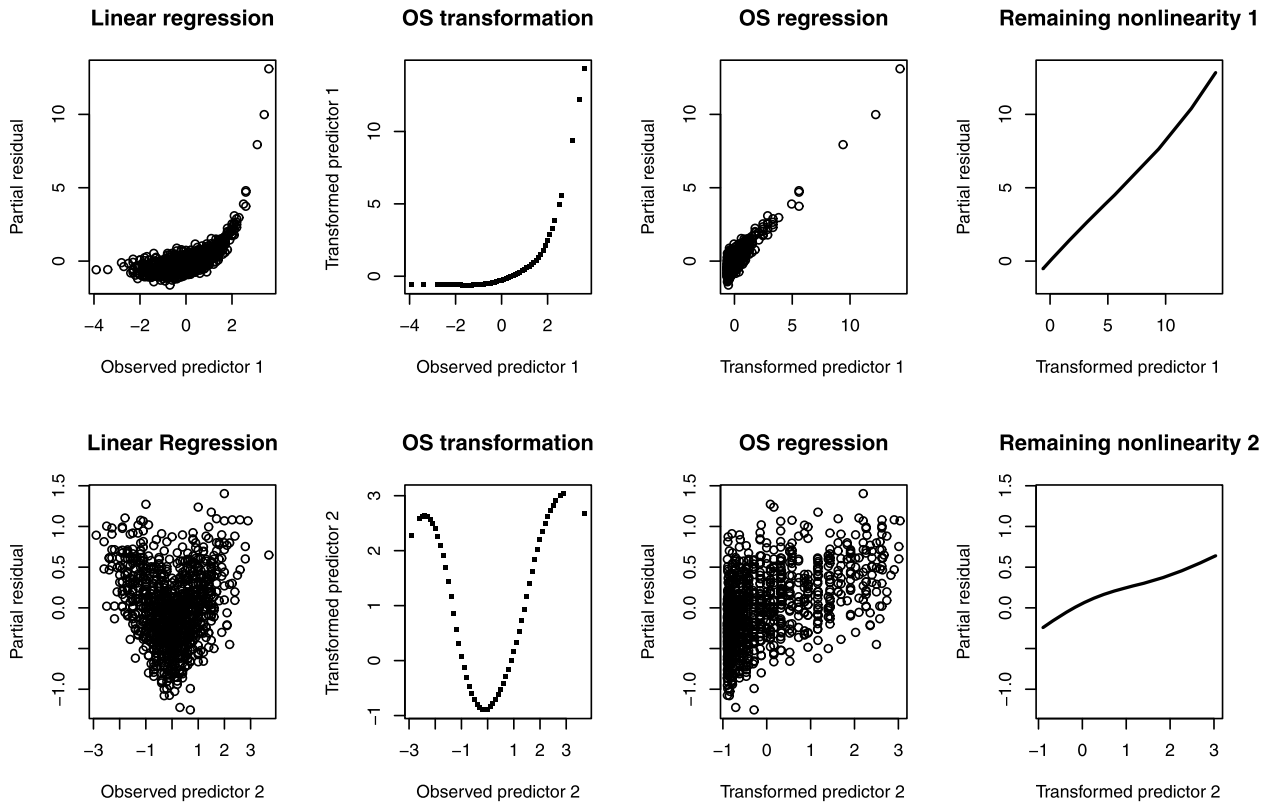

FIG. 1. *Two predictors*: *scatter plots of partial residuals versus observed predictors* (*left panels*) *and versus transformed predictors* (*right middle panels*). *Transformations* $\varphi_1(\mathbf{x}_1)$ *and* $\varphi_2(\mathbf{x}_2)$ *versus observed predictors displayed in the left middle panels*. *Linearization of partial residuals shown in far right panels*, *curves obtained by fitting smoothing splines with four knots to the scatter plots in the right middle panels*.

3.3 Mixed Nominal and Categorical Predictors with an Ordinal Outcome: The Marketing Data

We revisit the Marketing Data from Hastie, Tibshirani and Friedman ((2009), pages 492–494), introduced in Section 1 to demonstrate the optimal scaling features of the analysis. As was described previously, the data consist of an ordinal outcome variable $\mathbf{y}$ (*Annual Income per Household*) and 13 predictor variables ($\mathbf{x}_k$) for customers in a San Francisco shopping mall. The predictor set consists of a mixture of ordinal (ordered) and nominal (unordered) categorical variables. For the optimal scaling regression analysis, we analyzed the data matrix of 6876 objects and 13 predictors directly, without creating dummy variables, and with appropriate optimal scaling level for the nominal and ordinal variables, respectively. The apparent prediction error has been minimized by optimal scaling and is 0.483, with 49 degrees of freedom; the estimate of the expected prediction error obtained by 10-fold cross-validation is 0.492, with standard error 0.011.

In Section 2 it was shown how the transformations in optimal scaling are found. We obtain a transformed data matrix with the same dimensions as the original data matrix. First we use the transformed predictor matrix, with columns $\mathbf{G}_k\mathbf{v}_k$, to display the various transformations in Figures 2 and 3 (the transformation for the binary variable *sex* is omitted). We describe the most important predictor variables. The first transformation plot is for *marital status*, a nominal variable with five categories (married, living together, divorced, widowed, single). The quantifications are represented as dots, and are connected by a nonmonotonic stepfunction. The standardized regression coefficient is 0.189, with an estimated standard error of 0.026, obtained by a bootstrap with 1000 bootstrap samples. The transformation plot shows a basically decreasing function, which combined with the positive coefficient indicates that married couples have the highest income on average, while widowed people have the lowest. The category for single obtains the same quantification as the category for divorced/separated. This information can be used in further analyses, where we would use the transformed predictor, which has numeric properties, instead of the original one. In other words, us-

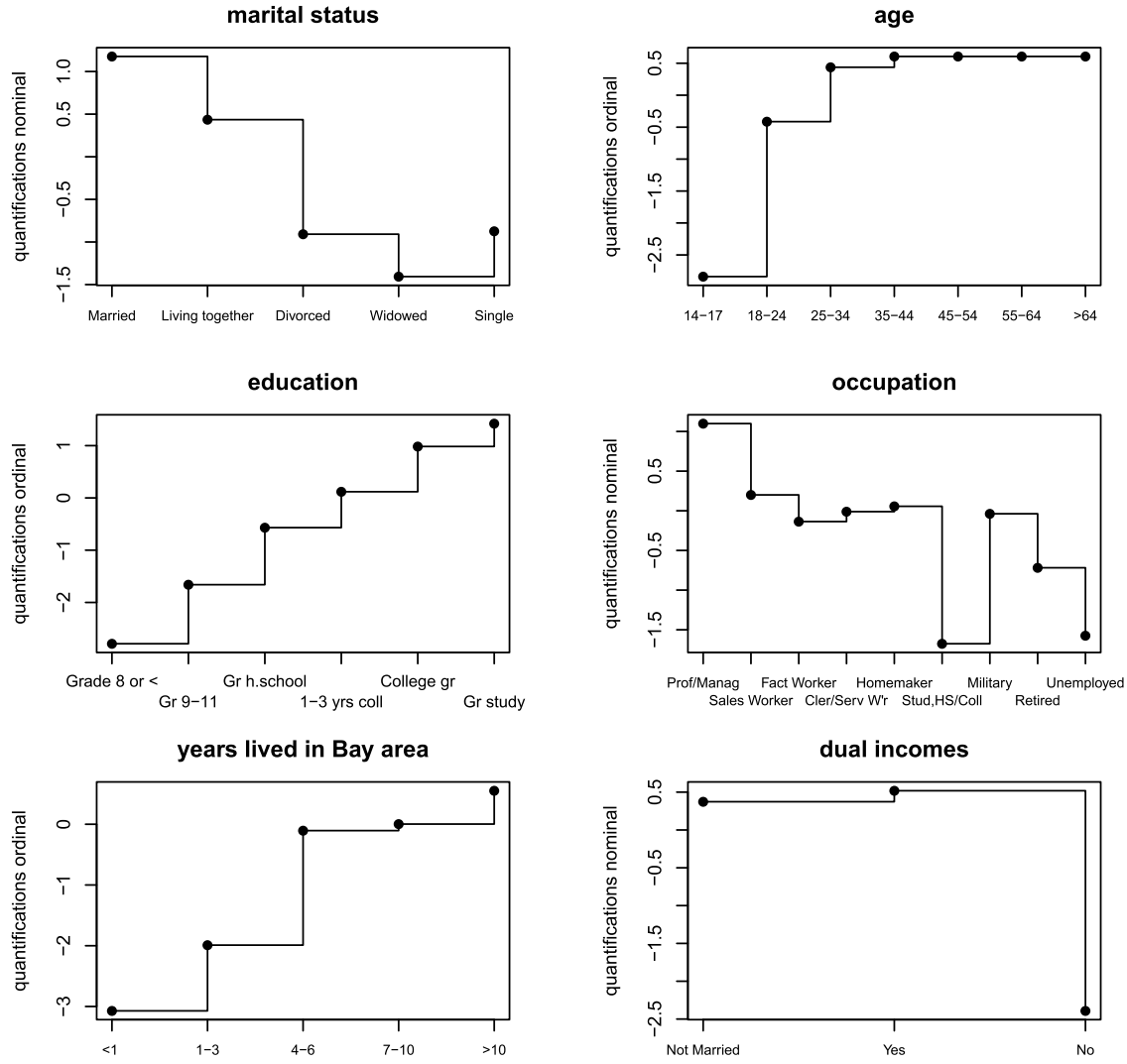

FIG. 2. *Marketing Data. Transformations of predictors* 2–7 *in predicting Annual Income of Household. Nominal or ordinal quantifications on vertical axis versus original categorical values on horizontal axis.*

ing the transformed predictors subsequently in a linear multiple regression analysis, would give the same results as obtained in the optimal scaling analysis.

The predictor, *age*, shows a monotonically increasing function, a positive regression weight, 0.279 (0.022), but we see that the four highest categories are tied (obtaining the same quantifications). The transformation of *education* ($\beta = 0.122$ (0.16)) is very regular. Looking at *occupation* ($\beta = 0.252$ (0.14)), we see that category 6 (Student, HS or College) obtains the lowest quantification, similar to category 9 (unemployed). The categories 3, 4, 5 and 7 have very similar quantifications. The full order of the categories obtained by optimal scaling is: Professional/Managerial, Sales Worker, Military, Clerical/Service Worker, Homemaker, Factory Worker, Retired, Unemployed, Student. Even with 8 degrees of freedom, the standard error is small. The predictor *householder status* ($\beta = 0.124$ (0.14)) gives the new order as Own, Live with Parents/Family, Rent.

Figure 4 contains two panels. In the panel on the left, predicted values for each outcome category (of *Annual Income*) are represented by a boxplot, with the usual range from the lower hinge (the 25th percentile) to the upper hinge (the 75th percentile). The predicted values for the outcome variable ($\hat{\mathbf{y}}$) are given by the vector with the weighted sum of nominal and ordinal, quantified, predictors in $\mathbf{G}_k\mathbf{v}_k$, thus

$$\hat{\mathbf{y}} = \sum_{k=1}^{K} \beta_k \mathbf{G}_k \mathbf{v}_k. \tag{3.2}$$

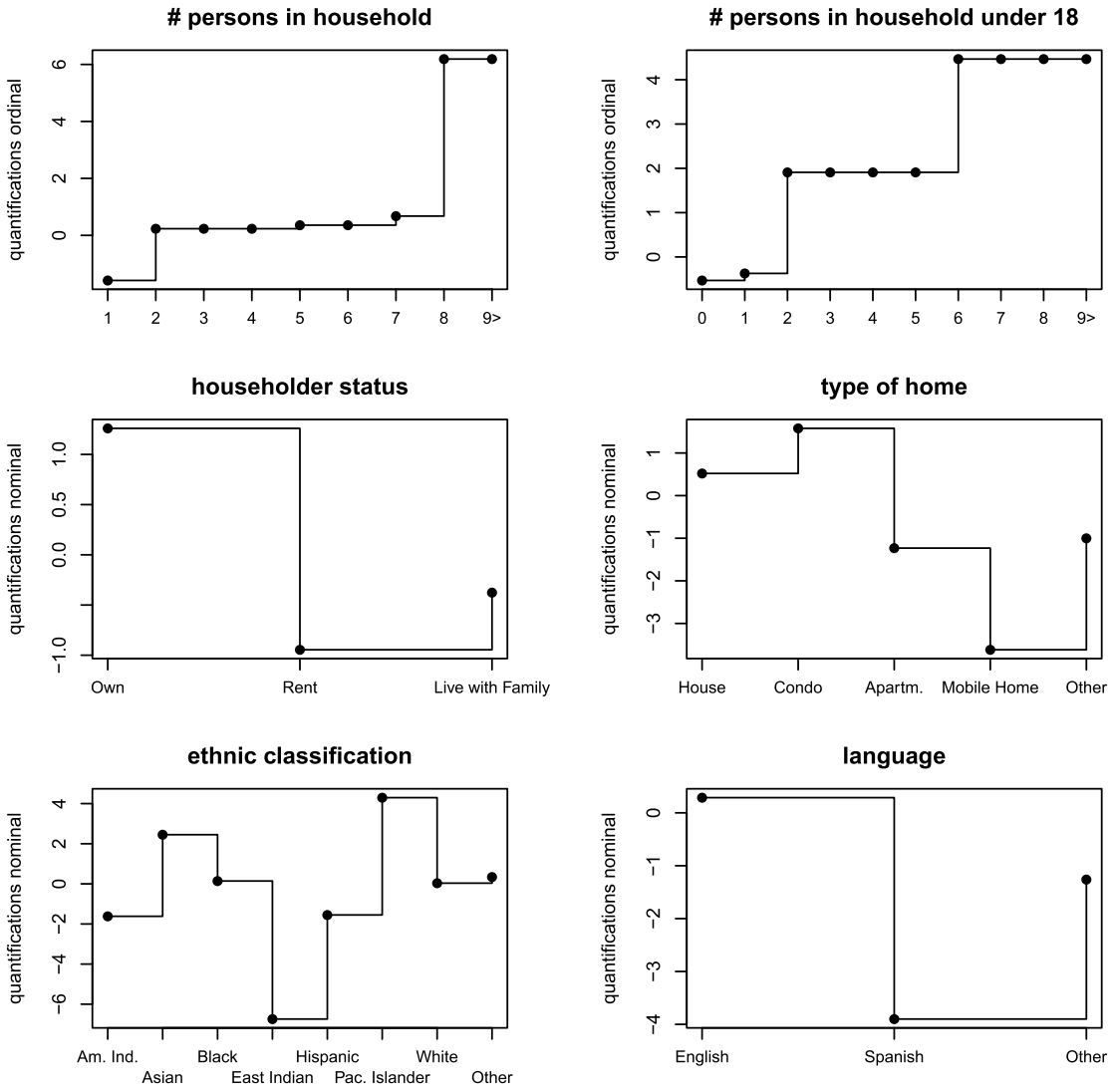

FIG. 3. *Marketing Data. Transformations of predictors* 8–13 *in predicting Annual Income of Household. Nominal or ordinal quantifications on vertical axis versus original categorical values on horizontal axis.*

