



# A new algorithm and a discussion about visualization for logistic reduced rank regression

Mark de Rooij<sup>1</sup>

Received: 21 November 2022 / Accepted: 26 June 2023 / Published online: 12 July 2023  
© The Author(s) 2023

## Abstract

Logistic reduced rank regression is a useful data analysis tool when we have multiple binary response variables and a set of predictors. In this paper, we describe logistic reduced rank regression and present a new majorization minimization algorithm for the estimation of model parameters. Furthermore, we discuss Type I and Type D triplots for visualizing the results of a logistic reduced rank regression model, compare them, and then develop a hybrid triplot using elements of both types. Two empirical data sets are analyzed. This analysis is used to (1) compare the new algorithm to an existing one in terms of speed; and (2) to show the hybrid triplot and its interpretation.

**Keywords** Constrained PCA · MM algorithm · Generalized SVD · Visualization

## 1 Introduction

In many scientific disciplines, quantitative information is collected on a set of variables for a number of objects or participants. When the set of variables can be divided into a set of  $P$  predictors and a set of  $R$  responses, we need regression type of models. Researchers often neglect the multivariate nature of the response set and consequently univariate multiple regression models are fitted, one for each response variable separately. Such an approach does not take into account that the response variables might be correlated and does not provide insight into the relationships between the response variables. In this paper, the interest lies in the case where the  $R$  response variables are binary.

For binary response variables, the typical regression model is a logistic regression (Agresti 2013), that is, a generalized linear model (McCullagh and Nelder

---

Communicated by Alfonso Iodice D'Enza.

✉ Mark de Rooij  
rooijm@fsw.leidenuniv.nl

<sup>1</sup> Institute of Psychology, Department Methodology and Statistics, Leiden University, Leiden, The Netherlands

1989) with logit link function and a Bernoulli or binomial distribution. In logistic regression models, the probability that person  $i$  answers yes (or 1) on the response variable  $Y$  is defined as  $\pi_i = P(Y_i = 1)$ . These probabilities (or estimated values) are commonly defined in terms of the log-odds form (in GLMs the “linear predictor’ ’), denoted by  $\theta_i$ , that is,

$$\pi_i = \frac{\exp(\theta_i)}{1 + \exp(\theta_i)} = \frac{1}{1 + \exp(-\theta_i)},$$

and similarly

$$1 - \pi_i = \frac{1}{1 + \exp(\theta_i)} = \frac{\exp(-\theta_i)}{1 + \exp(-\theta_i)}.$$

Finally, the log-odds form is a (linear) function of the  $P$  predictor variables, that is

$$\theta_i = m + \sum_{p=1}^P x_{ip} a_p,$$

where  $x_{ip}$  is the observed value for person  $i$  on predictor variable  $p$  and  $m$  and  $a_p$  are the parameters which need to be estimated. The intercept  $m$  is the expected log-odds when all predictor variables equal zero, and the regression weights  $a_p$  indicate the difference in log-odds for two observations that differ one unit in predictor variable  $p$  and have equal values for all other predictor variables.

Having  $R$  outcome variables ( $Y_r$ ,  $r = 1, \dots, R$ ) a multivariate model can be defined with probabilities  $\pi_{ir} = P(Y_{ir} = 1)$  that are parameterized as

$$\pi_{ir} = \frac{\exp(\theta_{ir})}{1 + \exp(\theta_{ir})} = \frac{1}{1 + \exp(-\theta_{ir})},$$

and the log-odds form is

$$\theta_{ir} = m_r + \sum_{p=1}^P x_{ip} a_{pr}.$$

The intercepts can be collected in a vector  $\mathbf{m}$  and the regression weights can be collected in a matrix  $\mathbf{A}$  of size  $P \times R$ .

For multivariate outcomes, Yee and Hastie (2003) proposed reduced rank vector generalized linear models, that are multivariate models with a rank constraint on the matrix with regression weights, that is,

$$\mathbf{A} = \mathbf{B}\mathbf{V}'$$

where  $\mathbf{B}$  is a matrix of size  $P \times S$  and  $\mathbf{V}$  a matrix of size  $R \times S$ . The rank of the matrix  $\mathbf{A}$  is  $S$ , a number in the range 1 to  $\min(P, R)$ , and when  $S < \min(P, R)$  the rank is reduced, hence the name reduced rank regression. The matrix  $\mathbf{B}$  has elements  $b_{ps}$  for  $s = 1, \dots, S$  and  $\mathbf{V}$  has elements  $v_{rs}$ . The  $S$  elements for the predictor variable

$p$  (i.e., the  $p$ -th row of  $\mathbf{B}$ ), are collected in the column vector  $\mathbf{b}_p$ ; similarly, the  $S$  elements for response variable  $r$  are collected in the column vector  $\mathbf{v}_r$ .

At this point, it is instructive to take a step back to models for continuous outcomes. Reduced rank regression (Anderson 1951; Izenman 1975; Tso 1981; Davies and Tso 1982), also called redundancy analysis (Van den Wollenberg 1977), has been proposed as a multivariate tool for simultaneously predicting the responses from the set of predictors. Reduced rank regression can be motivated from a regression point of view, but also from a principal component point of view.

From the regression point of view of reduced rank regression, the goal is to predict the response variables using the set of predictor variables. We, therefore, set up a multivariate regression model

$$\mathbf{Y} = \mathbf{1}\mathbf{m}' + \mathbf{X}\mathbf{A} + \mathbf{E},$$

where  $\mathbf{m}$  denote the vector with the  $R$  intercepts,  $\mathbf{A}$  is a  $P \times R$  matrix with regression weights, and  $\mathbf{E}$  is a matrix with residuals. In the usual multivariate regression model the matrix of regression coefficients is unrestricted. In reduced rank regression, the restriction  $\mathbf{A} = \mathbf{B}\mathbf{V}'$  is imposed.

Reduced rank regression can also be cast as a constrained principal component analysis (PCA, Takane (2013)). In PCA, the matrix with multivariate responses ( $\mathbf{Y}$ ) of size  $N \times R$  is decomposed into a matrix with *object scores* ( $\mathbf{U}$ ) of dimension  $N \times S$ , and a matrix with *variable loadings* ( $\mathbf{V}$ ) of size  $R \times S$ . The rank, or dimensionality, should be smaller or equal to  $\min(N, R)$ . We can write PCA as the following expression

$$\mathbf{Y} = \mathbf{1}\mathbf{m}' + \mathbf{U}\mathbf{V}' + \mathbf{E},$$

where usually, identifiability constraints are imposed such as  $\mathbf{V}'\mathbf{V} = \mathbf{I}$  or  $\mathbf{U}'\mathbf{U} = \mathbf{N}\mathbf{I}$ . To estimate the PCA parameters, usually the means of the responses are computed, that is  $\hat{\mathbf{m}} = N^{-1}\mathbf{Y}'\mathbf{1}$ . Then, the centered response matrix  $\mathbf{Y}_c = \mathbf{Y} - \mathbf{1}\hat{\mathbf{m}}'$  is computed, and subsequently  $\mathbf{U}$  and  $\mathbf{V}$  are estimated by minimizing the least squares loss function

$$L(\mathbf{U}, \mathbf{V}) = \|\mathbf{Y}_c - \mathbf{U}\mathbf{V}'\|^2,$$

where  $\|\cdot\|^2$  denotes the squared Frobenius norm of a matrix. Eckart and Young (1936) show that this can be achieved by a singular value decomposition. With predictor variables, the scores ( $\mathbf{U}$ ) can be restricted to be a linear combination of these predictor variables, sometimes called external variables (i.e.,  $\mathbf{X}$ ), that is,  $\mathbf{U} = \mathbf{X}\mathbf{B}$ , with  $\mathbf{B}$  a  $P \times S$  matrix, to be estimated. Ten Berge (1993) shows that  $\mathbf{B}$  and  $\mathbf{V}$  can be estimated using a *generalized* singular value decomposition in the metrics  $\mathbf{X}'\mathbf{X}$  and  $\mathbf{I}$  (see Appendix A for details), that is, we decompose  $(\mathbf{X}'\mathbf{X})^{-\frac{1}{2}}\mathbf{X}'\mathbf{Y}_c$  with a singular value decomposition, that is

$$(\mathbf{X}'\mathbf{X})^{-\frac{1}{2}}\mathbf{X}'\mathbf{Y}_c = \mathbf{P}\mathbf{\Phi}\mathbf{Q}' \tag{1}$$

where  $P'P = Q'Q = I$  and  $\Phi$  is a diagonal matrix with singular values. Subsequently, define the following estimates

$$\hat{B} = (X'X)^{-\frac{1}{2}}P \quad (2)$$

$$\hat{V} = Q\Phi. \quad (3)$$

The  $m$  are subsequently estimated as

$$\hat{m} = N^{-1}(Y - XB'V')'1.$$

Reduced rank regression can thus be understood from two different points of views: a multivariate regression model with constraints on the regression weights or a principal component analysis with constraints on the object scores. Yee and Hastie (2003) approached logistic reduced rank regression as a multivariate regression model with a rank constraint on the regression weights. We could approach logistic reduced rank regression also from the PCA point of view. PCA for binary variables has received considerable attention lately (Collins et al. 2001; Schein et al. 2003; De Leeuw 2006; Landgraf and Lee 2020).

