Time-to-event latent variable models for the statistical analysis of clinical data

Thesis submitted in accordance with the requirements of
the University of Liverpool for the degree of Doctor in Philosophy
by

Daniel T. Lythgoe

5th March 2020

For Gem and Elsie.

For Mum and Dad.

Thesis summary

In clinical research, interest sometimes lies in analysing variables which are not measured directly. Instead, information about these 'latent variables' can be inferred from surrogates or other imperfect indicators, using latent variable models. Common examples of 'hypothetical' latent variables in clinical research include quality of life (QoL), anxiety and depression. Another type of latent variable is a variable used as a device for dimension reduction, for example, a principal component. The aim of this thesis is to explore and develop latent variable methods for the statistical analysis of clinical data, with an emphasis on including latent variables in time-to-event models.

In Chapter 1, latent variables and their utility in clinical research are discussed. Overviews of two multivariate latent variable methods fundamental to this thesis, latent class analysis (LCA) and multidimensional scaling (MDS), are given, and time-to-event analysis is introduced.

In Chapter 2, several statistical models for estimating the effect of a latent class on a time-to-event outcome are described, including a joint or 'one-step' model in which latent classes and time-to-event data are modelled simultaneously. A simulation study is then used to evaluate the empirical properties of latent class effect estimates from several different models on a time-to-event distal outcome. Research in this area has previously been restricted to continuous and dichotomous outcome variables. Additionally, a solution to the problem of class label switching in latent class simulation studies is proposed. This work was published in the Journal of Structural Equation Modeling (Lythgoe et al., 2019).

In Chapter 3, a general joint latent class and time-to-event model is presented. It is shown how the model can be fitted and standard errors obtained. An author-written `R` function is presented. Various versions of the joint model are applied to a prostate cancer clinical trial data set in which the effect of treatment is found to differ across the identified latent subgroups.

In Chapter 4, MDS and the concept of dissimilarity are introduced. It is shown how MDS can be used with clinical data in which variables are usually of mixed type (nominal, ordinal, continuous) using Gower's general coefficient (Gower, 1971). Gower's method for adding test points to an MDS configuration is also detailed.

In Chapter 5, the use of accelerated failure time (AFT) models is proposed to fit MDS biplot axes for time-to-event variables. In particular, the Weibull AFT model is considered since it can be formulated in both AFT and proportional hazards form, the latter representation being far more common in clinical research. A time-to-event biplot is constructed for a hepatocellular carcinoma (HCC) data set in which in the relationships between observations, variables and a time-to-event outcome are illustrated simultaneously.

In Chapter 6, it is shown how MDS can be used for covariate dimension reduction in regression modelling using distance-based regression (DBR). Two supervised versions of DBR are proposed in order to reduce covariate dimensionality further than standard DBR, and are compared against conventional DBR using simulated and real clinical data.

In Chapter 7, two simple extensions to MDS are presented in which an MDS configuration is supervised by an outcome variable. It is shown how these methods can be used for visualisation and prediction, and both methods are shown to be competitive with existing classification methods in a simulation study. These proposed MDS-based methods are compared with an established clinical diagnostic tool for HCC.

The thesis concludes with a discussion of the developed methods and suggestions for further work.

Contents

List of Figures

Acknowledgements

I would like to thank Dr Trevor Cox and Prof. Marta Garcia-Fiñana at the University of Liverpool for their invaluable support and guidance throughout this process. Thanks also go to the Liverpool Cancer Trials Unit for supporting my application and funding the first year of this PhD programme. Additionally, thanks go to Prof. Philip Johnson for kindly granting access to the HCC data set and Dr Sarah Berhane for her assistance with the data. Also at the University of Liverpool I would like to thank Dr Ian Smith for help and assistance with use of the high throughput Condor system.

It would not have been possible to complete this work without the support of my current employer, PHASTAR, who funded me and generously granted me study leave at various points. A big thank you in particular goes to Kevin Kane, Dr Susan Lovick, Andrew Lloyd and the late Prof. Sally Hollis.

Last but not least, a huge thank you to my family who made a number of sacrifices to help me complete this thesis. In particular, I would like to thank my amazing wife Gem for getting me through this process. Without her love, support and unwavering patience this would not have been possible.

Abbreviations

AF	Acceleration factor
AFP	Alpha feto-protein
AFT	Accelerated failure time
AIC	Akaike's information criterion
ALT	Alanine aminotransferase
AST	Aspartate aminotransferase
BIC	Bayesian information criterion
CLD	Chronic liver disease
COPD	Chronic obstructive pulmonary disease
DBR	Distance-based regression
DCP	Des-gamma-carboxy prothombin
EM	Expectation maximisation (algorithm)
HCC	Hepatocellular carcinoma
HR	Hazard ratio
INR	International normalised ratio
L3	Lens culinaris agglutinin
LCA	Latent class analysis
LCR	Latent class regression
MAE	Mean absolute error
MCAR	Missing completely at random
MDS	Multidimensional scaling
OCMDS	Outcome-constrained multidimensional scaling
OTMDS	Outcome-transformed multidimensional scaling
PCR	Principal components regression
PLS	Partial least-squares
PH	Proportional hazards
QoL	Quality of life
RCT	Randomised controlled trial
RMSE	Root mean square error
ScDBR	Variable screened distance-based regression
SMDS	Supervised multidimensional scaling
WCMDS	Weakly-constrained multidimensional scaling
WDBR	Variable weighted distance-based regression
WHO	World Health Organisation

Chapter 1

Introduction

In clinical research, interest sometimes lies in analysing variables which are not measured directly. Instead, information about these 'latent variables' can be inferred from surrogates or other imperfect indicators, referred to as 'manifest variables', using latent variable models.

One type of latent variable is a hypothetical construct, for which a classic example is intelligence. A common example in clinical research is quality of life (QoL), where a subject completes a series of questions which each represent some measure of QoL, since QoL cannot be measured directly. A conceptual model for the relationship between this type of latent variable and the manifest variables is shown in Figure 1.1(a), where it is assumed that the observed variables are determined by the latent variable.

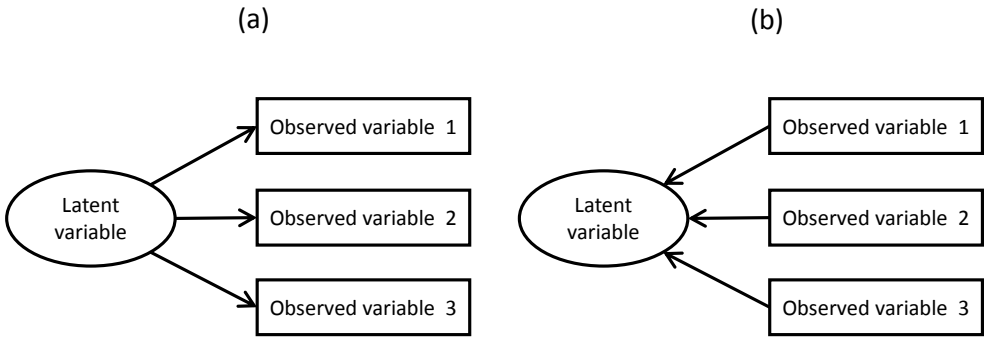

Figure 1.1: Two types of latent variable: (a) observed/manifest variables are dependent on an underlying variable such as QoL, (b) a latent variable determined by the observed data, e.g. a principal component.

Another type of latent variable is a variable used as a device for dimension reduction, for example, a principal component (Bollen, 2002). As shown in Figure 1.1(b), in this case, the latent variable is determined by the observed variables. One use of such latent

variables is to reduce the dimensionality of covariates in regression modelling, where large numbers of covariates relative to the number of observations or events can be problematic. Another possible use is in summarising many observed variables, possibly in the form of a low-dimensional visualisation to aid interpretation of the data.

Broadly, a latent variable can be defined as a random variable whose realisations are hidden (Skrondal and Rabe-Hesketh, 2004). This definition is wide-ranging, only requiring that a variable is unmeasured, not *unmeasurable*, and includes variables that are not usually referred to, or regarded, as latent variables, such as those used to model unexplained heterogeneity (i.e. random and frailty effects). Many alternative definitions exist (see e.g. Bollen, 2002, for a detailed discussion).

In this thesis, the analysis of two types of latent variable are considered: 1) 'latent classes', unobserved subgroups of subjects underlying observed clinical data, and 2) 'latent dimensions', continuous latent variables which capture key features of the observed data. Respectively, this thesis is based around the development and application of two latent variable methods: latent class analysis (LCA) and multidimensional scaling (MDS). In particular, there is an emphasis on using these latent variable methods in the context of time-to-event modelling. Time-to-event variables are common outcome measures in clinical research, for example, overall and progression free survival times in oncology, time to first severe exacerbation (attack) in asthma and chronic obstructive pulmonary disease (COPD), and time to first seizure recurrence in epilepsy.

1.1 Areas of research

In this thesis, latent variable methods are explored and developed for the statistical analysis of clinical data, with an emphasis on including latent variables in time-to-event models. For the analysis of latent classes, latent class models with a time-to-event outcome (manifest variable) are considered, and the two areas of research are:

1. Estimating the effect of latent classes on a time-to-event outcome variable.
2. Development and application of `R` code for various latent class models with a time-to-event outcome variable.

For the analysis of latent dimensions, several MDS-based methods are developed for the purpose of visualisation, modelling and prediction using clinical data. The three areas of research are:

1. Visualisation of the relationship between observations, predictor variables and a time-to-event outcome.
2. Dimensionality reduction of covariate data in regression modelling.
3. Classification, prediction and visualisation of clinical data using MDS.

