Comparing Singular Value Regularization Methods in The MELODIC Family for Simultaneous Binary Logistic Regression in a Reduced Space

Lieshout, Kenny van

Universiteit Leiden

Faculteit der Sociale Wetenschappen

Comparing Singular Value Regularization Methods in The MELODIC Family for Simultaneous Binary Logistic Regression in a Reduced Space

Thesis at Leiden University

Kenny van Lieshout

Master Thesis Methodology and Statistics Master

Methodology and Statistics Unit, Institute of Psychology,

Faculty of Social and Behavioral Sciences, Leiden University

Date: August 2021

Student number: 1765752

Supervisor: Prof. Dr. M.J. de Rooij

Abstract

As the availability of data becomes more widespread and computational technology develops, the need to model several outcome variables at once increases. To do this for multiple dichotomous outcome variables, De Rooij and Groenen (2021) proposed the MELODIC family for simultaneous binary logistic regression in a reduces space. As an added feature, 4 different forms of regularization on the singular values were proposed to let the algorithm itself select the true dimensionality of the data set. In this paper a simulation study was performed to provide empirical evidence for the functionality of the regularization features on data sets with differing numbers of subjects, predictor variables, and outcome variables. The results show that the hard-thresholding regularization consistently estimates the correct dimensionality with regularization on the logarithm of the singular values slightly outperforming the regularization on the unaltered singular values.

Table of Content

Introduction

Nowadays, more and more data is available to researchers (Holst, 2021). Due to advances in computation technology, it has become possible to analyze and gain insights from this data (Dehuri & Sanyal, 2015). Many of the available variables, especially in the social- and biological sciences, are dichotomous in nature (Mayya et al., 2017; Shreffler & Huecker, 2021). Dichotomous variables are categorical variables with 2 categories where each subject belongs to either of the two. Examples of these variables are healthy/sick, male/female, high risk/low risk, etc.

In analyses of large data sets, multivariate techniques can give insights into dependencies of the variables where separate analyses cannot. Multivariate techniques have shown to outperform single variable analyses in predictive accuracy (Huberty & Morris, 1992), and are therefore preferred when possible. To accommodate the need for multivariate dichotomous analyses, De Rooij and Groenen (2021) proposed the MELODIC family for simultaneous binary logistic regression in a reduced space. MELODIC stands for *Multivariate Logistic Distance to Categories* and the model uses a combination of mathematical concepts to improve predictive accuracy including multivariate shrinkage (Fourdrinier et al., 2018), dimensionality reduction (Krishnaiah & Kanal, 1982), and multidimensional unfolding (Busing, 2010).

Logistic distance models have been proposed before (Takane et al., 1987; Takane, 1987; De Rooij, 2009). These models have been the basis for the multivariate logistic distance model developed by Worku and De Rooij (2018) of which the MELODIC multivariate distance model is an extension. In the model, the categories and subjects are placed in a low dimensional Euclidean space using multidimensional unfolding. The Euclidean distance (a straight line through space) is then measured between the subject and the 2 possible categories. The subject is assigned to the category with the smallest distance. The dimensionality of the solution should have a theoretical basis.

The relationship of the predictor variables to the position of the subject is assumed to be linear and is defined as

$$\mathbf{u}_i = \mathbf{x}_i^T\mathbf{B}, \tag{1.1}$$

where $\mathbf{u}_i$ is the location of subject *i* and **B** is a matrix of regression weights for the predictor variables (P) times the number of dimensions (M).