The median values (lines) in the boxplots are very close to the associated, unstandardized, category quantifications ($\tilde{\mathbf{v}}_y$) for the outcome variable in the right panel, indicated by dots, and computed as

$$\tilde{\mathbf{v}}_y = \mathbf{D}_y^{-1}\mathbf{G}_y' \sum_k^K \beta_k \mathbf{G}_k \mathbf{v}_k = \mathbf{D}_y^{-1}\mathbf{G}_y'\hat{\mathbf{y}}. \tag{3.3}$$

Thus, the unstandardized quantifications of the categories of the outcome variable in $\tilde{\mathbf{v}}_y$ are obtained as averages ($\mathbf{D}_y$ contains the marginals of the outcome categories) of the appropriate values of $\hat{\mathbf{y}}$, as coded in $\mathbf{G}_y$. The dots are connected by a (monotonic) step function. The most remarkable feature of the transformation of the outcome variable is the big jump from category 1 (less than \$10,000) to category 2 (\$10,000 to \$14,999). The remaining steps are very similar.

3.4 Mixed Ordered Categorical and Continuous Predictors for Five Ordered Categories of Cervical Cancer

The data used in this example were collected at the Leiden Cytology and Pathology Laboratory, and concern characteristics of cells obtained from patients with various grades of cervical preneoplasia and neoplasia. To obtain the samples, taken from the ectocervix as well as the endocervix, special sampling and preparation techniques were used. The correct histological diagnosis was known by a subsequently taken biopsy. A subset of the data has been previously analyzed

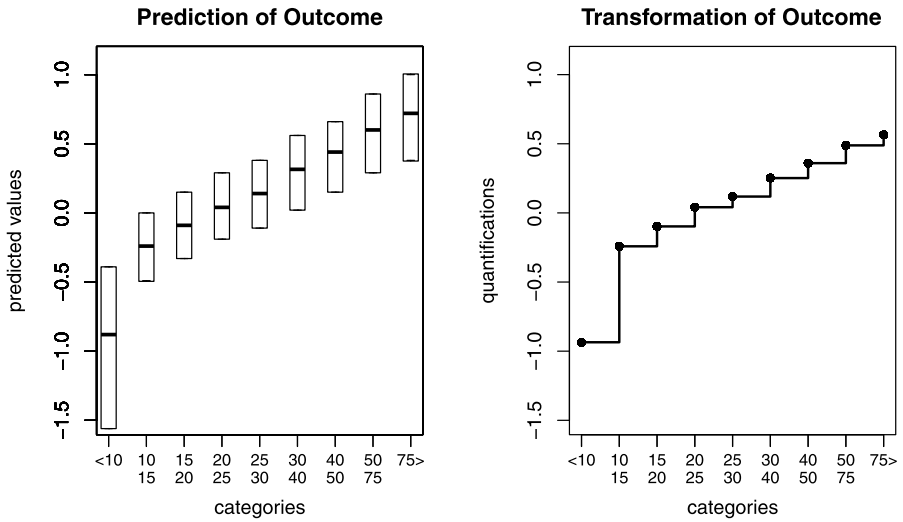

FIG. 4. *Outcome in Marketing Data. Boxplots for predicted values* ($\hat{\mathbf{y}}$) *for each category of Annual Income* (*in the left panel*). *Quantifications*, *indicated by dots*, *of the nine categories of Income in the panel on the right. Dots are connected by a stepfunction.*

in Meulman et al. (1992) and Friedman and Meulman (2003), and contains, according to the histological diagnosis, 50 cases with mild dysplasia (histological group 1), 50 cases with moderate dysplasia (histological group 2), 50 cases with severe dysplasia (histological group 3), and 50 cases with carcinoma in situ (histological group 4). The number of cases with invasive squamous cell carcinoma (histological group 5) is 42. For each of the 242 cases, seven qualitative features of the cells were determined. The features were rated by a pathologist on a scale ranging from 1 (normal) to 4 (very abnormal); so these seven variables are ordered categorical. The features under consideration are *Nuclear Shape, Nuclear Irregularity, Chromatin Pattern, Chromatin Distribution, Nucleolar Irregularity, Nucleus/Nucleolus Ratio*, and *Nucleus/Cytoplasm Ratio*. In addition, four quantitative features of each sample were established: *Number of Abnormal Cells per Fragment* (mean values), *Total Number of Abnormal Cells, Number of Mitoses* and *Number of Nucleoli* (mean values). From the earlier analyses mentioned above, it is known that this data set is noisy, and accurate prediction of the outcome is thereby difficult.

The complete analysis design can be described as follows, and is depicted in Figure 5. The full data set contains 242 objects, and 20 objects are set apart as supplementary validation data in each of 12 steps in the outer loop (the validation phase). The remaining objects form the active data set, used in the modeling phase. Thus, in the inner loop, we analyze the remaining set of $242 - 20 = 222$ objects; the analysis of the active data set gives us the so-called apparent prediction error (APE), the mean squared error loss (MSE). On the active data set, we apply an 11-fold cross-validation, so 202 objects are used as training data, and 20 objects as test data. In this step, we repeatedly compute estimates for both the regression coefficients and the transformations for the training data, and these are subsequently applied to the test data. This gives us the estimate of the expected prediction error ($\widehat{\text{EPE}}$), by averaging over the MSE for the 11 test data sets, and its standard error (s.e.). In the validation phase (outer loop), the regression coefficients and the transformations for each of the 12 active data sets (of size

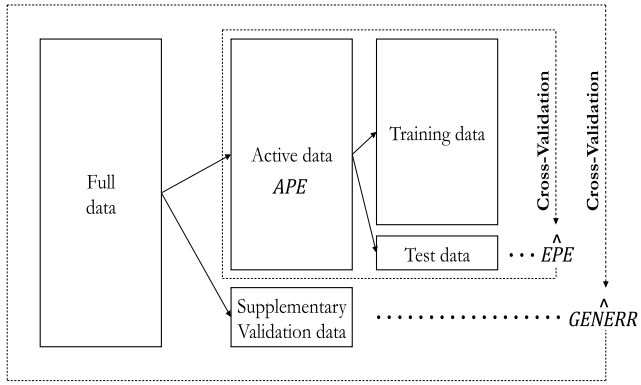

FIG. 5. *Schematic representation of the estimation of the apparent prediction error* (APE), *expected prediction error* ($\widehat{\text{EPE}}$) *and generalization error* ($\widehat{\text{EPE}}$). *Solid arrows represent data splits by resampling routines*, *which are repeated* (*dotted arrows*). APE *is obtained directly from the Active Data. In the example presented*, $\widehat{\text{EPE}}$ *estimation is done by* 11*-fold cross-validation* (*CV*) *in the inner loop*, *and* $\widehat{\text{GENERR}}$ *estimation is done by* 12*-fold CV in the outer loop.*

TABLE 4

Prediction error (MSE) *for different sets of transformations of the Cervix Cancer Data. Values are averages over the* 12 *steps in the validation phase. The* MSE *for models labeled by an* $*$ *are used to display results in Figure* 6

Models	Predictors		Outcome	Average	Average	
	Qual	Quant		APE	$\widehat{\text{EPE}}$ (s.e.)	$\widehat{\text{GENERR}}$ (s.e.)
1*	linear	linear	linear	0.254	0.288 (0.026)	0.279 (0.076)
2*	nominal	linear	linear	0.216	0.274 (0.027)	0.269 (0.080)
3*	ordinal	linear	linear	0.217	0.270 (0.027)	0.266 (0.079)
4*	ordinal	spl(nmon, 2, 2)	linear	0.148	0.203 (0.018)	0.195 (0.055)
5	ordinal	spl(nmon, 2, 1)	linear	0.150	0.197 (0.017)	0.189 (0.052)
6	ordinal	spl(mono, 2, 1)	linear	0.153	0.191 (0.017)	0.183 (0.051)
7*	ordinal	spl(mono, 2, 2)	linear	0.150	0.189 (0.017)	0.182 (0.051)
8*	ordinal	spl(nmon, 2, 2)	ordinal	0.125	0.179 (0.021)	0.179 (0.060)
9	ordinal	spl(nmon, 2, 1)	ordinal	0.128	0.175 (0.020)	0.174 (0.059)
10	ordinal	spl(mono, 2, 1)	ordinal	0.134	0.175 (0.019)	0.174 (0.055)
11	ordinal	spl(mono, 2, 3)	ordinal	0.127	0.171 (0.019)	0.170 (0.058)
12*	ordinal	spl(mono, 2, 2)	ordinal	0.128	0.169 (0.019)	0.167 (0.056)

222) are applied to the 20 objects in the associated supplementary validation set, and the resulting 12 values of the MSE for the validation set are used to obtain the estimate of the generalization error ($\widehat{\text{GENERR}}$) and its standard error. Since in the outer loop we also obtain 12 values for (APE, and 12 values for $\widehat{\text{EPE}}$ and their standard error for the test data, we average those as well. These are the values are given in Table 4.

This table shows the results for three different sets of transformations; the rows are ordered according to the average of the estimated expected prediction error ($\widehat{\text{EPE}}$). In the first three rows (models 1 to 3), transformations of the outcome and the quantitative predictors are linear; the qualitative predictors obtain a linear, nominal, and ordinal transformation, respectively. Because the results indicate that ordinal transformations for the qualitative variables are most appropriate, we also fit those in the second set of models. In addition, we fit nonlinear spline functions for the quantitative variables, both nonmonotonic and monotonic, and varying the number of interior knots. In the third set, we apply the same transformations, but now also an ordinal transformation of the outcome variable (diagnosis) has been fitted.

Conclusions from this extended example, based on both $\widehat{\text{EPE}}$ and $\widehat{\text{GENERR}}$, are as follows.

- The results show that a model with all transformations linear is least successful.
- Better results are obtained when nonlinear transformations are applied, first for the qualitative predictors, allowing for ordinal transformation), and next also for the quantitative predictors.
- For the latter, models that were fitted include both nonmonotonic and monotonic cubic splines, with a varying number of interior knots. The results show that monotone functions are to be preferred over nonmonotone functions.
- Best results are obtained when the outcome variable (Diagnosis) is given an ordinal transformation. Increasing the number of knots for the monotonic splines is hardly worthwhile.
- The average values for $\widehat{\text{EPE}}$ and $\widehat{\text{GENERR}}$ are very similar, but the average values for the standard errors is about three times as large for $\widehat{\text{GENERR}}$. This is caused by the fact that in the validation phase, estimates for $\widehat{\text{GENERR}}$ are based on 20 values, while estimates for $\widehat{\text{EPE}}$ are based on 222 values.

From these results, a selection ($*$) is depicted in Figure 6 to emphasize the similarities and differences in the results. The first seven boxplots show the apparent prediction error for the active data.

- The differences between linear transformations for the quantitative predictors (a1, a2, a3) versus nonlinear transformations (a4, a7, a8, a12) are large.
- Those between nonmonotonic (a4 and a7) and monotonic transformations (a8 and a12) are small.
- There is a difference, however, between linear (a4 and a7) versus ordinal transformation of the outcome (a8 and a12).

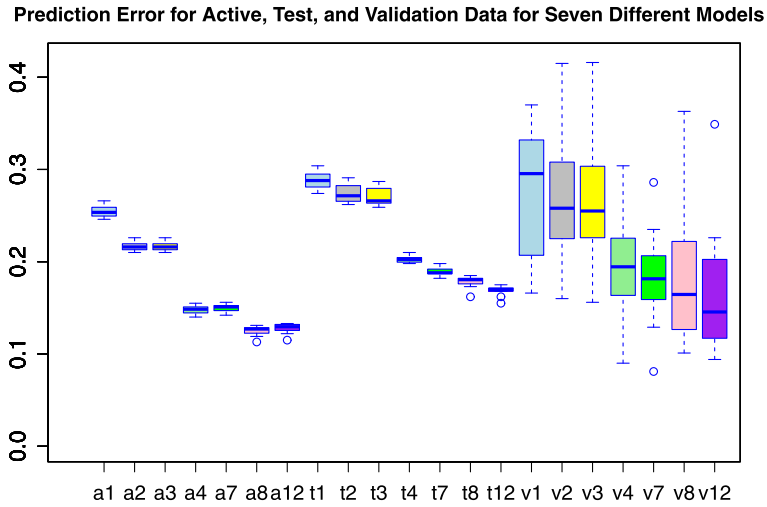

FIG. 6. *Comparing different models for the Cervix data with respect to prediction error*: APE *for active data* $(a1, \ldots, a12)$, $\widehat{\text{EPE}}$ *for test data* $(t1, \ldots, t12)$, *and* $\widehat{\text{GENERR}}$ *for validation data* $(v1, \ldots, v12)$; *numbers refer to the models in Table* 4.

The next seven boxplots give the estimated expected prediction error, based on the test data.