De Leeuw (2006) defined the log-odds term  $\theta_{ir}$  in terms of a principal component analysis  $\theta_{ir} = m_r + u_i'v_r$ , where  $u_i$  is the vector with object scores for participant  $i$ . Similar to standard PCA, object scores and variable loadings are obtained. These scores and loading, however, reconstruct the log-odds term, not the response variables themselves. De Leeuw (2006) also proposed a Majorization Minimization (MM) algorithm (Heiser 1995; Hunter and Lange 2004; Nguyen 2017) for maximum likelihood estimation of the parameters. In the majorization step, the negative log-likelihood is majorized by a least squares function. This majorizing least squares function is minimized by applying a singular value decomposition on a matrix with *working responses*, that are variables that function as the response variables in each iteration of the algorithm but are updated from iteration to iteration.

Yee and Hastie (2003, Sect. 3.4) note that “*in general, minimization of the negative likelihood cannot be achieved by use of the singular value decomposition*” and therefore proposed an alternating algorithm, where in each iteration first  $B$  is estimated considering  $V$  fixed, and subsequently  $V$  is estimated considering  $B$  fixed. For both steps a weighted least squares update is derived, where in every iteration the weights and the responses need to be redefined based on the current set of parameters. In the next section, we develop an MM algorithm for logistic reduced rank regression (note, not *all* reduced rank generalized linear models) based on the work of De Leeuw (2006), where in each of the iterations a generalized singular value decomposition is applied. We compare the two algorithms in terms of speed of computation in Sect. 4.

For interpretation of (logistic) reduced rank models, a researcher can inspect the estimated coefficients  $A = BV'$ . Coefficients in this matrix can be interpreted like the usual regression weights in (logistic) regression models. Because the number of coefficients is usually large (i.e.,  $P \times R$ ) it is difficult to obtain a holistic interpretation from this matrix. Visualization can help to obtain such a holistic interpretation

of the reduced rank model. PCA solutions can be graphically represented by biplots (Gabriel 1971; Gower and Hand 1996; Gower et al. 2011). Biplots are generalizations of usual scatterplots for multivariate data, where the observations ( $i = 1, \dots, N$ ) and the variables ( $r = 1, \dots, R$ ) are represented in low dimensional visualizations. The observations are represented by points, whereas the variables are represented by variable axes. These biplots have been extended for reduced rank regression (Ter Braak and Looman 1994), that not only represent observations and response variables but also predictor variables by variable axes, that is, they represent three different types of information, and therefore we call them *triplots*.

For visualization of the logistic reduced rank regression model, two types of triplots have been proposed. Vicente-Villardón et al. (2006) and Vicente-Villardón and Vicente-Gonzalez (2019) modified the usual biplots/triplots for the representation of binary data based on logistic models. Another type of triplot was proposed by Poole and Rosenthal (1985), Clinton et al. (2004), Poole et al. (2011), and De Rooij and Groenen (2023) that represents the logistic reduced rank model in a distance framework. In these triplots, each of the response variables is represented by two points instead of a variable axis, one for each response category. In Sect. 3, we discuss the two types of triplots, show advantages and disadvantages of both, and propose a hybrid type of triplot that has elements of both types.

In Sect. 4, we compare our new algorithm in terms of speed against that of Yee and Hastie (2003) with two empirical data sets. With these data sets, we also show the hybrid triplot for both data sets and provide an interpretation. We end this paper with a brief discussion.

## 2 An MM-algorithm for logistic reduced rank regression

Logistic models are often fitted by maximizing the likelihood, or equivalently minimizing the negative log-likelihood, that is

$$\mathcal{L}(\theta) = - \sum_{i=1}^N \sum_{r=1}^R y_{ir} \log(\pi_{ir}) + (1 - y_{ir}) \log(1 - \pi_{ir}),$$

where the probabilities ( $\pi_{ir}$ ) are functions of  $\theta_{ir}$ , with  $\theta_{ir} = m_r + \mathbf{x}_i \mathbf{B} \mathbf{v}_r$ . Let us define  $q_{ir} = 2y_{ir} - 1$ , so that the loss function can be written as

$$\mathcal{L}(\theta) = - \sum_{i=1}^N \sum_{r=1}^R \log \frac{1}{1 + \exp(-q_{ir} \theta_{ir})}.$$

To estimate the parameters, below we derived an MM-algorithm following the ideas of De Leeuw (2006). We implemented this algorithm in the R-package `lmap` (De Rooij and Busing 2022).

## 2.1 General theory about MM algorithms

The idea of MM for finding a minimum of the function  $\mathcal{L}(\theta)$ , where  $\theta$  is a vector of parameters, is to define an auxiliary function, called a *majorization function*,  $\mathcal{M}(\theta|\vartheta)$  with two characteristics

$$\mathcal{L}(\vartheta) = \mathcal{M}(\vartheta|\vartheta)$$

where  $\vartheta$  is a supporting point, and

$$\mathcal{L}(\theta) \leq \mathcal{M}(\theta|\vartheta).$$

The two equations tell us that  $\mathcal{M}(\theta|\vartheta)$  is a function that lies above (i.e., majorizes) the original function and touches the original function at the support point. Because of the above two properties, an iterative sequence defines a convergent algorithm because by construction

$$\mathcal{L}(\theta^+) \leq \mathcal{M}(\theta^+|\vartheta) \leq \mathcal{M}(\vartheta|\vartheta) = \mathcal{L}(\vartheta),$$

where  $\theta^+$  is

$$\theta^+ = \operatorname{argmin}_{\theta} \mathcal{M}(\theta|\vartheta),$$

the updated parameter. A main advantage of MM algorithms is that they always converge monotonically to a (local) minimum. The challenge is to find a parametrized function family,  $\mathcal{M}(\theta|\vartheta)$ , that can be used in every step.

In our case, the original function equals the negative log-likelihood. We majorize this function with the least squares function, that is,  $\mathcal{M}(\theta|\vartheta)$  is a least squares function. We use *majorization via a quadratic upper bound*, that is, for a twice differentiable function  $\mathcal{L}(\theta)$  and for each  $\vartheta$  the function

$$\mathcal{M}(\theta|\vartheta) = \mathcal{L}(\vartheta) + \mathcal{L}'(\vartheta)(\theta - \vartheta) + \frac{1}{2}(\theta - \vartheta)'A(\theta - \vartheta)$$

majorizes  $\mathcal{L}(\theta)$  at  $\vartheta$  when the matrix  $A$  is such that

$$A - \partial^2 \mathcal{L}(\theta),$$

is positive semi definite. We also use the property that majorization is closed under summation, that is, when  $\mathcal{M}_1$  majorizes  $\mathcal{L}_1$  and  $\mathcal{M}_2$  majorizes  $\mathcal{L}_2$ , then  $\mathcal{M}_1 + \mathcal{M}_2$  majorizes  $\mathcal{L}_1 + \mathcal{L}_2$ .