The coordinates of the categories c (where $c = \{0, 1\})$ within the Euclidean space is denoted as v_{qcm} (with q being the indicator of response variables up to the total number of response variables Q) and are stored in vector $\mathbf{v}_{qc}$ of M dimensions. From the distance between the subject to the categories of the response variable, the conditional probability of this subject belonging to either category 0, or category 1 is calculated by

$$\pi_{qc}(\mathbf{x}_i) = \frac{\exp\left(-\delta(\mathbf{u}_i, \mathbf{v}_{qc})\right)}{\exp\left(-\delta(\mathbf{u}_i, \mathbf{v}_{q0})\right) + \exp\left(-\delta(\mathbf{u}_i, \mathbf{v}_{q1})\right)}, \tag{1.2}$$

with $\delta(.,.)$ representing half the squared Euclidean distance

$$\delta(\mathbf{u}_i, \mathbf{v}_{qc}) = \frac{1}{2}\sum_{m=1}^{M}(u_{im} - v_{qcm})^2 = \frac{1}{2}\sum_{m=1}^{M}(u_{im}^2 + v_{qcm}^2 - 2u_{im}v_{qcm}). \tag{1.3}$$

To interpret the influence of the variables on the prediction, the log odds in favor of category 1 over category 0 for each independent variable can be found using

$$log\frac{\pi_{q1}(\mathrm{x}_i)}{1 - \pi_{q1}(\mathrm{x}_i)} = \delta(\mathbf{u}_i, \mathbf{v}_{q0}) - \delta(\mathbf{u}_i, \mathbf{v}_{q1}). \tag{1.4}$$

The log odds are then defined as

$$log\frac{\pi_{q1}(x_i)}{1 - \pi_{q1}(x_i)} = \sum_{m=1}^{M}\left[\frac{1}{2}(v_{q0m}^2 - v_{q1m}^2) + \mathbf{x}_i^T\mathbf{b}_m(v_{q1m} - v_{q0m})\right]. \tag{1.5}$$

So, the effect of predictor variable x on response variable q is determined by the distance between the categories, and by the regression coefficients $\mathbf{b}_m$. The model can be interpreted as a regular univariate logistic model by writing the log odds as

$$log\frac{\pi_{qc}(\mathbf{x}_i)}{1 - \pi_{qc}(\mathbf{x}_i)} = a_q^* + \mathbf{x}_i^T\mathbf{b}_q^*, \tag{1.6}$$

where a_q^* and $\mathbf{b}_q^*$ ($\mathbf{b}_q^*$ being the implied coefficients) are defined as

$$a_q^* = \frac{1}{2}\Sigma_{m=1}^{M}(v_{q0m}^2 - v_{q1m}^2) \tag{1.7}$$

and

$$\mathbf{b}_q^* = \Sigma_{m=1}^{M} \mathbf{b}_m\left(v_{q1m} - v_{q0m}\right). \tag{1.8}$$

The ability for the predictor variables to predict the outcome variable can be derived from the distance between the 2 categories of the outcome variable. The larger the distance, the better the predictor variables are able to distinguish between the categories. So, if the categories fall on the same location in the Euclidean space, the predictor variables have no predictive value for the outcome variable (Anderson, 1984).

To improve the predictive accuracy of the MELODIC model, multivariate shrinkage is used by modelling several logistic regressions in a reduced space (De Rooij & Groenen, 2021). Shrunken averages in terms of mean squared error have been shown to outperform simple averages of a multivariate distribution by Stein et al. (1956). Also, shrinkage of coefficients (towards zero) has been shown to improve predictive accuracy in a number of multivariate models (Breiman & Friedman, 1997).

Dimensionality reduction has several advantages, for example removing redundant or noisy features from a data set, discovering hidden correlations and easier visualization. Dimension reduction can be done by finding patterns in subspaces within a larger multidimensional space and projecting these subspaces onto a smaller number of dimensions (Carreira-Perpinán, 1997). Two dominant ways of dimensionality reduction are *feature selection* (Pudil & Novovičová, 1998; Jain & Zongker, 1997) and *feature extraction* (Guyon et al., 2008; Nevatia