- Pattern of differences between (t1, t2, t3) versus (t4, t7, t8, t12) is confirmed by cross-validation.
- Transformation of quantitative predictors by monotonic splines (t7 and t12) is beneficial with respect to $\widehat{\text{EPE}}$ compared to nonmonotonic splines (t4 and t8).

The last six boxplots show the estimated generalization error for the validation data.

- The median values of $\widehat{\text{GENERR}}$ are similar to the median values of $\widehat{\text{EPE}}$.
- The variation is obviously much larger, as was the standard error in Table 4.
- The overall pattern shows again that monotonic transformations should be preferred throughout.

We display the optimal quantifications for model 12 in the transformation plots in Figure 7. The red dots represent the category quantifications from the analyses for the 12 separate active data sets. The black lines connect the averages of the quantifications in the 12 different analyses. We observe the following.

- Overall, the quantifications for the 12 active data sets are remarkably stable.
- With respect to the transformation of Diagnosis, the biggest step is between the categories 3 and 4. This has a very clear clinical counterpart, since it is the difference between severe dysplasia and the first class of cancer (carcinoma in situ).
- It turns out that this departure from linearity has a positive effect on the prediction accuracy.
- Steps have about equal size for Nucl_Shape, and Chrom_Pat, but not for the five other qualitative predictors.
- Transformations for #Abn_Cells, Tot#_Abn, #Mitoses are smooth, and the one for #Nucleoli is flat at the upper end.

To conclude this section, we display in Figure 8 properties of the diagnostics for a hierarchy of models with increasing number of degrees of freedom due to different sets of transformation. We display the smallest eigenvalues (left panel), and the corresponding log-determinant divergence from independence (middle panel). The values for both diagnostics differ considerably among the six models, where the size differences between model 1 and model 6 are (almost) of the order 3 (0.136 and 0.393, respectively, for SMEV, and 4.52 and 1.81, respectively, for DLD). In the panel at the right, we display the average Variance Inflation Factor (VIF), indicating the dependence of a predictor on the other predictors. The average dependence can easily be computed from the inverse of the eigenvalues of the predictor correlation matrix. If we write $\mathbf{R} = \mathbf{L}\boldsymbol{\Lambda}\mathbf{L}'$ then $\mathbf{R}^{-1} = \mathbf{L}\boldsymbol{\Lambda}^{-1}\mathbf{L}'$, and $\text{tr}(\mathbf{R}^{-1}) = \text{tr}(\boldsymbol{\Lambda}^{-1})$; the diagonal of $\boldsymbol{\Lambda}^{-1}$ contains the eigenvalues of $\mathbf{R}^{-1}$ in reversed order. The three diagnostics show overall the same pattern for the different sets of transformations. The smallest eigenvalues (left panel) increase in each step, and divergence from independence (middle panel) and average predictor dependency (right panel) decrease. The largest step is taken when going from the first to the second model in the hierarchy (which are models 1 (predictors linear) and 7 (op-

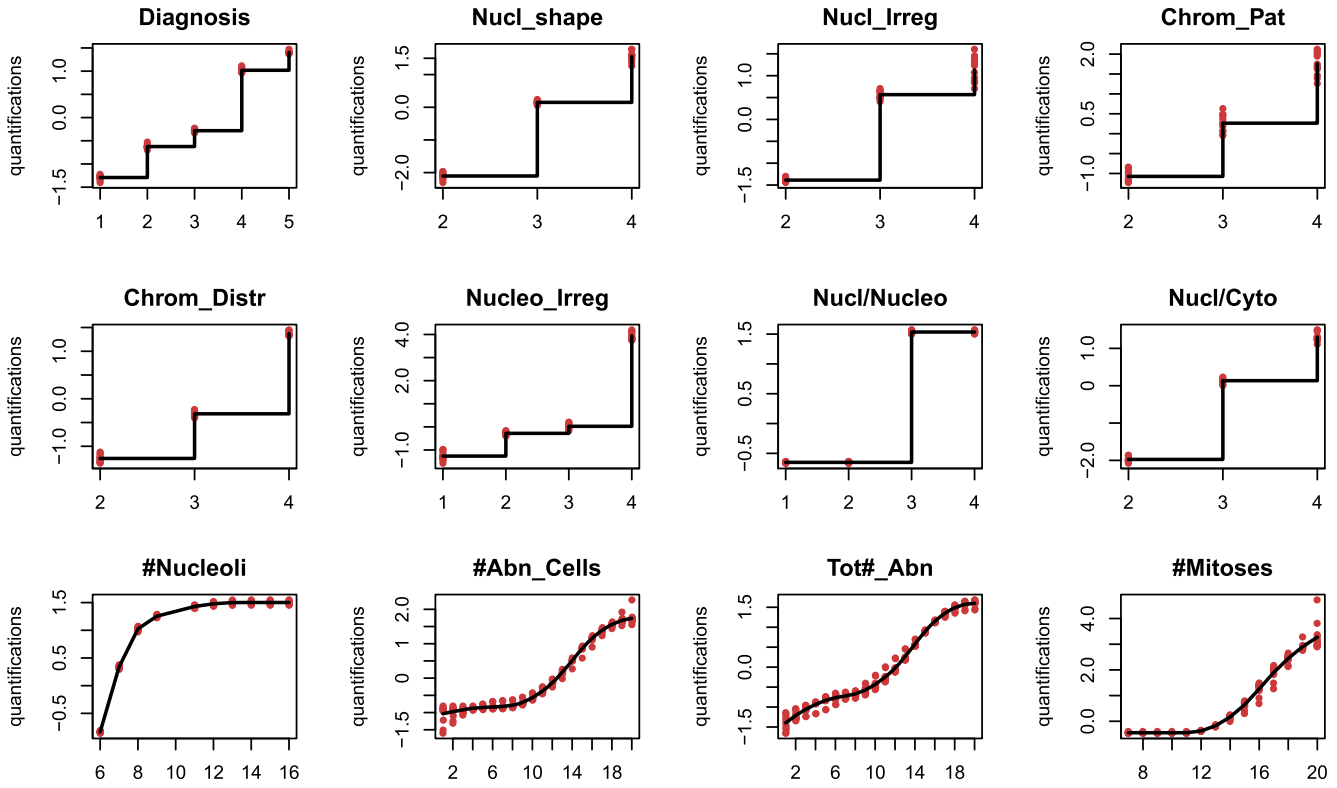

FIG. 7. *Ordinal transformations for diagnosis* (*outcome*, *five categories*), *and seven categorical*, *qualitative predictors*. *Monotonic quadratic spline transformations with two interior knots for the four quantitative predictors. The red dots indicate* 12 *different transformations for the predictors in the active data in the outer cross-validation loop* (*see Figure* 5); *their averages have been connected.*

timally scaled predictors) in Table 4, respectively). The smallest eigenvalue plot shows a substantial increase between 5 and 6, which models are identical except for the fact that in the sixth model the outcome is ordinal instead of linear. The divergence from independence and average predictor dependency both show a

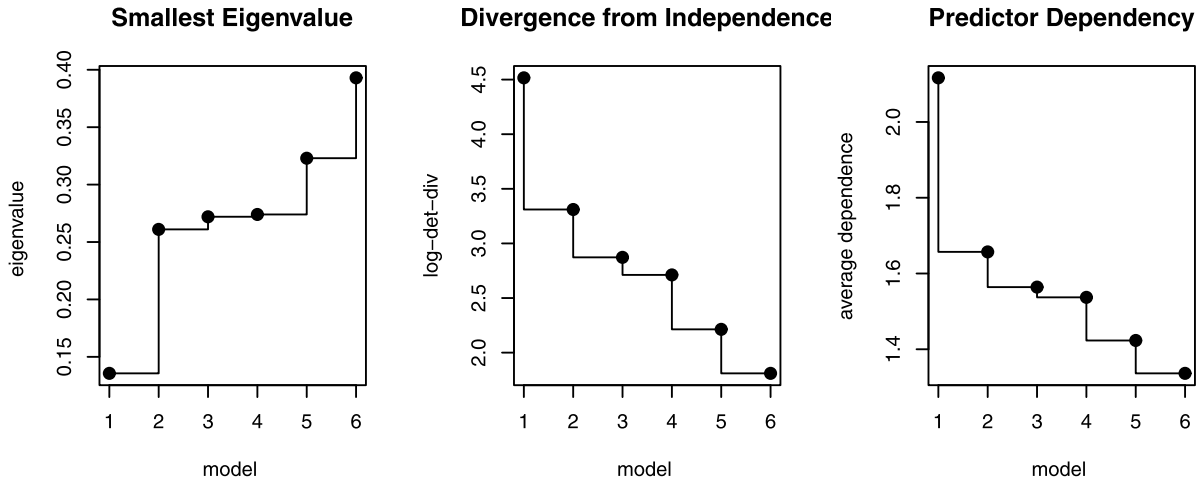

FIG. 8. *Smallest eigenvalues of six predictor correlation matrices* (*left panel*), *difference between* **R** *and* **I** *by log determinants* (*middle panel*), *and average dependence*, *the average multiple correlation between each predictor and the other predictors*. *The correlation matrices are for* 1: *All predictors and outcome linear.* 2: *Outcome linear, categorical predictors ordinal, quantitative predictors monotonic spline* (2, 2). 3: *Like* 2, *but quantitative predictors nonmonotonic spline* (2, 2). 4: *Like* 3, *but with quantitative predictors nonmonotonic spline* (3, 3). 5: *Like* 4, *but quantitative predictors nominal transformation*. 6: *Like* 5, *but outcome ordinal*.

drop when the outcome is transformed. We conclude that ordinal transformation of the outcome (instead of a linear one) has a positive effect.

4. REGULARIZED OPTIMAL SCALING REGRESSION WITH LASSO, RIDGE, AND ELASTIC NET PENALTIES

Regularization addresses the prediction accuracy problem. It is well known that in certain circumstances OLS regression may result in highly variable estimates of the (unbiased) regression coefficients, and this high variance leads to poor predictions, especially when data require complex models. In those cases, it is beneficial to add a penalty term to the loss function that controls the variance of the regression coefficients, hereby decreasing the standard error of the estimates, at the cost of usually a small increase of the bias (the bias-variance tradeoff), which overall leads to improved prediction accuracy. First, we shall give some background to the most popular regularization methods, notably Ridge regression, the Lasso, and the Elastic Net, and the associated computational issues. Next, we rewrite the penalized loss functions in the optimal scaling framework, to show that regularized optimal scaling regression estimates can be obtained very easily, both for the regularization methods mentioned above, as well as for the Group-Lasso and Blockwise Sparse Regression. The latter methods expand categorical variables to blocks of dummy variables, and continuous variables to blocks of basis functions, and apply regularization to these blocks of variables by joint shrinkage of the dummy coefficients. We will show that these methods are equivalent to particular choices of transformations within regularized optimal scaling regression, and subsequently can be extended with transformations that are restricted to be monotonic.

4.1 Background

Over the years, several methods for regularized regression have been developed. Without any claim to be complete, regularization methods in statistics began with Ridge regression (Hoerl and Kennard, 1970a; Hoerl and Kennard, 1970b), adapting Tikhonov regularization (Tikhonov, 1943), followed by Bridge regression (Frank and Friedman, 1993), the Garotte (Breiman, 1995), and the Lasso (Tibshirani, 1996), also known as Basis Pursuit (Chen, Donoho and Saunders, 1998) in the signal processing literature, and were followed somewhat later by LARS (Efron et al., 2004), Pathseeker (Friedman and Popescu, 2004), and the Elastic Net (Zou and Hastie, 2005). Since then the number of references especially to the Lasso and its extensions has grown exponentially.

The oldest regularization method, Ridge regression, reduces the variability by shrinking the coefficients, resulting, as mentioned above, in less variance at the cost of usually only a small increase of bias. The coefficients are shrunken towards each other and to zero, but will never become exactly zero. So, when the number of predictors is large, Ridge regression will not provide a sparse model that is easy to interpret. Subset selection, on the other hand, does provide interpretable models, but assumes more sparseness. The Lasso was developed by Tibshirani (1996) to improve both prediction accuracy and model interpretability by combining the nice features of Ridge regression and subset selection. Thus, the Lasso reduces the variability of the estimates by shrinking the coefficients, and at the same time produces interpretable models by shrinking some coefficients to exactly zero. The Elastic Net (Zou and Hastie, 2005) combines Ridge regression and the Lasso, obtaining sparse models due to the use of a Lasso penalty, and encouraging grouping of variables due to the use of a Ridge penalty. Where the Lasso would only select one variable of the group, the Elastic Net tends to select groups of highly correlated variables together.

The original Lasso algorithm uses a quadratic programming strategy that is complex and computationally demanding; hence it is not feasible for large values of P, and moreover, it can not be used when $P > N$. Since the Lasso paper, various less complex and/or more efficient lasso algorithms were proposed. For example, Osborne, Presnell and Turlach (2000a) developed a homotopy method that can handle $P > N$ predictors, but it is still computationally demanding when P is large. The same method was discussed in Efron et al. (2004) in a different framework, and became known as the LARS-Lasso. These methods provide efficient algorithms to find the entire Lasso regularization path. The "Grafting" algorithm of Perkins, Lacker and Theiler (2003), the "Pathseeker" algorithm of Friedman and Popescu (2004), and the "boosting" algorithm of Zhao and Yu (2007) are gradient descent algorithms that can deal with $P > N$ predictors in a computationally less demanding way. However, in the $P > N$ case, none of these Lasso algorithms can select more than N predictors. The Elastic Net algorithm that is based on the LARS-Lasso algorithm is capable of selecting more than N predictors due to the use of the additional Ridge penalty.