## 2.2 The algorithm

Lets recap our loss function

$$\mathcal{L}(\theta) = \sum_{i=1}^N \sum_{r=1}^R \mathcal{L}_{ir}(\theta_{ir}) = \sum_{i=1}^N \sum_{r=1}^R -\log \frac{1}{1 + \exp(-q_{ir}\theta_{ir})}.$$

Because of the summation property, we can focus on a single element,  $\mathcal{L}_{ir}(\theta_{ir})$ . The first derivative of  $\mathcal{L}_{ir}(\theta_{ir})$  with respect to  $\theta_{ir}$  is

$$\xi_{ir} \equiv \frac{\partial \mathcal{L}_{ir}(\theta_{ir})}{\partial \theta_{ir}} = -(y_{ir} - \pi_{ir})$$

Filling in the derivative and using the upper bound  $A = \frac{1}{4}$  (Böhning and Lindsay 1988; Hunter and Lange 2004), we have that

$$\begin{aligned} \mathcal{L}_{ir}(\theta_{ir}) &\leq \mathcal{L}_{ir}(\vartheta_{ir}) + \xi_{ir}(\theta_{ir} - \vartheta_{ir}) + \frac{1}{8}(\theta_{ir} - \vartheta_{ir})(\theta_{ir} - \vartheta_{ir}) \\ &\leq \mathcal{L}_{ir}(\vartheta_{ir}) + \xi_{ir}\theta_{ir} - \xi_{ir}\vartheta_{ir} + \frac{1}{8}(\theta_{ir}^2 + \vartheta_{ir}^2 - 2\theta_{ir}\vartheta_{ir}) \\ &\leq \mathcal{L}_{ir}(\vartheta_{ir}) + \frac{1}{8}\theta_{ir}^2 + \xi_{ir}\theta_{ir} - 2\frac{1}{8}\theta_{ir}\vartheta_{ir} - \xi_{ir}\vartheta_{ir} + \frac{1}{8}\vartheta_{ir}^2. \end{aligned}$$

Let us now define  $z_{ir} = \vartheta_{ir} - 4\xi_{ir}$  to obtain

$$\begin{aligned} \mathcal{L}_{ir}(\theta_{ir}) &\leq \mathcal{L}_{ir}(\vartheta_{ir}) + \frac{1}{8}\theta_{ir}^2 - 2\frac{1}{8}\theta_{ir}z_{ir} + \left(\frac{1}{8}z_{ir}^2 - \frac{1}{8}z_{ir}^2\right) - \xi_{ir}\vartheta_{ir} + \frac{1}{8}\vartheta_{ir}^2 \\ &\leq \mathcal{L}_{ir}(\vartheta_{ir}) + \frac{1}{8}(\theta_{ir} - z_{ir})^2 - \frac{1}{8}z_{ir}^2 - \xi_{ir}\vartheta_{ir} + \frac{1}{8}\vartheta_{ir}^2 \\ &\leq \frac{1}{8}(\theta_{ir} - z_{ir})^2 + c_{ir}, \end{aligned}$$

where  $c_{ir} = \mathcal{L}_{ir}(\vartheta_{ir}) - \frac{1}{8}z_{ir}^2 - \xi_{ir}\vartheta_{ir} + \frac{1}{8}\vartheta_{ir}^2$  is a constant.

Now as

$$\mathcal{L}(\theta) = \sum_{i=1}^N \sum_{r=1}^R \mathcal{L}_{ir}(\theta_{ir})$$

we have that

$$\mathcal{L}(\theta) \leq \sum_{i=1}^N \sum_{r=1}^R \frac{1}{8}(\theta_{ir} - z_{ir})^2 + c,$$

a least squares majorization function, with  $c = \sum_i \sum_r c_{ir}$ .

For logistic principal component analysis, De Leeuw (2006) defined  $\theta_{ir} = m_r + \mathbf{u}'_i \mathbf{v}_r$ . Collecting the elements  $z_{ir}$  in the matrix  $\mathbf{Z}$ , in every iteration of the MM-algorithm he minimizes

$$\|\mathbf{Z} - \mathbf{1}\mathbf{m}' - \mathbf{UV}'\|^2.$$

For logistic reduced rank regression, we define  $\theta_{ir} = m_r + \mathbf{x}'_i \mathbf{B}\mathbf{v}_r$ . In every iteration of the MM-algorithm we minimize

$$\|\mathbf{Z} - \mathbf{1}\mathbf{m}' - \mathbf{XBV}'\|^2,$$

which can be done by computing the mean and a generalized singular value decomposition as in Expressions 1, 2, and 3. In detail, in every iteration we compute  $\mathbf{Z} = \mathbf{1}\mathbf{m}' + \mathbf{X}\mathbf{B}\mathbf{V}' + 4(\mathbf{Y} - \mathbf{\Pi})$ , where  $\mathbf{\Pi}$  is the matrix with elements  $\pi_{ir}$ , using the current parameter values and update our parameters as

- $\mathbf{m}^+ = N^{-1}(\mathbf{Z} - \mathbf{X}\mathbf{B}\mathbf{V}')\mathbf{1}$
- $(\mathbf{X}'\mathbf{X})^{-\frac{1}{2}}\mathbf{X}'(\mathbf{Z} - \mathbf{1}\mathbf{m}') = \mathbf{P}\mathbf{\Phi}\mathbf{Q}'$
- $\mathbf{B}^+ = \sqrt{N}(\mathbf{X}'\mathbf{X})^{-\frac{1}{2}}\mathbf{P}_S$
- $\mathbf{V}^+ = (\sqrt{N})^{-1}\mathbf{Q}_S\mathbf{\Phi}_S$

where  $\mathbf{P}_S$  are the singular vectors corresponding to the  $S$  largest singular values, similarly for  $\mathbf{Q}_S$ , and  $\mathbf{\Phi}_S$  is the diagonal matrix with the  $S$  largest singular values.

### 3 Visualization

A rank 2 model can be visualized with a two-dimensional representation in a so-called *triplot*, that shows simultaneously three types of information: the predictor variables, the response variables, and the participants (objects). When a higher rank model is fitted, triplots can be constructed for any pair of dimensions.

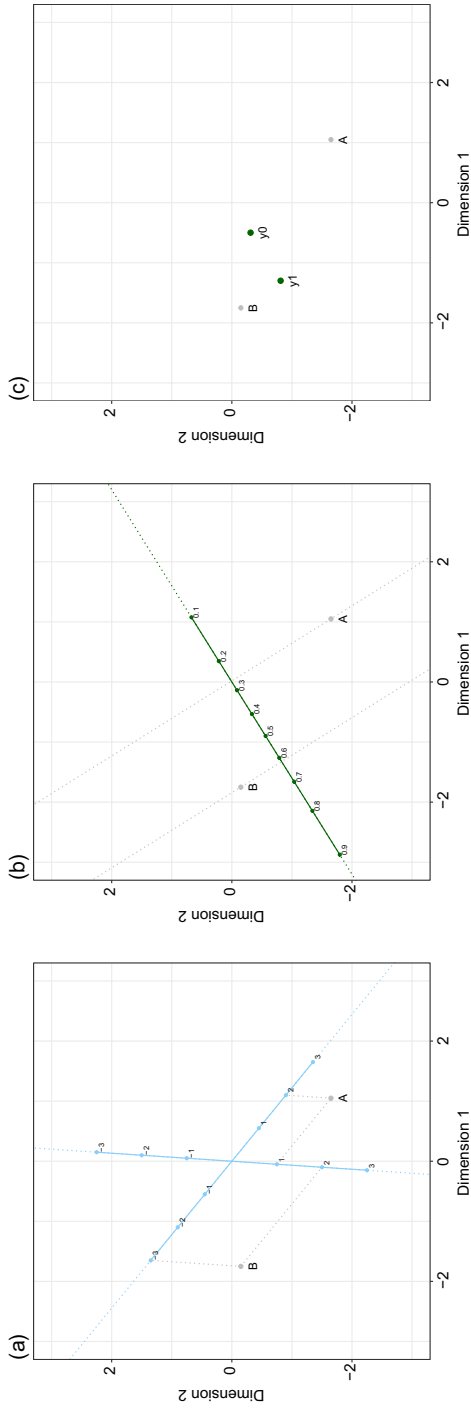
We discuss two types of triplots for logistic reduced rank regression. The first is based on the inner product relationship where we project points, representing the participants, on response variable axes with markers indicating probabilities of responding with yes (or 1). We call this a triplot of type I. Another type of triplot was recently described in detail by De Rooij and Groenen (2023) and uses a distance representation. We call this a triplot of type D. The two triplots are equivalent in the sense that they represent the same information but in a different way. In this section, we describe the two triplots in detail, make a comparison, and propose a new hybrid type of triplot that combines the advantages of the two types of triplots.