& Babu, 1980). Feature selection selects a subset of the relevant data to make a model more parsimonious and/or remove noisy features which in turn can lower its prediction error. Examples of feature selection methods are *Lasso* (which select features through regularization), *Best Subset Selection* and *Forward and Backward Stepwise Selection* (James et al, 2013). Feature extraction uses the inner product of a matrix (a scalar resulting from the summation of the resulting numbers when 2 matrices are multiplied) to select a low-dimensional set of features out of a higher dimensional data set. The most commonly used form of feature extraction is *Principal component analysis* (Pearson, 1901; Jolliffe, 2002; Abdi & Williams, 2010) which extracts features through a Singular Value Decomposition. However, unlike the usual feature extraction, which is applied to the predictor variables, the MELODIC methods extracts the dimensionality of its solution from the outcome variables instead. To find the solution, the algorithm uses a singular value decomposition in an iterative majorization algorithm (see De Rooij & Groenen, 2021, p. 14-21) to estimate the regression weights of the discrimination parameters.

Singular value decomposition represents any matrix by the product of 3 matrices usually denoted as **U**, **Σ,** and $\mathbf{V}^T$ (Van Loan, 1976; Klema & Laub, 1980). These separate matrices are the left singular vectors (**U**), a diagonal matrix **Σ** of the singular values (which has its values sorted in decreasing), and the right singular vectors **V**. From these matrices, the underlying structure or “concepts” can be distilled. For example, if a data set with viewer scores of a group of movies were to be decomposed, one could expect movies to be clustered into genre (science fiction or comedies getting higher or lower scores due to viewer preference) and the viewers into their preference (viewers who love science fiction or comedies). Each of these groups would be represented by one column in the singular vectors matrix and the number of columns in the matrix would be the total number of discernible concepts. However, as the vectors are orthogonal, later columns might just represent noise and no clear concept will discernable. The singular values associated with these columns will be small.

One of the characteristics of the MELODIC algorithm is that a pre-determined dimensionality is required. This pre-determined dimensionality reflects the researchers' idea of how many concepts the data is expected to map onto based on the research hypothesis. However, the dimensionality of the solution might not always be what the researcher expects and post-hoc confirmation might be desirable. Also, exploratory investigations might want to find the dimensionality of data set and not apply a pre-determined expectation. Therefore, an added feature is proposed to the MELODIC algorithm which lets the algorithm find the dimensionality of the data set by itself. The determination of the dimensionality of the data set for the MELODIC algorithm is done by regularization of the Singular Values (Groenen, & Josse, 2016; Gavish & Donoho, 2017; Candes et al., 2013; Josse & Sardy, 2016).

Penalty terms are used to constrain or shrink estimates towards zero. This shrinking can significantly reduce the variance of the estimates and can also be used for feature selection. Ridge regression and Lasso are well-known techniques that make use of penalties (James et al, 2013). In the MELODIC algorithm, regularization will be used to find the optimal number of dimensions or "concepts" hidden within the data set. The optimal number of dimensions will be found by applying a penalty term to the singular values, slowly reducing them to zero and, by extension, reducing the size of the matrices $\mathbf{U}$, $\mathbf{\Sigma}$ and $\boldsymbol{V}^T$. The 4 proposed methods are split into soft-thresholding or hard-thresholding and are defined as follow:

1) Soft-threshold:

$$\Sigma_s = (\Sigma_s \ - \ \lambda)_+, \tag{1.9}$$

where Σ_s are the singular values, λ the penalty parameter, and taking only the results that are positive, setting the rest to zero:

$$(\Sigma_s \ - \ \lambda)_+ = \begin{cases} \Sigma - \lambda & if \ \Sigma > \lambda \\ 0 & if \ \Sigma \leq \lambda \end{cases}. \tag{1.10}$$

2) Soft-threshold of the logarithm of the singular values:

$$\Sigma_s = \exp\left((\log(\Sigma_s + 1) - \ \lambda)_+ - 1\right). \tag{1.11}$$

3) Hard-threshold:

$$\Sigma_s = \Sigma_s \mathbb{1}(\Sigma_s > \lambda), \tag{1.12}$$

where $\mathbb{1}$ is the indicator function.