4.2 Computation of Regression Coefficients in Regularized Least Squares

Ridge regression, the Lasso, and the Elastic Net constrain the size of the regression coefficients by setting a maximum on the sum of the squared coefficients (Ridge), on the sum of absolute values of the coefficients (Lasso), or on both these sums (Elastic Net). Ridge regression uses an L_2 restriction, which is written

$$\sum_{k=1}^{P} \beta_k^2 \leq t_2, \tag{4.1}$$

with t_2 a tuning parameter with respect to the sum of squares of the β_k, and its value has to be determined in the optimization process. The Lasso uses an L_1 restriction, constraining the sum of the absolute values of the regression coefficients:

$$\sum_{k=1}^{P} |\beta_k| \leq t_1. \tag{4.2}$$

The Elastic Net combines the Ridge and Lasso constraints:

$$\sum_{k=1}^{P} \beta_k^2 \leq t_2 \quad \text{and} \quad \sum_{k=1}^{P} |\beta_k| \leq t_1. \tag{4.3}$$

It is well known that the restrictions on the regression coefficients in (4.1)–(4.3) can be rewritten using an additional penalty term to the loss function, because there is a one-to-one relation between the value of the sum constraint and the value of the penalty. Throughout, we denote λ_1 as the strength of the Lasso penalty, and λ_2 that of the Ridge penalty. The Elastic Net loss function is written (in matrix notation) as

$$L^{\text{e-net}}(\beta) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_1 \mathbf{w}'\boldsymbol{\beta} + \lambda_2 \boldsymbol{\beta}'\boldsymbol{\beta}, \tag{4.4}$$

where the elements w_k of $\mathbf{w}$ are either $+1$ or -1, depending on the sign of the corresponding coefficient $\hat{\beta}_k$. The loss function for Ridge or Lasso is obtained by setting λ_1 or λ_2 to zero.

For least squares loss, the solution for the Ridge regression coefficient has a well-known, analytic expression:

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = (\mathbf{X}'\mathbf{X} + \lambda_2 \mathbf{I})^{-1}\mathbf{X}'\mathbf{y}. \tag{4.5}$$

Provided $\mathbf{X}'\mathbf{X}$ is nonsingular, the regression coefficients for the Lasso are

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} = (\mathbf{X}'\mathbf{X})^{-1}\left(\mathbf{X}'\mathbf{y} - \frac{\lambda_1}{2}\mathbf{w}\right), \tag{4.6}$$

where again the elements w_k of $\mathbf{w}$ are either $+1$ or -1, depending on the sign of the corresponding coefficient $\hat{\beta}_k^{\text{lasso}}$. Since the elements of $\mathbf{w}$ depend on the estimates of the coefficients, obtaining the Lasso coefficients is a least squares problem with 2^P inequality constraints (there are 2^P possible sign patterns for the coefficients), and was efficiently solved for by the LARS algorithm. For the Elastic Net, the regression coefficients are estimated as

$$\hat{\boldsymbol{\beta}}^{\text{e-net}} = (\mathbf{X}'\mathbf{X} + \lambda_2 \mathbf{I})^{-1}\left(\mathbf{X}'\mathbf{y} - \frac{\lambda_1}{2}\mathbf{w}\right), \tag{4.7}$$

and minimization of this loss function is much like minimizing the Lasso loss function, and the entire Elastic Net regularization paths can be estimated almost as efficiently as the Lasso paths with the LARS-ENet algorithm (Zou and Hastie, 2005).

4.3 Computation of Regression Coefficients in ROS: One-Variable-at-a-Time

Generalizing regularized least squares loss to include optimal scaling is straightforward:

$$\begin{aligned} L^{\text{e-net}}(\beta, \vartheta, \varphi) = &\left\| \vartheta(\mathbf{y}) - \sum_{k=1}^{P} \beta_k \varphi_k(\mathbf{x}_k) \right\|^2 \\ &+ \lambda_1 \sum_{k=1}^{P} |\beta_k| + \lambda_2 \sum_{k=1}^{P} \beta_k^2, \end{aligned} \tag{4.8}$$

with $\lambda_1 = 0$ for $L^{\text{ridge}}(\beta, \vartheta, \varphi)$ and $\lambda_2 = 0$ for $L^{\text{lasso}}(\beta, \vartheta, \varphi)$.

As was shown in Section 2, the OS algorithm estimates the transformations and regression coefficients one at a time, and it removes the effect of the other predictors from the outcome when estimating a particular coefficient using (2.8), (2.9) and (2.14). In this section, we will show that the "one-variable-at-a-time" approach of optimal scaling results in straightforward estimation of coefficients and enables the Lasso to select more than N predictors.

The very same strategy to find the Lasso solution in linear regression problems was already applied in Fu (1998), who used the name "shooting algorithm". However, the fact that this algorithm worked was not fully appreciated at the time, or not fully understood. For example, Osborne, Presnell and Turlach (2000b) state that it is not applicable in the $P > N$ case. The same approach was independently re-invented in Daubechies, Defrise and De Mol (2004) and in the optimal scaling research in Leiden in 2006, as reported in Van der Kooij (2007), where it was shown that the one-variable-at-a-time algorithm made the computation of

regularized coefficients for the Lasso and thus also for the Elastic Net trivially simple. Friedman et al. (2007) subsequently showed that the algorithm was also very fast and convergence properties were established building further on work by Tseng (1988) and Tseng (2001). Also, see Wu and Lange (2008) for an additional convergence proof. The crux is that least squares loss functions that are extended with a penalty term can be separated in a convex, differentiable part (the L_2-norm) and a part consisting of convex penalties summed over the variables. Tseng's convergence results can be trivially generalized with transformations $\vartheta(\mathbf{y})$ and $\{\varphi(\mathbf{x}_k)\}$, as in (4.8), without loss of generality, since transformations are considered fixed in (4.9), (4.10) and (4.11). Convergence of the ROS regression algorithm is then implied by independently combining two algorithms that are known to work (coordinatewise penalized regression and optimal scaling, with convergence results given in Section 2.4).

The important consequence of the one-variable-at-a-time approach is that the estimates of the regularized coefficients $\hat{\beta}_k^{\text{ridge}}$ and/or $\hat{\beta}_k^{\text{lasso}}$ can be computed in the setting of simple univariate regression of $\mathbf{u}_k$ on $\varphi(\mathbf{x}_k)$, as follows:

$$L^{\text{ridge}}(\beta_k) = \|\mathbf{u}_k - \beta_k\varphi_k(\mathbf{x}_k))\|^2 + \lambda_2\sum_{l\neq k}\beta_l^2 + \lambda_2\beta_k^2, \tag{4.9}$$

$$L^{\text{lasso}}(\beta_k) = \|\mathbf{u}_k - \beta_k\varphi_k(\mathbf{x}_k))\|^2 + \lambda_1\sum_{l\neq k}|\beta_l| + \lambda_1|\beta_k|, \tag{4.10}$$

$$L^{\text{e-net}}(\beta_k) = \|\mathbf{u}_k - \beta_k\varphi_k(\mathbf{x}_k))\|^2 + \lambda_1\sum_{l\neq k}|\beta_l| + \lambda_2\sum_{l\neq k}\beta_l^2 + \lambda_1|\beta_k| + \lambda_2\beta_k^2, \tag{4.11}$$

where $\mathbf{u}_k = \vartheta(\mathbf{y}) - \sum_{l\neq k}\beta_l\phi_l(\mathbf{x}_l)$ analogous to (2.8). If we define $\tilde{\beta}_k$ as the simple update $\tilde{\beta}_k = N^{-1}\mathbf{u}_k'\varphi_k(\mathbf{x}_k)$ analogous to (2.14), then incorporating regularization in the OS regression loss function only requires a simple adjustment of the regression coefficients for Ridge, the Lasso and the Elastic Net, and this amounts to

$$\hat{\beta}_k^{\text{ridge}} = \tilde{\beta}_k/(1+\lambda_2), \tag{4.12}$$

$$\hat{\beta}_k^{\text{lasso}} = \begin{cases} \tilde{\beta}_k - \dfrac{\lambda_1}{2} & \text{if } \tilde{\beta}_k > 0 \text{ and } \dfrac{\lambda_1}{2} < \tilde{\beta}_k, \\ \tilde{\beta}_k + \dfrac{\lambda_1}{2} & \text{if } \tilde{\beta}_k < 0 \text{ and } \dfrac{\lambda_1}{2} < |\tilde{\beta}_k|, \\ 0 & \text{otherwise,} \end{cases} \tag{4.13}$$

$$\hat{\beta}_k^{*\text{e-net}} = \begin{cases} \dfrac{\tilde{\beta}_k - \frac{\lambda_1}{2}}{1+\lambda_2} & \text{if } \tilde{\beta}_k > 0 \text{ and } \dfrac{\lambda_1}{2} < \tilde{\beta}_k, \\ \dfrac{\tilde{\beta}_k + \frac{\lambda_1}{2}}{1+\lambda_2} & \text{if } \tilde{\beta}_k < 0 \text{ and } \dfrac{\lambda_1}{2} < |\tilde{\beta}_k|, \\ 0 & \text{otherwise} \end{cases} \tag{4.14}$$

for Ridge, the Lasso and the Elastic Net, respectively. Equation (4.14) has a direct connection with the shooting algorithm by Fu (1998). As suggested by Zou and Hastie (2005), when reporting the coefficients for the Elastic Net, we correct for the double amount of shrinkage in the estimation by rescaling the coefficients $\hat{\beta}_k^{*\text{e-net}}$ after convergence of the algorithm

$$\hat{\beta}_k^{\text{e-net}} = \hat{\beta}_k^{*\text{e-net}}(1+\lambda_2). \tag{4.15}$$

4.4 Selection of the Penalty Parameter(s)

To select the optimal value of the penalty parameter(s), λ_1 and/or λ_2, the expected prediction error ($\widehat{\text{EPE}}$) has to be estimated for each (combination of) penalty value(s). To estimate the $\widehat{\text{EPE}}$, well-known analytic methods like Generalized Cross Validation (GCV; Golub, Heath and Wahba, 1979), AIC, or BIC cannot be used, because we include optimal transformations, which complicates the computation of degrees of freedom, and we wish our method to work for the case where $P \gg N$. So, we resort to resampling methods, such as cross-validation and bootstrapping. The latter methods can be very time consuming, especially when we have to find the optimal combination of penalty values λ_1 and λ_2 in the Elastic Net, requiring a full grid search. However, application of the bootstrap or cross-validation can be made much less time-consuming by not assessing the estimate of the expected prediction error for all combinations of penalty parameter values. We have observed in many examples that estimates of the expected prediction error usually show regular curves when we plot the prediction error for increasing values of λ_1, the Lasso parameter, and repeat this for different values of λ_2, the Ridge parameter. An example will be shown in Figure 12. Thus, the model selection procedure can be made much more efficient by conducting the analysis in two phases. In the first phase, the region of the optimal values on the path is determined by using a rather big step size for consecutive values of the penalty parameters, and using warm starts. In the second phase, the search is limited to this region (that contains the minimum), and the optimal values are determined by taking much smaller steps, again using warm starts.

In most applications, we apply the one-standard-error rule, as originally proposed in Breiman et al.

(1984): we first determine the optimal model according to the lowest prediction error of a combination of penalty parameters, and then select the most parsimonious model within one standard error of the minimum. In our applications, we either use K-fold cross-validation or the 0.632 bootstrap method (Efron, 1983), since the latter theoretically gives a better estimate of the estimate of the expected prediction error than the *standard* bootstrap. Usually, the 0.632 bootstrap gives a somewhat higher estimate of the expected prediction error than K-fold cross-validation; however when different models are compared using the respective resampling methods, the same conclusions are obtained. Details on how to use the 0.632 bootstrap are extensively described in Van der Kooij (2007).

4.5 The Group Lasso and Regularized Optimal Scaling of Categorical Variables

In standard linear regression, it is common practice to deal with a categorical variable by replacing it by a set of dummy variables. Each dummy variable is a binary predictor, and a coefficient is sought; we will call these dummy coefficients. Applying regularization straightforwardly in this situation would amount to regularizing these dummy coefficients. The Group Lasso method of Yuan and Lin (2006) treats a set of dummy variables as a group, and applies a norm restriction to the vector of dummy coefficients in the group.

If a categorical variable is given a nominal scaling level in ROS regression, it can be shown that the category quantifications $\tilde{\mathbf{v}}_k$ (2.13) before standardization are equal to the dummy coefficients. The category quantifications are subsequently normalized by $\mathbf{v}_k^* = N^{1/2}\tilde{\mathbf{v}}_k(\tilde{\mathbf{v}}_k'\mathbf{D}_k\tilde{\mathbf{v}}_k)^{-1/2}$ so that $\mathbf{v}_k^{*\prime}\mathbf{D}_k\mathbf{v}_k^* = N$, where $(\tilde{\mathbf{v}}_k'\mathbf{D}_k\tilde{\mathbf{v}}_k)^{1/2}$ is the norm of the coefficients. This norm is equal to the absolute value of the unregularized regression coefficient $\tilde{\beta}_k$.

It follows that applying the adjustment for the Lasso as in (4.13) is equivalent to penalizing the norm of the dummy coefficients. Thus, the Group Lasso is equivalent to Lasso regularization in ROS regression, provided that we use the nominal scaling level for a categorical variable.

For continuous variables, the analogy is similar. In the Blockwise Sparse Regression (BSR) method of Kim, Kim and Kim (2006), for example, a continuous predictor is represented by a group/block of basis functions, such as polynomials. In the optimal scaling approach, continuous predictors can be smoothly transformed by applying nonmonotonic regression splines, using an I-spline basis $\mathbf{S}_k$, fitting spline coefficients $\tilde{\mathbf{b}}_k$, and standardizing the resulting $\mathbf{S}_k\tilde{\mathbf{b}}_k$, giving a transformed predictor. As in the Group Lasso, regularization is then applied to the associated regression weights $\tilde{\beta}_k$. Concluding, the regularized regression weights associated with the step functions or the regression splines in optimal scaling are equivalent to the results obtained in the Group Lasso and Blockwise Sparse Regression, but only when nominal quantifications are used. Otherwise, our approach is more general. For example, we can apply ordinal restrictions instead of nominal quantifications, and Ridge/E-net penalties instead of Lasso penalties. And an important consequence of separating β_k from $\mathbf{v}_k$, is a much more attractive representation/interpretation for categorical predictors. We can interpret β_k as in standard regression, and obtain transformed predictors in $\mathbf{G}_k\mathbf{v}_k$.