1.2 Thesis layout

The content of this thesis is in two main parts: latent class-related (Chapters 2 and 3) and MDS-related (Chapters 4 to 7) statistical models. Statistical notation is kept as consistent as possible within parts, but inevitably some changes of notation are required. For clarity, notation is generally redefined within chapters. Similarly, whilst some of the statistical literature covered is applicable to multiple chapters, generally the areas covered are sufficiently diverse to warrant their own discussion of the relevant literature. In this chapter, LCA, MDS and time-to-event data are introduced. Only brief overviews of LCA and MDS are given, however, with more technical information provided in later chapters. The area of time-to-event analysis is common to both latent class and MDS-based chapters and is therefore considered in more detail here. The thesis concludes with a discussion of the developed methods and ideas and suggestions for further work.

1.3 Latent class analysis

LCA was introduced by Lazarsfeld (1959) and is a statistical method for finding latent subgroups that drive observed data. In LCA, both the latent and observed variables are categorical. Related methods are factor analysis, latent trait analysis and latent profile analysis, which are applicable to different combinations of data types (Table 1.1). Any model which includes a latent categorical variable can be regarded as a latent class model and, as will be shown in Chapter 3, these models are not limited to manifest variables of one type.

		Manifest	
		Continuous	Categorical
Latent	Continuous	Factor Analysis	Latent Trait Analysis
	Categorical	Latent Profile Analysis	Latent Class Analysis

Table 1.1: Classical latent variable methods for the four categorical/continuous combinations.

A simple simulated example is now used to illustrate LCA. Figure 1.2 depicts a model for the relationship between a latent variable and three manifest variables, which are QoL questions. The questions are taken from a QoL questionnaire, EORTC QLQ-C30 (Aaronson et al., 1993), and are related to the underlying 'physical function' domain. In this case, the purpose of LCA is to find underlying subgroups of patient which differ in their physical function, using the observed responses to the three questions.

The questions are "Do you have trouble taking a long walk?" (*Walk*), "Do you need to stay in bed or a chair during the day?" (*Chair*) and "Do you need help with eating, dressing, washing yourself or using the toilet?" (*Help*). For simplicity, the responses are

limited to "Yes" (1) or "No" (0). There are eight possible response patterns, as shown in Table 1.2 and at least some respondents' answers correspond to each pattern. LCA was applied, assuming two latent classes. Table 1.3 gives the point estimates of various probabilities (parameters) returned by the LCA model. Firstly, Class 1 appears to be a more prevalent class than Class 2 (73% vs. 27%). The remaining estimates are for the probability of responding "Yes" to each question, depending on the class. Patients belonging to Class 1 are characterised as having a lower probability of trouble taking a long walk and staying in a bed or chair during the day than Class 2, however they are more likely to require help with other activities.

Two important features of LCA in general are: 1) manifest variables are assumed to be independent given latent class, and 2) the underlying class for a respondent is not 'found'; it can only be inferred with some probability. This second property differs from, for example, some cluster analysis techniques where observations are assigned absolutely to a group.

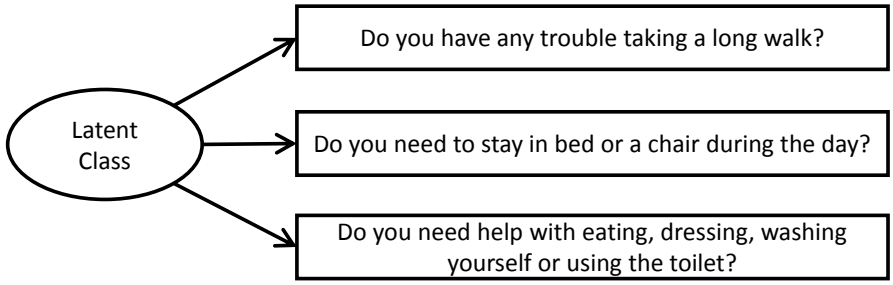

Figure 1.2: An assumed latent class model for the answer to three QoL questions in the simulated example.

Walk	Chair	Help	Observed
0	0	0	29
0	0	1	64
0	1	0	10
0	1	1	12
1	0	0	16
1	0	1	26
1	1	0	30
1	1	1	13

Table 1.2: Frequency of observed response patterns for the response to three QoL questions from the simulated example.

	Class 1	Class 2
Prevalence	0.73	0.27
Walk	0.28	0.84
Chair	0.14	0.84
Help	0.69	0.24

Table 1.3: Estimated probabilities for a two-class LCA model fitted to the simulated QoL example data set. Estimated prevalences are unconditional and sum to one. The remaining probabilities are the probability of answering "Yes" given membership to a latent class.

1.4 Multidimensional scaling

MDS originated in the 1930s (see e.g. Cox and Cox, 2000), and encompasses a broad range of methods for finding a configuration of points in low-dimensional space, where the distance between points represents their 'proximity'. Proximity data can either be in the form of dissimilarities or similarities between objects, variables, subjects, etc. Dissimilarities/similarities are found between every pair of objects and are then used in MDS analysis. For example, the 'dissimilarity' between several countries could be measured using the Euclidean (straight-line) distance. MDS can then be used to find a low (usually 2 or 3) dimensional configuration where the *distance* between points, representing the countries, approximates the *dissimilarities* between the countries. In clinical research, the observations would be subjects rather than countries and another proximity measure, Gower's general coefficient (Gower, 1971), might be more appropriate as it can accommodate different manifest variable types (nominal, ordinal, continuous).

An example of MDS is presented in Figure 1.3. A random sample of ten subjects with hepatocellular carcinoma (HCC, circles) and ten subjects with chronic liver disease (CLD, triangles) have been analysed using classical MDS with two dimensions. The full data set is described in more detail in Section 1.6. Gower's coefficient was used to measure the dissimilarity between subjects on five variables: three cancer biomarkers and two measures of liver function. It is clear from the plot that the CLD subjects form a small cluster, suggesting they have similar values for the five variables (except one subject in the bottom left corner who has an atypically high measure for one of the biomarkers). The HCC subjects are generally very spread out, suggesting a wide spread of values within this subgroup and, barring one subject, the HCC subjects appear to be quite dissimilar to the CLD subjects. The example demonstrates that MDS can be a valuable tool for obtaining a low-dimensional approximation of a multidimensional clinical data set.

An important feature of MDS is that the orientation of the configuration is essentially arbitrary since the relative distance between points is unaffected by rotation,

reflection or translation of the points (Cox and Cox, 2000). The latent dimensions may or may not have a clear substantive meaning, or some rotation may be required in order for the dimensions to be interpretable.

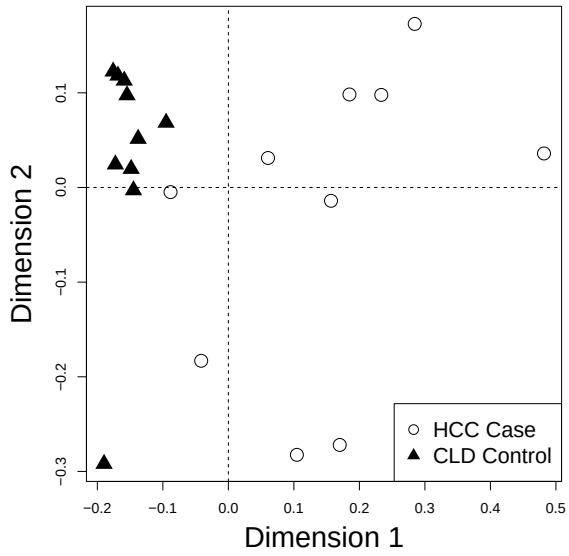

Figure 1.3: A random sample of ten HCC subjects and ten CLD subjects were analysed using classical scaling with two dimensions.

1.5 Time-to-event analysis

Throughout this thesis there is an emphasis on analysing latent variables in time-to-event models. An overview of time-to-event data is now given, but informative introductions can be found in Hougaard (1999), Kalbfleisch and Prentice (2011) and Collett (2015). Time-to-event analysis is often referred to as 'survival analysis' as it naturally arises in the analysis of survival times. However, time-to-event analysis is used widely in clinical research across many therapeutic areas, for example, time-to-first severe exacerbation (attack) in asthma and COPD, and time-to-first seizure recurrence in epilepsy. In engineering applications the term 'failure time' is often used. The event of interest does not need to have negative connotations, for example, the event time of interest could be time to remission of some disease.

Event times are usually highly skewed. In clinical trials, the distribution of event times is typically right-skewed, whereas human lifetimes tend to be exhibit left skewness (Hougaard, 1999). Perhaps the most important feature of time-to-event data is that typically the event time is not known for all subjects, and this is referred to as 'censoring'. Suppose in a clinical trial that a subject is randomised and is known to have not experienced the event of interest by 60 days but is then lost to follow-up.

This subject's event time is said to be right-censored as it is known that the event time was not experienced up to day 60, but it is not known if the event occurred the next day, the next year, or perhaps it never occurred. If a subject *completes* a trial without experiencing the event of interest, this is another type of right-censoring referred to as administrative censoring. Other types of censoring are left-censoring (the event time is known to have occurred prior to some time, but the actual time is unknown) and interval censoring (the event time is known to have occurred in some interval, but it is not known exactly when). Many statistical models make the assumption that the probability of experiencing the event and probability of censoring are independent (see e.g. Kalbfleisch and Prentice, 2011, Chapter 6). Ignoring censored event times can lead to biased and possibly misleading analyses and moreover, doing so is wasteful since at least some information on a subject's event time is known and can usefully contribute to an analysis using time-to-event methods.