#### 3.1 The type I triplot

This type of logistic triplot was proposed by Vicente-Villardón et al. (2006) and Vicente-Villardón and Vicente-Gonzalez (2019). The objects, or participants, are depicted as points in a two-dimensional Euclidean space with coordinates  $\mathbf{u}_i = \mathbf{B}'\mathbf{x}_i$ .

Each of the predictor variables is represented by a variable axis through the origin of the Euclidean space with direction  $b_{p2}/b_{p1}$ . Markers can be added to the variables axis representing units  $t = \pm 1, \pm 2, \dots$  with coordinates  $t \frac{b_{p2}}{b_{p1}}$ . We use the convention that the variable axis has a dotted and a solid part. The solid part represents the observed range of the variable in the data, the dotted part extends the variable axis to the border of the display. The variable label is printed on the side with the highest value of the variable.





**Fig. 1** Visualization of Logistic Reduced Rank Models. **a** graphical representation of two predictor variables and the process for interpolation for two observations A and B; **b** type I representation of one response variable and the process of prediction for the two observations; **c** type D representation of one response variable, where the distance between the two observations A and B and the two response classes determines the probabilities

Object coordinates ( $u_i$ ) are a direct result of these predictor variable axis by the process of *interpolation*, as described in Gower and Hand (1996) and Gower et al. (2011).

We illustrated in Fig. 1a, where two predictor variables are represented one with regression weights 0.55 and -0.45 (diagonal variable axis), the other with regression weights -0.05 and -0.75 (almost vertical variable axis). Markers for values -3 till 3 are added to both variable axes. Also included are two observations, A and B, who have values on the two predictor variables 2 and 1 (A), and -3 and 2 (B) respectively. The process of interpolation is illustrated by the grey dotted lines, that is, we have to add two vectors to obtain the coordinates for the two observations.

Each of the response variables is also represented by a variable axis through the origin. The direction of the variables axis is  $v_{r2}/v_{r1}$ . Markers can be added to these variables axis as well. We add markers that represent the probabilities of responding 1 equal to  $\pi = \{0.1, 0.2, \dots, 0.9\}$ . The location of these markers is given by  $\lambda(v'_r v_r)^{-1} v_r$ , where  $\lambda = \log(\pi/(1 - \pi)) - \hat{m}_r$  (based on Gower et al. 2011, page 24). To obtain the predicted probabilities for an object or participant for response variables  $r$ , the process called *prediction* (Gower and Hand 1996; Gower et al. 2011) has to be used, where the point representing this object is projected onto the variable axis.

This is illustrated in Fig. 1b for a single response variable with  $v_{r1} = -0.8$  and  $v_{r2} = -0.5$ . The value of  $m = -1$ . On the variable axis there are markers indicating the expected probabilities. By projecting the points of the observations, A and B, onto this variable axis we obtain the expected probabilities. For observation A this is approximately 0.28, while for observation B the expected probability is approximately 0.61.

The two-dimensional space can be partitioned into two parts by a *decision line* for response variable  $r$ . These decision lines are perpendicular to the response variable axis and through the  $\pi = 0.5$  marker. As we have a decision line for each response variable, we have  $R$  of these lines in total, together partitioning the space in a maximum of  $\sum_{s=0}^S \binom{R}{s}$  regions (Coombs and Kao 1955), each region having a favourite response profile.

### 3.2 The type D triplot

These types of triplots are described in detail in De Rooij and Groenen (2023). Similar work has been proposed earlier mainly in the context of political vote casts by Poole and Rosenthal (1985); Clinton et al. (2004), and Poole et al. (2011).

In Type D triplots, the response variables are represented in a different manner, while the object points and the variable axes for the predictor remain the same (see Fig. 1a). Each response variable is represented by two points, one for the no-category (with coordinates  $w_{r0}$ ) and one for the yes-category (with coordinates  $w_{r1}$ ). The squared distance between an object location and these two points defines the probability, that is

$$\pi_{ir} = \frac{\exp\left(-\frac{1}{2}d^2(\mathbf{u}_i, \mathbf{w}_{r1})\right)}{\exp\left(-\frac{1}{2}d^2(\mathbf{u}_i, \mathbf{w}_{r0})\right) + \exp\left(-\frac{1}{2}d^2(\mathbf{u}_i, \mathbf{w}_{r1})\right)}, \tag{4}$$

where  $d^2(\mathbf{u}_i, \mathbf{w}_{r1})$  is the squared Euclidean two-mode distance

$$d^2(\mathbf{u}_i, \mathbf{w}_{r1}) = \sum_{s=1}^S (u_{is} - w_{r1s})^2.$$

The coordinates  $\mathbf{w}_{r0}$  and  $\mathbf{w}_{r1}$  for all  $r$  can be collected in the  $2R \times S$  matrix  $\mathbf{W}$ . This matrix can be reparametrized as

$$\mathbf{W} = \mathbf{A}_l \mathbf{L} + \mathbf{A}_k \mathbf{K}$$

with  $\mathbf{A}_l = \mathbf{I}_R \otimes [1, 1]'$  and  $\mathbf{A}_k = \mathbf{I}_R \otimes [1, -1]'$ , where  $\otimes$  denotes the Kronecker product, and where  $\mathbf{L}$  is the  $R \times S$  matrix with response variable *locations* and  $\mathbf{K}$  the  $R \times S$  matrix representing the *discriminatory power* for the response variables. Elements of  $\mathbf{K}$  and  $\mathbf{L}$  are denoted by  $k_{rs}$  and  $l_{rs}$ , respectively.

To obtain the coordinates  $\mathbf{W}$  from the parameters of the logistic reduced rank regression we take the following steps (for details, see De Rooij and Groenen 2023). The matrix  $\mathbf{K}$  is defined as  $\mathbf{K} = -\mathbf{V}/2$ . The matrix  $\mathbf{L}$  can be obtained from  $\mathbf{m}$  and  $\mathbf{K}$ . For every response variable we have that

$$m_r = \sum_{s=1}^S k_{rs} l_{rs}.$$

The solution that is closest to the origin of the Euclidean space is

$$l_{rs} = \frac{m_r k_{rs}}{2 \sum_s k_{rs}^2}.$$

Classification of an object is straightforward by choosing the response category that is closest by the object location. Probabilities are more difficult to obtain from this Type D triplot, as the mental operation described in Eq. 4 has to be performed. Because the model is defined in squared Euclidean distances, it is not the distance to the category points that is needed per se, the distance towards the decision line suffices to derive the probabilities (De Rooij and Groenen 2023). The decision line is defined as the line orthogonal to the line joining the two category points and through their midpoint. As discussed by De Rooij and Groenen (2023), the distance between the two category points relates to the discriminatory power: points far away can be discriminated well, while points close to each other can not be discriminated on the basis of the predictor variables.

This is illustrated in Fig. 1c for the same response variable and the same two observations. It is directly clear from this figure that observation A is closer to category 0, and observation B closer to category 1. Therefore, A has a higher probability for responding no (0) and B a higher probability for responding yes (1). The distance

between the two points indicates the discriminative power, the further the two points of a response variable are apart the better these two classes are distinguished by the predictor variables.

### 3.3 Comparison

Both triplots are equal on the predictor side of the model, that is, the object positions and the variable axes for the predictor variables. The response side of the model, however, has a different geometric representation. For the Type I triplot, the points of the objects need to be projected on the variable axes to obtain the probability of answering yes. For the Type D triplot, the distance of the object point towards the two category points of a response variable has to be inspected. The probability is highest for the response category closest to the object point. Determining the exact probability for the Type D biplot is, however, more involved. On the other hand, for classification the Type D plot has advantages, as it is simply the closest class to which an object is classified. Another advantage of the Type D plot is that the discriminatory power of a response variable can be directly inspected as it is dependent on the distance between the two class points, that is when the two class points of a response variable are further apart, they are overall better distinguished by the predictor variables. The discriminatory power can also be obtained from the Type I plot, but as the markers are not evenly spread over the variable axis this is usually cumbersome. We would need to inspect the distance between, say, the  $\pi = 0.5$  marker and the  $\pi = 0.6$  marker for two variables to make such a comparison, where smaller distances indicate higher discriminatory power. Which predictor variables are responsible for the discrimination of the two categories of a given response variable is most easily judged by the Type I plot, as it corresponds to the angle between the variable axis for a predictor and that of a response. The information can also be obtained from the Type D plot, but first we would need to draw the line connecting the two category points and subsequently inspect the angle.