5. APPLICATIONS OF REGULARIZED OPTIMAL SCALING

This section contains three examples that have been chosen for their particular properties. First, we shall apply regularization to the simple model with nonlinear relationships. Next, we shall introduce a new data set, with a mixture of predictor variables related to test failure in the United States. Finally, we shall show that ROS regression can deal with high-dimensional data.

5.1 A Simple Example, Regularized

We apply the three forms of regularization to the small data example in Section 3.2, and from those we choose the model that has the smallest estimate of the expected prediction error within 1 standard error from the optimal model (results are given in Table 5). For the

TABLE 5
Simple model: *Best regularized model for three combinations of transformations*

Transformation	r^2	β_1 (s.e.)	β_2 (s.e.)	$\widehat{\text{EPE}}$ (s.e.)	$r(x_1, x_2)$	TOL	λ_1	λ_2
1. lin(x_1), lin(x_2)	0.379	0.215 (0.022)	– (–)	0.826 (0.232)	0.706	0.502	0.80	0.00
2. lin(x_1), spl(x_2)	0.538	0.255 (0.009)	0.237 (0.014)	0.607 (0.199)	0.340	0.885	0.00	1.10
3. spl(x_1), spl(x_2)	0.853	0.746 (0.029)	0.225 (0.024)	0.152 (0.007)	0.354	0.874	0.00	0.10

analysis with linear transformations, this turns out to be the Lasso regularization, where the Lasso penalty is 0.80, and where the second predictor variable is left out of the analysis. If we include transformation of the second predictor only (since it was omitted from the first analysis), the Ridge regularization is selected, with a penalty of 1.10, and both predictors in the model, with regression coefficients 0.255 and 0.237, respectively. Regression coefficients are very similar, the expected prediction error decreases, and so is the correlation between the two predictors. The tolerance increases. Next, if we allow both predictors to be transformed, the first predictor becomes dominant again, the expected prediction error becomes very small, as well as its standard error, while the dependence between transformed predictors and the tolerance are comparable to the previous analysis. However, the ridge penalty in the chosen model is merely 0.10, thus results are very similar to those of the OS analysis without regularization in Table 3.

5.2 The United States Data

The United States Data example concerns data per state ($N = 50$) analyzed in Meulman (1986), with the predictor variables taken from Wainer and Thissen (1981) who used seven social indicator statistics in order to re-examine the Angoff and Mencken (1931) search for "The Worst American State.". The outcome variable gives the percentage of failure on a nation-wide test (and was taken from Walberg and Rasher, 1977). The description of the variables is given in Table 6.

To combine transformation with estimation of the expected prediction error using the 0.632 bootstrap, the 50 values in the original variables were binned into 15 categories, following a uniform distribution as closely as possible. It was already shown in Meulman (1986) that the original data contain some serious nonlinearities, for example, the relation between POPUL on the one hand, and INCOME and ILLIT on the other hand. We summarize the analysis as follows.

- The first model option is the base analysis, since it uses neither optimal scaling nor regularization, and the expected prediction error (estimated with 50 samples for the 0.632 bootstrap) is 0.191 (with standard error 0.036).
- Next, regularization was applied using the Elastic Net. The optimal model (with the smallest estimate of the expected prediction error) turns out to be the unregularized analysis; if we choose the model that has the smallest estimate of the expected prediction error within 1 standard error, we obtain a sparse model with both POPUL and INCOME omitted from the predictor set, resulting in an estimate of the expected prediction error of 0.216 (0.044). The values for the Ridge and Lasso penalties are 9.00 and 0.900, respectively.
- The third option uses spline transformations on the basis of the partial residual plots from the second analysis (not shown).
- In the fourth analysis, the predictor POPUL has been omitted from the predictor set, but compared to the first Elastic Net model (option 2), the transformed variable INCOME remains in the model, with corresponding values for the Elastic Net penalties 3.00 for the Ridge penalty and 0.800 for the Lasso penalty, respectively.

TABLE 6
Test failure (outcome) and social indicator variables (predictors) for the United States

Label	Outcome
FAIL	Failure on nation-wide test
	Predictors
POPUL	1975 population in thousands
INCOME	Per capita income in dollars
ILLIT	Illiteracy rate in percent of population
LIFE	Life expectancy in years
HOMIC	1976 homicide and nonnegligent manslaughter (per 1000)
SCHOOL	Percent of population over age 25 who are high school graduates
FREEZE	Average number of days of the year with temperatures below zero

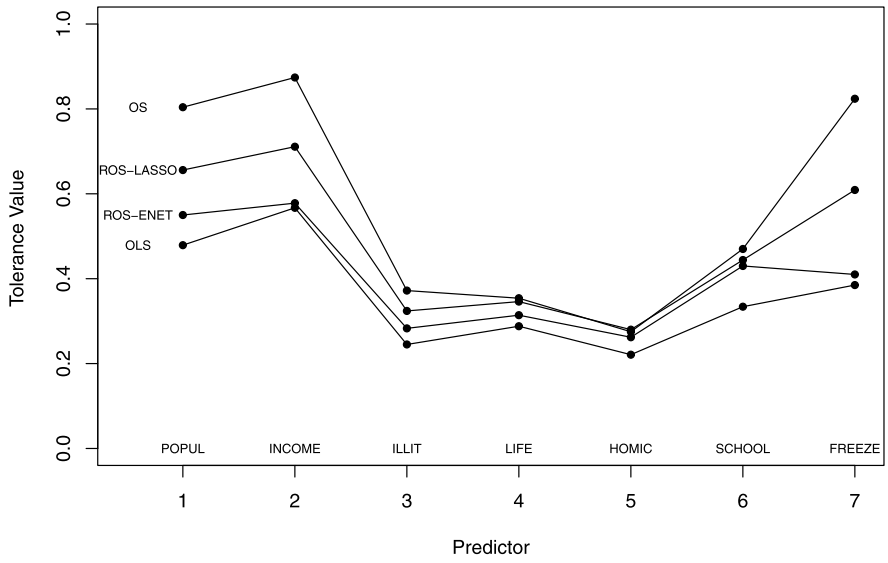

FIG. 9. *Tolerance values for the seven predictors in the United States Data for four different analyses.*

- Concluding this example, although we obtain different values for $\widehat{\text{EPE}}$, none of the differences are significant due to the small sample size. Optimal scaling decreases the value of DLD, but when regularization is added, this effect is diminished. This can be explained as follows: because coefficients are shrunken in the regularization, the contribution of the other predictors is not optimally removed.

We can depict the effect on DLD by plotting the values for *tolerance*, being the inverse of the diagonal elements of the inverse predictor correlation matrix (Figure 9). It is clear that optimal scaling without regularization OS (upper curve) produces the largest values for *tolerance* when compared to the two other curves at the bottom of the panel (ROS-ENET and OLS, respectively), especially for predictor 1, 2 and 7 that obtained a nonmonotonic transformation. If we fit an additional curve for OS combined with the Lasso, we obtain the value 0.20 for the Lasso penalty, which is slightly higher than the optimal model for ROS-ENET $(0.0, 0.10)$ in Table 7, with $\widehat{\text{EPE}} = 0.173$ (0.033). The corresponding curve for the tolerance values is perfectly in between ROS-ENET and OS. The corresponding value for DLD equals 3.217.

The transformations of POPUL, INCOME, ILLIT and FREEZE from analysis 4 in Table 7 are shown in Figure 10. As we use these to depict the partial residuals in Figure 11, we notice a close to linear relationship between the partial residuals $\vartheta(\mathbf{y}) - \sum_{l \neq k} \beta_l \varphi_l(\mathbf{x}_l)$ on the vertical axis and the transformed predictor $\varphi_k(\mathbf{x}_k)$ on the horizontal axis, except for POPUL. The latter predictor, however, has been omitted from the model by the Lasso penalty. The plot at the bottom right

TABLE 7
Four model options for United States data, with/without Elastic Net regularization and/or Optimal Scaling. $\lambda_1 =$ *Lasso penalty*, $\lambda_2 =$ *Ridge penalty*

Transformation	Regularization	r^2	$\widehat{\text{EPE}}$ (s.e.) optimal	λ_1	λ_2	$\widehat{\text{EPE}}$ (s.e.) selected	λ_1	λ_2	df	DLD
1. No	No	0.876	0.191 (0.036)	–	–	0.191 (0.036)	–	–	7	4.121
2. No	Yes	0.825	0.191 (0.036)	0.00	0.00	0.216 (0.044)	0.90	9.00	5	4.121
3. Yes[1]	No	0.933	0.210 (0.045)	–	–	0.210 (0.045)	–	–	18	2.603
4. Yes[1]	Yes	0.865	0.159 (0.031)	0.10	0.00	0.176 (0.037)	0.80	3.00	14	3.965

[1]POPUL, INCOME and FREEZE transformed with cubic nonmonotonic spline, one interior knot, ILLIT transformed with quadratic monotonic spline, one interior knot, FAIL, LIFE, HOMIC and SCHOOL with numeric transformation.

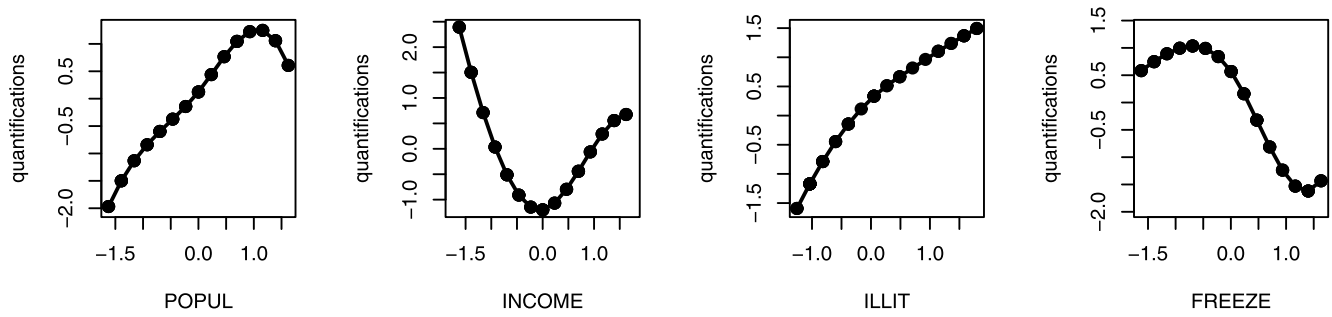

FIG. 10. *United States Data. Optimal scaling transformations for four predictors from ROS-ENET regression.*

shows $\hat{\mathbf{y}} = \sum_{k=1}^{P} \beta_k \varphi_k(\mathbf{x}_k)$ on the vertical axis versus the transformed outcome $\vartheta(\mathbf{y})$ on the horizontal axis.

Figure 12 shows all the paths for Ridge penalties ranging from 10 (at the top) to 0.0 (at the bottom), with a stepsize of 1.0. The horizontal axis represents the value of the Lasso penalty, ranging from 0.0 to 1.7, and the vertical axis gives the estimates of the prediction error, obtained with the 0.632 bootstrap ($\widehat{\text{EPE}}$). As was mentioned in Section 4.3, the different curves for the Lasso penalty for increasing values of the Ridge penalty are quite regular. The figure shows that even for very large Ridge penalties, the different paths cross each other for Lasso penalties around 0.8.

In Figure 13, we focus on the paths for Ridge penalties ranging from 0.0 to 3.0. The curve on the bottom gives the $\widehat{\text{EPE}}$ for the Lasso penalty 0.0, and shows that the smallest value for $\widehat{\text{EPE}}$ is obtained for the Lasso penalty 0.10. From this point, the curve is monotoni-

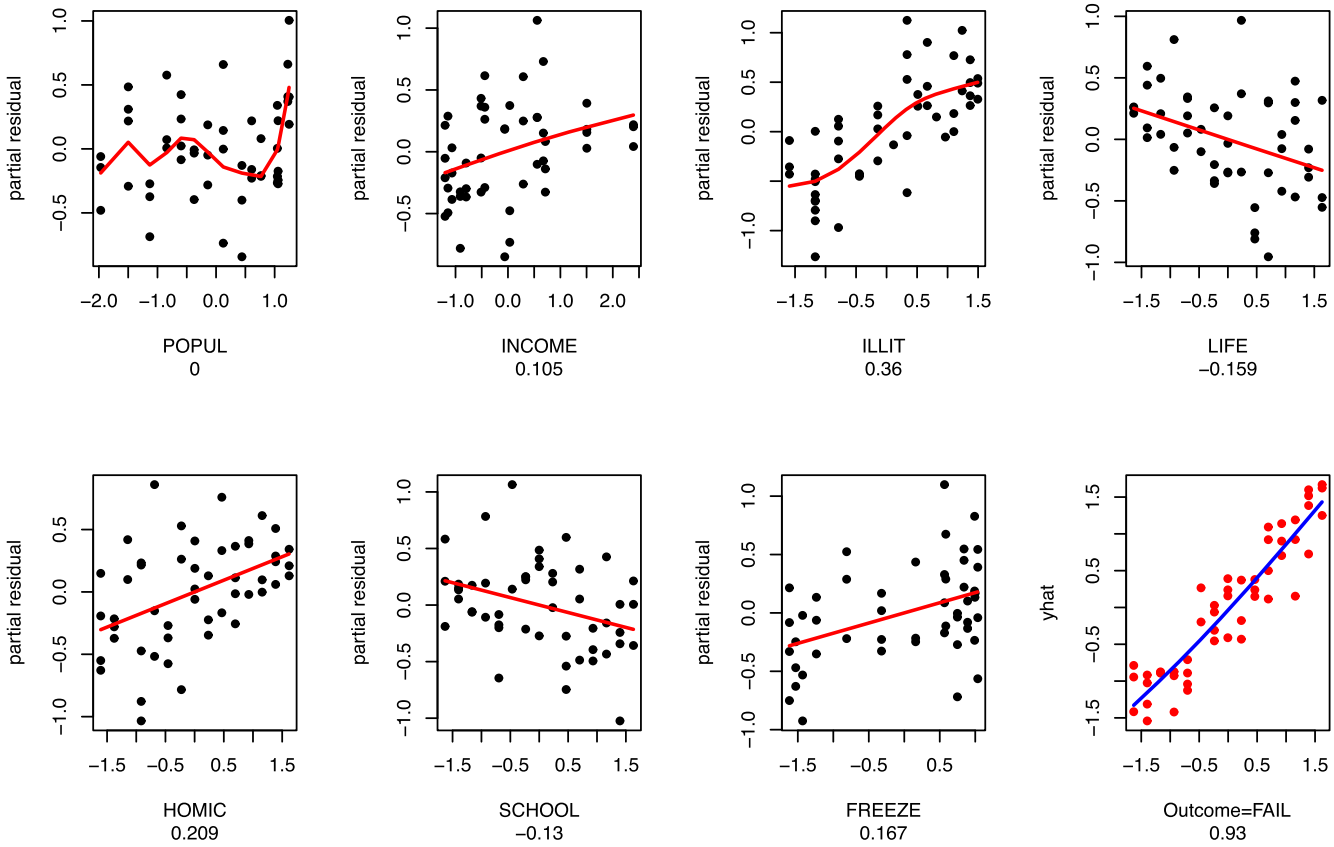

FIG. 11. *Partial residuals for each predictor obtained in Regularized Optimal Scaling regression (ROS-ENET), and the residuals of the model (linear combination of transformed predictors) versus outcome variable FAIL). Curves, obtained by fitting smoothing splines with four knots, show remaining nonlinearities, which are neglectable, except for POPUL.*