Time to event analysis is now presented in more detail. Let T represent a random variable which is a non-negative event time, with observed value t. The random variable T has a probability distribution with underlying density function $f(t)$. The cumulative distribution function is then

$$F(t) = P(T < t) = \int_0^t f(v)\, dv,$$

i.e. the probability of experiencing the event before time t. Sometimes it is required to know the probability of experiencing the event at or beyond time t, and this is given by the survivor function

$$S(t) = P(T \geq t) = 1 - F(t).$$

Usually the survivor function is not modelled directly, and instead the hazard function is used. The hazard function is the instantaneous failure rate, obtained from the probability of failing at time t given that the subject has not experienced the event up to time t. Formally, the hazard function is given by

$$h(t) = \lim_{\Delta t \to 0} \left[\frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \right],$$

where Δt denotes a change in t. The approximate probability of experiencing the event in the interval $(t, t + \Delta t)$, given that it has not occurred by time t, is $h(t)\Delta t$. Defining the cumulative hazard function as $H(t) = \int_0^t h(v)\, dv$, useful relations between the aforementioned functions are:

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt} \left\{ \log \left[S(t) \right] \right\},$$

and

$$S(t) = -\exp\left[H(t)\right].$$

The functions mentioned so far have not depended on subject-level covariates. For example, the survivor function for a subject with late stage cancer would be expected to differ from a subject with early stage cancer. The two most common type of statistical models for modelling the effect of covariates on event times are 'proportional hazards' (PH) and 'accelerated failure time' (AFT) models. A PH model for an individual is given by

$$h(t|\mathbf{x}) = h_0(t)\exp(\boldsymbol{\beta}^\intercal \mathbf{x}),$$

where $h_0(t)$ is the 'baseline' hazard function, $\mathbf{x}$ is a vector of covariates for an individual, e.g. randomised treatment, cancer stage and age, and $\boldsymbol{\beta}$ is a vector of corresponding parameters for the effects of the covariates. The quantity $\exp(\beta_p)$ is the 'hazard ratio' for the pth variable, with values greater than one corresponding to a reduced expected event time. It is assumed that the hazard function increases or decreases proportionally across different values of x_p.

The baseline hazard function can be regarded as the reference, i.e. the hazard function for a subject with all covariates coded to equal zero. Parametric models assume that the baseline hazard function follows from an assumed probability distribution for T (e.g. Weibull), whilst semi-parametric models, including the most widely applied time-to-event model, the Cox model (Cox, 1972), make few or no distributional assumptions about the form of the baseline hazard function.

In contrast to PH models, AFT models instead model the effect of covariates directly on the time-scale so that the effect of a covariate can be interpreted as the speeding up (or slowing down) of the event time, for example, the progression of a disease. Let $S_0(t)$ represent the survival function for a subject with all covariates coded to zero, then a general survivor function according to an AFT model is given by

$$S_0\left[\frac{t}{\exp(\boldsymbol{\alpha}^\intercal \mathbf{x})}\right], \tag{1.1}$$

where $\boldsymbol{\alpha}$ is a vector of covariate effects. The quantity $\exp(-\alpha_p)$ is the 'acceleration factor' for the pth variable, where values less than one correspond to a shortening of the event time. AFT models can also be expressed in a linear form and, as with PH models, there are parametric and semi-parametric versions. Both of these features of AFT models are discussed in detail in Chapter 5. One advantage of AFT models over PH models is that they are more robust to the influence of covariates which are not included in the model (see e.g. Hougaard, 1999).

1.6 Data sets

Three clinical data sets are used for statistical analyses in this thesis. Whilst all three are oncology-based, the methods described are not limited to oncology and are in fact equally relevant to other therapeutic areas.

1.6.1 Hepatocellular carcinoma

HCC is the most common type of primary liver cancer and can be a complication of liver cirrhosis (Tan and Huang, 2018). Depending on tumour-related, liver function and patient-related factors, the prognosis for a patient diagnosed with HCC may be extremely poor (Cancer Research UK, 2018). The HCC data set was collected for a case-control diagnostic study (Johnson et al., 2014) at the Queen Elizabeth Hospital (Birmingham, UK), and contains four key subgroups:

I HCC patients prospectively recruited between 2007 and 2012 ($N = 315$). As part of the original study, this subgroup was split into two cohorts for statistical model training ($N = 218$, recruited 2007-11) and internal validation ($N = 97$, recruited 2011-12).

II Prospectively recruited control patients with CLD recruited from patients attending outpatient clinics ($N = 339$), split into training ($N = 247$) and internal validation ($N = 92$) cohorts.

III Prospectively recruited liver-healthy control subjects ($N = 92$).

IV HCC patients for which data were collected retrospectively from the hospital data base ($N = 409$). An important feature of this last group, in particular, is that there is a considerable amount of missing data as not all of the variables available for other subgroups were recorded in the hospital database.

To suit different analysis purposes, various combinations of these subsets are used in analyses presented in this thesis. The original study was largely based around the collection and analysis of three serum biomarkers: Alpha-fetoprotein (AFP), lens culinaris agglutinin (L3) and des-gamma-carboxy prothombin (DCP), for which elevated levels are suggestive of HCC. Additionally, the data set contains a number of other demographic and disease-related characteristics. These variables are discussed in more detail as part of the analyses.

1.6.2 Pancreatic cancer

Pancreatic cancer is one of the most deadly cancers and one of the most common causes of cancer mortality in the UK (Cancer Research UK, 2017). The ESPAC3v2

trial was an open-label randomised controlled trial (RCT) in patients with pancreatic ductal adenocarcinoma who had undergone cancer resection (Neoptolemos et al., 2010). Patients were randomised to either fluorouracil plus folinic acid or gemcitabine (the standard of care). This data set is not analysed in detail but is used to simulate realistic survival times in Chapter 2.

1.6.3 Prostate cancer

The Byar prostate cancer data set is a publicly available data set which can be downloaded from http://lib.stat.cmu.edu/datasets/Andrews, Table 46.1. The data set is from an RCT comparing different doses of diethylstilbestrol in patients with Stage III and Stage IV prostate cancer (Bailar III et al., 1970; Byar and Corle, 1977). This data set was analysed using latent class models by Hunt and Jorgensen (1999), where latent classes were found which differed from the clinically assigned tumour stages. To assess the survival prospects for these latent classes however, Hunt and Jorgensen (1999) used simple cross-tabulations against a landmark survival time of 48 months, rather than modelling survival times directly. The data set is used in this thesis to analyse the effect of latent classes on survival for prostate cancer patients using patient-level survival data.

1.7 Summary

In this chapter, latent variables were discussed and specifically the ideas of underlying latent classes and latent dimensions, which are relevant to the statistical methods of LCA and MDS, were discussed. An overview of time to event analysis was also provided. Finally, the three clinical data sets used in this thesis were described.

Chapter 2

Latent class modelling with a time-to-event distal outcome: A comparison of one, two and three-step approaches

2.1 Introduction

Latent class methods encompass a broad range of models which can be used to identify and characterise unobserved subgroups which differ in their observed or 'manifest' data. These models have been widely applied in many scientific disciplines including medicine (e.g. Downing et al., 2010; Rahbar et al., 2015), social and behavioural science (e.g. Chung et al., 2006; Stapinski et al., 2016) and education (e.g. Denson and Ing, 2014; Auer et al., 2016). For example, Stapinski et al. (2016) used latent class analysis (LCA) of a large cohort study to identify four groups of adolescents who differed in their motives for alcohol use.

So far, only simple LCA models have been introduced, however, a common objective of latent class methods is to assess the relationship between the identified latent classes and a distal outcome variable. In some disciplines, time-to-event variables are common outcome measures, for example, overall and progression-free survival times in oncology. Time-to-event variables differ from other variable types since they are typically highly skewed and subject to censoring (see Chapter 1). Applications of various latent class models with a time-to-event distal outcome can be found in Snuderl et al. (2008), Muthén et al. (2009), Zhang and Wang (2010), Desantis et al. (2012) and Leigh et al.

(2015).

When model assumptions are met, a preferred statistical approach is to jointly model the latent classes and distal outcome in one step (Bakk et al., 2013). Larsen (2004) introduced a one-step latent class model with a time-to-event distal outcome variable and a framework for continuous time latent class models was set out by Asparouhov et al. (2006). A general criticism of one-step approaches, however, is that the distal outcome variable can influence the composition of the latent classes (Vermunt, 2010; Asparouhov and Muthén, 2014). Moreover, one-step approaches may be impractical if there are many distal outcome variables, or if the outcome data are collected at a different stage of a trial and/or by different researchers (Vermunt, 2010).

A simple and frequently applied alternative approach to incorporating a distal outcome variable into a latent class model is the 'Classify-Analyze' (Clogg, 2013) or 'standard three-step approach': Step 1) a latent class model is fitted, Step 2) subjects are assigned to a latent class, and Step 3) the distal outcome is regressed on the assigned class. Whilst intuitive, the standard three-step approach has two important drawbacks. Firstly, estimates of the relationship between latent class and the distal outcome variable can be attenuated due to misclassification in Step 2 (Bolck et al., 2004). Secondly, standard errors in Step 3 can be underestimated since class is treated as known in the regression model, potentially misleading statistical inference (Clark and Muthén, 2009). Bray et al. (2015) identified that non-inclusion of the distal outcome variable in the classification model (Step 1) as a further cause of bias in Step 3 and proposed an 'inclusive' approach to correct for this bias, where the distal *outcome* variable is included as a latent class *predictor* variable in Step 1, along with other covariates.

Bakk and Kuha (2018) proposed a two-step alternative to address the aforementioned issues with one and standard three-step approaches. In this approach, a latent class model is fitted in Step 1, as in the three-step approach. Then, in Step 2, the full joint latent class and distal outcome model is fitted, as in the one-step approach, but the parameters for the latent class part of the model are held fixed at their estimates from Step 1. A correction is then applied to account for additional uncertainty in the second step.

Research into estimating the effect of latent class on distal outcomes has so far been restricted to categorical or continuous outcome variables (Clark and Muthén, 2009; Bakk et al., 2013; Lanza et al., 2013; Asparouhov and Muthén, 2014; Bray et al., 2015; Bakk and Vermunt, 2016; Collier and Leite, 2017; Bakk and Kuha, 2018).