### 3.4 The hybrid triplot

We propose to use a type of hybrid triplot, combining the features of both the Type I and the Type D plots. As the objects and predictor variables are represented in exactly the same way in the two triplots, these remain the same. The hybrid plot, uses both variables axes with markers as well as the information of the category points in its representation.

The two points  $w_{r,0}$  and  $w_{r,1}$  in the Type D triplot, derived as outlined above, lie on the variable axis for response variable  $r$  of the Type I triplot. The midpoint of these two points coincides with the marker for  $\pi = 0.5$ . These two properties allow us to combine the two types of triplots in a hybrid visualization, where the response variables axis are printed by dotted lines with a solid part from  $w_{r,0}$  to  $w_{r,1}$ . The endpoints of this solid part of the variable axis represent the two category points. The length of this solid part indicates the discriminatory power.

## 4 Empirical examples

### 4.1 Drug consumption data set

The drug consumption data (Fehrman et al. 2017) has records for 1885 respondents and has been analyzed by De Rooij and Groenen (2023) before. Here we replicate that analysis, but focus on the new, hybrid, triplot. For each respondent there is information about age and gender plus measurements for the big five personality traits, neuroticism (N), extraversion (E), openness to experience (O), agreeableness (A), and conscientiousness (C), and another personality characteristic, namely sensation seeking (SS).

In addition, participants were questioned concerning their use of 18 legal and illegal drugs. For each drug, participants indicated whether they used that drug in the last year (yes or no). We focus on the 11 drugs that had a minimum percentage of 10% and a maximum of 90%, which are amphetamine (Am), benzodiazepine (Be), cannabis (Ca), cocaine (Co), ecstasy (Ex), ketamine (Ke), legal highs (Le), LSD, methadone (Me), mushrooms (Mu), and nicotine (Ni) ( $R = 11$ ).

#### 4.1.1 Comparison of algorithms

In this section, we compare the two algorithms in terms of speed. The IRLS algorithm of Yee and Hastie (2003) been implemented in the VGAM package (Yee 2022), the MM algorithm is implemented in the lmap package (De Rooij and Busing 2022). An exact comparison is difficult, because in the implementation in VGAM the call starts with a formula, so the design matrices have to be built, whereas the lmap packages starts with the two matrices. Furthermore, the VGAM implementation uses several criteria for convergence, while the lmap convergence criterion is solely based on deviance. In the comparison, we first checked using the complete data set which convergence criterion in lmap leads to the same solution. Nevertheless, the comparison is not completely fair. We estimate the same rank 2 model using these two algorithms. The data set has  $N = 1885$ ,  $P = 9$ , and  $R = 11$ . To compare the speed of the two algorithms we use the microbenchmark package (Mersmann 2021) where the two algorithms are applied ten times. Results are shown in Table 1, where it can be seen that MM algorithm is much faster.

**Table 1** Timing of rrvglm-algorithm in the VGAM package and the lpca-algorithm in the lmap package for the drug consumption and NCD data sets

data	algorithm	min	lq	mean	median	uq	max	neval
drugs	rrvglm	3.075	3.216	3.336	3.328	3.433	3.655	10
	lpca	0.029	0.030	0.040	0.038	0.047	0.059	10
ncd	rrvglm	15.274	16.570	16.930	16.804	17.728	18.038	10
	lpca	1.378	1.406	1.464	1.452	1.537	1.559	10

For our MM algorithm,  $(X'X)^{-\frac{1}{2}}$  and  $(X'X)^{-\frac{1}{2}}X$  need only to be evaluated once. During the iterations a SVD of a  $P \times R$  matrix has to be solved which is a relatively small matrix. Therefore, the computational burden during the iterations is small. As is usual for MM algorithms convergence is relatively slow in terms of the number of iterations needed (Heiser 1995). For the drug consumption data, 42 iterations were needed.

The IRLS algorithm alternates between an update of  $B$  and  $V$  (Yee 2015, Sect. 5.3.1). For both updates, a matrix with working weights needs to be inverted. As this matrix depends on the current estimated values, the inverse needs to be re-evaluated in every iteration. These matrix inversions are computationally heavy. In terms of the number of iterations the IRLS algorithm is faster, that is, only 6 iterations were needed for the drug consumption data.

Yee (2015) (Sect. 3.2.1) points out that, apart from some matrix multiplications, the number of floating point operations (flops) for the IRLS algorithm is  $2NS^3(P^2 + R^2)$  per iteration, which for this application is 6092320. The SVD in the MM algorithm takes  $2PR^2 + 11R^3 = 16819$  flops per iteration (cf., Trefethen and Bau 1997). These computations exclude the intercepts ( $m$ ). Also with respect to storage our algorithm has some advantages, that is, the MM algorithm works with the matrix  $X$  while the IRLS algorithm works with  $X \otimes I_S$ , a much larger matrix.

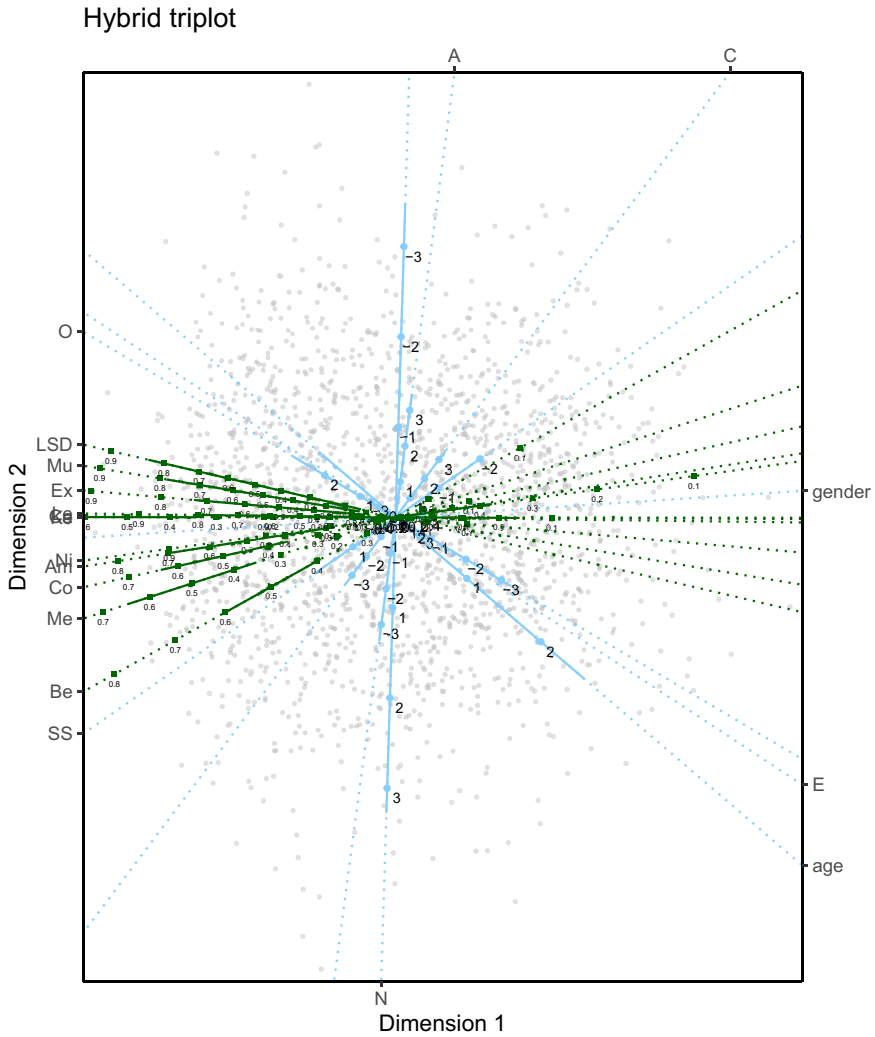
#### 4.1.2 Visualization

De Rooij and Groenen (2023) introduced a quality of representation measure for the response variables. The quality of representation measure,  $Q_r$ , compares the fit of a single response variable in the rank 2 model with the fit obtained in a separate logistic regression based on the same predictor variables.  $Q_r$  is defined by

$$Q_r = (\mathcal{D}_{(0,r)} - \mathcal{D}_r) / (\mathcal{D}_{(0,r)} - \mathcal{D}_{lr}),$$

where  $\mathcal{D}_{(0,r)}$  is the deviance of the intercept-only logistic regression model for a response variable  $r$ ,  $\mathcal{D}_r$  is part of the deviance (i.e.,  $2\mathcal{L}(\theta)$ ) corresponding to response variable  $r$ , and  $\mathcal{D}_{lr}$  is the deviance from a logistic regression of response variable  $r$  with the same predictor variables as in the reduced rank model.  $Q_r$  ranges between 0 (very bad fit) to 1 (no loss due to rank restriction). As such, we see how much fit is lost for every response variable, by reducing the rank of the model. For this analysis, the quality measures are all very high, ranging between 0.90 and 1.00, with the worst fit for Cocaine.