4) Hard-threshold of the logarithm of the singular values:

$$\Sigma_s = \exp\left(\Sigma_s \mathbb{1}(\log(\Sigma_s + 1) > \lambda) - 1\right). \tag{1.13}$$

In this Thesis, a simulation study will be performed to test the efficacy of the 4 proposed methods of regularization under different data characteristics. The study will try to answer the following research questions:

1) Does the MELODIC algorithm with the regularization of the singular values select the correct dimensionality of the data?
2) Which of the proposed regularization methods works best to select the correct dimensionality under differing numbers of predictor variables and outcome variables?
3) If the algorithm does not select the correct dimensionality, does it select a dimension that is too high, or too low?

Methods

A Monte-Carlo simulation will be used to compare the new regularization features of the MELODIC method algorithm to each other. The answer to the question of which of the four proposed regularizations works best may depend on the data characteristics. These characteristics include the number of participants, the number of response variables, their correlational structure, signal-to-noise ratio (meaningful output to background noise (data with no predictive power)) and collinearity of the predictor variables (Breiman & Friedman 1997). This study, however, will focus only on varying the sample size, number of predictor variables and outcome variables. As the 1997 study of Breiman and Friedman compared multivariate data analysis techniques to each other, just as this study will do, 9 populations will be generated through the method described in the Breiman and Friedman (1997) paper. However, as the method described is for continuous outcome variables and this study investigates dichotomous variables, an adjustment was made to turn the outcome variables into dichotomous outcome variables.

The data was generated through the formula

$$y_{iq} = \sum_{p=1}^{P} (\beta_{pq} x_{ip}) + \epsilon_{iq}, \tag{2.1}$$

where y_{iq} is the response variable for subject i, β_{pq} the coefficient for predictor variable x_{ip} and ϵ_{iq} the error term.

Each of the predictors was independently drawn from a normal distribution with mean zero

$$\mathbf{x}_i \sim N(\mathbf{0}, \mathbf{\Theta}), \tag{2.2}$$

with $\mathbf{\Theta}$ being the P by P covariance matrix and was randomized once (and kept the same for all populations) through

$$\theta_{oh} = r^{|o-h|}, \tag{2.3}$$

where o is the row number and h is the column number and the absolute result was taken from their subtraction. r represents the actual correlation and was randomly drawn once (and used for all the generated data sets) from a uniform distribution between -1 and 1.

$$r \sim U[-1, 1]. \tag{2.4}$$

In this specific study r was randomly chosen to be 0.3581868.

To determine the coefficients , the following formula was used

$$\beta_{pq} = \sum_{k=1}^{10} d_{ok}\, g(j, k), \tag{2.5}$$

with d_{ok} being randomly sampled coefficients from a normal distribution with Q dimensions

$$\{d_{ok}\}_{q=1}^{Q} \sim N(\mathbf{0}, \mathbf{\Gamma}). \tag{2.6}$$

So, a normal distribution with mean **0** and Q by Q covariance matrix **Γ** that determines the degree of correlation through parameter ρ in formula

$$\gamma_{oh} = \rho^{|o-h|}. \tag{2.7}$$

For this study, ρ was set to 0.1, to a low value, so a high correlation would not interfere with the effect of the number of predictor variable on the algorithm to find the correct dimensionality. $g(j, k)$ was used to randomize "peaks" in the coefficient matrix and is defined as

$$g(j, k) = h_k(l_k - |j - j_k|)_+^2. \tag{2.8}$$

j_k and l_k are randomly sampled integers from a uniform distributions of ranges [1, 50] and [1, 6], respectively these "peaks" are centered around point j_k (the highest point) and have a slope distance of l_k in which the peak is reduced to 0 again. The peaks, from $l_k - j_k$ to $j_k + l_k$ were normalized so that it's sum was equal to 1, that is,

$$\sum_{j=1}^{50} g(j,k) = 1. \tag{2.9}$$

$g(j,k)$ was selected only once and kept equal over all data sets.