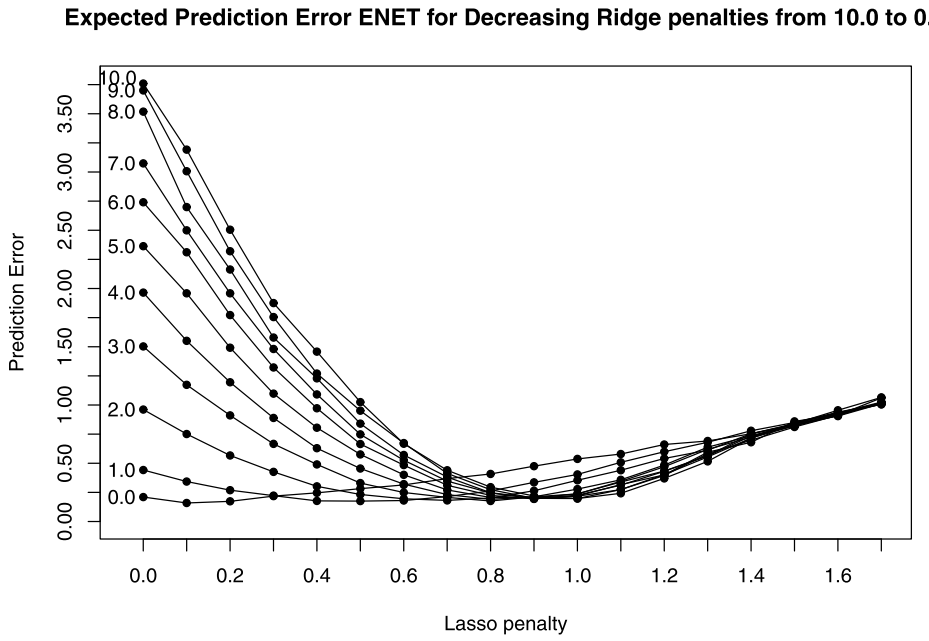

FIG. 12. *United States Data. Estimates of the Expected Prediction Error obtained by Elastic Net Regularized Optimal Scaling Regression. Paths from top to bottom represent decreasing values for the Ridge parameter from* 10.0–0.0. *The graph shows that values for* $\widehat{\text{EPE}}$ *are very similar for Lasso penalties around the value* 0.8, *no matter the value of the Ridge penalty.*

cally increasing. The picture for the three other curves (Lasso penalties from 1.0 up to 3.0) show a different picture. The different paths cross each other close to the value 0.80 for the Lasso penalty. The two large dots indicate the smallest overall value, which is 0.159, and the smallest value within one standard error (0.031), which is 0.176. The latter is on the curve for the Ridge penalty 3.0.

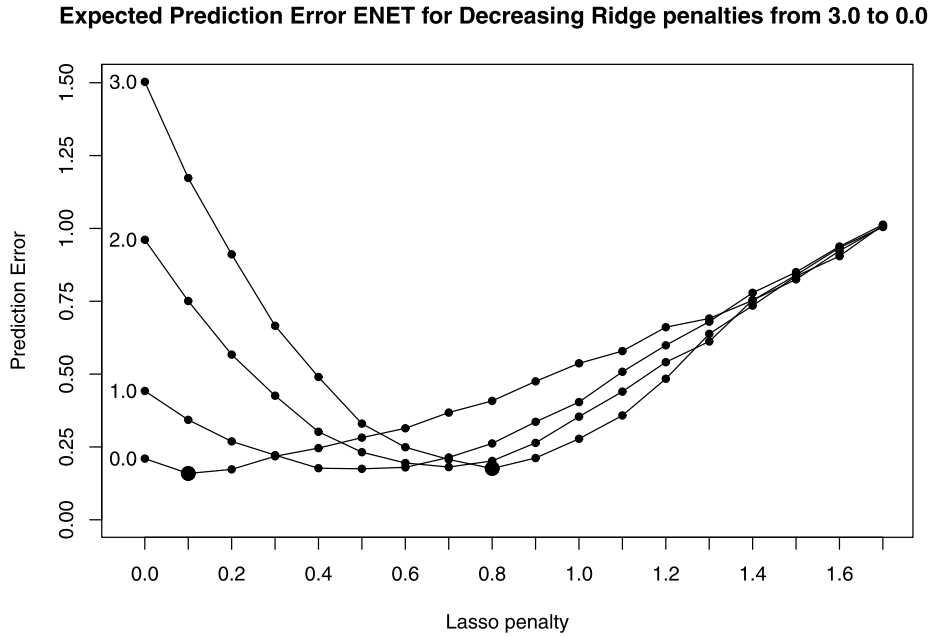

FIG. 13. *United States Data. Expected Prediction Error obtained for Elastic Net Regularized Optimal Scaling Regression. Paths from top to bottom represent decreasing values for the Ridge parameter from* 3.0 *to* 0.0. *The optimal value* (*with* $\lambda_1 = 0.1$, $\lambda_2 = 0.0$) *and the selected value, chosen within one standard error from the optimal value* (*with* $\lambda_1 = 0.8$, $\lambda_2 = 3.0$) *are indicated by large dots.*

5.3 High-Dimensional Metabolomics Data: The Leiden ApoE3 Data

The data were provided by the Systems Biology group at Leiden University (Thomas Hankemeier and colleagues), and concern 1550 LC-MS (Liquid Chromatography Mass Spectrometry) measurements of plasma lipids. LC-MS is an exceedingly sensitive and specific analytical technique that can precisely determine the identities and concentration of compounds in plasma. The biochemical background is as follows. ApoE3 stands for Apolipoprotein E3, which makes up cholesterol particles, such as LDL, VLDL, HDL. A strongly increased lipoprotein level in plasma results in artherosclerosis, and if blocking a blood vessel, might lead to stroke or heart attack. The objects are two samples of 10 mice: one of an (untreated) wildtype, and another of transgenic mice that contain the Human Leiden ApoE3 variety. The main question is whether differences in metabolomic profiles can be detected, and which predictors are important distinguishing the wildtype from the transgenic mice. The latter were not on a high, but on a low-fat diet, and the LC-MS data were collected after nine weeks, while artherosclerosis usually becomes manifest after 20 weeks. For each mouse, we have two measurement vectors available, resulting in a data matrix with 38 rows and 1550 columns (one transgenic mouse died during the experiment). ROS regression was performed with four different options: two forms of regularization (the Lasso and the Elastic Net), combined with two types of transformation (linear and quadratic spline). The outcome variable has four categories (two types of mice, two replications of the measurements) and we assume the categories to be ordered (an ordinal optimal scaling level was applied. The results for the 2×2 analysis plan are given in Table 8. The table shows that optimal scaling regression (with monotonic quadratic spline transformations) outperforms linear regression: the cross-validated prediction error $\widehat{\text{EPE}}$ is diminished by a factor 2. This is true for both the Lasso and the Elastic Net.

We first show the full paths of the estimates of the expected prediction error ($\widehat{\text{EPE}}$, determined by 13-fold cross-validation since we have only 38 objects), for 12 different values for the Lasso parameter (λ_1) on the horizontal axis and 10 different values for the Ridge parameter (λ_2), from 0.1 to 1.0, in the Elastic Net (Figure 14). Note that we do not give results for $\lambda_1 = 0.1$, since this value for the Lasso parameter is too small to give an admissible solution. The optimal model is found for spline transformations, e-net regularization, $\lambda_1 = 0.3$ and $\lambda_2 = 0.2$. The paths for different values of the Ridge penalty seem to cross at the point where the Lasso penalty (horizontal axis) is around 0.7. One standard error from the $\widehat{\text{EPE}}$ for the optimal model gives the selected model (spline transformations, e-net regularization, with $\lambda_1 = 0.7$ and $\lambda_2 = 0.4$). In Table 8, we see that also the Lasso with quadratic spline transformations does very well, with only 8 out of 1550 predictors, from the middle of the LC-MS spectrum. The Elastic Net uses the same predictors, but adds predictors (with smaller coefficients) that are adjacent on the spectrum (and thus are correlated), and uses a total of 26 predictors. Choosing the Elastic Net model might give more stable results for future data: by using correlated variables, a weighted average would be applied, and this might diminish the effect of uncorrelated measurement error compared to the error introduced by the use of a single variable.

Figure 15 shows the five different ordinal quantifications for the outcome variable for the models given in Table 8. The quantifications are very similar for the five different options, especially when the same transformations (spline or linear transformations) have been applied. The black squares show the quantifications for outcome in the chosen model (spline transformations, e-net regularization, with $\lambda_1 = 0.7$ and $\lambda_2 = 0.4$).

TABLE 8
Prediction error for different sets of transformations

Transformation	Regularization	$\widehat{\text{EPE}}$ (optimal)	λ_1	λ_2	# pred's
1. Linear$^{(1,0)}$	lasso	0.122 (0.028)	0.20	0.00	12
2. Linear$^{(1,0)}$	e-net	0.117 (0.030)	0.30	0.10	16
3. Spline$^{(2,0)}$	lasso	0.059 (0.016)	0.20	0.00	8
4. Spline$^{(2,0)}$	e-net	0.054 (0.020)	0.30	0.20	26
		$\widehat{\text{EPE}}$ (selected)			
5. Spline$^{(2,0)}$	e-net	0.069 (0.021)	0.70	0.40	26

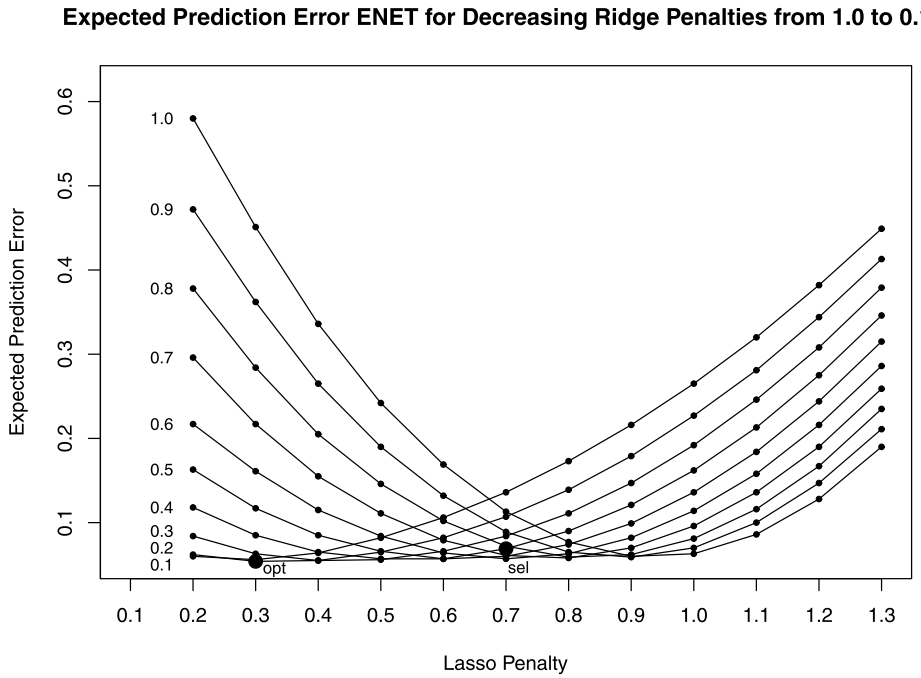

FIG. 14. *ApoE3 Data. Expected Prediction Error obtained by Elastic Net Regularized Optimal Scaling Regression. Paths from top to bottom represent decreasing values for the Ridge parameter from* 1.0–0.1. *The optimal value* (*opt*, $\lambda_1 = 0.3$, $\lambda_2 = 0.2$) *and the selected value, chosen within one standard error from the optimal* (*sel*, $\lambda_1 = 0.7$, $\lambda_2 = 0.4$) *are indicated by large dots. The graph shows that the values for* $\widehat{\text{EPE}}$ *around* $\lambda_1 = 0.7$ *are about the same no matter the value of the Ridge penalty.*

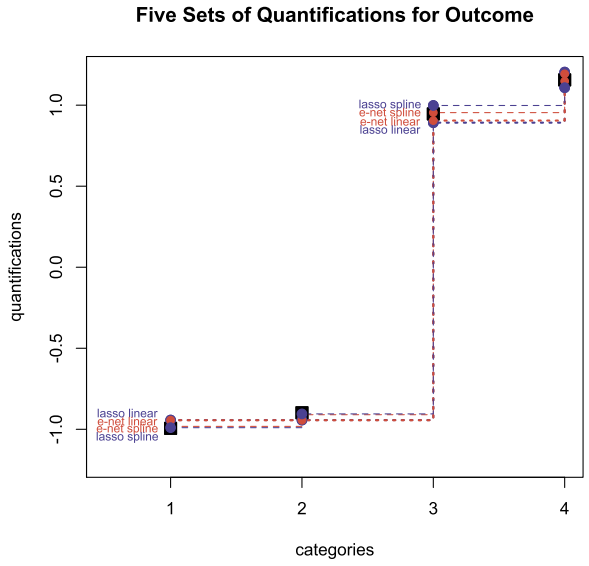

FIG. 15. *Five monotonic step functions for the categorical outcome variable in the ApoE3 Data. Dashed lines show stepfunctions for the outcome when the predictors are transformed by spline functions, dotted lines when the latter have linear transformations. Red lines refer to quantifications found by the Elastic Net, and blue lines are shown for the Lasso. The large squares* (*no line drawn*) *show the quantifications of outcome for the chosen model* ($\widehat{\text{EPE}}$ *one standard error from the optimal model*: *elastic net with splines*, $\lambda_1 = 0.7$, $\lambda_2 = 0.4$).

