A Dimensional Approach for Multivariate Discrete Time Survival Data

Fatma ÖZBAY

Thesis advisor: Professor Dr. Mark de Rooij

Defended on 18th *August*, 2023

MASTER THESIS

STATISTICS AND DATA SCIENCE
UNIVERSITEIT LEIDEN

Abstract

The dimensional approach to data sets aims to provide a more accurate interpretation of the relationships between variables and to reveal the underlying patterns among them.

In this thesis, the survival data set, consisting of childhood adversities (CAs) and correlated adult psychiatric disorders, was analyzed using the Logistic Reduced Rank Regression (LRRR) approach. This method allows dimensional investigation of CAs and disorders while considering the correlations between the disorders.

The data set was also analyzed using the traditional Logistic Regression (LR) approach, and the results showed that the LRRR outperforms LR by providing more information about the data set. It was observed that parental mental illness (PMI) and physical illness (PIllness) adversities experienced during childhood strongly influence the development of disorders, as evident with both LR and LRRR approaches. However, since LR assumed the same effect of these adversities for each disease, it failed to identify which disorders were more affected by them. The effects of PMI and PIllness were found to have a greater impact on the development of posttraumatic stress disorder (PTS) when analyzed using the LRRR approach.

This dimensional approach, which provides more information on the data set, does have limitations. Biplots, used for dimensional analysis, are easier to interpret in two-dimensional models. However, when dealing with high-dimensional models, constructing the biplots with pairwise dimensions becomes necessary, which in turn makes the simultaneous examination of biplots challenging. Furthermore, the missing data in the data set must follow the assumption of being missing completely at random (MCAR). If the missing data does not meet this assumption, it needs to be addressed through advanced techniques such as multiple imputation. Failure to handle the missing data appropriately may lead to limitations in the effectiveness of this method. In future research, cross-validation can be employed to assess the generalization ability of the model and enhance related analyses.

Keywords: Dimensional approach, LRRR, LR, CAs, adult psychiatric disorders

Contents

1 Introduction

“The statistical method is more than an array of techniques. The statistical method is a mode of thought; it is sharpened thinking; it is power,” said the famous American engineer and statistician Deming (1953). This quote highlights the significance of statistical methods. It emphasizes that statistical methods require more than merely employing algorithms or formulas but also developing an analytical approach to comprehend and interpret given data. Furthermore, this quote points out that using a method allows individuals to gain deeper insights into patterns and even reveal hidden patterns, which is a great power.

In implementing a statistical method, it is critical to select an appropriate analytical approach to benefit from its effectiveness and power; otherwise, improper selection of approach can lead to incorrect conclusions, inaccurate interpretation of results, and so manipulate or even lead to loss of information (Khusainova et al., 2016). Therefore, employing suitable statistical techniques to extract meaningful results and uncover knowledge from data is highly topical and vital.

Data analysis techniques differ according to both research objectives and the nature of the variables (Hastie et al., 2009). When analysing time-to-event (survival) data, where the time variable indicates the time until occurring a specific event, the Kaplan-Meier Estimator (Kaplan and Meier, 1958), a nonparametric method, or Cox Proportional Hazards Model (Cox, 1972), a semiparametric method, are commonly used. However, when exploring the relationship between predictors and the risk of occurring the event at any certain time, the Kaplan-Meier method is not appropriate, and instead, the Cox model is preferred.

Furthermore, when the time variable is discrete, partial logistic regression (pLR) can be used to fit parametric survival curves to the data (Efron, 1988) to investigate the associations between the predictors and the probability of event occurrence. This parametric method makes parameter estimates more precise by using the data more effectively. On the other hand, it is referred to as partial logistic regression because it has a connection with partial likelihood estimation (Cox, 1975), which indicates that the likelihood function is determined by the order of the event times instead of the distribution of event times. However, since the application of pLR is the same as standard logistic regression (LR), the term LR will be used instead of pLR henceforth throughout this thesis.

LR is also widely used when multiple response variables or events exist in the data set. However, De Rooij (2023) notes that many researchers fail to consider the multivariate nature of the response variables, leading them to use univariate multiple models that analyse each response variable separately—for instance, independently applying LR for each response variable. This approach ignores the possibility of response variable correlations and offers no insight into the connection between the response variables. To gain a more complete understanding of the underlying relationships, the multivariate method known as Reduced Rank Regression (RRR) has been suggested as a technique for predicting all responses from a set of predictors simultaneously (Anderson, 1951; Izenman, 1975; Tso, 1981; Davies and Tso, 1982).

In the case of multiple correlated response variables, RRR is used to provide a more parsimonious model (Ter Braak and Looman, 1994) by restricting the rank of the regression coefficient matrix (Anderson, 2003) while maintaining the relationship between responses. Moreover, if response variables are dichotomous, Logistic Reduced Rank Regression (LRRR) analysis is used instead of RRR. While this method describes the relationship between variables, it also allows for dimensional interpretation of the associations. There is a gap in the literature where a multivariate discrete-time survival data set has yet to be evaluated dimensionally using LRRR, creating a research opportunity to address this limitation.

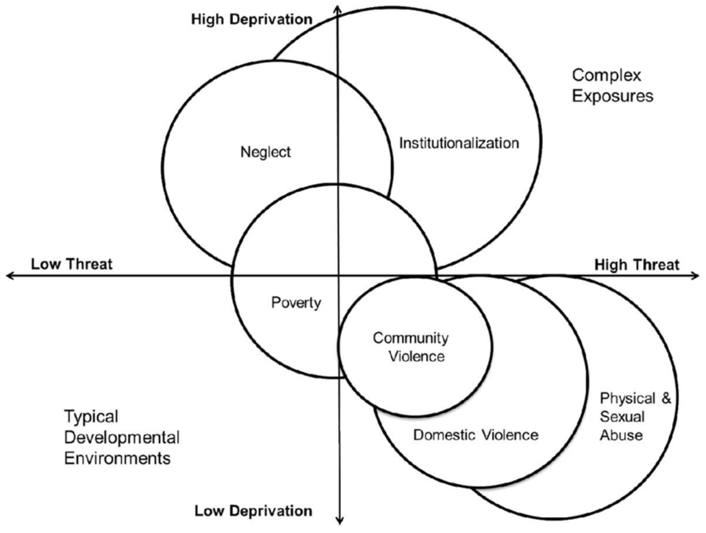
Therefore, the main goal of this thesis is to assess the multivariate discrete-time survival data set dimensionally using LRRR. The primary motivation in this study is the work of McLaughlin and Sheridan (2016) on the dimensional approach to childhood adversities (CAs), which has not been statistically tested.

In the work of McLaughlin and Sheridan (2016), childhood experiences involving threats, such as community violence, domestic violence, and physical/sexual abuse, as well as childhood conditions that cause the absence of expected cognitive inputs, social stimulation, and consistent interactions, such as poverty, neglect, and institutional rearing were examined to assess their effects on the developmental process of children.

As illustrated in Figure 1, these CAs are represented in two dimensions whose axes are threat and deprivation. This dimensional approach model was developed by analysing the frequency and severity of these CAs, and obtained results were reflected in the dimensions with circles. Larger circles show more significant variation in how well a CA fit the underlying dimensions (deprivation and threat). Thus, it can be said that institutionalization and physical/sexual abuse had the most significant influence on deprivation and threat, respectively, although poverty did not intrinsically entail aspects of both threat and deprivation (McLaughlin et al., 2014).

Figure 1

Dimensional Approach to the Childhood Adversities



Note. A two-dimensional model of childhood adversity with two primary dimensions: threat and deprivation. Adversities are classified across these dimensions depending on how frequently they include threat and deprivation. Adapted from “Beyond Cumulative Risk: A Dimensional Approach to Childhood Adversity,” by McLaughlin and Sheridan (2016).

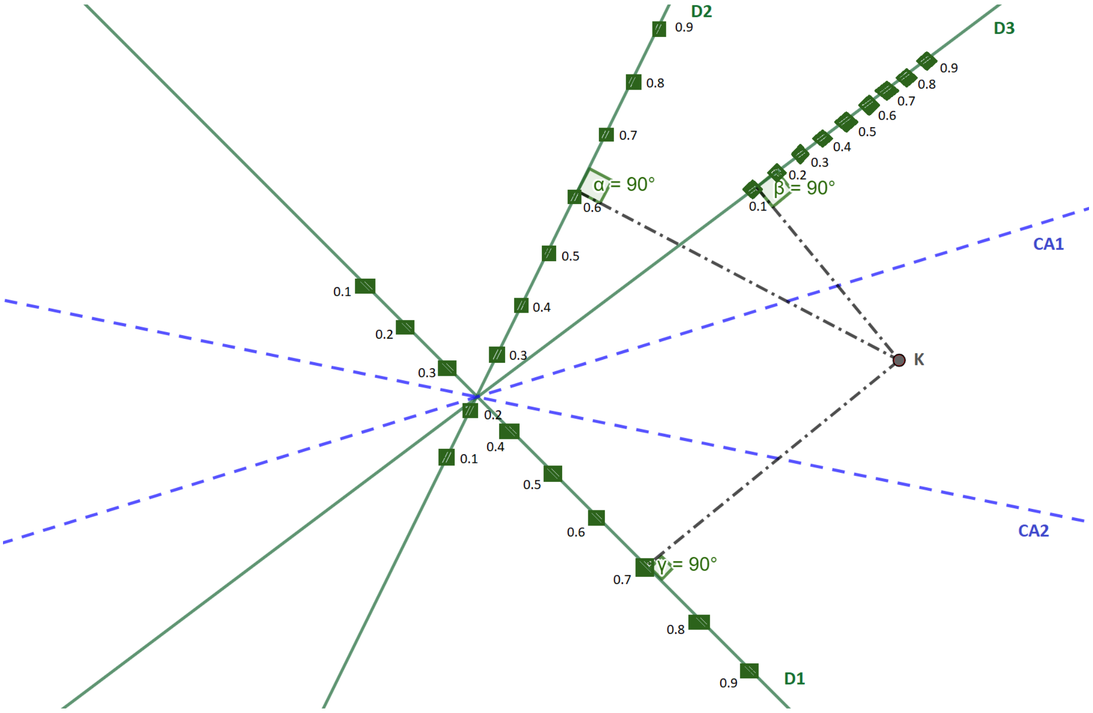
The dimensional approach by McLaughlin and Sheridan (2016) attempted to distinguish the CAs from each other and to determine the extent to which CA impact developmental outcomes like psychopathology and chronic diseases. This identification is critical for creating effective strategies to reduce the negative developmental effects caused by CAs.