To determine the signal to noise ratio of the data set, the covariance matrix **F** was set created

$$f_{oh} = \sqrt{\sigma_{pq}^2}, \tag{2.10}$$

with σ_{pq}^2 being the variance for each predictor variable for each value of q. **F** was then used to sample the errors from a normal distribution

$$\{\epsilon_i\}_1^Q \sim N(\mathbf{0}, \mathbf{F}). \tag{2.11}$$

The signal-to-noise ratio was set to 1.

To turn the outcome variables into dichotomous variables, the outcome variables were split into quintiles and a number was randomly selected from a uniform distribution between .2 and .8. for each of the outcome variables. The values of the outcome variables that fell into the quintile equal to or higher than the randomly selected number were then converted to 1 and the values falling in the lower quintiles were converted to 0.

Nine populations of 100.000 subjects were simulated with a signal-to-noise ratio of 1, and a correlation among the predictor variables of 0.35. The populations had differing number of predictor variables (10, 20 and 30) and outcome variables (7, 15 and 30), giving 9 different populations in total.

To pre-set the dimensionality of the populations, the MELODIC algorithm was run on each of the 9 populations, using dimensionality 2. Through these runs, probabilities of a subject belonging to either group 1 or 0 were estimated. The estimated probabilities were then used to generate new outcome variables for each population. These new outcome variables would set the solution for the regularization algorithm for each of the 9 populations equal to the pre-selected dimensionality of 2. In the study the original predictor variables were used in combination with the new outcome variables.

Finally, the MELODIC algorithm with the regularization features was run on subject samples of 3 differing sizes (500, 100, 1500). They were run 100 times for each of the 27 situations starting from a dimensionality of 7 as that was the maximum number of possible dimensions for the population with the lowest number of outcome variables. Each iteration included a 5-fold cross validation, dividing the sample in 5 equal parts and leaving one of those parts as a validation set. This was repeated 10 times for each run and the final dimensionality selected by the algorithm is the dimensionality with which the algorithm found the lowest total deviance. These final dimensionalities were registered to check how often the algorithm selects the correct dimensionality out of the 100 runs.

Results

To compare the results, tables were made for each of the situations and each situation was then split into the final result of the algorithm (R) and a 1 standard error more parsimonious solution (1SE). The results are shown in tables 1 to 3. Each table shows the result of a different sample size. As is clear in all 3 tables, the hard-threshold outperforms the soft-threshold over all situations.

Within the soft-thresholding, in all three tables, the thresholding on the singular values outperforms the soft-threshold on the logarithm of the singular values. And for the soft-threshold on the singular values, taking the 1 standard error seems to perform slightly better. In some of the situations the 1SE was able to get the correct dimensionality, but more often than not, a higher dimensionality was selected by the algorithm. Neither of the 2 forms of soft-thresholding is able to consistently reduce the dimensionality all the way to the correct dimensionality 2 and they seem to get stuck in higher dimensionalities. The soft-thresholding on the logarithm is not able to get lower than the starting value of dimensionality 7. The soft-thresholding also seems to perform worse as the value of Q goes up whereas the value of P appears to have little impact on the performance of the soft-thresholding algorithm.

Table 1

Selected dimensions with n = 500. The first letter represents either the regularization on the singular values (S) or on the logarithm of the singular values (L). The second letter represents either soft-thresholding (S) of hard-thresholding (H). Finally either the selected dimensionality (R) was taken or a model selecting 1 standard error more parsimonious (1SE).