In this chapter, Monte Carlo simulation is used to compare one, two and three-step approaches to latent class modelling with a time-to-event distal outcome. For the one and two-step approaches, joint latent class models with piecewise constant baseline hazard functions are used (Asparouhov et al., 2006; Muthén et al., 2009). For the three-step models, four approaches to class assignment are compared and the impact of the

inclusive approach for bias-correction (Bray et al., 2015) with a time-to-event distal outcome variable is assessed. This work was published in the Journal of Structural Equation Modeling: A Multidisciplinary Journal (Lythgoe et al., 2019).

2.2 Latent class modelling with a time-to-event distal outcome

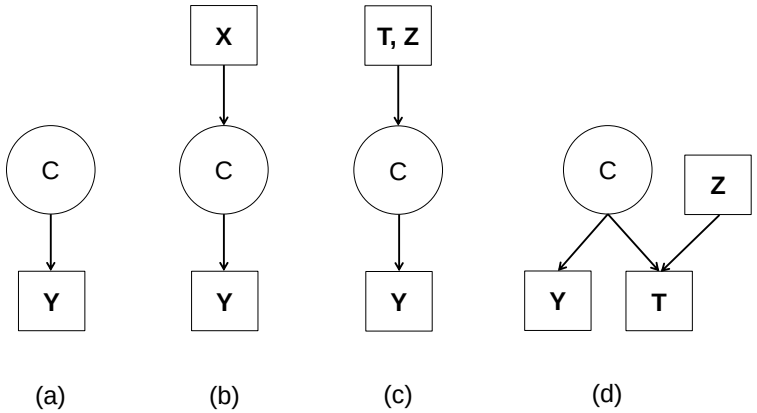

Figure 2.1: Schematics for the latent class models discussed in this chapter: (a) latent class model, (b) latent class regression model, (c) inclusive latent class regression model and (d) one-step latent class model with a distal outcome. Circles and squares are used to identify unobserved (i.e. latent class) and observed variables, respectively. C latent class variable, $\mathbf{Y}$ manifest variables, $\mathbf{X}$ latent class predictors, $\mathbf{T}$ distal outcome(s) and $\mathbf{Z}$ covariates possibly related to $\mathbf{T}$.

2.2.1 The one-step approach

In this section, latent class models are introduced and a full one-step latent class model with a time-to-event distal outcome, as introduced by Larsen (2004), is developed.

The latent class model

LCA was introduced by Lazarsfeld (Lazarsfeld, 1959) and is used to identify and characterise unobserved and mutually exclusive subgroups using multiple imperfect indicators known as manifest variables. The basic latent class model is depicted in Figure 2.1(a). The latent class variable, C, is assumed to consist of J categories with prevalences

$$P(C = j) = \eta_j,$$

for j $= 1, \ldots, J$ and $\sum_{j=1}^{J} \eta_j = 1$. Let $\mathbf{Y} = (Y_1 \ldots, Y_M)^\intercal$ denote a vector of manifest variables with observed values $\mathbf{y} = (y_1, \ldots, y_M)^\intercal$ for a given subject. Typically each Y_m $(m = 1, \ldots, M)$ is categorical with $g = 1, \ldots, G_m$ categories, so that the probability of observing category g on the mth manifest variable for subjects in the jth class is given by

$$P(Y_m = g|C = j) = \pi_{mgj} = \prod_{g=1}^{G_m} \pi_{mgj}^{I\{y_m = g\}},$$

where $\sum_{g=1}^{G_m} \pi_{mgj} = 1$, and $I\{y_m = g\}$ is an indicator function which equals 1 if y_m takes the value g and 0 otherwise, for a given subject. Other distributions for the manifest variables are discussed in Chapter 3. The distribution of the responses for an individual is given by

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}) &= \sum_{j=1}^{J} P(C = j) f_{\mathbf{Y}|C}(\mathbf{y}|j) \\ &= \sum_{j=1}^{J} \eta_j \prod_{m=1}^{M} f_{Y_m|C}(y_m|j), \end{aligned} \tag{2.1}$$

where the manifest variables are assumed to be independent conditional on class and $f(.)$ is used to denote a probability density or mass function as required. Some options for introducing dependencies between manifest variables can be found in Hunt and Jorgensen (1999) and Desantis et al. (2012), and are discussed in Chapter 3. The posterior probability that a subject belongs to class j given $\mathbf{Y} = \mathbf{y}$ is obtained using Bayes theorem, so that

$$P(C = j|\mathbf{Y} = \mathbf{y}) = \frac{\eta_j f_{\mathbf{Y}|C}(\mathbf{y}|j)}{\sum_{k=1}^{J} \eta_k f_{\mathbf{Y}|C}(\mathbf{y}|k)}. \tag{2.2}$$

Latent class regression

A natural extension to the latent class model (equation 2.1) is the concomitant-variable or 'latent class regression' (LCR) model (Dayton and Macready, 1988; Formann, 1992; van der Heijden et al., 1996; Bandeen-Roche et al., 1997; Chung et al., 2006), as depicted in Figure 2.1(b). In the LCR model the class prevalences, η_j, are allowed to vary as a function of a vector of 'latent class predictors' $\mathbf{X}$, with observed values $\mathbf{x}$. Following on from equation 2.1 the distribution function for a given subject is

$$f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) = \sum_{j=1}^{J} \eta_j(\mathbf{x}) \prod_{m=1}^{M} f_{Y_m|C}(y_m|j),$$

where the latent class predictors and manifest variables are assumed to be conditionally independent given latent class. Huang and Bandeen-Roche (2004) showed how depen-

dencies between latent class predictors and categorical manifest variables can be added to the model. A generalised linear model with a logit link function is used to model the relationship between the latent class predictors and class prevalences, so that the inverse of the logit link function is

$$P(C = j|\mathbf{X} = \mathbf{x}) = \eta_j(\mathbf{x}) = \frac{\exp(\mathbf{x}^\intercal \boldsymbol{\kappa}_j)}{\sum_{k=1}^{J} \exp(\mathbf{x}^\intercal \boldsymbol{\kappa}_k)}, \tag{2.3}$$

for $j = 1, \ldots, J$ and where $\boldsymbol{\kappa}_j$ is a vector of log odds ratios for the jth class, $\boldsymbol{\kappa}_J = \mathbf{0}$ for identifiability and the first element of $\mathbf{x}$ is set to 1 in order to include an intercept. An intercept only LCR model is equivalent to the latent class model. Other suitable link functions can be used. To obtain posterior probabilities equation 2.2 is updated to $P(C = j|\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{X} = \boldsymbol{x})$.

A time-to-event distal outcome model

Time-to-event variables are typically highly skewed and subject to censoring, since the event of interest is not always observed (see Chapter 1). Note that it is assumed throughout that the risk of censoring and experiencing the event are independent.

Larsen (2004) extended the latent class model to include a time-to-event submodel in order to model latent class and a time-to-event distal outcome in one step, as depicted in Figure 2.1(d). An extensive framework for these continuous time one-step or 'joint' models is presented in Asparouhov et al. (2006) and software functionality is available in M-Plus (Muthén and Muthén, 2011). Let T denote a time-to-event variable with observed value t. One option for the time-to-event submodel is a proportional hazards model extended to include a latent class effect

$$\alpha(t|\mathbf{Z} = \mathbf{z}, C = j) = \alpha_0(t) \exp(\mathbf{z}^\intercal \boldsymbol{\beta} + \gamma_j), \tag{2.4}$$

for $j = 1, \ldots, J$ where $\alpha(t|\cdot)$ represents the hazard for a given subject at time t, $\alpha_0(t)$ is the baseline hazard at time t, $\boldsymbol{\beta}$ is a vector of log hazard ratios for the corresponding covariates $\mathbf{z}$ and γ_j represents the log hazard ratio for the effect of latent class j on the baseline hazard, with $\gamma_J = 0$ for identifiability. In this model, both the covariate and class effects are assumed to act proportionally on the baseline hazard and independently of time. Options for assessing the suitability of the proportionality assumption are discussed in Section 2.5.

A useful approach to modelling the baseline hazard function is the piecewise exponential model (Friedman, 1982), where the baseline hazard function is assumed to be piecewise constant. For a piecewise exponential time-to-event submodel, let time be partitioned into $s = 1, \ldots, S$ intervals and let $\boldsymbol{\alpha_0} = (\alpha_{01}, \ldots, \alpha_{0S})^\intercal$ denote a vector of

baseline hazard parameters. To complete the required notation, let Δ represent the censoring indicator with observed value δ, where δ equals 1 if the event is observed and 0 otherwise. The required density function of the event time for a given subject is

$$f_{T,\Delta|\mathbf{z},C}(t,\delta|\mathbf{z},j) = \prod_{s=1}^{S} \left[\alpha_{0s}\exp(\mathbf{z}^\intercal\boldsymbol{\beta}+\gamma_j)\right]^{\delta\psi_s} \times \exp\left\{ -\psi_s\left[\alpha_{0s}(t-a_{s-1}) + \sum_{h=1}^{s-1}\alpha_{0h}(a_h-a_{h-1})\right]\exp(\mathbf{z}^\intercal\boldsymbol{\beta}+\gamma_j)\right\},$$

for $j = 1,\ldots,J$ and where ψ_s denotes an indicator variable which equals 1 if the event occurs in the sth interval and 0 otherwise, a_s denotes the upper boundary for the sth interval on the time grid and a_0 equals 0. The joint density for the manifest variables and time-to-event distal outcome for a given subject is then

$$f_{\mathbf{Y},T,\Delta|\mathbf{z}}(\mathbf{y},t,\delta|\mathbf{z}) = \sum_{j=1}^{J}\eta_j\prod_{m=1}^{M} f_{Y_m|C}(y_m|j)f_{T,\Delta|\mathbf{z},C}(t,\delta|\mathbf{z},j), \tag{2.5}$$

where the distributions of the manifest variables and time-to-event distal outcome are assumed to be conditionally independent given class. Latent class predictors can also be included, as in equation 2.3, as will be shown in Chapter 3, but will not be considered in this chapter. The log-likelihood of the observed data is given by

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N}\log\left[f_{\mathbf{Y}_i,T_i,\Delta_i|\mathbf{z}_i}(\mathbf{y}_i,t_i,\delta_i|\mathbf{z}_i)\right], \tag{2.6}$$

where N is the total number of subjects indexed by i and $\boldsymbol{\theta} = (\boldsymbol{\eta}^\intercal,\boldsymbol{\pi}^\intercal,\boldsymbol{\alpha}^\intercal,\boldsymbol{\beta}^\intercal,\boldsymbol{\gamma}^\intercal)^\intercal$ denotes the full vector of parameters to be estimated. Note that for estimation purposes, Lagrange multipliers are required to implement the constraint $\sum_{g=1}^{G_m}\pi_{mgj} = 1$ (see e.g. Bartholomew et al., 2011, Chapter 6), but further Lagrange multipliers are not required for the constraint $\sum_j^J \eta_j = 1$ if equation 2.3 is used. Further details on model estimation are given in Chapter 3.