In this thesis, the publicly released survey data set, which was also used in the research Green et al. (2010) focusing on CAs and adult psychiatric disorders, is employed to evaluate the proposed dimensional approach to CAs statistically. The data is initially analyzed using the traditional LR method, as done by Green et al. Subsequently, a dimensional analysis is performed using the LRRR method.

Biplots are utilised to visually interpret the LRRR analysis results, as they effectively summarize essential components and facilitate the dimensional investigation (Ter Braak and Looman, 1994). This study uses biplots of the type represented in Figure 2 to indicate results.

Figure 2

Biplot Representation for Two-Dimensional LRRR Model



Note. Blue dashed lines depict childhood adversities (CAs) labelled as CA1 and CA2, while disorders assumed to be influenced by these CAs are represented by green solid lines labelled as D1, D2 and D3. The blue and green lines are also called the predictor and response lines, respectively. The grey “K” point indicates the observation at the interpolated location, obtained as the convergence point of the predictor lines. Green markers represent the probabilities of disorders with “yes” corresponding to the perpendicular lines from the “K” point. This figure is just a representation and does not reflect the actual data and values.

In Figure 2, the probability values corresponding to the perpendiculars drawn from the observation to the response lines are examined to determine the probabilities of this observation experiencing each disorder. Disorders with higher probability values are more likely to be present.

Furthermore, it is possible to determine which CAs are associated with which disorders by evaluating the angle between predictor and response lines. The lines with acute angles indicate a strong relationship between CAs and disorders. Conversely, when the angle is right, they may not be related to each other, and if the angle is obtuse, they could have a negative relationship. Additionally, the disorders can be evaluated among themselves based on the situation between the angles, allowing for categorising disorders with similar or different characteristics (De Rooij, 2023).

As a result, using the dimensional approach obtained by using LRRR, experts can determine the association between specific disorders and CAs and develop treatment approaches accordingly. This statistical method enables them to take targeted action against these disorders based on the strength of their association with CAs.

In summary, this thesis aims to test a proposed dimensional approach on the multivariate discrete-time survival data set, which includes CAs and adult psychiatric disorders, using LRRR. In addition, the traditional LR approach, commonly used for such data sets, will be conducted and compared with the LRRR approach. The analysis also aims to examine the differences between these two approaches and assess whether the new approach, LRRR, offers a more comprehensive explanation of observed patterns and findings by retaining the multivariate nature of correlated response variables.

Considering the thesis's structure, Section 2 will provide details regarding the source of the data set, its structural characteristics, and the formats in which it is converted for the analyses. Additionally, the analysis methods employed in this study will be explained theoretically. In Section 3, the conducted analyses will be presented step by step, and the results will be interpreted. Finally, in Section 4, a general conclusion and discussion will be provided to summarize the key findings and implications of the study.

2 Materials and Methods

2.1 Data Source

The Collaborative Psychiatric Epidemiology Surveys (CPES, Alegria et al., 2016) were created to meet the demand for current and thorough epidemiological data on the prevalence, risk factors for, and variables associated with, mental diseases in the general population, particularly among minority groups.

The primary goal of CPES was to compile data on the prevalence of mental illnesses, how they affect people's functioning, and the treatment trends across representative samples of adult populations in majority and minority groups in the United States. Additionally, CPES examined language use, racial inequalities, support networks, prejudice, and assimilation to research the links between mental health illnesses and social and cultural aspects.

Three nationally representative surveys—the National Comorbidity Survey Replication (NCS-R), the National Survey of American Life (NSAL), and the National Latino and Asian American Study (NLAAS)—were integrated by CPES to accomplish these goals.

In this thesis, the publicly released NCS-R survey data set was used, which was collected around ten years after the first NCS-1 survey in 1992. The NCS-R expanded the scope of inquiry to include assessments in line with the diagnostic criteria stated in the Diagnostic and Statistical Manual - IV (DSM-IV) of 1994 and also repeated many questions from the NCS-1. The data set and further information for the NCS-R can be obtained from Alegria et al. (2016).

2.2 Data Structure

The target population of the NCS-R survey was English-speaking adults of at least 18 years old living in the contiguous United States. The survey was divided into two parts: Part 1 featured a core diagnostic evaluation of each of the 9,282 respondents, and Part 2 asked questions concerning risk factors, repercussions, other diseases, linked variables, and associated factors. Only 5,692 of the Part 1 candidate received the Part 2 assessment. Between February 2001 and April 2003, the interviews were performed mainly at the respondents' homes utilising laptop computer-assisted personal interview techniques (Alegria et al., 2016).

The dataset from the Part 2 evaluation, which included 5,692 individuals, was used in this thesis. It consists of 20 DSM-IV correlated disorders, 9 childhood adversities (CAs), and participants' age and gender information.

Additionally, since some observations had missing values in the data, these observations were excluded, and only complete cases, which are observations with no missing values, were considered for analyses. The total number of observations remaining was 5,381.

These 20 DSM-IV disorders are as follows: Major depressive disorder, dysthymic disorder, bipolar I, bipolar II, subthreshold bipolar, panic disorder, agoraphobia without a history of panic disorder, generalised anxiety disorder, specific phobia, social phobia, posttraumatic stress disorder, separation anxiety disorder, intermittent explosive disorder, attention-deficit/hyperactivity disorder, oppositional-defiant disorder, conduct disorder, alcohol abuse, alcohol dependence with abuse, drug abuse, and drug dependence with abuse.

Additionally, 9 dichotomous CAs experienced before the age of 18 years are as follows: Separation from parents or caregivers, excluding cases of parental death or parental divorce (respondent under foster care), parental mental illness, parental substance abuse, parental criminality, parental violence, physical abuse, neglect, life-threatening childhood physical illness and extreme childhood family economic adversity.

Table 1 illustrates the representation of the data set in a person-level format with a multivariate structure. Twenty disorders serve as response variables, with their values representing each patient's onset age of the related disorders. Additionally, 9 CAs are considered predictor variables, indicating whether participants experienced these CAs (i.e., 1) or not (i.e., 0). In the data set, there are also the gender (i.e., 1 for male; 0 for female) and age at the interview of the participants, which are also used as predictor variables.

Table 1

Person-Level Data Set

CaseID	**Gender**	**Age**	D_1	D_2	**...**	D_{20}	CA_1	CA_2	**...**	CA_9
1	1	40	6	7	...	NA	1	0	...	1
2	0	72	NA	NA	...	NA	0	0	...	1

Note. D_1, ..., D_{20} abbreviations refer to 20 disorders, and CA_1, ..., CA_9 refer to 9 childhood adversities. It should be noted that the values in the table do not reflect real values; they are just a representation to understand the data structure.

The person-level data set, where each person is represented in one row, needed to be converted into a suitable format because this structure was insufficient to analyse the outcome changes over time (Singer and Willett, 2003). Therefore, the data set was transformed into a person-period data set as indicated in Table 2 and Table 3 for LR and LRRR, respectively.

Table 2

Stacked Person-Period Data Set for LR

CaseID	**Gender**	**Age**	**Event**	**Disorder**	CA_1	CA_2	**...**	CA_9
1	1	4	0	D_1	1	0	...	1
1	1	5	0	D_1	1	0	...	1
1	1	6	1	D_1	1	0	...	1
1	1	7	NA	D_1	1	0	...	1
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
1	1	40	NA	D_1	1	0	...	1
1	1	4	0	D_2	1	0	...	1
1	1	5	0	D_2	1	0	...	1
1	1	6	0	D_2	1	0	...	1
1	1	7	1	D_2	1	0	...	1
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
1	1	40	NA	D_2	1	0	...	1
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
1	1	4	0	D_{20}	1	0	...	1
1	1	5	0	D_{20}	1	0	...	1
1	1	6	0	D_{20}	1	0	...	1
1	1	7	0	D_{20}	1	0	...	1
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
1	1	40	0	D_{20}	1	0	...	1
2	0	4	0	D_1	0	0	...	1
2	0	5	0	D_1	0	0	...	1
2	0	6	0	D_1	0	0	...	1
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
2	0	72	0	D_1	0	0	...	1
2	0	4	0	D_2	0	0	...	1
2	0	5	0	D_2	0	0	...	1
2	0	6	0	D_2	0	0	...	1
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
2	0	72	0	D_2	0	0	...	1
.	.	.	.	.	.	...	.	.
.	.	.	.	.	.	...	.	.
.	.	.	.	.	.	...	.	.
2	0	4	0	D_{20}	0	0	...	1
2	0	5	0	D_{20}	0	0	...	1
2	0	6	0	D_{20}	0	0	...	1
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
2	0	72	0	D_{20}	0	0	...	1

Note. D_1, ...,D_{20} abbreviations refer to 20 disorders used as the labels of the “Event” response variable, and CA_1, ...,CA_9 refer to 9 childhood adversities used as predictor variables.

Table 3

Person-Period Data Set for LRRR

CaseID	**Gender**	**Age**	D_1	D_2	**...**	D_{20}	CA_1	CA_2	**...**	CA_9
1	1	4	0	0	...	0	1	0	...	1
1	1	5	0	0	...	0	1	0	...	1
1	1	6	1	0	...	0	1	0	...	1
1	1	7	NA	1	...	0	1	0	...	1
.	.	.	.	.	...	.	.	.	...	.
.	.	.	.	.	...	.	.	.	...	.
.	.	.	.	.	...	.	.	.	...	.
1	1	40	NA	NA	...	0	1	0	...	1
2	0	4	0	0	...	0	0	0	...	1
2	0	5	0	0	...	0	0	0	...	1
2	0	6	0	0	...	0	0	0	...	1
.	.	.	.	.	...	.	.	.	...	.
.	.	.	.	.	...	.	.	.	...	.
.	.	.	.	.	...	.	.	.	...	.
2	0	72	0	0	...	0	0	0	...	1

Note. D_1, ..., D_{20} abbreviations refer to 20 disorders used as response variables, and CA_1, ..., CA_9 refer to 9 childhood adversities used as predictor variables.

In creating these longitudinal formats, the age of 4 was considered the earliest age for the beginning of the disorder, which the person-period indicator started. Then, the person-period was censored and classified as "0" up to the age of onset of the disorder. It was also coded as "1" at the beginning of the disorder encountered by the participant. Subsequently, the indicator was recorded as "NA" until the current age. Moreover, if individuals had never experienced any disorder, these 20 response variables were encoded with zero for them.