P = 10

	Q = 7							*Q* = 15							*Q* = 30						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7
SS R				2	3	2	93					1	2	97							100
SS1SE		2	3	2	93						3	6	91						2	4	94
LS R							100							100							100
LS1SE							100						2	98							100
SH R		100						1	98	1						100					
SH1SE		100						1	98	1						100					
LH R		99	1						99		1					100					
LH1SE		100							100							100					

P = 20

	Q = 7							*Q* = 15							*Q* = 30						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7
SS R					3	95	2							100							100
SS1SE			1	4	94	1					1	2	96	1						2	98
LS R							100							100							100
LS1SE							100							100							100
SH R		100							100							100					
SH1SE		100							100							100					
LH R		100							100							100					
LH1SE		100							100							100					

P = 30

	Q = 7							*Q* = 15							*Q* = 30						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7
SS R					5	5	90							100							100
SS1SE			1	5	94						1	3	96						1		99
LS R							100							100							100
LS1SE							100							100							100
SH R		100							100							100					
SH1SE		100							100							100					
LH R		100							100							100					
LH1SE		100							100							100					

Table 2

Selected dimensions with n = 1000. The first letter represents either the regularization on the singular values (S) or on the logarithm of the singular values (L). The second letter represents either soft-thresholding (S) of hard-thresholding (H). Finally either the selected dimensionality (R) was taken or a model selecting 1 standard error more parsimonious (1SE).

P = 10

	Q = 7							*Q* = 15							*Q* = 30						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7
SS R			2	1		4	93			1				99							100
SS1SE		6	2	90	2			1	1			1	5	92						3	97
LS R							100							100							100
LS1SE							100							100							100
SH R		100							99		1					100					
SH1SE		100						1	99							100					
LH R		100							99		1					100					
LH1SE		100							100							100					

P = 20

	Q = 7							*Q* = 15							*Q* = 30						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7
SS R						2	98							100							100
SS1SE			1	92	4	3						3	93	4						1	99
LS R							100							100							100
LS1SE							100							100							100
SH R		100							100							100					
SH1SE		100							100							100					
LH R		100							100							100					
LH1SE		100							100							100					

P = 30

	Q = 7							*Q* = 15							*Q* = 30						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7
SS R					1	94	5							100							100
SS1SE			1	93	6					1	1	92	4	2						93	7
LS R							100							100							100
LS1SE							100							100							100
SH R		100							100							100					
SH1SE		100							100							100					
LH R		100							100							100					
LH1SE		100							100							100					

Table 3

Selected dimensions with n = 1500. The first letter represents either the regularization on the singular values (S) or on the logarithm of the singular values (L). The second letter represents either soft-thresholding (S) of hard-thresholding (H). Finally either the selected dimensionality (R) was taken or a model selecting 1 standard error more parsimonious (1SE).

P = 10

	Q = 7							*Q* = 15							*Q* = 30						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7
SS R			1		1	2	96							100							100
SS1SE		1	91	1	4	3						1	92	7				1	2	94	3
LS R							100							100							100
LS1SE							100							100							100
SH R		100						2	98							100					
SH1SE		100						2	98							100					
LH R		100						1	99							100					
LH1SE		100						1	99							100					

P = 20

	Q = 7							*Q* = 15							*Q* = 30						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7
SS R					1	5	94							100							100
SS1SE		2	2	3	90	2	1				2	2	91	5							100
LS R							100							100							100
LS1SE							100							100							100
SH R		100							100							100					
SH1SE		100							100							100					
LH R		100							100							100					
LH1SE		100							100							100					

P = 30

	Q = 7							*Q* = 15							*Q* = 30						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7
SS R						96	4							100							100
SS1SE			2	91	6	1			1		92	5	1	1					2	1	97
LS R							100							100							100
LS1SE							100							100							100
SH R		100							100							100					
SH1SE		100							100							100					
LH R		100							100							100					
LH1SE		100							100							100					

For the hard-thresholding, all variants managed to consistently select the correct dimensionality. Only a few times did the algorithm fail to do so. The best performer overall is the hard-thresholding on the logarithm of the singular values and selecting the 1SE result. This variant had only 1 instance that it did not select the correct dimensionality but instead selected a dimensionality lower (dimensionality 1). Because of the low number of wrongly selected dimensionalities it is hard to discern a direction to which mistakes would be made by the algorithm. Also, the few mistakes that were made are split in both directions and are never more than 1 dimension away from the correct one. All the mistakes made by the hard-thresholding algorithm however, seem to have been made in the situations with lower values of *P* and none of the mistakes have been made with the largest number of *Q*.