The hybrid triplot for the rank 2 model of the drug consumption data is shown in Fig. 2. First, let us inspect the response variables. All the labels are printed on the left-hand side of the triplot, indicating that the probability of drug use is highest on that side of the display and lowest on the right-hand side of the triplot. We see that the variable axes have solid parts of different lengths, that is, the solid part for benzodiazepine (Be) and methadone (Me) is much smaller than the solid part of cannabis (Ca) or LSD. Therefore, overall cannabis use and LSD are better discriminated by the predictor variables than benzodiazepine or methadone use. The endpoints of



**Fig. 2** The hybrid triplot for the drug consumption data. The grey points represent the participants. The blue lines are the variable axes for the predictor variables (age, gender, neuroticism (N), extraversion (E), openness (O), conscientiousness (C), agreeableness (A), sensation seeking (SS)). The green lines are the variable axes for the responses (amphetamine (Am), benzodiazepine (Be), cannabis (Ca), cocaine (Co), ecstasy (Ex), ketamine (Ke), legal highs (Le), LSD, methadone (Me), mushrooms (Mu), and nicotine (Ni)). The variable labels are printed on the positive side of the variable, that is, the highest values of the predictor variables and the highest probabilities for the response variables (color figure online)

each solid line correspond to the positions of the yes and no point for each response variable in the type D plot, where in this particular triplot the yes marker is always on the lefthand side of the display.

Some predictor variable axes for the predictors have sharp angles with the variable axes of the responses, whereas others have obtuse angles. Openness,

extraversion, age, sensation seeking, and gender have sharp angles indicating that these predictor variables discriminate well between drug use and abstinence.

The variable axes for openness, extraversion, and age are close together. Therefore these variables have similar effects on the response variables. More open participants, younger participants, and more introvert participants tend to have a higher probability for using LSD, mushrooms (Mu), and ecstasy (Ex). Participants who score higher on sensation seeking have a higher probability of using benzodiazepine (Be), methadone (Me), and cocaine (Co). Female participants (i.e., gender = 1) have a lower probability of drug use than male participants.

Neuroticism (N) and Agreeableness (A), on the other hand, are almost orthogonal to the response variable axes indicating that the relationship between these two predictor variables and drug use are weak.

The variable axis for conscientiousness has a sharp angle with, for example, benzodiazepine (Be) and methadone (Me), but an obtuse angle with the variable axes for LSD, mushrooms, and ecstasy. Less conscientious participants, therefore, have a higher probability to use benzodiazepine and methadone while more conscientious participants have a lower probability to use benzodiazepine and methadone. The degree of conscientiousness, however, does not result in large differences in the probability of, say, LSD use.

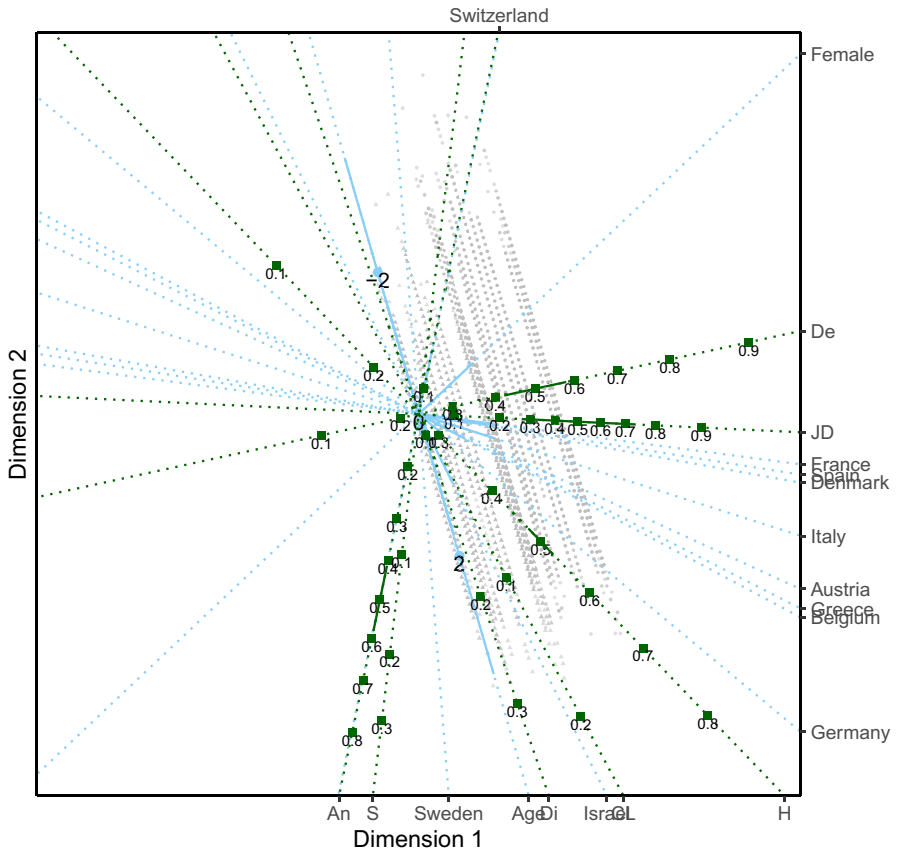
## 4.2 Non communicable diseases data set

Noncommunicable diseases (NCDs) such as diabetes, chronic lung disease, stroke, and depression tend to be of long duration and are the result of a combination of genetic, physiological, environmental and behavioural factors. The 2030 Agenda for Sustainable Development recognizes NCDs as a major challenge for sustainable development. Here, data (wave 1) of the Survey of Health, Ageing and Retirement in Europe (SHARE; Börsch-Supan et al. 2013) were used. In this data set, we have for  $N = 29207$  subjects information on seven NCDs, diabetes (Di), hypertension (H), chronic lung disease (CL), joint disorders (JD), angina (An), stroke (S), and depression (De). Overall, the prevalence of NCDs is not large, percentages range from 3.5% for stroke till 37.8% for depression. However, relatively many people develop multiple NCDs, that is, in the current data set 6021 participants have two NCDs, 2451 participants developed three NCDs, 894 have four, 214 have five, 40 have six, and 3 participants suffer from all seven NCDs.

Participants of this study live in various European countries, Austria (AT, 1538), Germany (DE, 2902), Sweden (SE, 2937), The Netherlands (NL, 2805), Spain (ES, 2226), Italy (IT, 2494), France (FR, 2890), Denmark (DK, 1665), Greece (EL, 2800), Switzerland (CH, 957), Belgium (BE, 3735), and Israel (IL, 2258). In the current analysis we are interested in the relationship between country of origin and NCDs, taking into account the gender and age of the participants. As, country is coded by 11 dummy variables, with The Netherlands as a reference category, the number of predictor variables in this analysis equals  $P = 13$ .



### Hybrid triplot



**Fig. 3** The hybrid triplot for the NCD data. The grey points represent the participants. The blue lines are the variable axes for the predictor variables (age, gender, Austria, Germany, Sweden, Spain, Italy, France, Denmark, Greece, Switzerland, Belgium, and Israel). The green lines are the variable axes for the responses (diabetes (Di), hypertension (H), chronic lung disease (CL), joint disorders (JD), angina (An), stroke (S), and depression (De)). The variable labels are printed on the positive side of the variable, that is, the highest values of the predictor variables and the highest probabilities for the response variables (color figure online)

#### 4.2.1 Comparison of algorithms

Again, we compare the two algorithms in terms of speed. This data set has  $N = 29207$ ,  $P = 13$ , and  $R = 7$ . To compare the speed of the two algorithms we use the `microbenchmark` package (Mersmann 2021). Results are shown in Table 1, where it can be seen that again the MM algorithm is much faster.