Additionally, the time-invariant CAs predictors remained constant across all the recordings of a participant, and they were included in the person-period dataset by being coded according to the participant's state of exposure (i.e., 1) or absence (i.e., 0) of these adversities.

The only difference between Table 2 and Table 3 lies in the arrangement of the disorders. In Table 2, the disorders were stacked vertically and presented in a single column, while in Table 3, the disorders were shown in different columns.

2.3 Data Analyses

A person-period data set with multiple binary response variables and discrete periods, such as age, is often analysed using LR (McCullagh and Nelder, 1989, Willett and Singer, 1993, Diggle et al., 2002), one of the traditional methods. Since the data set utilised in this thesis was previously analysed with the LR approach in the work by Green et al. (2010), it was reanalyzed with LR. However, such analyses do not account for the correlations between the response variables. Therefore, the LRRR method was employed, a new approach for this data set, incorporating the association between the variables.

During the data set analysis with LR and LRRR, several models were developed with various selection options. Akaike Information Criterion (AIC, Akaike (1974)) and Bayesian Information Criterion (BIC, Schwarz (1978)) values were calculated to decide on the optimal model among the others. These criteria are defined as

$$\text{AIC} = \text{deviance} + 2 * \text{number of parameters} \tag{1}$$

and

$$\text{BIC} = \text{deviance} + \log(N) * \text{number of parameters}, \tag{2}$$

where deviance shows the value obtained by multiplying the log-likelihood function by -2 and N indicates the sample size. The calculation of the number of parameters differs for LR and LRRR (see Section 2.3.1 and Section 2.3.2). The model with the lowest AIC and BIC values and indicated by both criteria simultaneously was chosen as the optimal model. When these criteria pointed out the different models, the BIC value was taken into because BIC prevents overfitting at large sample sizes (Burnham and Anderson, 2002).

In addition, the predictors were removed individually to determine which predictors should be included in the model. After being removed from the model, the variables causing the most significant decrease in AIC and BIC values were identified, and these predictors were completely excluded from the model.

Following the optimal model selection, odds ratios (ORs) were obtained to evaluate the associations between the predictor and response variables, and confidence intervals (CIs) were generated to offer a sense of the precision or accuracy of the estimated values.

Although similar procedures were applied for both LR and LRRR approaches, different steps were followed while analysing them as they are two different methodologies. Section 2.3.1 and Section 2.3.2 provide detailed explanations for these approaches.

All analyses were performed using R (version 4.1.2), and the corresponding codes are provided in Appendix A.

2.3.1 Traditional Approach: LR for Multivariate Discrete-Time Survival Data

In a time-to-event (survival) data set, LR can be used to model the probability of a binary event based on the predictor variables if the time is discrete (Efron, 1988). If the probability of an event occurrence for the participant i at time j is defined as $\pi_{ij}(\mathbf{x}_{ij}) = P(y_{ij} = 1|\mathbf{x}_{ij})$, the general LR model can be expressed as

$$\text{logit}(\pi_{ij}) = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \mathbf{x}_{ij}^{'}\boldsymbol{\beta},$$

where $\mathbf{x}_{ij}$ is the vector of predictor variables for participant i at time j, $\boldsymbol{\beta}$ is the vector of parameters, and logit is the link function. Further, $\frac{\pi_{ij}}{1-\pi_{ij}}$ is known as the odds of the event occurring, and the logit function is written as the natural logarithm of the odds. $\boldsymbol{\beta}$ parameters are typically estimated by the method of maximum likelihood estimation (MLE), and the logarithm of the likelihood function is defined by (Collett, 2003)

$$\log \text{L}(\boldsymbol{\beta}) = \sum_{i=1}^{N}\sum_{j=1}^{n_i}\left\{y_{ij}\log\pi_{ij} + (1-y_{ij})\log(1-\pi_{ij})\right\},$$

where N is the number of participants, and n_i represents the number of time points for the individual i.

The LR analyses in this thesis were performed using the "glm()" function, with the family set to binomial(link = "logit").

The polynomial structure of the age variable was first explored to determine the optimal LR model. Then, feature selection was performed using the fixed polynomial degree acquired in the previous phase. To identify which predictors should be included in the model, the predictors were systematically removed one by one.

While checking the models to decide on the optimal one, the information criteria values in equations 1 and 2 were computed, with the number of parameters specified by the sum of the number of predictor variables and intercept.

After obtaining the final LR model, ORs were obtained to measure the strength and direction of the association between the predictor variables and the odds of an event occurring. They indicate how much the odds of a binary outcome change when the predictor variable changes from its reference category (e.g., coded as 0) to the alternative category (e.g., coded as 1). An odds ratio of more than one implies an increase in the odds of the event occurring, whilst an odds ratio of less than one points to a drop in the odds of the event happening (Hosmer et al., 2013).

In cases where CIs around the ORs are calculated, a range of possible values is generated. The CI contributes to the accuracy of the odds ratio estimate. A smaller confidence range indicates a more accurate estimate, whereas a bigger interval indicates higher uncertainty. Furthermore, CIs are used to evaluate the odds ratio's statistical significance. As a result, CI values were obtained to assess the significance of estimations.

2.3.2 New Approach: LRRR for Multivariate Discrete-Time Survival Data

LRRR enables a lower-dimensional subspace of variables that captures the shared information among the correlated responses (Reinsel and Velu, 1998). This approach reduces the dimensionality by imposing constraints on the regression coefficients matrix while modelling the multivariate structures.

LRRR is also a modified version of the RRR, used when response variables are dichotomous. Therefore, the RRR model is first examined below to comprehend the LRRR model fully from the regression perspective. As Schmidli (1995) and De Rooij (2023) stated, the RRR models are considered multivariate regression models in that the regression coefficient matrix has reduced rank rather than full rank. Using these two references, the following explanations are provided.

Consider the multivariate linear regression model defined by

$$\mathbf{Y} = \mathbf{1}\mathbf{m}^{'} + \mathbf{X}\mathbf{A} + \mathbf{E}, \tag{3}$$

where $\mathbf{Y}(N \text{ x } R)$ and $\mathbf{X}(N \text{ x } P)$ are matrices of R response and P predictor variables of each of N participants. It is assumed that $\mathbf{X}$ matrix has full rank and is centered, i.e. $\mathbf{X}'\mathbf{1}_N = \mathbf{0}_P$, where $\mathbf{1}$ denotes a vector with ones, and $\mathbf{0}$ denotes a vector with zeros. Additionally, in equation 3, $\mathbf{m}$ symbolises the intercept for R response variables, $\mathbf{A}$ is a P x R matrix containing regression weights, and $\mathbf{E}$ refers to a N x R matrix containing the residuals. The rank of the regression weight matrix is equal to S. This involves applying the constraint, namely $\mathbf{A} = \mathbf{B}\mathbf{V}'$, where $\mathbf{B}$ is P x S matrix and $\mathbf{V}$ is a R x S matrix. Thus, equation 3 gets into

$$\mathbf{Y} = \mathbf{1}\mathbf{m}' + \mathbf{X}\mathbf{B}\mathbf{V}' + \mathbf{E}. \tag{4}$$

If the probability of an event r (response variable) occurrence for the participant i at time j is defined as $\pi_{ijr}(\mathbf{x}_{ijr}) = P(y_{ijr} = 1|\mathbf{x}_{ijr})$, equation 4 becomes

$$\log\left(\frac{\pi_{ijr}}{1-\pi_{ijr}}\right) = m_r + \mathbf{x}_{ijr}^{'}\mathbf{B}\mathbf{v}_r, \tag{5}$$

which is the LRRR model representation.

The parameters of the equation 5 are obtained using the Majorization Minimization (MM) algorithm developed by De Rooij (2023), which is similar to the algorithm of De Leeuw (2006). This method is based on minimising the negative log-likelihood (log-loss) function given by

$$\begin{aligned}\mathcal{L}(\theta) = \sum_{i=1}^{N}\sum_{j=1}^{n_i}\sum_{r=1}^{R}\mathcal{L}_{ijr}(\theta_{ijr}) &= -\sum_{i=1}^{N}\sum_{j=1}^{n_i}\sum_{r=1}^{R}\{y_{ijr}\log\pi_{ijr} + (1-y_{ijr})\log(1-\pi_{ijr})\} \\ &= -\sum_{i=1}^{N}\sum_{j=1}^{n_i}\sum_{r=1}^{R}\log\frac{1}{1+\exp(-q_{ijr}\theta_{ijr})},\end{aligned} \tag{6}$$

where $\theta_{ijr} = m_r + \mathbf{x}_{ijr}\mathbf{B}v_r$, $q_{ijr} = 2y_{ijr} - 1$ and n_i shows the number of time points for the individual i.

The LRRR analyses in this thesis were performed using the “lpca()” function in the “lmap” package, developed by De Rooij and Busing (2022).

When deciding on the optimal model, the first step was to determine the rank of the regression coefficient matrix (S), which corresponds to the rank of the solution. To achieve this, several models were estimated for different S values while controlling the other attributes, and the model with the lowest information criteria values was accepted as the optimal one. The number of parameters is $(P + R - S)S + R$ (Mukherjee et al., 2015) when computing the information criteria values in equations 1 and 2.

In the second step, the different polynomial degrees of the age variable were evaluated using the determined S value in the first step and keeping the other features constant; then, the optimal model was chosen among others.

In the last step, feature selection was performed using the polynomial degree of the age variable and the S value acquired in the previous stages while holding the other components constant. To determine which predictors should be included in the model, the predictors were systematically removed individually.

After deciding on the final model, Quality of Representation (Q_r, De Rooij and Groenen (2021)) values were calculated to assess how well the model explains or captures the variance in response variables. This measurement is defined as

$$Q_r = (L_{(0,r)} - L_r)/(L_{(0,r)} - L_{lr}),$$

where the deviation of the intercept-only LR model for response variable r is denoted by $L_{(0,r)}$, the loss function for response variable r is indicated by L_r, and the deviation from the LR model with the same predictors is referred to as L_{lr}.

The implied regression coefficients (De Rooij and Groenen, 2021) were then computed to assess the estimated effects or relationship between the predictor and the response variables. These coefficients corresponded to the $\mathbf{BV'}$ in equation 4, and the exponential of these values were obtained to represent the OR values.