Discussion

In this paper four different methods for regularization within the MELODIC family of models were tested. From the results we can conclude that the two soft-thresholding variants do not consistently select the correct dimensionality of the data set under the tested conditions. The hard-thresholding, on the other hand, performed well and was able to select the correct dimensionality in over 99% of the runs. The hard-thresholding on the logarithm of the singular values performed best overall, which was only marginally better than the thresholding directly on the singular value, and based on these results is recommended for future use. However, the algorithm was not able to give a clear distinction between regularization on the singular values or regularization on the logarithm of the singular values, as was the case in the soft-thresholding.

The soft-thresholding was in many cases not able to lower the dimensionality from the starting point (dimensionality of 7) and was only rarely able to reduce it as far as the correct dimensionality of 2. In the hard-thresholding no specific direction in which it consistently selects the wrong dimensionality was observed. Although one dimension higher (3) is slightly more common in the results than one dimension lower (1). However due to the small number of wrongly chosen dimensionalities, concluding that the hard-thresholding overestimates the dimensionality seems unwarranted.

The differing sample sizes, numbers of predictors, and number of outcome variables seem to have a larger effect on the results of the soft-thresholding. The algorithm seems to perform better on data sets with low numbers of outcome variables. As the number of outcome variables in the data set increased, the less likely it was for the algorithm to select a dimensionality lower than the starting point of 7. For the hard-thresholding, however, the number of outcome variables seemed to have less of an impact. Here, the number of predictor variables seems to slightly influence the ability to select the correct dimension as all mistakes were made in the samples with the lowest number of predictor variables.

The method used to generate the data was advantageous for the study as it allowed for specific adjustments within the characteristics of the data, keeping many characteristics of the data the same. This allowed us to observe only the effect of the tested data characteristics of sample size, number of predictor variables and number of outcome variables. To further investigate the regularization features, other aspects of a data set could be investigated. Some examples of possible data aspects are the formerly mentioned signal-to-noise ratio, the collinearity, correlational structure and the true dimensionality. A higher collinearity could for example cause less accurate model estimations and a higher model variance, making it harder for the model to accurately select the correct dimensionality. Varying these data features could potentially create a distinction between the regularization of the singular values or their logarithm. The way the data was generated, however, might not represent real data that could be encountered in the field and testing on real data could further show the effectiveness of the tested regularization methods. For now, however, the hard-thresholding on the logarithm of the singular values and taking a 1 standard error more parsimonious solution appears to be the most effective method, a result that could potentially be generalizable to other situations as the hard-thresholding consistently outperformed the soft-thresholding over all tested situations.

Acknowledgment

I would like to thank prof. Dr. M. de Rooij and the Methodology & Statistics department at Leiden University for the support in writing this thesis.

References

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, *2*(4), 433-459.

Anderson, J. A. (1984). Regression and ordered categorical variables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *46*(1), 1–22.

Breiman, L. and Friedman, J. H. (1997). Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *59*(1), 3–54.

Busing, F. M. T. A. (2010). *Advances in multidimensional unfolding*. Doctoral thesis, Leiden University.

Candes, E. J., Sing-Long, C. A., & Trzasko, J. D. (2013). Unbiased risk estimates for singular value thresholding and spectral estimators. *IEEE transactions on signal processing*, *61*(19), 4643-4657.