2.2.2 The two-step approach

For the two-step approach of Bakk and Kuha (2018), the required parameters from the one-step model are partitioned into those to be estimated in Steps 1 and 2 so that $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\intercal,\boldsymbol{\theta}_2^\intercal)^\intercal$. In the first step a latent class model is fitted (equation 2.1) so that $\boldsymbol{\theta}_1 = (\boldsymbol{\eta}^\intercal,\boldsymbol{\pi}^\intercal)^\intercal$ and therefore $\boldsymbol{\theta}_2 = (\boldsymbol{\alpha}^\intercal,\boldsymbol{\beta}^\intercal,\boldsymbol{\gamma}^\intercal)^\intercal$. Let $\tilde{\boldsymbol{\theta}}_1$ denote the estimates from Step 1 and then in Step 2 maximise the log-likelihood for the observed data conditional on the Step 1 estimates, i.e. $\ell(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1 = \tilde{\boldsymbol{\theta}}_1)$. The required log-likelihood is given in

equation 2.6.

Clearly the uncertainty of the estimates obtained in Step 2, $\tilde{\boldsymbol{\theta}}_2$, will be underestimated since the Step 1 parameters have been held fixed during estimation. Xue and Bandeen-Roche (2004) and Bakk and Kuha (2018) demonstrate how to obtain corrected standard errors in the two-step approach, as is discussed in detail in Section 3.7.2.

2.2.3 Standard and inclusive three-step approaches

Three-step approaches proceed as follows: Step 1) a latent class model is fitted, Step 2) an assignment rule is used to classify subjects according to their class conditional posterior probabilities, Step 3) the assigned classes are used as a covariate in a regression model to estimate the relationship between the latent classes and the external variable. These steps are now considered in more detail in the context of modelling with a time-to-event distal outcome.

Step 1: Fit the latent class model

For Step 1 in a standard three-step approach a latent class model is simply fitted, as in equation 2.1 and depicted in Figure 2.1(a). In an inclusive three-step approach an LCR model is fitted with the distal *outcome* variable as a latent class *predictor* variable, along with other covariates related to the outcome, as depicted in Figure 2.1(c). How then might we incorporate an event time subject to censoring as a latent class predictor? For the purposes of multiple imputation of baseline covariate data in proportional hazards models, White and Royston (2009) recommended using the estimated cumulative hazard function (notably in preference to the observed survival time or its natural logarithm), the event indicator and other covariates related to the event time in the model. Expressing equation 2.3 in logit form and replacing $\mathbf{x}$ with the required elements the inclusive model is given by

$$\text{logit}\, P(C = j|H(t), \Delta = \delta, \mathbf{Z} = \mathbf{z}) = \kappa_{j0} + \kappa_{j1}H(t) + \kappa_{j2}\delta + \kappa_{j3}z,$$

for $j = 1, \ldots, J$ and where $H(t)$ is the (non-parametric) Nelson-Aalen estimate of the unconditional cumulative hazard, which is estimated separately. For illustration purposes only a single covariate, z, has been included but additional covariates can be incorporated easily.

Step 2: Class assignment

In Step 2, subjects are assigned to a latent class according to an assignment rule. The simplest and most commonly used assignment rule is modal assignment (MA) in which

each subject is assigned to the latent class for which they have the highest posterior probability. MA ensures that all subjects with the same response pattern are allocated to the same class.

Another commonly used method is random assignment, also known as the 'pseudo class' method (PC). For PC, class is imputed once for each subject by randomly drawing from a multinomial distribution with probabilities equal to the subject's posterior probabilities from the latent class model (Bolck et al., 2004; Bandeen-Roche et al., 1997). Consequently not all subjects with the same response pattern are guaranteed to be assigned to the same class. Wang et al. (2005) introduced multiple pseudo class draws (mPC) to improve estimation efficiency over a single random draw. With mPC, class is imputed multiple times for each subject, with the authors recommending at least 20 random draws. Note that mPC is distinct from multiple imputation since the estimated posterior probabilities are effectively treated as known (Wang et al., 2005).

Finally, partial assignment (PA) and proportional assignment (PrA) are highlighted. In these methods, each subject is assigned partially rather than absolutely to a latent class. In PA, no assignment is made and posterior probabilities are used in further analyses. In PrA each subject is assigned to all classes simultaneously with case-weights equal to their corresponding class-specific probabilities, and as a result each subject will enter any further analyses J times.

Step 3: Estimate the effect of latent class on the distal outcome

In Step 3, the distal outcome variable is regressed on the assigned class from Step 2, possibly in addition to other relevant covariates. For a time-to-event distal outcome the Cox proportional hazards model (Cox, 1972) is a natural choice and is utilised in the subsequent simulation study.

For MA and PC, $J-1$ dummy variables are used to represent the assigned/imputed class in the regression model. For mPC this process is repeated for each class imputation and parameter estimates are combined across regression models using Rubin's rules (Rubin, 2004). For PA, $J-1$ posterior probabilities are included as covariates in the regression model. For PrA each subject is included in the regression model J times with case-weights equal to the posterior probabilities from the latent class model. One consequence of PrA in a time-to-event setting is that tied event times are introduced.

2.2.4 Entropy

The extent to which latent classes can be distinguished by the data and the latent class model can be assessed using the principle of entropy (Muthén and Muthén, 2004; Bakk et al., 2013). The Ramaswamy entropy statistic (Ramaswamy et al., 1993; Muthén and

Muthén, 2004; Dziak et al., 2014) is defined as

$$E = 1 - \frac{\sum_{i=1}^{N}\sum_{j=1}^{J} -\hat{p}_{ij}\log(\hat{p}_{ij})}{N\log(J)},$$

for a sample of $i = 1, \ldots, N$ subjects and where $\hat{p}_{ij}$ is the estimated posterior probability of the ith subject belonging to class j from a latent class model. E can take values between 0 and 1, where 0 indicates that the model contains no information on class assignment and 1 indicates that all subjects are estimated to belong to a class with 100% probability. Lower entropy implies that classes are less well distinguished and corresponds to greater classification error being introduced in Step 2 for three-step methods. Note that this is contrary to classical entropy measures for which low values correspond to better classification (see e.g. Dziak et al., 2014).

2.2.5 A label switching solution

Latent class models are only identifiable up to a permutation of class labels (McLachlan and Peel, 2004). Whilst this is not an issue in standalone applications, it is a problem for simulation studies since it is not always straightforward to establish, for a particular simulated data set, the class label that corresponds to the true class. A useful discussion of this issue in latent class models is given in Tueller et al. (2011), and the same labelling problem can arise in Bayesian estimation of mixture distributions using Monte Carlo Markov Chain simulations (Celeux et al., 2000; Grün and Leisch, 2009; Sperrin et al., 2010).

A number of solutions have been proposed (e.g. Tueller et al., 2011; Yao, 2015; Celeux et al., 2000). For the simulation study described in the next section, a clustering and relabelling strategy based on Euclidean distances was used, where the distances between the true parameter values and their estimates were calculated for each simulated data set. A similar idea is presented in Celeux et al. (2000) and the proposed solution is now presented and justified.

Assume that data are simulated according to a particular latent class model with P 'true' parameter values $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_P)^\intercal$. There are $J!$ possible permutations of the class labels, $l = 1, \ldots, J!$, and the last permutation is assumed to represent the correct labelling. In a simulation study, $d = 1, \ldots, D$ data sets are simulated according to the true model. For each data set a latent class model of the same form as the true model is fitted. Let $\hat{\boldsymbol{\theta}}_d = (\hat{\theta}_{d1}, \ldots, \hat{\theta}_{dP})^\intercal$ represent a vector of parameter estimates from the latent class model fitted to the d^{th} data set. It is assumed that $\hat{\boldsymbol{\theta}}_d$ contains unbiased estimates of the true values but possibly labelled incorrectly. If $\hat{\boldsymbol{\theta}}_d$ are labelled 'correctly', then

$$\frac{1}{\mathrm{se}(\hat{\theta}_{dp})}\left(\hat{\theta}_{dp}-\theta_p\right)\sim N(0,1)$$

and

$$\left[\frac{1}{\mathrm{se}(\hat{\theta}_{dp})}\left(\hat{\theta}_{dp}-\theta_p\right)\right]^2\sim\chi^2_{(1)},$$

for $d=1,\ldots,D$ and $p=1,\ldots,P$. It is then assumed that the parameter estimates are independent, which in practice will be determined by the form of the model fitted (this issue is discussed in further detail below). Summing over P,

$$\sum_{p=1}^{P}\left[\frac{1}{\mathrm{se}(\hat{\theta}_{dp})}\left(\hat{\theta}_{dp}-\theta_p\right)\right]^2\sim\chi^2_{(P)},$$

with mean P, for $d=1,\ldots,D$. The standardised Euclidean distance, τ_d, between the estimates from a model fitted to the dth data set and the vector of true parameter values is