6. DISCUSSION

When confronted with categorical variables, nonlinear relationships or multicollinearity (for instance when $P > N$), the widely used linear regression framework requires adjustments. In an attempt to overcome these issues, most methods involve transformations of predictors and/or regularization. In this paper, we show how optimal scaling regression can be integrated with popular regularization methods (Ridge Regression, Lasso and Elastic Net) in a very general algorithm that can deal with both continuous and categorical variables. Categorical variables may have either ordered (ordinal) or unordered (nominal) values. Transformation of continuous variables is called for when nonlinear relationships exist between predictor variables and the outcome. optimal scaling linearizes these relationships, as can be seen from the partial residual plots. Furthermore, OS allows for transformations of both the outcome as well as the predictors. When compared with alternative methods, regularized optimal scaling has the power to generalize many existing procedures such as Group Lasso, Blockwise Sparse Regression, GAM and ACE, thanks to its large flexibility in choices of transformations and penalty functions. Combined with an efficient one-variable-at-a-time (coordinate descent) algorithm, it is able to handle

large data sets of mixed nature encountered in modern-day applied statistics. The option that only uses numeric optimal scaling levels awaits comparison with the algorithms proposed in Friedman, Hastie and Tibshirani (2010, 2012), and Mazumder, Friedman and Hastie (2011), while splines have also been applied in Chouldechova and Hastie (2015).

If the predictor correlation matrix is ill-conditioned, a good property of optimal scaling is that it improves upon this condition, as measured by the value of the smallest eigenvalue. We also proposed the Divergence of Log Determinants to quantify the conditionality of the predictor correlation matrix in a single diagnostic. As for the predictors, optimal scaling tends to increase their conditional independence (on average), as measured by so-called tolerance values (described in Section 3.1). In some cases, large Ridge and/or Lasso penalties may be required to prevent overfitting when allowing for optimal transformations.

In the context of a regularized analysis, there are two goals: model selection and assessment of the selected model. To achieve these goals, the best approach is a *three-way* data split, dividing the full data into an active data set and a validation set, where the active data set itself is divided into a training set and a test set (as was shown in Figure 5). In that case, the generalization error can be estimated by applying the values obtained for the parameters in the active data set to the validation set. This was done in the analysis of the Cervix Cancer Data in Section 3.4. When there are not enough data for a three-way split, the validation phase is omitted and the estimate of the generalization error is approximated by the estimate of the expected prediction error instead. Of course, the latter will be too optimistic since we use the cross-validation phase as well to select the optimal values for the regularization parameters. (Note that in this phase, we only select a model on the basis of the values of the regularization parameters, and not on the basis of the transformations.) If we are mainly interested in comparing the prediction accuracy for different choices such as with/without optimal scaling, and/or regularization, we may assume that the optimism will not differ very much between models, thus not affecting the conclusions too much.

The coordinate descent approach has a very exciting history (e.g., see Tibshirani, 2011). It was already shown in Van der Kooij (2007) that the alternating least squares (ALS) approach that is applied in OS to find the optimal transformations and regression weights (one-variable-at-a-time), automatically leads to very simple and efficient estimates for regularized regression coefficients in the Lasso and the Elastic Net. We may conclude that the ALS framework that was kept alive all these years in optimal scaling, gave rise to renewed interest and exciting research using coordinate descent optimization.

NOTE

The algorithm described in this paper has been implemented in a user-friendly procedure called CATREG that has been developed by the first two authors in the CATEGORIES module of IBM/SPSS Statistics (Meulman, Heiser and SPSS, 2010). This procedure contains all the different features of ROS regression that are mentioned in this paper. A somewhat more limited version in R (R Core Team, 2017) is also available, and can be obtained upon request from the second author.

APPENDIX A: COPULA REGRESSION

The use of copula methodology to describe dependence between variables is popular in for example finance and economics (Kolev and Paiva, 2009). The merit of Sklar's theorem reveals itself in practical scenarios where it may be more intuitive to state marginal CDFs and a copula $\mathcal{C}$ than to specify the joint distribution F directly (Parsa and Klugman, 2011; Trivedi and Zimmer, 2005). In its original form (Sklar, 1959), the function $\mathcal{C} : [0,1]^{P+1} \mapsto [0,1]$ maps the marginal CDFs of an outcome $\mathbf{y}$ and P predictors $\mathbf{x}_1, \ldots, \mathbf{x}_P$ to a joint distribution function F as follows:

$$F(\mathbf{y}, \mathbf{x}_1, \ldots, \mathbf{x}_P) = \mathcal{C}\big(F_0(\mathbf{y}), F_1(\mathbf{x}_1), \ldots, F_P(\mathbf{x}_P)\big). \tag{A.1}$$

In particular, following Cai and Zhang (2018), we say random data pairs $(y_i, X_i) \in \mathbb{R}^{P+1}$ satisfy a Gaussian copula regression (Song, 2000) if there exist strictly increasing functions $f_0, \ldots, f_P$ such that $(f_0(y_i), f_1(x_{i1}), \ldots, f_P(x_{iP}))$ is i.i.d. multivariate normally distributed. For $i = 1, \ldots, N$ and $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \boldsymbol{\Sigma})$, it follows that

$$f_0(y_i) = \sum_{k=1}^{P} \beta_k f_k(x_{ik}) + \epsilon_i. \tag{A.2}$$

At first glance, it seems model (A.2) bears a close resemblance to optimal scaling with transformations $\theta = f_0$ and $\phi_k = f_k$. However, we are currently unaware of a copula regression model that allows for high-dimensional continuous and categorical data with

adaptivity to both monotone and nonmonotone transformations in the way ROS regression does. Four essential distinctions may be pointed out.

Transformation and Computation

The optimal scaling transformations are conceptually different from (strictly increasing) marginal distribution functions, since they are (alternatingly) optimized over partial residuals. In addition to the (Gaussian) copula dependence parametrization (Σ), the marginals f in (A.2) must be specified to obtain estimates of the (nonlinear) regression parameters $\boldsymbol{\beta}$. Since the f's are unknown, they must either be estimated based on a (strictly monotone) functional form, a kernel or their empirical CDFs, in which case the problem becomes semiparametric. Pitt, Chan and Kohn (2006) propose Bayesian MCMC based estimation with dependent error structure $\boldsymbol{\Sigma}$.

Categorical Data

Secondly, it is generally not easy to deal with categorical predictors in the copula estimation framework (Parsa and Klugman, 2011; Genest and Nešlehová, 2007). Masarotto and Varin (2012) extend the Gaussian copula to nonnormal outcomes (continuous, discrete or categorical), requiring numerical integration. Hoff (2007) suggests a Bayesian method that allows for predictors of mixed type, but focuses on association analysis and only treats the low-dimensional setting.

Distributional Assumptions

Copula regression is usually likelihood driven, while ROS regression is least squares based. An advantage of the normality assumption on $\boldsymbol{\epsilon}$ exploited by Cai and Zhang (2018) is that it enables theory towards parametric statistical inference of debiased, L_1 regularized $\boldsymbol{\beta}$ parameters (van de Geer et al., 2014). However, such benefits come with strong assumptions on X, sparsity of $\boldsymbol{\beta}$ and the size of the penalty parameter. Moreover, these mathematical results are specific to the L_1 case; inference is lost for the Elastic Net generalization.

Dimensionality

For the model in (A.2), Cai and Zhang (2018) treat the high-dimensional case $P \gg N$ only for real-valued variables $(Y_i, X_i) \in \mathbb{R}^{P+1}$, while ROS regression allows for categorical outcomes and/or predictors. Furthermore, a recent contribution to (low-dimensional) Gaussian copula regression by Noh, El Ghouch and Bouezmarni (2013) led to a critical review paper by Dette, Van Hecke and Volgushev (2014). Not only is copula-based regression sensitive to misspecification of the dependence structure; it can also perform poorly when the regression function itself is not monotone. Performance was claimed to deteriorate further when the number of predictors grows.

APPENDIX B

OS Algorithm

The OS algorithm consists of the following steps.

1. Initialize ϑ, φ and $\boldsymbol{\beta}$ (standardized $\mathbf{y}$ and X, and OLS coefficients, or random values).
2. If scaling level outcome not linear update transformation:
   (a) $\tilde{\mathbf{v}}_{\mathrm{y}} = \mathbf{D}_{\mathrm{y}}^{-1}\mathbf{G}_{\mathrm{y}}' \sum_k \beta_k \varphi_k$. If scaling level not nominal restrict $\tilde{\mathbf{v}}_{\mathrm{y}}$.
   (b) $\vartheta(\mathbf{y}) =$ standardized $\mathbf{G}_{\mathrm{y}}\tilde{\mathbf{v}}_{\mathrm{y}}$.
3. For $k = 1, \ldots, P$:
   (a) $\mathbf{u}_k = \vartheta(\mathbf{y}) - \sum_{l \neq k} \beta_l \varphi_l(\mathbf{x}_l)$.
   (b) Update coefficient: $\beta_k = N^{-1}\mathbf{u}_k' \varphi(\mathbf{x}_k)$.
   (c) If scaling level predictor k not linear update transformation: $\tilde{\mathbf{v}}_k = \beta_k^{-1}\mathbf{D}_k^{-1}\mathbf{G}_k'\mathbf{u}_k$. If scaling level not nominal restrict $\tilde{\mathbf{v}}_k$.
   (d) $\varphi(\mathbf{x}_k) =$ standardized $\mathbf{G}_k\tilde{\mathbf{v}}_k$.
4. Compute loss and check convergence.
   If not converged return to step 2.

Regularization is incorporated by applying equation (4.12), (4.13), or (4.14) to β_k after step 3(d).

OS Transformations

For scaling levels other than nominal the cone $\mathbb{C}_k$ that contains all admissible transformations of X_k is defined by

$$\mathbb{C}_k(\mathbf{x}_k) \equiv \{\varphi_k(\mathbf{x}_k) | \varphi_k(\mathbf{x}_k) = \mathrm{trans}(\mathbf{x}_k)\},$$

and the metric projection is written as

$$P_{\mathbb{C}_k(\mathbf{x}_k)} \equiv \min_{\mathrm{trans}(\mathbf{x}_k)} \|\mathbf{u}_k - \beta_k \,\mathrm{trans}(\mathbf{x}_k)\|^2,$$

with $\mathbf{u}_k$ as defined in equation (2.9) and $\mathrm{trans}(\mathbf{x}_k)$ stands for $\mathrm{mon}(\mathbf{x}_k)$, denoting a least squares monotonic transformation of X_k, or $\mathrm{spl}(\mathbf{x}_k)$, denoting a spline transformation, or $\mathrm{lin}(\mathbf{x}_k)$, denoting a linear transformation, amounting to a standardized version of $\mathbf{x}_k$.

- In the case of ordinal transformation, the metric projection amounts to applying *monotonic (isotonic) regression* of $\mathrm{sign}(\beta_k^{-1})\mathbf{u}_k$ onto $\mathbf{x}_k$, written as $\mathrm{mon}(\mathrm{sign}(\beta_k^{-1})\mathbf{u}_k, \mathbf{x}_k)$, and standardizing the result. The monotonic regression can either be increasing or

decreasing, whichever gives the smaller loss value; if applicable, the sign of β_k has to be adjusted. The restricted quantification is obtained by applying the up-and-down-blocks minimum violators algorithm (Kruskal, 1964; Barlow et al., 1972) to $\mathbf{v}_k^{\text{nom}}$.

- In the case of spline transformation, the metric projection amounts to a smooth transformation of the predictor X_k using splines. One possibility is to construct an I-spline basis matrix $S_k(\mathbf{x}_k)$ (see Ramsay, 1988 for details), and having $\mathbf{S}_k = S_k(\mathbf{x}_k)$, we minimize

$$L(\mathbf{b}_k) = \|\mathbf{u}_k - \beta_k \mathbf{S}_k \mathbf{b}_k\|^2, \tag{B.1}$$

over $\mathbf{b}_k = \{b_t^k\}_{t=1}^{T_k}$, the T_k-vector with spline coefficients that have to be estimated, and where T_k 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_k, we can compute the analytical solution for $\mathbf{b}_k$ directly, since (B.1) is a straightforward regression problem, with the columns of $\mathbf{S}_k = \mathbf{s}_{t=1}^{T_k}$ as independent variables. If, however, the I-spline transformation is required to be monotonic with $\mathbf{x}_k$, we have to minimize (B.1) under the restriction that the vector $\mathbf{b}_k$ with spline coefficients contains only nonnegative elements. This constrained optimization problem can be solved by applying the one-variable-at-a-time strategy here as well. Thus, the problem is further partitioned by isolating the tth column of the spline basis matrix $\mathbf{S}_k$ (denoted by $\mathbf{s}_t^k$) and the tth element (b_t^k) of the spline coefficient vector $\mathbf{b}_k$ from the remaining elements $\{b_r^k\}_{r\neq t}$. Next, we minimize iteratively

$$L(b_t^k) = \left\| \left(\mathbf{u}_k - \beta_k \sum_{r\neq t} b_r^k \mathbf{s}_r^k \right) - \beta_k b_t^k \mathbf{s}_t^k \right\|^2 \tag{B.2}$$

over $b_t^k \geq 0$, for $t = 1, \ldots, T_k$. (There is a complication if we take the normalization condition $\mathbf{b}_k' \mathbf{S}_k' \mathbf{S}_k \mathbf{b}_k = N$ into account that ensures that the transformed variable is standardized; how this problem is solved can be found in Groenen, van Os and Meulman, 2000.)