### 4.2.2 Visualization

The quality of representation is not as good as in the drug consumption data. Some response variables are well represented (H, 0.86; JD, 0.93; An, 0.96; S, 0.82; De, 0.81), but diabetes and chronic lung disease are not well represented, the quality of representation are 0.66 and 0.57 respectively. The two-dimensional triplot is shown in Fig. 3. As can be seen, not for every response variable the solid green part is drawn (Di, CL, and S). The reason is that this part of the variable axis is far outside the range of the participant points. Extending the boundaries of the triplot to include this part, would clutter the center of the display. An important conclusion for these three variables is that for all participants the probability of these diseases is small, that is smaller than 0.3 for stroke and diabetes, and smaller than 0.2 for chronic lung disease. The decision lines fall outside the triplot, indicating that all participants are classified as not having those noncommunicable diseases. For the other response variables, we see that joint disorders have a relatively high discriminatory power (long solid part), whereas hypertension has a relatively low discriminatory power (small solid part). Angina and depression are in between.

Concerning the relationships between the predictor variables and response variables we see that being female is associated with higher depression and lower angina. Higher ages are associated with angina, stroke, diabetes, and hypertension, but age has no relationship to depression and only a small relationship with joint disorders.

Considering the countries, the effects of the different countries are compared against the profile of The Netherlands (i.e., The Netherlands is the baseline category). Austria and Sweden have very small vectors and Germany and Greece small vectors, indicating that their average profiles of non-communicable diseases are very similar to that of The Netherlands. The countries that differ the most from The Netherlands are Spain, Italy, France, Israel, and Switzerland.

Interpretation goes again by inspecting the angles of vectors corresponding to the countries to the variable axes of the responses. A small angle suggests that a country has, on average, a higher probability for that response, a  $90^\circ$  angle suggest a similar probability, while an obtuse angle indicates a lower probability for a noncommunicable disease. Switzerland scores lower on angina, stroke, diabetes, chronic lung disease, and hypertension, but higher on joint disorders and depression. France, Spain and Italy, and to a lower degree also Denmark and Belgium score higher on depression, joint disorders, and hypertension (small angles), but similar on angina, stroke, diabetes, and chronic lung disease ( $90^\circ$  angles). Israel has higher probabilities for angina, stroke, diabetes, chronic lung disease, and hypertension, but similar probabilities as The Netherlands on joint disorders and depression.

## 5 Discussion

Logistic reduced rank regression is a useful tool for regression analysis of multivariate binary response variables. We developed a new algorithm based on previous work by De Leeuw (2006) and implemented it in the `lmap`-package (De Rooij and Busing 2022). The algorithm uses the majorization inequality of De Leeuw (2006)

to move from the negative log-likelihood to a least squares function and the generalized singular value decomposition for minimizing the least squares function. Each of these two steps is known and as such the algorithm is a straightforward combination of the two. The inequality of De Leeuw (2006) to find a least-squares majorization function for the negative binomial or multinomial log-likelihood can more commonly be used to develop logistic models for categorical data.

We compared the new algorithm to the IRLS algorithm (Yee and Hastie 2003; Yee 2015) on two empirical data sets; our new algorithm is about ten times faster but uses more iterations. MM algorithms are known for slow convergence (see Heiser 1995) in terms of a number of iterations needed. Yet, the updates within the iterations are computationally cheap. A singular value decomposition of a  $P \times R$  matrix needs to be computed, where usually  $P$  and  $R$  are relatively small. Our MM algorithm is only applicable for logistic reduced rank models, whereas the IRLS algorithm of Yee (2015) is designed for the family of reduced rank generalized linear models.

In the VGAM-package (Yee 2022, 2015), in a second step also the standard errors of the model parameters can be obtained. In an MM algorithm, not the negative log-likelihood is itself minimized but iteratively the majorization function is minimized. Therefore, our algorithm does not automatically give an estimate of the Hessian matrix nor the standard errors. Is this a bad thing? Buja et al. (2019a, 2019b) recently argued that if we know statistical models are approximations, we should also carry the consequences of this knowledge. The computation of standard errors assumes the model to be true, that is, not an approximation. Therefore, assuming the model is true while knowing it is not true results in estimated standard errors that are biased. A better approach to obtain standard errors or confidence intervals is to use the so-called pairs bootstrap, where the predictors and responses are jointly re-sampled. For the bootstrap it is useful if the algorithm is fast.

Two types of triplots were discussed and compared, the Type I and Type D triplots. The two types of triplots are equal on the predictor side of the model, but differ in the representation of the response variables. Whereas the Type I triplot uses an inner product relationship where object points have to be projected onto response variable axes, the Type D uses a distance relationship, where the distance of an object point towards the yes and no point for each response variable determines the probabilities. We discussed advantages and disadvantages of both approaches, and were then able to develop a new, hybrid, type of triplot by combining the two.

In the Type D triplot we make use of the two-mode Euclidean distance. This distance is often used to model single-peaked response functions and the object points are then called *ideal points*. The single-peaked response function is usually contrasted with the dominance response function (see, for example, Drasgow et al. 2010, and references therein). Whereas in the latter the probability of answering yes is a monotonic function of the position on the latent trait, in the former this probability is a single-peaked function defined by distances. In the Type D triplot, however, the relationship is still of the dominance type, where the probability of answering yes goes monotonically up or down for objects located on any straight line in the Euclidean representation. The main reason is that the model is in terms of the distance towards the categories of the response variable, not the response variable

itself. De Rooij et al. (2022) recently developed a distance model where the distance between the object and the response variable (i.e., not the categories) determines the probability of answering yes. Such a representation warrants an interpretation in terms of single peaked relationships. Logistic reduced rank regression models assume a monotonic predictor-response relationship, no matter how the model is visualized.

## Appendix A: The generalized singular value decomposition

The singular value decomposition is a well known decomposition of a matrix and can be considered a special case of the generalized singular value decomposition in the metrics  $\mathbf{I}$  and  $\mathbf{I}$ . The generalized singular value decomposition of the matrix  $\mathbf{Y}$  in the metrics  $\mathbf{G}$  and  $\mathbf{H}$  is given by

$$\mathbf{Y} = \mathbf{P}\mathbf{\Lambda}\mathbf{Q}'$$

where  $\mathbf{P}$  is a matrix of left generalized singular vectors,  $\mathbf{Q}$  is a matrix of right generalized singular vectors, and  $\mathbf{\Lambda}$  a diagonal matrix with singular values. The matrices  $\mathbf{P}$  and  $\mathbf{Q}$  satisfy  $\mathbf{P}'\mathbf{G}\mathbf{P} = \mathbf{Q}'\mathbf{H}\mathbf{Q} = \mathbf{I}$ , that is, they are columnwise orthogonal with respect to the metric matrices  $\mathbf{G}$  and  $\mathbf{H}$  (Takane 2013, page 65). We assume that both  $\mathbf{G}$  and  $\mathbf{H}$  are square, symmetric, positive definite matrices.

The generalized singular value decomposition is computed as follows (Takane 2013; Abdi 2007). Do a usual SVD on the matrix  $\mathbf{Y}_*$  defined as  $\mathbf{Y}_* = \mathbf{G}^{\frac{1}{2}}\mathbf{Y}\mathbf{H}^{\frac{1}{2}}$ , that is

$$\mathbf{Y}_* = \mathbf{G}^{\frac{1}{2}}\mathbf{Y}\mathbf{H}^{\frac{1}{2}} = \mathbf{P}_*\mathbf{\Lambda}_*\mathbf{Q}_*'$$

and then

$$\begin{aligned}\mathbf{P} &= \mathbf{G}^{-\frac{1}{2}}\mathbf{P}_* \\ \mathbf{\Lambda} &= \mathbf{\Lambda}_* \\ \mathbf{Q} &= \mathbf{H}^{-\frac{1}{2}}\mathbf{Q}_*\end{aligned}$$

**Author Contributions** As this is a single-author paper, everything was done by the main author.

**Funding** The author declares that no funds, grants, or other support were received during the preparation of this manuscript.