Furthermore, CIs were constructed to discover more about the precision and uncertainty of these coefficients. Concerning this, using one of the resampling methods, block-bootstrap (Sherman and le Cessie, 1997), standard deviations were determined first, and then confidence intervals were generated. Bootstrap algorithm consisted of four main steps (Efron and Tibshirani, 1986):

(i) Independently b bootstrap samples were selected with replacement;

(ii) $\mathbf{B}_b^* \mathbf{V}_b^{*'} = \hat{\mathbf{A}}_b^*$ was calculated for each bootstrap sample (b = 1, 2, ..., 100);

(iii) The standard deviation of $\hat{\mathbf{A}}_b^*$ was computed;

(iv) $\hat{\mathbf{A}}_b^* \pm z * sd(\hat{\mathbf{A}}_b^*)$ was obtained, where z was the confidence level value.

Since the sample selection is dependent on the correlation structure within each block, it involves randomly selecting entire blocks for the sampling process.

As a final step, biplots (Gabriel, 1971) were constructed to assess the dimensional relationship between predictor variables and response variables.

3 Results

3.1 Developing the Model with Traditional Approach (LR)

LR analysis was employed to fit the model to the stacked person-period data set shown in Table 2. Firstly, the polynomial degrees of the age variable were examined to establish a flexible model that could better describe the relationship between the variables. Furthermore, all predictor variables were eliminated from the model individually except for the age variable. Then, the optimal model that best represents the data set was obtained. These related steps and results are explained in the following sections.

3.1.1 Optimal polynomial degree for the age variable

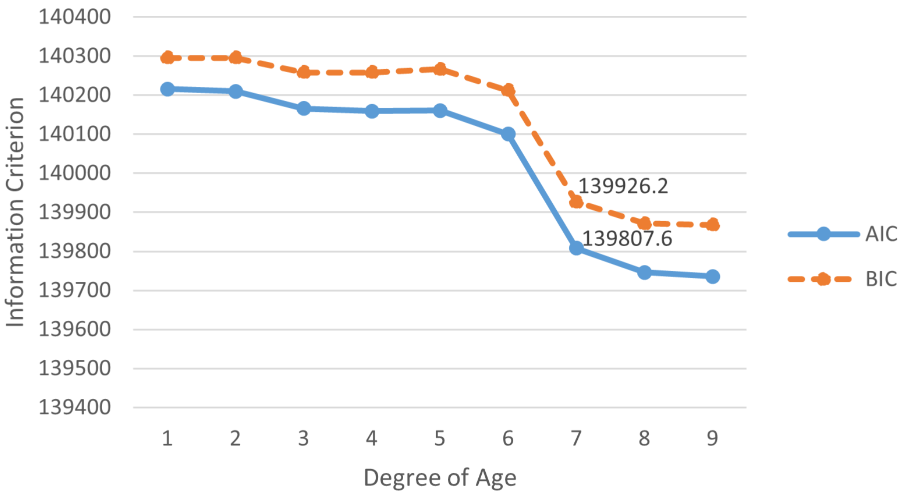
The form of the age variable was investigated to discover the model that fits the data set better. While deciding the degree of the age variable, the polynomial degrees of this variable were examined from 1 to 9, and AIC and BIC values were obtained for each degree.

Figure 3 shows the line chart constructed using the relevant values. Although the AIC values were lower than the BIC values, they followed a similar pattern. The information criteria values were decreasing while increasing polynomial degrees. However, the values remained relatively stable after the seventh degree.

As a consequence, it was decided that the model should contain the age variable with a polynomial degree of seven.

Figure 3

Information Criteria Corresponding to the Polynomial Degrees of the Age Variable



Note. AIC and BIC values of 139807.6 and 139926.2 were used as references when comparing models after predictors were removed from the model.

3.1.2 Feature selection

Predictor variables, except for the age variable with a polynomial degree of 7, were individually removed from the model to check the model quality.

According to Table 4, the AIC and BIC for "Parental criminality" predictor variable were lower than the reference values (see Figure 3); hence this predictor variable was excluded from the model.

In conclusion, the final LR model included an age variable with a seven-degree polynomial structure, eight CAs and gender predictor variables.

Table 4

Information Criteria Values Corresponding to the Models Obtained After Individual Removal of Predictor Variables from LR model

Predictor Variables	**Deviance**	**# Parameters**	**AIC**	**BIC**
Gender	139779.1	17	139813.1	139925.1
Parental mental illness	140102.4	17	140136.4	140248.5
Parental substance abuse	139874.4	17	139908.4	140020.4
Parental criminality	139772.0	17	139806.0*	139918.1*
Family violence	139926.4	17	139960.4	140072.4
Physical abuse	139821.1	17	139855.1	139967.2
Neglect	139834.4	17	139868.4	139980.4
Other parental loss	139812.2	17	139846.2	139958.3
Physical Illness	140133.4	17	140167.4	140279.4
Economic adversity	139856.9	17	139890.9	140003.0

Note. The * index indicates the lowest AIC/BIC value.

3.1.3 Associations between the variables

After obtaining the final LR model that optimally represents the data set, ORs were calculated to examine the relationship between the CAs and the disorders.

Table 5 shows that the model had ORs ranging from 1.20 to 1.52. While all CAs were associated with the disorders, some had a more significant impact. For instance, parental mental illness and physical illness had stronger effects on the disorders' occurrence than other CAs. Conversely, neglect and other parental loss had less pronounced influences on the disorders.

In the case of parental mental illness, which had the highest impact, the odds of acquiring each disorder were 1.52 times higher in individuals living with parents having mental problems than in those without such issues. This result was statistically significant because the 95 % confidence interval did not include the value 1 for this adversity (see Appendix B).

In conclusion, it was observed that each CA had different effects on disorders, but the specific effects of these CAs on certain disorders remained unclear. Nevertheless, these effects appeared to be consistent across all disorders.

Table 5

ORs values corresponding to predictor variables in the LR model

Predictor Variables	Odds Ratio
Parental mental illness	1.52
Parental substance abuse	1.27
Family violence	1.32
Physical abuse	1.25
Neglect	1.20
Other parental loss	1.22
Physical illness	1.47
Economic adversity	1.31

3.2 Developing the Model with New Approach (LRRR)

LRRR analysis was employed to fit the model to the person-period data set shown in Table 3. During the analysis, various dimensions were tested to determine the optimal dimensionality for the model, and the value that adequately represented the model was chosen.

Additionally, the polynomial degrees of the age variable were examined to establish a flexible model that could better describe the relationship between the variables. Moreover, except for the age variable, all predictor variables were eliminated from the model one by one to choose the informative predictors for the model. As a result of these processes, the optimal LRRR model was developed.

As a final step, the results obtained from the LRRR model were used to calculate the quality of representation scores and exponentiated implied regression coefficients. Biplots were then drawn to interpret the findings dimensionally. These related steps and results are explained in the following sections.

3.2.1 Optimal dimensionality

Analyses were performed for dimensionalities ranging from 1 to 9 to determine the optimal model. The polynomial degree of the age variable was set to 9, and this value was used for each dimension.

According to Table 6, while the six-dimensional model was optimal based on the AIC value, the three-dimensional model was considered optimal based on the BIC value. As mentioned in the method section, the BIC helps prevent overfitting, so the three-dimensional model with the lowest BIC value was selected as an optimal model.

Table 6

Information Criteria Values Corresponding to the Models with Various Dimensions

Dimensionality	**Deviance**	**# Parameters**	**AIC**	**BIC**
1	444685.7	58	444801.7	445183.9
2	439518.7	94	439706.7	440326.2
3	438388.0	128	438644.0	439487.6*
4	438214.4	160	438534.4	439588.9
5	437887.9	190	438267.9	439520.1
6	437792.3	218	438228.3*	439665.1
7	437741.2	244	438229.2	439837.3
8	437706.8	268	438242.8	440009.1
9	437677.5	290	438257.5	440168.8

Note. The * index indicates the lowest AIC/BIC values.

3.2.2 Optimal polynomial degree for the age variable

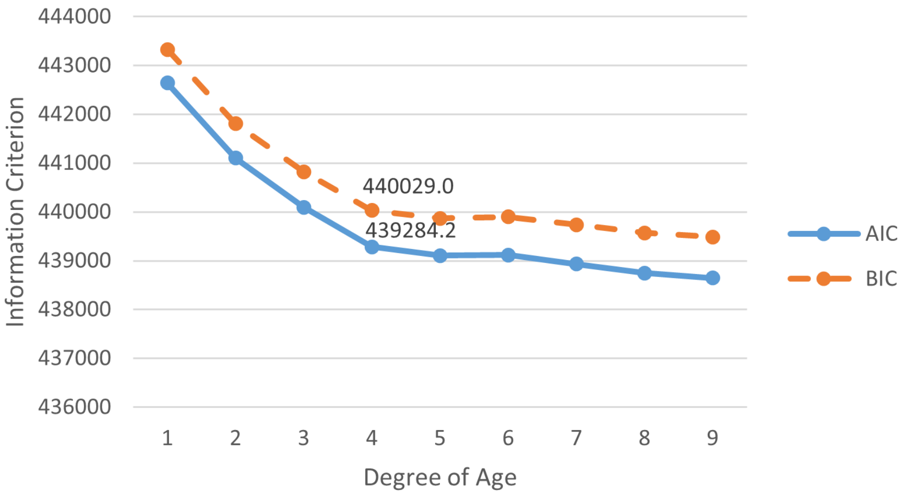
The form of the age variable was investigated to find the model that best fits the data set. The age variable was examined for each polynomial degree from 1 to 9, with S (dimension) adjusted to 3.

Figure 4 shows the line chart constructed using the information criteria values. Although the AIC values were lower than the BIC values, they followed a similar pattern. The information criteria values were decreasing while increasing polynomial degrees. However, the values remained relatively stable after the fourth degree.

As a consequence, it was decided that the three-dimensional model should contain the age variable, which has a polynomial degree of four.

Figure 4

Information Criteria Corresponding to the Polynomial Degrees of the Age Variable



Note. AIC and BIC values of 439284.2 and 440029.0 were used as references when comparing models after predictors were removed from the model.

3.2.3 Feature selection

Predictor variables, except for the age variable with a polynomial degree of 4, were removed from the three-dimensional model one by one to check the model quality.

According to Table 7, the AIC and BIC of "Parental criminality" predictor variable were lower than the reference values (see Figure 4); hence this predictor was excluded from the model.

In conclusion, the final three-dimensional LRRR model included an age variable with a four-degree polynomial structure, eight CAs and gender predictor variables.