$$\tau_d=\left\{\sum_{p=1}^{P}\left[\frac{1}{\mathrm{se}(\hat{\theta}_{dp})}(\hat{\theta}_{dp}-\theta_p)\right]^2\right\}^{1/2}\sim\chi_{(P)},$$

i.e. a central χ distribution. If $\hat{\boldsymbol{\theta}}_d$ are labelled 'incorrectly', then

$$\frac{1}{\mathrm{se}(\hat{\theta}_{dp})}\left(\hat{\theta}_{dp}-\theta_p\right)\sim N\left(\mu_p,1\right),$$

$$\sum_{p=1}^{P}\left[\frac{1}{\mathrm{se}(\hat{\theta}_{dp})}\left(\hat{\theta}_{dp}-\theta_p\right)\right]^2\sim\chi^2_{(P)}(\lambda),$$

i.e. a non-central χ^2 distribution with non-centrality parameter $\lambda=\sum_{p=1}^{P}\mu_p^2$ and mean $P+\lambda$. It therefore follows that

$$\tau_d\sim\chi_{(P)}\left(\lambda\right).$$

Letting $\boldsymbol{\tau}=(\tau_1,\ldots,\tau_D)^{\intercal}$, and assuming that the random starting values for the parameter estimates do not favour one label permutation over another,

$$\boldsymbol{\tau}\sim\frac{1}{J!}\sum_{l=1}^{J!}f_l(\boldsymbol{\tau}),$$

i.e. a $J!$ component mixture distribution with one central χ distribution, $f_{J!}\sim\chi_{(P)}$, and $J!-1$ non-central χ distributions, $f_l\sim\chi_{(P)}\left(\lambda_l\right)$, for $l=1,\ldots,(J!-1)$. A histogram of $\boldsymbol{\tau}$ should therefore yield a mixture distribution of $J!$ (hopefully distinct) probability distributions for which the component with the lowest mean is labelled correctly. Larger

differences in the true parameter values for the latent classes and greater numbers of class distinct parameters to estimate will result in clearer separation of the mixture components, making clustering and relabelling easier. An example of such a histogram for 2000 simulations from a latent class model with $J = 2$ and $P = 21$ parameters is depicted in Figure 2.2. Assuming sufficient separation between components, either by introducing some threshold or by clustering $\boldsymbol{\tau}$ (e.g. K-means clustering), estimates that have been labelled incorrectly can be easily identified and relabelled accordingly.

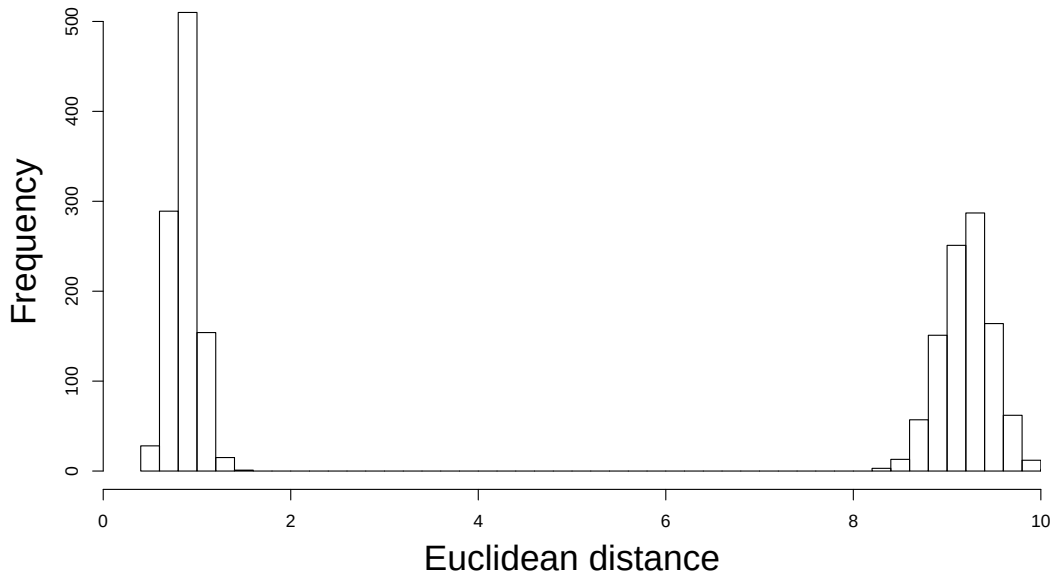

Figure 2.2: Example of Euclidean distances for 2000 simulations from a latent class model with 2 classes, before relabelling. The distribution on the left contains the models for which the class is correctly labelled.

A label switching procedure is therefore as follows

1. Fit latent class models to each of $d = 1, \ldots, D$ data sets.
2. For each of the $d = 1, \ldots, D$ data sets calculate the standardised Euclidean distances, τ_d, between each set of parameter estimates, $\hat{\boldsymbol{\theta}}_d$, and the true parameter values, $\boldsymbol{\theta}$.
3. Inspect a histogram of $\boldsymbol{\tau}$ for distinct component densities.
4. Use e.g. K-means clustering to assign each τ_d (and hence $\hat{\boldsymbol{\theta}}_d$) to a cluster (/component density). The cluster with the lowest mean corresponds to the cluster of correctly labelled parameter estimates.
5. Relabel those $\hat{\boldsymbol{\theta}}_d$ which do not belong to the correctly labelled component density.

As a check, the first three steps can be repeated using the relabelled estimates and the new histogram should reveal a unimodal (and central) χ distribution. If $J > 2$

it may be necessary to repeat this process using a permutation of the true parameter values in the place of the true values in order to distinguish between two or more incorrectly labelled clusters.

The histogram of $\boldsymbol{\tau}$ also serves as a useful diagnostic tool, since any outlying values, perhaps exceeding a selected critical threshold, can be identified and investigated further. These may represent local maximum and/or boundary solutions.

In practice, whilst theoretical justification for the distribution of the standardised Euclidean distances in the case of independent parameters has been provided, if dependencies are included in the model, then the procedure can still be used. In this case, these parameters can be included or excluded, as long as the histogram of the standardised Euclidean distances reveals distinct clusters. If the entropy is low, unfortunately the component distributions may overlap in which case it may not be possible to relabel the estimates with 100% accuracy using this method.

2.3 Monte Carlo Simulation Study

2.3.1 Aims

The purpose of this Monte Carlo simulation study was to investigate the empirical properties of latent class effect estimates on a time-to-event distal outcome using a number of different models and simulated scenarios. In particular, it was aimed to compare one, two, standard three and inclusive three-step approaches. For both the standard and inclusive three-step approaches, subjects were assigned to classes using four different approaches: MA, mPC, PA and PrA.

2.3.2 Software

Data were simulated using `R` (R Core Team, 2017, Version 3.5.2). Step 1 latent class models for the three-step approaches were fitted using R package `poLCA` (Linzer and Lewis, 2011) and Step 3 Cox regression models were fitted using the `coxph()` function in the `survival` package (Therneau, 2015), with the default of Efron's method for tied survival times. Reported standard errors and Wald 95% confidence intervals are those returned from these packages. Robust standard errors were used for three-step models with PrA to account for observations entering the analysis model twice. One and two-step models were fitted using an author-written R function, `LCSM()`, which uses an adapted version of the estimation routine detailed in Larsen (2004) to include a piecewise exponential time-to-event submodel. Both `LCSM()` and `poLCA` use the expectation-maximisation (EM) algorithm (Dempster et al., 1977) with Newton-Raphson steps to obtain maximum likelihood estimates (Larsen, 2004; Linzer and Lewis, 2011). The

`LCSM()` function and model fitting process are described in detail in Chapter 3. Standard errors for one and two-step models were obtained using Louis's method (Louis, 1982, see Chapter 3). Standard errors in the two-step models were corrected to account for Step 1 parameter fixing as described previously (Xue and Bandeen-Roche, 2004; Bakk and Kuha, 2018). Simulations were conducted using the Advanced Research Computing Condor high throughput environment at the University of Liverpool (Smith, 2017).

2.3.3 Data simulation

Two-class models with equal prevalences and ten independent Bernoulli distributed manifest variables were simulated. The factors manipulated were (a) sample size, $N \in \{500, 1000\}$, (b) approximate entropy statistic values, E of 0.35, 0.50, 0.70 and (c) the hazard ratio for the latent class effect, $\exp(\gamma_1) \in \{1, 1.5, 2, 3\}$ (note that $\exp(\gamma_2) = 1$, i.e. no effect, for identifiability), giving 24 simulation scenarios in total.

The entropy statistic values are similar to those used previously (Clark and Muthén, 2009) and correspond to low, medium and high class separation respectively. The manifest variables were simulated from independent Bernoulli distributions according to a crossed-profile plot, as depicted in Figure 2.3 and used previously by Clark and Muthén (2009). The entropy settings were obtained by varying the class conditional response probabilities, $\pi_{(1)} \in \{0.60, 0.65, 0.70\}$ and $\pi_{(2)} = 1 - \pi_{(1)}$.

Simulated event times were based on observed data from the ESPAC3v2 trial (Neoptolemos et al., 2010). The ESPAC3v2 trial was an open-label randomised controlled trial in patients with pancreatic ductal adenocarcinoma who had undergone cancer resection. Patients were randomised to either fluorouracil plus folinic acid or gemcitabine (the standard of care). Survival times were generated using the Kaplan-Meier estimate of the overall survival curve from the gemcitabine arm as described in Appendix A.

The hazard ratio values for the latent class effect, $\exp(\gamma_1)$, were chosen to represent no effect and approximate small, medium and large effect sizes respectively (Azuero, 2016). In addition to the latent class effect, an independent Bernoulli distributed time-to-event covariate, z, with a probability of 0.5 was simulated for each subject. This covariate was included to mimic randomised treatment in a clinical trial setting and the effect on survival was fixed across simulations as $\exp(\beta) = 0.75$. Administrative censoring was applied at 60 months and uniform censoring was added by generating censoring times from an exponential distribution such that overall approximately 50% of survival times were right-censored in each scenario.