**Data Availability Non communicable diseases data set:** This paper uses data from SHARE Wave 1 (DOI: 10.6103/SHARE.w1.800), see Börsch-Supan et al. (2013) for methodological details. The SHARE data collection has been funded by the European Commission, DG RTD through FP5 (QLK6-CT-2001-00360), FP6 (SHARE-I3: RII-CT-2006-062193, COMPARE: CIT5-CT-2005-028857, SHARELIFE: CIT4-CT-2006-028812), FP7 (SHARE-PREP: GA No 211909, SHARE-LEAP: GA No227822, SHARE M4: GA No261982, DASISH: GA No283646) and Horizon 2020 (SHARE-DEV3: GA No676536, SHARE-COHESION: GA No870628, SERISS: GA No654221, SSHOC: GA No823782,

SHARE-COVID19: GA No101015924) and by DG Employment, Social Affairs & Inclusion through VS 2015/0195, VS 2016/0135, VS 2018/0285, VS 2019/0332, and VS 2020/0313. Additional funding from the German Ministry of Education and Research, the Max Planck Society for the Advancement of Science, the U.S. National Institute on Aging (U01\_AG09740-13S2, P01\_AG005842, P01\_AG08291, P30\_AG12815, R21\_AG025169, Y1-AG-4553-01, IAG\_BSR06-11, OGHA\_04-064, HHSN271201300071C, RAG052527A) and from various national funding sources is gratefully acknowledged (see [www.share-project.org](http://www.share-project.org)). Data can be obtained from [www.share-project.org](http://www.share-project.org). **Drug consumption data set:** The drug consumption data is available on UCI data repository ([https://archive.ics.uci.edu/ml/machine-learning-databases/00373/drug\\_consumption.data](https://archive.ics.uci.edu/ml/machine-learning-databases/00373/drug_consumption.data)). **R-code:** R-code for the analyses conducted in this paper can be found on the github page of the first author.

## Declarations

**Conflict of interest** The authors have no relevant financial or non-financial interests to disclose.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Abdi H (2007) Singular value decomposition (svd) and generalized singular value decomposition. *Encyclopedia of measurement and statistics* 907–912
- Agresti A (2013) *Categorical data analysis*, 3rd edn. John Wiley & Sons
- Anderson TW (1951) Estimating linear restrictions on regression coefficients for multivariate normal distributions. *The Annals of Mathematical Statistics* 327–351
- Böhning D, Lindsay BG (1988) Monotonicity of quadratic-approximation algorithms. *Ann Inst Stat Math* 40(4):641–663
- Buja A, Brown L, Berk R, George E, Pitkin E, Traskin M, Zhang K, Zhao L et al (2019) Models as approximations I: consequences illustrated with linear regression. *Stat Sci* 34(4):523–544
- Buja A, Brown L, Kuchibhotla AK, Berk R, George E, Zhao L et al (2019) Models as approximations II: a model-free theory of parametric regression. *Stat Sci* 34(4):545–565
- Börsch-Supan A, Brandt M, Hunkler C, Kneip T, Korbmayer J, Malter F, Schaen B, Stuck S, Zuber S (2013) Data Resource Profile: The Survey of Health, Ageing and Retirement in Europe (SHARE). *Int J Epidemiol* 42(4):992–1001
- Clinton J, Jackman S, Rivers D (2004) The statistical analysis of roll call data. *Am Polit Sci Rev* 98(2):355–370
- Collins M, Dasgupta S, Schapire RE (2001) A generalization of principal components analysis to the exponential family. *Advances in neural information processing systems* 14
- Coombs CH, Kao R (1955) *Nonmetric factor analysis*. University of Michigan. Department of Engineering Research. Bulletin
- Davies P, Tso MK-S (1982) Procedures for reduced-rank regression. *J R Stat Soc* 31(3):244–255
- De Leeuw J (2006) Principal component analysis of binary data by iterated singular value decomposition. *Comput Stat Data Anal* 50(1):21–39
- De Rooij M, Busing FMTA (2022) *lmap: Logistic Mapping*. R package version 0.1.1
- De Rooij M, Groenen PJF (2023) The melodic family for simultaneous binary logistic regression in a reduced space. In: Okada A, Shigemasa K, Yoshino R, Yokoyama S (eds) *Facets of behaviormetrics: the 50th anniversary of the behaviormetric society*, Springer. (preprint available at [arXiv:2102.08232](https://arxiv.org/abs/2102.08232))

- De Rooij M, Woestenburg D, Busing FMTA (2022) Supervised and unsupervised mapping of binary variables: a single-peaked perspective. Submitted paper
- Dragow F, Chernyshenko OS, Stark S (2010) 75 years after likert: Thurstone was right! *Ind Organizational Psychol* 3(4):465–476
- Eckart C, Young G (1936) The approximation of one matrix by another of lower rank. *Psychometrika* 1(3):211–218
- Fehrman E, Muhammad AK, Mirkes EM, Egan V, Gorban AN (2017) The five factor model of personality and evaluation of drug consumption risk. In: Palumbo F, Montanari A, Vichi M (eds) *Data science: innovative developments in data analysis and clustering*. Springer, Berlin, pp 231–242
- Gabriel KR (1971) The biplot graphic display of matrices with application to principal component analysis. *Biometrika* 58(3):453–467
- Gower J, Hand D (1996) *Biplots*. Taylor & Francis
- Gower J, Lubbe S, Le Roux N (2011) *Understanding Biplots*. Wiley
- Heiser WJ (1995) Convergent computation by iterative majorization: theory and applications in multidimensional data analysis. In: Krzanowski WJ (ed) *Recent advances in descriptive multivariate analysis*. Clarendon Press, pp 157–189
- Hunter DR, Lange K (2004) A tutorial on MM algorithms. *Am Stat* 58(1):30–37
- Izenman AJ (1975) Reduced-rank regression for the multivariate linear model. *J Multivariate Anal* 5(2):248–264
- Landgraf AJ, Lee Y (2020) Dimensionality reduction for binary data through the projection of natural parameters. *J Multivariate Anal* 180:104668
- McCullagh P, Nelder JA (1989) *Generalized linear models*. Chapman & Hall / CRC
- Mersmann O (2021) *microbenchmark: accurate Timing Functions*. R package version 1(4):9
- Nguyen HD (2017) An introduction to majorization-minimization algorithms for machine learning and statistical estimation. *Wiley Interdisciplinary Rev* 7(2):e1198
- Poole K, Lewis JB, Lo J, Carroll R (2011) Scaling roll call votes with wnominate in r. *J Stat Softw* 42:1–21
- Poole KT, Rosenthal H (1985) A spatial model for legislative roll call analysis. *Am J Political Sci* 357–384
- Schein AI, Saul LK, Ungar LH (2003) A generalized linear model for principal component analysis of binary data. In: Bishop CM, Frey BJ (eds) *Proceedings of the ninth international workshop on artificial intelligence and statistics*
- Takane Y (2013) *Constrained principal component analysis and related techniques*. CRC Press
- Ten Berge JM (1993) *Least squares optimization in multivariate analysis*. DSWO Press, Leiden University Leiden
- Ter Braak CJ, Looman CW (1994) Biplots in reduced-rank regression. *Biometrical J* 36(8):983–1003
- Trefethen LN, Bau D (1997) *Numerical linear algebra*, vol 181. Siam
- Tso M-S (1981) Reduced-rank regression and canonical analysis. *J R Stat Soc Ser B (Methodological)* 43(2):183–189
- Van den Wollenberg AL (1977) Redundancy analysis an alternative for canonical correlation analysis. *Psychometrika* 42(2):207–219
- Vicente-Villardón JL, Galindo-Villardón MP, Blázquez-Zaballos A (2006) Logistic biplots. *Multiple Correspondence Anal Related Methods* 503–521
- Vicente-Villardón JL, Vicente-Gonzalez L (2019) Redundancy analysis for binary data based on logistic responses. In: Chadjipadelis T, Lausen B, Markos A, Lee TR, Montanari A, Nugent R (eds) *Data analysis and rationality in a complex world*. Springer, Berlin, pp 331–339
- Yee TW (2015) *Vector generalized linear and additive models: with an implementation in R*. Springer, New York, USA
- Yee TW (2022) *VGAM: Vector generalized linear and additive models*. R package version 1.1-7
- Yee TW, Hastie TJ (2003) Reduced-rank vector generalized linear models. *Stat Model* 3(1):15–41

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.