Table 7

Information Criteria Values Corresponding to the Models Obtained After Individual Removal of Predictor Variables from LRRR model

Predictor Variables	Deviance	# Parameters	AIC	BIC
Gender	439944.6	110	440164.6	440889.6
Parental mental illness	439436.0	110	439656.0	440380.9
Parental substance abuse	439263.5	110	439483.5	440208.5
Parental criminality	439059.7	110	439279.7*	440004.7*
Family violence	439228.1	110	439448.1	440173.1
Physical abuse	439118.4	110	439338.4	440063.3
Neglect	439138.1	110	439358.1	440083.1
Other parental loss	439125.6	110	439345.6	440070.6
Physical Illness	439457.5	110	439677.5	440402.4
Economic adversity	439169.1	110	439389.1	440114.0

Note. The * index indicates the lowest AIC/BIC values.

3.2.4 Quality representation of response variables

The quality of the representation values was calculated to determine how well the LRRR model represents the response variables (disorders). According to Table 8, most of the disorders were represented by more than 70%. The top three disorders best represented by the three-dimensional model were specific phobia (98%), generalized anxiety disorder (98%), and major depressive disorder (97%). On the other hand, the three worst-represented disorders were bipolar-I disorder (10%), bipolar-II disorder (10%), and subthreshold bipolar disorder (46%).

Table 8

Quality of Representation of Response Variables in LRRR Model

Response Variables	Quality	Response Variables	Quality
Attention-deficit disorder	0.71	Social phobia	0.90
Conduct disorder	0.71	Alcohol abuse	0.91
Intermittent explosive disorder	0.91	Alcohol dependence with abuse	0.82
Oppositional-defiant disorder	0.74	Drug abuse	0.84
Panic disorder	0.90	Drug dependence with abuse	0.67
Posttraumatic stress disorder	0.93	Bipolar I disorder	0.01**
Generalized anxiety disorder	0.98*	Bipolar II disorder	0.01**
Separation anxiety disorder	0.78	Subthreshold bipolar disorder	0.46**
Agoraphobia	0.67	Dysthymic disorder	0.91
Specific phobia	0.98*	Major depressive disorder	0.97*

Note. The * index indicates the three disorders well-represented by the model, while the ** index identifies the three worst-represented disorders.

3.2.5 Associations between the variables

Exponentiated implied regression coefficients were calculated to provide a more detailed interpretation of the relationship between the CAs and the disorders. These coefficients indicate the extent to which the presence of CAs influences the odds of the disorders. These coefficient values were shown in Table 9.

The coefficient of a CA being greater than 1 implies that the presence of this adversity is associated with an increased likelihood of the disorder. On the other hand, the coefficient of a CA is less than 1 implies that the odds of the disorder are lower in the presence of the CA compared to its absence.

As seen in Table 9, the occurrence of many CAs increased the likelihood of developing disorders. For example, if parental mental illness (PMI) adversity and posttraumatic stress disorder were considered, it was stated that the presence of this adversity increased the odds ratio of posttraumatic stress disorder by a factor of 1.35. In other words, those with this mental illness in their childhood were 1.35 times more likely to develop posttraumatic stress disorder in later years. This result was statistically significant because the 95 % confidence interval did not include the value 1 for this adversity (see Appendix C).

Among the CAs, only parental substance abuse (PSA) was found to be associated with a specific phobia (SP) disorder, with an implied regression coefficient value of less than 1, suggesting a potential opposite effect on the development of this disorder. However, upon evaluating the CI for this CA, it includes the value 1, indicating that the effect is not statistically significant. Therefore, any interpretation or speculation regarding this CA's impact on SP disorder would be unfounded.

As a result, the presence of CAs in this data set was found to be effective in disorders occurrence, while their absence does not influence disorders development.

Table 9

Exponentiated Implied Regression Coefficients in LRRR Model

	Predictor Variables							
	PMI	PSA	FV	PA	N	OPL	PI	EA
Response Variables								
Attention-deficit disorder	1.14	1.02	1.10	1.06	1.08	1.04	1.15	1.10
Conduct disorder	1.18	1.17	1.19	1.10	1.08	1.10	1.23	1.13
Intermittent explosive disorder	1.10	1.15	1.14	1.08	1.05	1.08	1.15	1.10
Oppositional-defiant disorder	1.20	1.13	1.18	1.10	1.10	1.09	1.24	1.13
Panic disorder	1.23	1.05	1.12	1.06	1.11	1.03	1.20	1.07
Posttraumatic stress disorder	1.35	1.04	1.17	1.08	1.16	1.04	1.29	1.10
Generalized anxiety disorder	1.26	1.03	1.12	1.05	1.12	1.02	1.21	1.05
Separation anxiety disorder	1.20	1.00	1.11	1.06	1.11	1.03	1.19	1.10
Agoraphobia	1.18	1.03	1.10	1.05	1.09	1.03	1.17	1.06
Specific phobia	1.17	0.97	1.08	1.06	1.11	1.01	1.16	1.09
Social phobia	1.16	1.09	1.14	1.08	1.08	1.06	1.19	1.11
Alcohol abuse	1.09	1.29	1.17	1.07	1.01	1.12	1.15	1.06
Alcohol dependence with abuse	1.11	1.21	1.14	1.06	1.03	1.09	1.15	1.06
Drug abuse	1.13	1.29	1.19	1.09	1.04	1.12	1.19	1.08
Drug dependence with abuse	1.15	1.16	1.14	1.06	1.06	1.07	1.17	1.07
Bipolar I disorder	1.10	1.05	1.07	1.03	1.05	1.03	1.10	1.04
Bipolar II disorder	1.11	1.04	1.07	1.03	1.05	1.02	1.10	1.04
Subthreshold bipolar disorder	1.08	1.06	1.07	1.03	1.04	1.03	1.09	1.04
Dysthymic disorder	1.24	1.05	1.12	1.05	1.11	1.03	1.20	1.06
Major depressive disorder	1.18	1.04	1.09	1.03	1.08	1.02	1.14	1.03

Abbreviations. PMI, Parental mental illness; PSA, Parental substance abuse; FV, Family violence; PA, Physical abuse; N, Neglect; OPL, Other parental loss; PI, Physical illness; EA, Economic adversity.

3.2.6 Visualization

The previous section interpreted the relationship between CAs and disorders using the exponentiated implied regression coefficients. In this section, these relationships were examined and interpreted using the biplots.

The biplots presented a more straightforward interpretation of the relationship between CAs and disorders, and they also provided insights into how correlated disorders cluster within each other. Furthermore, biplots allowed us to test the dimensional approach to CAs, which motivated this thesis.

Since the final LRRR model was three-dimensional, the biplots were created as pairwise combinations of dimensions, facilitating graphical interpretation. In other words, the plots were drawn by considering the 1st and 2nd dimensions first, then the 1st and 3rd dimensions, and finally, the 2nd and 3rd dimensions coordinates. The related plots are shown below in Figure 5.

In Figure 5, solid green lines illustrate the disorders, dashed blue lines indicate the CAs and grey dots represent the individuals per time. The labels of the lines are located on the side where the variables take the yes (i.e. 1) value. The other end of the line represents the part where the variable takes the no (i.e. 0) value.

Representing a three-dimensional model using two-dimensional plots is not straightforward, so these plots should be carefully interpreted. Thus, it was considered that the following three figures reflect the three-dimensional model from different angles, which means that while some show the positions of the variables, others give information about their depths. In addition, variables that seem close to each other in a plot may appear far apart in another. Therefore, these three plots were evaluated together rather than separately, and their insights were supported by exponentiated implied regression coefficients shown in Table 9.

When comparing the figures, it was evident that some CAs had stronger relationships with certain disorders than other CAs. To ensure the statistical significance of the examined associations, the focus was placed on disorders with high-quality representation, that is, well-represented disorders by the model, such as specific phobia (SP), generalized anxiety disorder (GAD) and major depressive disorder (MDDH) (see Table 8).

Neglect (N), economic adversity (EA) and parental mental illness (PMI) adversities had a higher impact on the specific phobia (SP) disorder than other CAs when all three biplots were examined simultaneously. The angles between the adversities' lines and the SP line were less than 90 degrees, and the lines of these adversities were closer to the SP line than other adversities, suggesting that these adversities strongly impacted the development of the SP disorder.

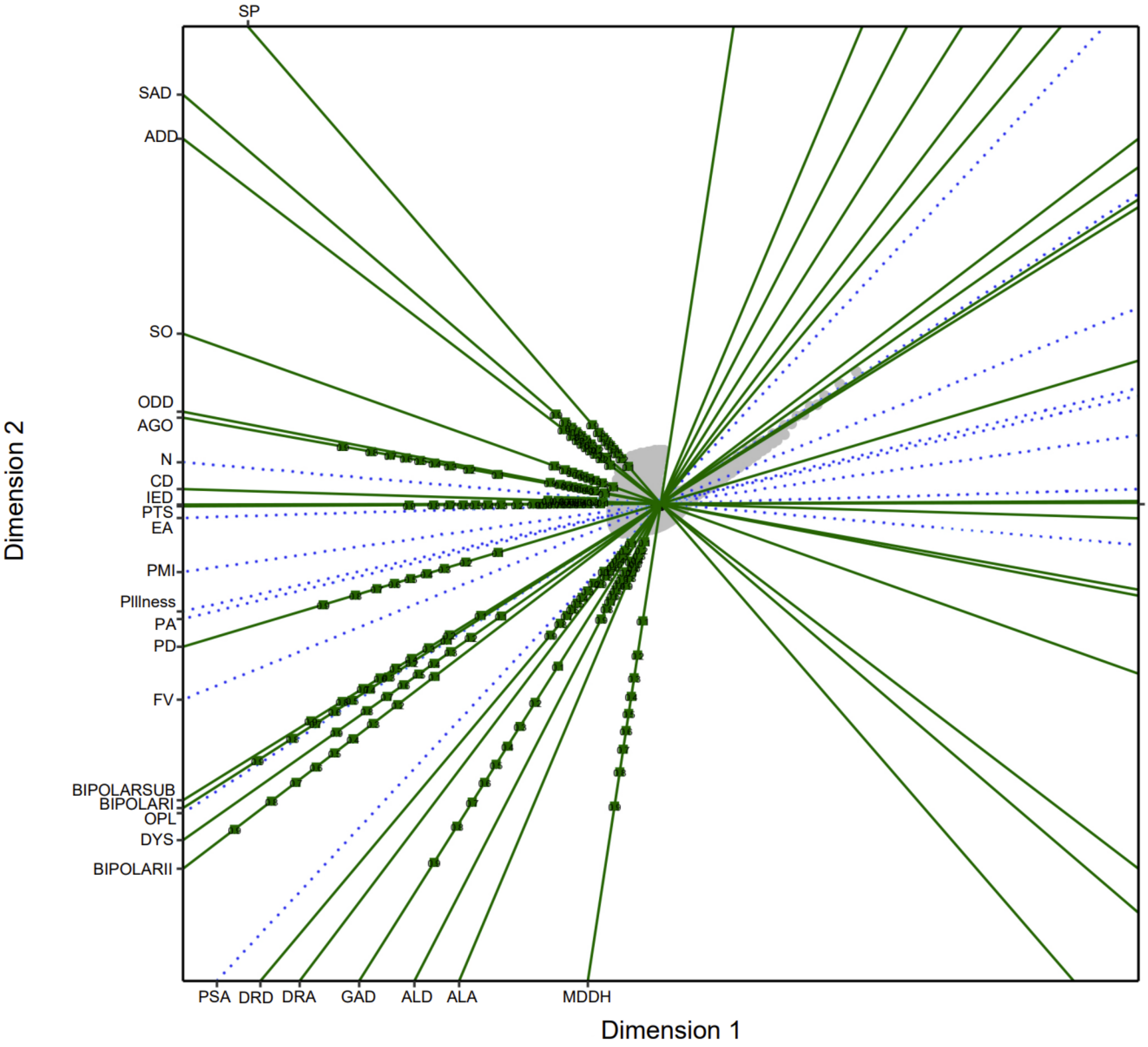
When generalized anxiety disorder (GAD) and major depressive disorder (MDDH) were examined, their lines were close to each other, indicating a similarity in their characteristics and suggesting that the same CAs influenced these disorders. In the analysis of three biplots, it was observed that family violence (FV), physical abuse (PA), physical illness (PIllness) and parental mental illness (PMI) had a stronger influence on these disorders, as the angles between these disorders and the CAs were acute. Additionally, the disorders' lines were much closer to these CAs' lines than other adversities.