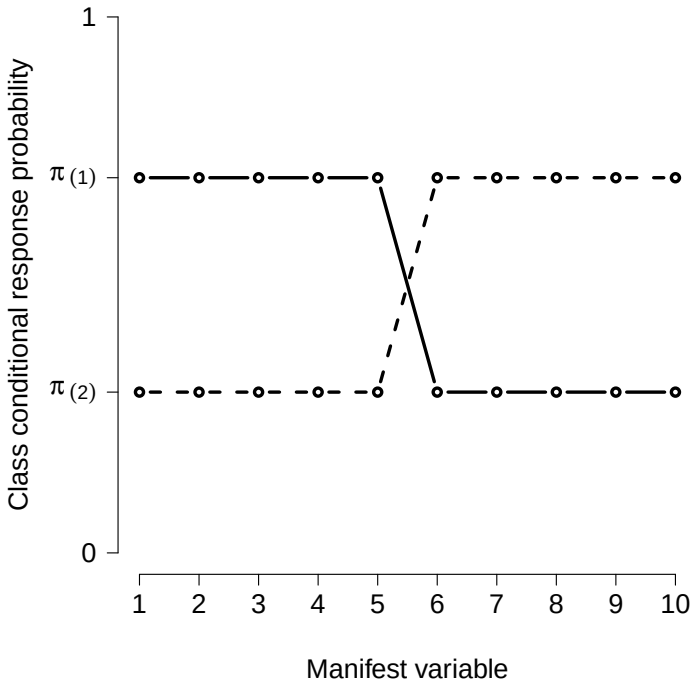

Figure 2.3: Class-conditional response probabilities used in the simulation study. Ten independent Bernoulli distributed manifest variables were simulated according to a crossed profile plot for the two latent classes, where $\pi_{(1)} \in \{0.60, 0.65, 0.70\}$ and $\pi_{(2)} = 1 - \pi_{(1)}$.

2.3.4 Model fitting

The ten fitted model types (1 × one-step, 1 × two-step, 4 × standard three-step and 4 × inclusive three-step) are detailed in Table 2.1. For each model type and simulation scenario, 50 sets of random starting values were used and the best fitting model was selected in order to avoid obtaining local maximum solutions. A tolerance of 10^{-9} was used for convergence and a maximum of 1000 iterations were permitted. For each scenario there were 2000 replications. Class labelling was evaluated using the method described in Section 2.2.5. Parameter estimates were to be evaluated in terms of bias, percentage bias, 95% confidence interval coverage and 95% confidence interval length (Burton et al., 2006, Table I).

2.4 Results

Simulation results for the estimated hazard ratios and corresponding performance measures are presented by true latent class effect and can be found in Tables 2.2 to 2.5. For simplicity, results aggregated over the small, medium and large effect sizes are presented in Table 2.1 (where the high and medium entropy scenarios have also been aggregated).

As an illustrative example, histograms of parameter estimates from Scenario 23 (low entropy, large effect, $N = 500$) are presented in Figure 2.4.

2.4.1 One and two-step approaches

Latent class effect estimates for the one-step model exhibited no or low bias and approximately nominal coverage in most scenarios, although in the low entropy and low sample size scenarios confidence interval coverage was slightly below the nominal level at 93% on aggregate (Table 2.1).

In the medium and high entropy scenarios two-step estimates were unbiased with nominal coverage. In the low entropy scenarios the two-step models exhibited some parameter attenuation, although this was less pronounced for the larger sample size (18% and 8% for small and large sample sizes respectively, on aggregate, Table 2.1). Confidence interval coverage was typically similar to the one-step approach.

2.4.2 Standard three-step approaches

Estimates from the standard three-step models were approximately unbiased in the no effect scenarios (Table 2.2). In these scenarios *MA* and *PA* exhibited nominal coverage, but coverage was generally too high for *mPC* and *PrA* at approximately 97-99%. For the small, medium and large effect scenarios (Tables 2.3 to 2.5), *MA*, *mPC* and *PrA* estimates exhibited considerable bias towards the null and poor coverage. Even in the high entropy and larger sample size scenarios these methods exhibited attenuation in the latent class effect of >19% and this became considerably worse in the low entropy scenarios.

PA estimates were approximately unbiased with nominal coverage in all of the high and medium entropy scenarios, irrespective of effect size. With $N = 500$ in the low entropy scenarios for the small, medium and large effect sizes (Tables 2.3 to 2.5) *PA* exhibited considerable attenuation (20% on aggregate, Table 2.1) and poor coverage, but the bias was far less than the other standard three-step procedures which exhibited >57% bias on aggregate. With $N = 1000$ in the low entropy scenarios, attenuation was improved (10% on aggregate) but coverage was below the nominal level (90% on aggregate).

2.4.3 Inclusive three-step approaches

Latent class effect estimates from the inclusive three-step approaches were further from the null than their counterpart standard three-step approaches, suggesting some reversal of attenuation as intended. This effect is illustrated in Figure 2.4. *Incl-PrA* and *Incl-mPC* produced no or low bias in all scenarios, and *Incl-MA* exhibited improved

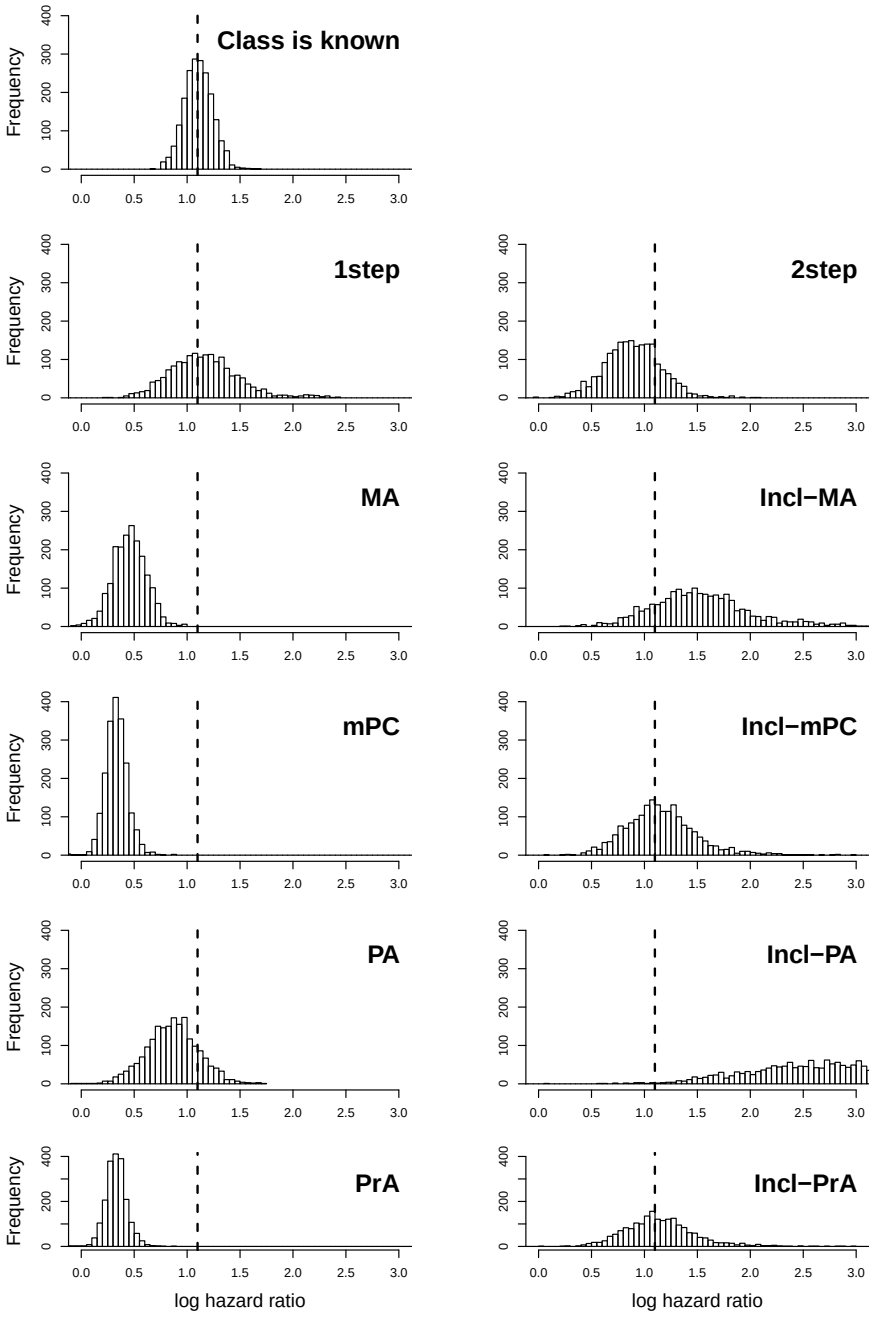

Figure 2.4: Histograms of simulation results taken from low entropy Scenario 23 ($N = 500$, $\pi_{(1)} = 0.4$, $\pi_{(2)} = 0.6$). The dashed vertical lines represent the true latent class effect, in this case $\log(3) \approx 1.10$, and deviations of the empirical distributions from the true value indicate bias. For corresponding confidence interval coverage and length see Table 2.5. 'Class is known' refers to results from a Cox regression model including the known underlying class and is included for demonstration purposes only. MA modal assignment, mPC multiple pseudo class draws, PA partial assignment, PrA proportional assignment, Incl inclusive.