ACKNOWLEDGMENTS

The authors would like to thank Bradley Efron, Jerome Friedman and Willem Heiser for their suggestions and comments, and Brad Efron for his encouragement to submit the paper to *Statistical Science*. Also, we would like to thank the Editor for his patience, and the Associate Editor and the anonymous reviewers for their elaborate and very useful suggestions and comments on the original and revised manuscript. Subsequent versions of this paper were written while the first author was in the Stanford Statistics Department.

REFERENCES

ANGOFF, C. and MENCKEN, H. L. (1931). The worst American state. *American Mercury* **24** 1–16. 175–188, 355–31.

BARLOW, R. E., BARTHOLOMEW, D. J., BREMNER, J. M. and BRUNK, H. D. (1972). *Statistical Inference Under Order Restrictions*: *The Theory and Application of Isotonic Regression*. Wiley, New York.

BOCK, R. D. (1960). *Methods and Applications of Optimal Scaling*. *Report* 25. L. L. Thurstone Lab, Univ. North Carolina, Chapel Hill.

BREIMAN, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* **37** 373–384. MR1365720

BREIMAN, L. and FRIEDMAN, J. H. (1985). Estimating optimal transformations for multiple regression and correlation (with discussion). *J. Amer. Statist. Assoc.* **80** 580–619. MR0803258

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees*. *Wadsworth Statistics/Probability Series*. Wadsworth, Belmont, CA. MR0726392

BUJA, A. (1990). Remarks on functional canonical variates, alternating least squares methods and ACE. *Ann. Statist.* **18** 1032–1069. MR1062698

BUJA, A., HASTIE, T. and TIBSHIRANI, R. (1989). Linear smoothers and additive models. *Ann. Statist.* **17** 453–555. MR0994249

CAI, T. T. and ZHANG, L. (2018). High-dimensional Gaussian copula regression: Adaptive estimation and statistical inference. *Statist. Sinica* **28** 963–993. MR3791096

CÉA, J. and GLOWINSKI, R. (1973). Sur des méthodes d'optimisation par relaxation. *Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge* **7** 5–31. MR0367765

CHEN, S. S., DONOHO, D. L. and SAUNDERS, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.* **20** 33–61. MR1639094

CHOULDECHOVA, A. and HASTIE, T. J. (2015). Generalized Additive Model Selection. Available at arXiv:1506.03850 [stat.ML].

DAUBECHIES, I., DEFRISE, M. and DE MOL, C. (2004). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure Appl. Math.* **57** 1413–1457. MR2077704

DE LEEUW, J., YOUNG, F. W. and TAKANE, Y. (1976). Additive structure in qualitative data. *Psychometrika* **41** 471–503.

DETTE, H., VAN HECKE, R. and VOLGUSHEV, S. (2014). Some comments on copula-based regression. *J. Amer. Statist. Assoc.* **109** 1319–1324. MR3265699

DHILLON, I. S. (2008). The log-determinant divergence and its applications. Paper presented at the Householder Symposium XVII, Zeuthen, Germany.

EFRON, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *J. Amer. Statist. Assoc.* **78** 316–331. MR0711106

EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. MR2060166

FRANK, I. E. and FRIEDMAN, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35** 109–148.

FRIEDMAN, J. H. (1991). Multivariate adaptive regression splines (with discussion). *Ann. Statist.* **19** 1–141. MR1091842

FRIEDMAN, J. HASTIE, T. J. and TIBSHIRANI, R. J. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **22** 1548–7660.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2012). glmnet: Lasso and elastic-net regularized generalized linear models. Available at http://CRAN.R-project.org/package=glmnet. R package version 1.9-5.

FRIEDMAN, J. H. and MEULMAN, J. J. (2003). Prediction with multiple additive regression trees with application in epidemiology. *Stat. Med.* **22** 1365–1381.

FRIEDMAN, J. H. and POPESCU, B. E. (2004). Gradient directed regularization for linear regression and classification Technical report, Dept. Statistics, Stanford Univ., Stanford, CA.

FRIEDMAN, J. H. and STUETZLE, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76** 817–823. MR0650892

FRIEDMAN, J., HASTIE, T., HÖFLING, H. and TIBSHIRANI, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Stat.* **1** 302–332. MR2415737

FU, W. J. (1998). Penalized regressions: The Bridge versus the Lasso. *J. Comput. Graph. Statist.* **7** 397–416. MR1646710

GENEST, C. and NEŠLEHOVÁ, J. (2007). A primer on copulas for count data. *Astin Bull.* **37** 475–515. MR2422797

GIFI, A. (1981). Nonlinear multivariate analysis. Unpublished Manuscript. Department of Data Theory, Leiden University, Leiden.

GIFI, A. (1990). *Nonlinear Multivariate Analysis*. *Wiley Series in Probability and Mathematical Statistics*: *Probability and Mathematical Statistics*. Wiley, Chichester. MR1076188

GOLUB, G. H., HEATH, M. and WAHBA, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21** 215–223. MR0533250

GROENEN, P. J. F., VAN OS, B.-J. and MEULMAN, J. J. (2000). Optimal scaling by alternating length-constrained nonnegative least squares, with application to distance-based analysis. *Psychometrika* **65** 511–524. MR1849280

GURIN, L. G., POLJAK, B. T. and RAĬK, È. V. (1967). Projection methods for finding a common point of convex sets. *Zh. Vychisl. Mat. Mat. Fiz.* **7** 1211–1228. MR0232225

HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models*. *Monographs on Statistics and Applied Probability* **43**. CRC Press, London. MR1082147

HASTIE, T., TIBSHIRANI, R. and BUJA, A. (1994). Flexible discriminant analysis by optimal scoring. *J. Amer. Statist. Assoc.* **89** 1255–1270. MR1310220

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2722294

HAYASHI, C. (1952). On the prediction of phenomena from qualitive data and the quantification of qualitative data from the mathematico-statistical point of view. *Ann. Inst. Statist. Math.* **1952** 93–96.

HOERL, A. E. and KENNARD, R. (1970a). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67.

HOERL, A. E. and KENNARD, R. (1970b). Ridge regression: Applications to nonorthogonal problems. *Technometrics* **12** 69–82.

HOFF, P. D. (2007). Extending the rank likelihood for semiparametric copula estimation. *Ann. Appl. Stat.* **1** 265–283. MR2393851

IBM CORP. (2010). *IBM SPSS Statistics* 19.0 *Algorithms*. IBM Corp., Amonk, NY.

JAMES, W. and STEIN, C. (1961). Estimation with quadratic loss. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I* 361–379. Univ. California Press, Berkeley, CA. MR0133191

KIM, Y., KIM, J. and KIM, Y. (2006). Blockwise sparse regression. *Statist. Sinica* **16** 375–390. MR2267240

KOLEV, N. and PAIVA, D. (2009). Copula-based regression models: A survey. *J. Statist. Plann. Inference* **139** 3847–3856. MR2553771

KRUSKAL, J. B. (1964). Nonmetric multidimensional scaling: A numerical method. *Psychometrika* **29** 115–129. MR169713

KRUSKAL, J. B. (1965). Analysis of factorial experiments by estimating monotone transformations of the data. *J. Roy. Statist. Soc. Ser. B* **27** 251–263. MR0195212

MASAROTTO, G. and VARIN, C. (2012). Gaussian copula marginal regression. *Electron. J. Stat.* **6** 1517–1549. MR2988457

MAZUMDER, R., FRIEDMAN, J. H. and HASTIE, T. (2011). SparseNet: Coordinate descent with nonconvex penalties. *J. Amer. Statist. Assoc.* **106** 1125–1138. MR2894769

MEULMAN, J. J. (1986). *A Distance Approach to Nonlinear Multivariate Analysis*. DSWO Press, Leiden.

MEULMAN, J. J., HEISER, W. J. and SPSS (1998). SPSS Categories 8.0, Chicago, IL SPSS Inc.

MEULMAN, J. J., HEISER, W. J. and SPSS INC. (2010). *IBM SPSS Categories* 19. IBM Corp., Amonk, NY.

MEULMAN, J. J., ZEPPA, P., BOON, M. E. and RIETVELD, W. J. (1992). Prediction of various grades of cervical neoplasia on plastic-embedded cytobrush samples. Discriminant analysis with qualitative and quantitative predictors. *Anal. Quant. Cytol. Histol.* **14** 60–72.

NELDER, J. A. and WEDDERBURN, R. W. M. (1972). Generalized linear models. *J. Roy. Statist. Soc. Ser. A* **135** 370–384.

NISHISATO, S. (1980). *Analysis of Categorical Data*: *Dual Scaling and Its Applications*. *Mathematical Expositions* **24**. Univ. Toronto Press, Toronto. MR0600656

NISHISATO, S. (1994). *Elements of Dual Scaling*: *An Introduction to Practical Data Analysis*. Lawrence Erlbaum, Hillsdale, NJ.

NOH, H., EL GHOUCH, A. and BOUEZMARNI, T. (2013). Copula-based regression estimation and inference. *J. Amer. Statist. Assoc.* **108** 676–688. MR3174651

OBERHOFER, W. and KMENTA, J. (1974). A general procedure for obtaining maximum likelihood estimates in generalized regression models. *Econometrica* **42** 579–590. MR0440805

OSBORNE, M. R., PRESNELL, B. and TURLACH, B. A. (2000a). A new approach to variable selection in least squares problems. *IMA J. Numer. Anal.* **20** 389–403. MR1773265

OSBORNE, M. R., PRESNELL, B. and TURLACH, B. A. (2000b). On the LASSO and its dual. *J. Comput. Graph. Statist.* **9** 319–337. MR1822089

PARSA, R. A. and KLUGMAN, S. A. (2011). Copula regression. *Proc. Casualty Actuar. Soc.* **5** 45–54.

PERKINS, S., LACKER, K. and THEILER, J. (2003). Grafting: Fast, incremental feature selection by gradient descent in function space. *J. Mach. Learn. Res.* **3** 1333–1356. MR2020763

PITT, M., CHAN, D. and KOHN, R. (2006). Efficient Bayesian inference for Gaussian copula regression models. *Biometrika* **93** 537–554. MR2261441

R CORE TEAM (2017). R Foundation for Statistical Computing, Vienna, Austria. Available at http://www.R-project.org/.

RAMSAY, J. O. (1988). Monotone regression splines in action (with discussion). *Statist. Sci.* **4** 425–441.

SAS/STAT (1990). *User's Guide*, *Version* 6 **2**. SAS Institute Inc., Cary NC.

SKLAR, M. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Stat. Univ. Paris* **8** 229–231. MR0125600

SONG, P. X.-K. (2000). Multivariate dispersion models generated from Gaussian copula. *Scand. J. Stat.* **27** 305–320. MR1777506

TIBSHIRANI, R. (1988). Estimating transformations for regression via additivity and variance stabilization. *J. Amer. Statist. Assoc.* **83** 394–405. MR0971365

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

TIBSHIRANI, R. (2011). Regression shrinkage and selection via the lasso: A retrospective. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 273–282. MR2815776

TIKHONOV, A. N. (1943). On the stability of inverse problems. *C. R. (Dokl.) Acad. Sci. URSS* **39** 176–179. MR0009685

TRIVEDI, P. K. and ZIMMER, D. M. (2005). Copula modeling: An introduction for practitioners. *Found. Trends Econom.* **1** 1–111.

TSENG, P. (1988). Coordinate Ascent for Maximizing Nondifferentiable Concave Functions. Technical Report LIDS-P, 1840, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA.

TSENG, P. (2001). Convergence of a block method for nondifferentiable minimization. *J. Optim. Theory Appl.* **109** 475–494. MR1835069

VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. MR3224285

VAN DER KOOIJ, A. J. (2007). Prediction accuracy and stability of regression with optimal scaling transformations. Thesis, Leiden Univ. Available at https://openaccess.leidenuniv.nl/handle/1887/12096.

WAINER, H. and THISSEN, D. (1981). Graphical data analysis. *Annu. Rev. Psychol.* **32** 191–241.

WALBERG, H. J. and RASHER, S. P. (1977). The ways schooling makes a difference. *Phi Delta Kappan* **58** 703–707.

WINSBERG, S. and RAMSAY, J. O. (1980). Monotonic transformations to additivity using splines. *Biometrika* **67** 669–674. MR0601105

WU, T. T. and LANGE, K. (2008). Coordinate descent algorithms for lasso penalized regression. *Ann. Appl. Stat.* **2** 224–244. MR2415601

YOUNG, F. W. (1981). Quantitative analysis of qualitative data. *Psychometrika* **46** 357–388. MR0668307

YOUNG, F. W., DE LEEUW, J. and TAKANE, Y. (1976). Regression with qualitative and quantitative variables: An alternating least squares method with Optimal Scaling features. *Psychometrika* **41** 505–529.

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 49–67. MR2212574

ZANGWILL, W. I. (1969/70). Convergence conditions for nonlinear programming algorithms. *Manage. Sci.* **16** 1–13. MR0302199

ZHAO, P. and YU, B. (2007). Stagewise lasso. *J. Mach. Learn. Res.* **8** 2701–2726. MR2383572

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. MR2137327