With the biplots presented in Figure 5, one can examine not only the association between CAs and disorders but also the relationships among disorders. Therefore, it can be said from the biplots that several disorders were closely positioned and formed a group in each plot, indicating acute angles between the disorder lines. The first group included alcohol abuse (ALA), alcohol dependence with abuse (ALD), drug abuse (DRA) and drug dependence with abuse (DRD); the second group contained dysthymic disorder (DYS), bipolar I (BIP-I), bipolar II (BIP-II) and subthreshold bipolar (BIP-SUB); the third group included agoraphobia (AGO), posttraumatic stress disorder (PTS) and panic disorder (PD); and final group contained intermittent explosive disorder (IED) and conduct disorder (CD).

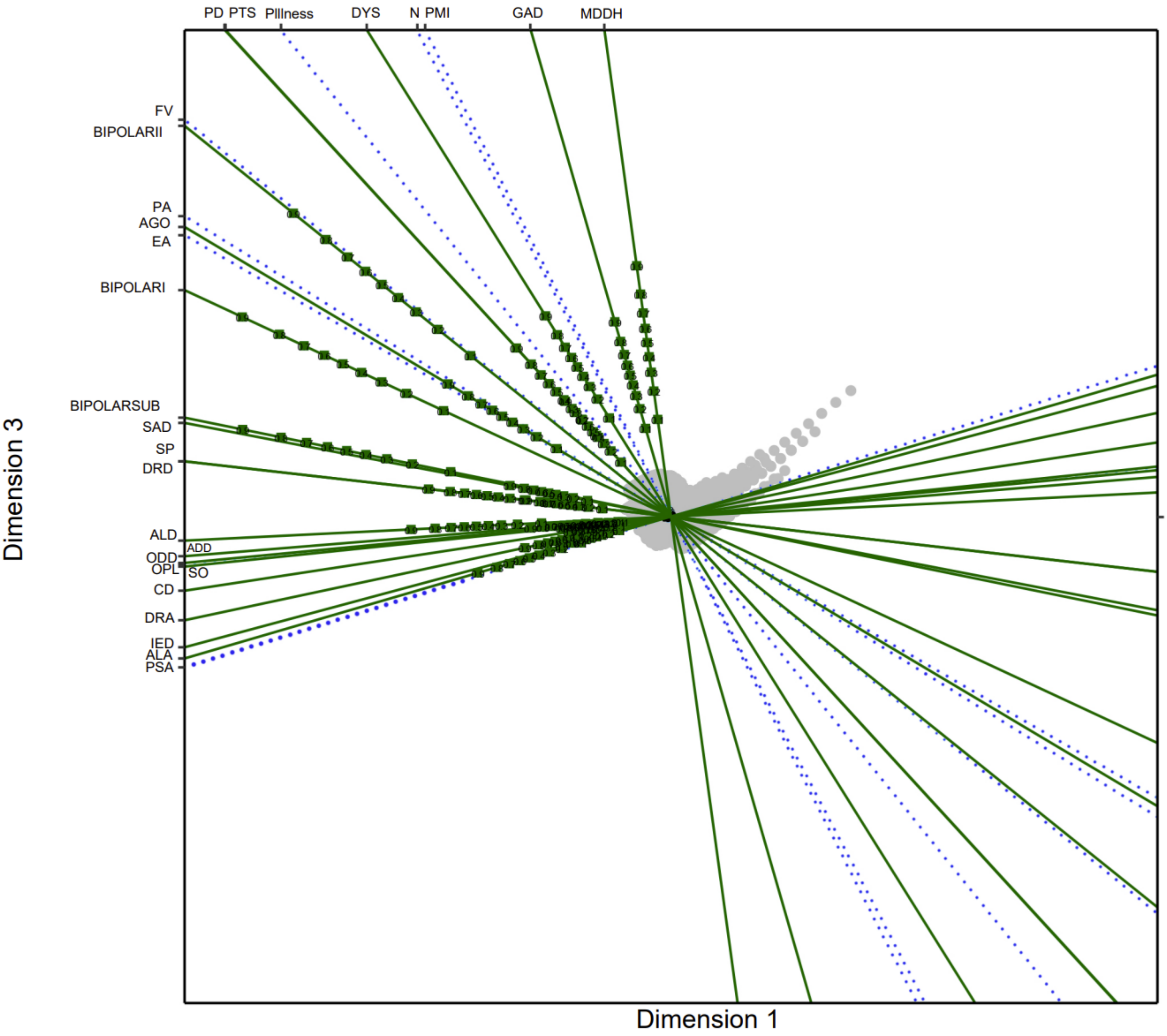
In conclusion, when interpreting the results obtained through LRRR analysis with biplots, a dimensional approach was applied to the multivariate data set. This approach allows us to examine the relationships between CAs and disorders, as well as the relationships among the disorders themselves.

Figure 5

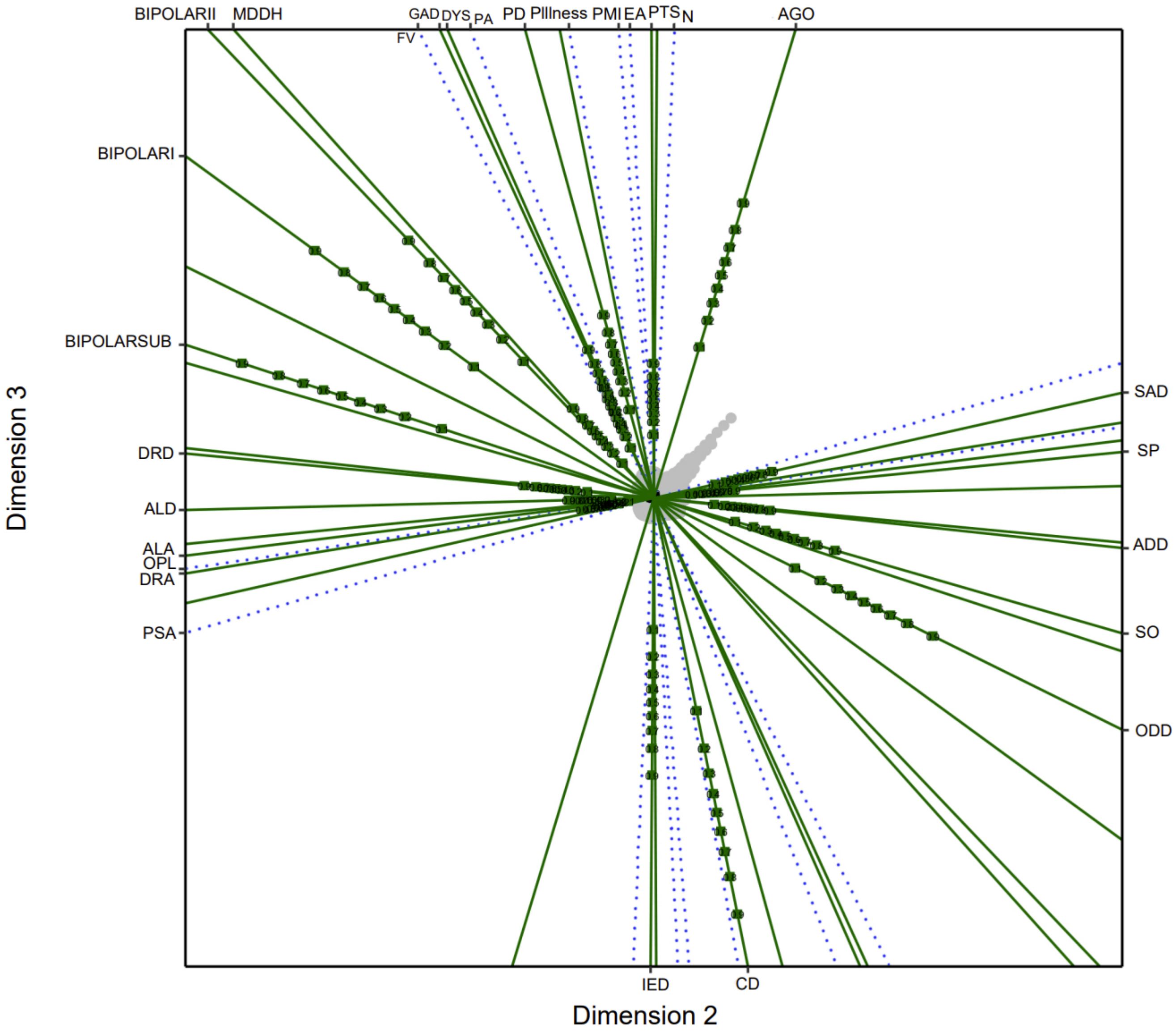
Dimensional Approach to the CAs and the Disorders with Three-Dimensional LRRR Model



(a) First and second dimensions of biplot



(b) First and third dimensions of biplot



(c) Second and third dimensions of biplot

Abbreviations. For response variables, ADD, Attention-deficit disorder; CD, Conduct disorder; IED, Intermittent explosive disorder; ODD, Oppositional-defiant disorder; PD, Panic disorder; PTS, Post-traumatic stress disorder; GAD, Generalized anxiety disorder; SAD, Separation anxiety disorder; AGO, Agoraphobia; SP, Specific phobia; SO, Social phobia; ALA, Alcohol abuse; ALD, Alcohol dependence with abuse; DRA, Drug abuse; DRD, Drug dependence with abuse; BIPOLARI, Bipolar I disorder; BIPOLARII, Bipolar II disorder; BIPOLARSUB, Subthreshold bipolar disorder; DYS, Dysthymic disorder; MDDH, Major depressive disorder. For predictor variables, PMI, Parental mental illness; PSA, Parental substance abuse; FV, Family violence; PA, Physical abuse; N, Neglect; OPL, Other parental loss; PIllness, Physical illness; EA, Economic adversity.