Dehuri, S., & Sanyal, S. (2015). *Computational Intelligence for Big Data Analysis*. Springer International Publishing.

de Rooij, M. (2009). Ideal point discriminant analysis revisited with a special emphasis on visualization. *Psychometrika*, *74*(2), 317 – 330.

de Rooij, M. and Groenen, P.J.F. (2020). The MELODIC family for simultaneous binary logistic regression in a reduced space. https://arxiv.org/pdf/2102.08232.pdf

de Rooij, M. and Heiser, W. J. (2005). Graphical representations and odds ratios in a distance-association model for the analysis of cross-classified data. *Psychometrika*, *70*(1), 99–122.

Fourdrinier, D., Strawderman, W. E., & Wells, M. T. (2018). *Shrinkage estimation*. Cham, Switzerland: Springer International Publishing.

Gavish, M., & Donoho, D. L. (2017). Optimal shrinkage of singular values. *IEEE Transactions on Information Theory*, *63*(4), 2137-2152.

Groenen, P. J. F., & Josse, J. (2016). *Multinomial multiple correspondence analysis*. https://arxiv.org/abs/1603.03174

Guyon, I., Gunn, S., Nikravesh, M., & Zadeh, L. A., 2008, *Feature extraction: foundations and applications* (Vol. 207). Springer.

Holst, A. (2021). *Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2025*. Retrieved from https://www.statista.com/statistics/871513/worldwide-data-created/

Huberty, C. J., & Morris, J. D. (1992). Multivariate analysis versus multiple univariate analyses. In A. E. Kazdin (Ed.), Methodological issues & strategies in clinical research (pp. 351–365). *American Psychological Association*. https://doi.org/10.1037/10109-030

Jain, A., & Zongker, D. (1997). Feature selection: Evaluation, application, and small sample performance. *IEEE transactions on pattern analysis and machine intelligence*, *19*(2), 153-158.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning,* New York: springer.

Jolliffe, I. T. (2002). *Principal Component Analysis.* Springer.

Josse, J., & Sardy, S. (2016). Adaptive shrinkage of singular values. *Statistics and Computing*, *26*(3), 715-724.

Klema, V., & Laub, A. (1980). The singular value decomposition: Its computation and some applications. *IEEE Transactions on automatic control*, *25*(2), 164-176.

Krishnaiah, P.R., & Kanal, L.N. (1982). Classification, pattern recognition and reduction of dimensionality*, Handbook of statistics; vol. 2*, Elsevier Science Pub.

Mayya, S. S., Monteiro, A. D., & Ganapathy, S. (2017). Types of biological variables. *Journal of thoracic disease*, *9*(6), 1730–1733.

Nevatia, R., & Babu, K. R. (1980). Linear feature extraction and description. *Computer Graphics and Image Processing*, *13*(3), 257-269.

Pearson, K. (1901). Principal components analysis. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *6*(2), 559.

Pudil, P., & Novovičová, J. (1998). Novel methods for feature subset selection with respect to problem knowledge. In *Feature extraction, construction and selection* (pp. 101-116). Boston, MA: Springer.

Shreffler, J. & Huecker, M.R. *Types of Variables and Commonly Used Statistical Designs*. (Updated 2021 Mar 1) Available from: https://www.ncbi.nlm.nih.gov/books/NBK557882/

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley Symposium on mathematical Statistics and Probabilities.* Univ. of California Press, pp. 197-206

Takane, Y. (1987). Analysis of contingency tables by ideal point discriminant analysis. *Psychometrika*, *52*(4), 493–513.

Takane, Y., Bozdogan, H., & Shibayama, T. (1987). Ideal point discriminant analysis. *Psychometrika*, *52*(3), 371–392.

Van Loan, C. F. (1976). Generalizing the singular value decomposition. *SIAM Journal on numerical Analysis*, *13*(1), 76-83.

Worku, H. M. and De Rooij, M. (2018). A multivariate logistic distance model for the analysis of multiple binary responses. *Journal of Classification*, *35*(1), 124– 146.