3.3 Comparison of Traditional (LR) and New (LRRR) Approaches

This study compared a traditional approach LR with a new approach LRRR for multivariate discrete-time survival data. The LR is a widely used method for modelling survival data sets with discrete time variables. However, the LRRR produces more effective results if the response variables are correlated.

While LRRR can examine the impact of each predictor on each response variable, LR cannot distinguish this situation. With the LR approach, it is considered that the influence of a predictor on each response variable is the same. For more detailed explanations, assume that C_1 and C_2 are any two CAs as the predictor variables, and that D_1, D_2, ..., D_{20} are twenty disorders as the response variables. In this case, it can be said that

$$\begin{gathered} C_1 \rightarrow D_1 = C_1 \rightarrow D_2 = ... = C_1 \rightarrow D_{20} \\ C_2 \rightarrow D_1 = C_2 \rightarrow D_2 = ... = C_2 \rightarrow D_{20}, \end{gathered}$$

is provided for the LR method.

In Table 5, it is seen that the OR value for parental mental illness (PMI) was 1.52, which was consistent for all twenty disorders. This suggests that individuals living with parents having a mental illness had 1.52 times higher odds of acquiring disorders compared to those without such circumstances. Hence, it is evident that PMI had a uniform impact on each disorder. The same pattern can also be applied to individuals with episodes of severe physical illness (PIllness). Individuals who experienced PIllness in childhood had 1.47 times higher odds of developing any disease.

According to the LR analyses, it was not possible to determine which disorders were impacted the most by these CAs. However, when Table 9 is evaluated, it becomes apparent that LRRR allows us to observe to what extent each CA impacts which disorders. As a result, it was discovered that PMI and PIllness were shown to have a significant influence on posttraumatic stress disorder (PTS), with odds ratios of 1.35 and 1.29, respectively. Furthermore, when examining the three biplots in Figure 5 simultaneously, it can be observed that the PMI and PIllness lines are closer to the PTS line compared to the other disorder lines.

On the other hand, while LR is insufficient to investigate the link between disorders, LRRR explores not only the association between CAs and disorders but also the relationships between different disorders. Figure 5 shows which disorders are close to each other, indicating clustering where the distance between lines is narrow. Section 3.2.6 provides detailed information about disorders categorized into four groups with narrower and closer angles between the disorder lines compared to other disorders.

In conclusion, in multivariate discrete-time survival data sets, LRRR outperforms LR in revealing more information about the data set and facilitating the interpretation.

4 Conclusion and Discussion

Time-to-event (survival) data sets are commonly analyzed using Kaplan-Meier Estimator or Cox Proportional Hazards Model techniques. Moreover, parametric modelling-based techniques such as LR are employed when the time variable is discrete, providing a more flexible approach to modelling the survival data set (Efron, 1988). However, if the data set includes multiple correlated response variables, the LR approach does not consider the relationship between the variables. Therefore, reduced rank-based analyses considering the relationship between variables should be preferred.

The main objective of this thesis was to propose an alternative approach for analysing multivariate discrete-time survival data sets using LRRR, which is one of the reduced rank-based analysis techniques, determines a lower-dimensional subspace that captures the essential correlations among the variables and detects underlying links between them while providing a trade-off between bias and variance (Breiman and Friedman, 2002).

This thesis was inspired by the proposal of a dimensional approach to CAs in the article written by McLaughlin and Sheridan (2016). This article underscored the significance of dimensional evaluation to understand how CAs impact developmental outcomes in children, enabling efficient interventions against the consequences of CAs. Therefore, this thesis aimed to test this dimensional approach statistically, providing an alternative approach for analyzing multivariate discrete-time survival data.

The publicly released NCS-R survey data set, containing information on CAs and adult psychiatric disorders, was utilized in the analyses. The data set was initially analyzed with LR, which follows the traditional approach. Subsequently, it was analyzed using LRRR, and the findings were compared to the results from the LR analysis. Both LR and LRRR analyses indicated that the parental criminality (PC) predictor variable should not be included in either the LR model or the LRRR model.

Furthermore, both analyses revealed that parental mental illness (PMI) and physical illness (PIllness) were the most significant CA predictors of disorders onsets. However, using the LR approach, it was impossible to discern precisely how these CAs impact the initiation of each specific disorder. The LR approach assumed that CAs had the same impact on all disorders. In contrast, with the LRRR, it was revealed the specific disorders impacted by these CAs and the extent of their effects.

When examining the exponentiated implied regression coefficient values and biplots obtained through LRRR analysis, it was evident that PMI had the strongest influence on posttraumatic stress disorder (PTS). This finding is consistent with the existing literature and is also corroborated in the studies conducted by Meiser-Stedman et al. (2005) and Trickey et al. (2012). Additionally, it was found that PIllness had a substantial impact on PTS disorder, as supported by the research conducted by Pinquart (2018).

On the other hand, the dimensional method revealed that several disorders exhibit clustering within themselves. These groupings were found to be organized in line with DSM-IV classifications. Specifically, ALA, ALD, DRD, and DRA disorders were classified as substance use disorders; DYS, BIP-I, BIP-II, and BIP-SUB disorders were classified as mood disorders; AGO, PTS and PD were classified as anxiety disorders; and IED and CD were classified as disruptive behaviour disorders.

According to these findings, supported by the literature, individuals who experienced PMI and PIllness during childhood are more likely to develop PTS later. These results highlight the importance of early interventions for individuals with a history of PMI and PIllness to potentially minimize the impact of these CA experiences on the development of PTS in adulthood.

Moreover, identifying subgroups of disorders that demonstrate clustering allows further research and focus on specific CAs that influence these disorder groupings. This may lead to more targeted and effective interventions tailored to the unique characteristics of each disorder subgroup, potentially improving treatment outcomes for individuals with different disorders.

As a result, the findings demonstrated that the dimensional approach to CAs proposed in Figure 1 is feasible using LRRR analysis. It allowed us to evaluate the relationships between CAs and disorders dimensionally, providing insights into which disorders were associated with CAs and how they were affected by them.

Although relevant analyses were conducted on a psychiatric-based dataset containing CAs and adult illnesses, the dimensional approach method can be extended to other data sets with correlated multivariate structures in various fields of investigation. For instance, in environmental science, researchers may investigate the impacts of variables like temperature, rainfall, and humidity on soil quality indicators. In finance, analysts can examine the effects of predictors like sales, interest rates, and prices on stock prices.

While this study demonstrated positive findings regarding a dimensional approach using LRRR, it has some limitations. Firstly, biplots are visually appealing and easily interpretable in two dimensions. When the model is high-dimensional, biplots can be created in binary combinations, but this makes it difficult to analyze all biplots simultaneously.

On the other hand, when data is missing in the LRRR approach, it should be completely random (MCAR) like in the LR approach, implying the missing is entirely independent of observed and unobserved values. The missing values in the utilized data set for this study were MCAR. Since the analyses were conducted by considering the complete cases, this situation did not pose any issues in the analysis. If the missing is not MCAR, it becomes a limitation for LRRR, and imputation techniques should be implemented to obtain a complete set of observations.

In the future, it is recommended to perform analyses using multiple imputation techniques to impute missing observations rather than eliminating them from the data. Multiple imputations can provide to keep the essential information and eliminate analytical bias. Additionally, exploring the number of CAs experienced and including them in the dimensional approach could provide valuable insights. Moreover, employing cross-validation techniques can be beneficial for evaluating model performance and assessing its generalization ability.

This thesis has demonstrated that by employing suitable methods that preserve the structural integrity of the data, one can obtain valuable and reliable information about the data and reveal hidden patterns. These emerging patterns enable us to predict future developments, implement appropriate safeguards, and devise effective strategies. We hope this study contributes to discovering patterns in data sets and leads to meaningful insights.

5 References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Alegria, M., Jackson, J. S., Kessler, R. C., & Takeuchi, D. (2016). Collaborative psychiatric epidemiology surveys (cpes), 2001-2003 [united states]. https://doi.org/https://doi.org/10.3886/ICPSR20240.v8

Anderson, T. W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *The Annals of Mathematical Statistics*, 327–351.

Anderson, T. W. (2003). *An introduction to the multivariate statistical analysis* (3rd). Wiley.

Breiman, L., & Friedman, J. H. (2002). Predicting Multivariate Responses in Multiple Linear Regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, *59*(1), 3–54. https://doi.org/10.1111/1467-9868.00054

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and inference: A practical information-theoretic approach* (2nd). Springer-Verlag. https://doi.org/10.1007/b97636

Collett, D. (2003). *Modeling binary data* (2nd). Chapman; Hall/CRC.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, *34*(2), 187–220. http://www.jstor.org/stable/2985181

Cox, D. R. (1975). Partial likelihood. *Biometrika*, *62*(2), 269–276. https://doi.org/10.1093/biomet/62.2.269

Davies, P. T., & Tso, M. K.-S. (1982). Procedures for Reduced-Rank Regression. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *31*(3), 244–255. https://doi.org/https://doi.org/10.2307/2347998

De Leeuw, J. (2006). Principal component analysis of binary data by iterated singular value decomposition. *Computational Statistics Data Analysis*, *50*(1), 21–39. https://EconPapers.repec.org/RePEc:eee:csdana:v:50:y:2006:i:1:p:21-39

De Rooij, M. (2023). A new algorithm and a discussion about visualization for logistic reduced rank regression. *Behaviormetrika*. https://doi.org/10.1007/s41237-023-00204-3

De Rooij, M., & Busing, F. (2022). *lmap: Logistic Mapping*, R package version 0.1.1.

De Rooij, M., & Groenen, J. F., Patrick. (2021). The melodic family for simultaneous binary logistic regression in a reduced space. https://doi.org/10.48550/arXiv.2102.08232

Deming, W. E. (1953). Paper presented at meeting of the international statistical institute.

Diggle, P., Heagerty, P., Liang, K.-Y., & Zeger, S. (2002). *Analysis of longitudinal data (2nd ed.)* Oxford University Press.

Efron, B. (1988). Logistic regression, survival analysis, and the kaplan–meier curve. *Journal of the American Statistical Association*, *83*(402), 414–425. https://doi.org/10.1080/01621459.1988.10478612

Efron, B., & Tibshirani, R. (1986). Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science*, *1*(1), 54–75. https://doi.org/10.1214/ss/1177013815

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, *58*(3), 453–467. https://doi.org/10.2307/2334381

Green, J. G., McLaughlin, K. A., Berglund, P. A., Gruber, M. J., Sampson, N. A., Zaslavsky, A. M., & Kessler, R. C. (2010). Childhood Adversities and Adult Psychiatric Disorders in the National Comorbidity Survey Replication I: Associations With First Onset of DSM-IV Disorders. *Archives of General Psychiatry*, *67*(2), 113–123. https://doi.org/10.1001/archgenpsychiatry.2009.186

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd). Springer Series in Statistics. Springer, New York, NY. https://doi.org/10.1007/b94608_2

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.

Izenman, A. J. (1975). Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, *5*(2), 248–264. https://doi.org/https://doi.org/10.1016/0047-259X(75)90042-1

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, *53*(282), 457–481. https://doi.org/10.2307/2281868

Khusainova, R. M., Shilova, Z. V., & Curteva, O. V. (2016). Selection of appropriate statistical methods for research results processing. *International Electronic Journal of Mathematics Education*, *11*(1), 303–315. https://doi.org/10.29333/iejme/334

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models (2nd ed.)* Routledge. https://doi.org/10.1201/9780203753736

McLaughlin, K. A., & Sheridan, M. A. (2016). Beyond cumulative risk: A dimensional approach to childhood adversity. *Current Directions in Psychological Science*, *25*(4), 239–245. https://doi.org/10.1177/0963721416655883

McLaughlin, K. A., Sheridan, M. A., & Lambert, H. K. (2014). Childhood adversity and neural development: Deprivation and threat as distinct dimensions of early experience. *Neuroscience and Biobehavioral Reviews*, *47*, 578–591. https://doi.org/10.1016/j.neubiorev.2014.10.012

Meiser-Stedman, R. A., Yule, W., Dalgleish, T., Smith, P., & Glucksman, E. (2005). The Role of the Family in Child and Adolescent Posttraumatic Stress Following Attendance at an Emergency Department. *Journal of Pediatric Psychology*, *31*(4), 397–402. https://doi.org/10.1093/jpepsy/jsj005

Mukherjee, A., Chen, K., Wang, N., & Zhu, J. (2015). On the degrees of freedom of reduced-rank estimators in multivariate regression. *Biometrika*, *102*(2), 457–477. https://doi.org/10.1093/biomet/asu067

Pinquart, M. (2018). Posttraumatic stress symptoms and disorders in children and adolescents with chronic physical illnesses: A meta-analysis. *Journal of Child Adolescent Trauma*, *13*(1), 1–10. https://doi.org/10.1007/s40653-018-0222-z

Reinsel, G. C., & Velu, R. P. (1998). *Multivariate reduced-rank regression: Theory and applications*. Springer.

Schmidli, H. (1995). Classical analysis of reduced rank regression. In *Reduced rank regression: With applications to quantitative structure-activity relationships* (pp. 49–102). Physica-Verlag HD. https://doi.org/10.1007/978-3-642-50015-2_5

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. http://www.jstor.org/stable/2958889

Sherman, M., & le Cessie, S. (1997). A comparison between bootstrap methods and generalized estimating equations for correlated outcomes in generalized linear models. *Communications in Statistics - Simulation and Computation*, *26*(3), 901–925. https://doi.org/10.1080/03610919708813417

Singer, J. D., & Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195152968.001.0001

Ter Braak, C. J. F., & Looman, C. W. N. (1994). Biplots in reduced-rank regression. *Biometrical Journal*, *36*(8), 983–1003. https://doi.org/https://doi.org/10.1002/bimj.4710360812

Trickey, D., Siddaway, A. P., Meiser-Stedman, R., Serpell, L., & Field, A. P. (2012). A meta-analysis of risk factors for post-traumatic stress disorder in children and adolescents. *Clinical Psychology Review*, *32*(2), 122–138. https://doi.org/https://doi.org/10.1016/j.cpr.2011.12.001

Tso, M. K.-S. (1981). Reduced-rank regression and canonical analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, *43*(2), 183–189. https://doi.org/https://doi.org/10.1111/j.2517-6161.1981.tb01169.x

Willett, J. B., & Singer, J. D. (1993). Investigating onset, cessation, relapse, and recovery: Why you should, and how you can, use discrete-time survival analysis to examine event occurrence. *Journal of consulting and clinical psychology*, *61*(6), 952–965. https://doi.org/10.1037//0022-006x.61.6.952

Appendix A

All codes and packages used for the analyses can be found on Github: https://github.com/FatmaOZBAY/master-thesis-statistical-analysis.git.

Appendix B

Table 10

Confidence Intervals for Estimated Parameters in LR Model

Predictor Variables	**Confidence Interval** (95 %)
Parental mental illness	1.46-1.59
Parental substance abuse	1.22-1.33
Family violence	1.26-1.38
Physical abuse	1.18-1.32
Neglect	1.14-1.25
Other parental loss	1.15-1.30
Physical illness	1.41-1.52
Economic adversity	1.24-1.38

Appendix C

Table 11

Confidence Intervals for Estimated Parameters in LRRR Model

	Predictor Variables							
	PMI	PSA	FV	PA	N	OPL	PI	EA
Response Variables								
ADD	1.08-1.19	0.98-1.06	1.05-1.15	1.03-1.11	1.04-1.13	1.01-1.07	1.11-1.21	1.06-1.13
CD	1.12-1.25	1.12-1.23	1.13-1.24	1.04-1.15	1.04-1.13	1.05-1.15	1.18-1.30	1.07-1.17
IED	1.05-1.14	1.11-1.20	1.09-1.19	1.03-1.12	1.01-1.08	1.04-1.12	1.11-1.20	1.05-1.13
ODD	1.15-1.26	1.09-1.17	1.13-1.24	1.04-1.16	1.06-1.14	1.04-1.13	1.19-1.30	1.08-1.18
PD	1.18-1.29	1.00-1.09	1.08-1.17	1.02-1.09	1.07-1.16	0.99-1.06	1.15-1.25	1.03-1.10
PTS	1.30-1.42	0.99-1.10	1.11-1.22	1.03-1.13	1.10-1.22	0.99-1.08	1.24-1.36	1.05-1.14
GAD	1.21-1.33	0.98-1.09	1.07-1.18	1.01-1.09	1.07-1.17	0.97-1.06	1.16-1.25	1.01-1.09
SAD	1.16-1.26	0.96-1.04	1.05-1.16	1.03-1.10	1.07-1.15	0.99-1.05	1.15-1.24	1.06-1.13
AGO	1.14-1.24	1.00-1.06	1.06-1.14	1.02-1.08	1.05-1.13	1.00-1.05	1.12-1.21	1.04-1.09
SP	1.13-1.22	0.93-1.01	1.03-1.14	1.02-1.09	1.06-1.15	0.98-1.04	1.12-1.21	1.06-1.13
SO	1.11-1.22	1.05-1.13	1.10-1.18	1.04-1.12	1.05-1.12	1.03-1.09	1.15-1.22	1.07-1.14
ALA	1.03-1.15	1.21-1.38	1.11-1.24	1.01-1.12	0.95-1.08	1.05-1.19	1.09-1.22	1.00-1.10
ALD	1.06-1.19	1.16-1.27	1.10-1.19	1.01-1.10	0.99-1.08	1.04-1.14	1.11-1.20	1.01-1.09
DRA	1.06-1.20	1.21-1.37	1.14-1.27	1.03-1.13	0.98-1.10	1.06-1.20	1.13-1.26	1.02-1.13
DRD	1.10-1.21	1.12-1.21	1.11-1.20	1.02-1.10	1.01-1.10	1.04-1.12	1.12-1.22	1.02-1.10
BIP-I	1.06-1.16	1.03-1.08	1.05-1.10	1.01-1.06	1.02-1.08	1.00-1.04	1.06-1.15	1.02-1.06
BIP-II	1.07-1.15	1.02-1.07	1.04-1.10	1.01-1.06	1.03-1.08	1.00-1.04	1.07-1.14	1.02-1.06
BIP-SUB	1.05-1.13	1.04-1.08	1.05-1.10	1.01-1.05	1.01-1.06	1.01-1.05	1.06-1.13	1.02-1.05
DYS	1.19-1.31	1.00-1.09	1.08-1.16	1.02-1.09	1.06-1.16	0.99-1.06	1.16-1.25	1.03-1.10
MDDH	1.14-1.22	1.00-1.08	1.05-1.14	1.00-1.06	1.04-1.12	0.99-1.05	1.11-1.18	1.00-1.06

Abbreviations. ADD, Attention-deficit disorder; CD, Conduct disorder; IED, Intermittent explosive disorder; ODD, Oppositional-defiant disorder; PD, Panic disorder; PTS, Posttraumatic stress disorder; GAD, Generalized anxiety disorder; SAD, Separation anxiety disorder; AGO, Agoraphobia; SP, Specific phobia; SO, Social phobia; ALA, Alcohol abuse; ALD, Alcohol dependence with abuse; DRA, Drug abuse; DRD, Drug dependence with abuse; BIP-I, Bipolar I disorder; BIP-II, Bipolar II disorder; BIP-SUB, Subthreshold bipolar disorder; DYS, Dysthymic disorder; MDDH, Major depressive disorder. PMI, Parental mental illness; PSA, Parental substance abuse; FV, Family violence; PA, Physical abuse; N, Neglect; OPL, Other parental loss; PI, Physical illness; EA, Economic adversity.