						Absolute % Bias				% CI Coverage			
						N=500		N=1000		N=500		N=1000	
Model No.	Name	Steps	LC predictors	Ass. method	Base haz.	High/Med.	Low	High/Med.	Low	High/Med.	Low	High/Med.	Low
1	1step	1	None	-	Piecewise constant	2	6	1	3	95	93	95	94
2	2step	2	None	-	Piecewise constant	1	18	1	8	95	93	95	94
3	MA	3	None	MA	Unspecified	28	57	28	55	62	30	44	13
4	mPC	3	None	mPC	Unspecified	38	69	38	70	50	22	30	2
5	PA	3	None	PA	Unspecified	3	20	2	10	95	87	94	90
6	PrA	3	None	PrA	Unspecified	38	69	38	70	40	10	23	0
7	Incl-MA	3	$H_0(t)$,δ,z	MA	Unspecified	13	43	14	45	77	41	73	32
8	Incl-mPC	3	$H_0(t)$,δ,z	mPC	Unspecified	1	8	0	3	89	67	88	68
9	Incl-PA	3	$H_0(t)$,δ,z	PA	Unspecified	56	196	56	186	43	16	28	6
10	Incl-PrA	3	$H_0(t)$,δ,z	PrA	Unspecified	1	7	0	2	82	52	82	54

Table 2.1: Details of models used in the simulation study and aggregated results for the estimated latent class effect. MA modal assignment, mPC multiple pseudo class draws, PA partial assignment, PrA proportional assignment, Incl inclusive. Absolute % bias and 95% confidence interval (CI) coverage are aggregate results over the small, medium and large effect sizes. High and medium entropy results have been aggregated. One-step models (1) exhibit little or no bias and approximately nominal coverage, two-step models (2) exhibit no bias and nominal coverage in High/Med entropy scenarios but are biased in Low entropy scenarios, standard three-step models (3-6) generally exhibit bias and poor coverage (excepting PA in High/Med entropy scenarios), inclusive three-step models (7-10) offer improved bias over standard three-step models (excepting Incl-PA) but coverage is generally poor.

bias compared with the counterpart standard three-step approaches. *Incl-PA* estimates, however, considerably exceeded the true values for small, medium and large effects for all entropy levels (Tables 2.3 to 2.5).

Coverage was below the nominal value for estimates from all of the inclusive three-step approaches in all scenarios. In the no effect scenarios, the low coverage of the inclusive estimates generally resulted in more than double the nominal Type I error rates, and this became far worse as the entropy decreased (Table 2.2).

Scenario	HR	log(HR)	Entropy	$\pi_{(1)}$	$\pi_{(2)}$	N	Measure	1step	2step	MA	mPC	PA	PrA	Incl-MA	Incl-mPC	Incl-PA	Incl-PrA
1	1	0	High	0.30	0.70	500	Estimate	-0.00	-0.00	-0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
							Bias	-0.00	-0.00	-0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
							Bias (%)	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC
							SE	0.15	0.15	0.13	0.13	0.15	0.12	0.13	0.13	0.15	0.12
							CI Coverage (%)	95.0	95.2	95.0	98.5	95.0	96.8	85.2	91.0	84.4	87.8
							CI Length	0.59	0.59	0.51	0.52	0.59	0.47	0.51	0.52	0.59	0.47
2	1	0	High	0.30	0.70	1000	Estimate	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
							Bias	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
							Bias (%)	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC
							SE	0.11	0.11	0.09	0.09	0.11	0.08	0.09	0.09	0.11	0.08
							CI Coverage (%)	94.9	94.9	94.7	98.0	95.2	96.5	81.5	91.2	85.2	87.7
							CI Length	0.42	0.42	0.36	0.37	0.42	0.33	0.36	0.37	0.42	0.33
3	1	0	Medium	0.35	0.65	500	Estimate	-0.00	-0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
							Bias	-0.00	-0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
							Bias (%)	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC
							SE	0.18	0.18	0.13	0.14	0.17	0.11	0.13	0.14	0.17	0.11
							CI Coverage (%)	94.0	94.4	94.1	99.2	94.3	97.4	75.0	85.5	69.2	77.4
							CI Length	0.70	0.70	0.51	0.54	0.68	0.45	0.51	0.54	0.68	0.45
4	1	0	Medium	0.35	0.65	1000	Estimate	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
							Bias	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
							Bias (%)	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC
							SE	0.12	0.13	0.09	0.10	0.12	0.08	0.09	0.10	0.12	0.08
							CI Coverage (%)	95.0	95.4	95.0	99.3	95.2	97.8	75.3	85.9	71.4	79.0
							CI Length	0.49	0.49	0.36	0.37	0.48	0.31	0.36	0.37	0.48	0.31
5	1	0	Low	0.40	0.60	500	Estimate	-0.01	-0.01	-0.00	-0.00	-0.01	-0.00	-0.01	0.00	-0.22	0.00
							Bias	-0.01	-0.01	-0.00	-0.00	-0.01	-0.00	-0.01	0.00	-0.22	0.00
							Bias (%)	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC
							SE	0.25	0.25	0.14	0.15	0.22	0.11	0.14	0.15	0.22	0.11
							CI Coverage (%)	91.8	96.6	94.9	99.9	93.9	98.4	52.6	68.7	45.9	55.2
							CI Length	0.95	0.96	0.54	0.58	0.87	0.43	0.56	0.59	0.88	0.44
6	1	0	Low	0.40	0.60	1000	Estimate	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.01	-0.00
							Bias	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.01	-0.00
							Bias (%)	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC
							SE	0.17	0.17	0.09	0.10	0.16	0.08	0.10	0.10	0.16	0.08
							CI Coverage (%)	93.1	96.0	94.9	99.9	94.4	99.2	53.1	71.5	44.9	58.7
							CI Length	0.69	0.69	0.37	0.39	0.63	0.29	0.37	0.39	0.63	0.30

Table 2.2: Simulation results for the effect of latent class in scenarios with a hazard ratio of 1. Estimates are presented on the log scale. NC not calculable, MA modal assignment, mPC multiple pseudo class draws, PA partial assignment, PrA proportional assignment, Incl inclusive. All models are unbiased. One-step, two-step, MA and PA models exhibit approximately nominal coverage. mPC and PrA coverage is too high and for all inclusive methods coverage is too low, implying increased Type I error rates.

Scenario	HR	log(HR)	Entropy	$\pi_{(1)}$	$\pi_{(2)}$	N	Measure	1step	2step	MA	mPC	PA	PrA	Incl-MA	Incl-mPC	Incl-PA	Incl-PrA
7	1.5	0.41	High	0.30	0.70	500	Estimate	0.41	0.41	0.33	0.30	0.40	0.30	0.45	0.41	0.55	0.41
							Bias	0.00	-0.00	-0.08	-0.11	-0.00	-0.11	0.04	0.01	0.15	0.00
							Bias (%)	0.98	-0.09	-19.33	-26.61	-0.53	-26.71	10.73	1.29	36.31	1.07
							SE	0.15	0.15	0.13	0.13	0.15	0.12	0.13	0.13	0.15	0.12
							CI Coverage (%)	94.8	95.2	90.1	91.2	95.1	86.7	85.4	91.6	73.6	87.5
							CI Length	0.59	0.59	0.50	0.52	0.58	0.46	0.50	0.52	0.58	0.47
8	1.5	0.41	High	0.30	0.70	1000	Estimate	0.41	0.41	0.33	0.30	0.40	0.30	0.47	0.41	0.55	0.41
							Bias	0.00	-0.00	-0.08	-0.11	-0.00	-0.11	0.07	0.00	0.15	0.00
							Bias (%)	0.44	-0.08	-19.13	-26.70	-0.18	-26.75	17.05	0.53	36.19	0.54
							SE	0.11	0.11	0.09	0.09	0.10	0.08	0.09	0.09	0.11	0.08
							CI Coverage (%)	94.7	94.9	86.7	82.3	94.5	76.0	80.2	91.5	66.8	88.5
							CI Length	0.41	0.41	0.35	0.36	0.41	0.33	0.36	0.36	0.41	0.33
9	1.5	0.41	Medium	0.35	0.65	500	Estimate	0.41	0.40	0.27	0.22	0.40	0.22	0.48	0.42	0.74	0.41
							Bias	0.01	-0.01	-0.14	-0.19	-0.01	-0.19	0.07	0.01	0.33	0.01
							Bias (%)	1.98	-1.62	-33.69	-45.84	-2.48	-46.11	18.25	2.49	82.42	2.17
							SE	0.18	0.18	0.13	0.14	0.17	0.11	0.13	0.14	0.17	0.11
							CI Coverage (%)	94.5	95.0	80.7	78.5	94.7	62.4	72.3	84.5	49.0	75.9
							CI Length	0.69	0.69	0.50	0.53	0.67	0.44	0.51	0.54	0.68	0.44
10	1.5	0.41	Medium	0.35	0.65	1000	Estimate	0.41	0.40	0.27	0.22	0.40	0.22	0.49	0.41	0.74	0.41
							Bias	0.00	-0.01	-0.14	-0.19	-0.01	-0.19	0.08	0.00	0.34	0.00
							Bias (%)	0.53	-1.33	-33.80	-46.49	-1.50	-46.48	19.97	1.07	83.07	0.71
							SE	0.12	0.12	0.09	0.09	0.12	0.08	0.09	0.09	0.12	0.08
							CI Coverage (%)	94.6	95.2	66.8	46.8	94.8	30.6	70.5	85.7	34.0	77.3
							CI Length	0.49	0.49	0.35	0.37	0.48	0.31	0.36	0.37	0.48	0.31
11	1.5	0.41	Low	0.40	0.60	500	Estimate	0.43	0.33	0.17	0.13	0.33	0.13	0.54	0.41	0.97	0.40
							Bias	0.02	-0.08	-0.23	-0.28	-0.08	-0.28	0.14	0.00	0.56	-0.00
							Bias (%)	4.94	-18.91	-56.94	-68.80	-19.29	-68.98	33.63	0.69	138.42	-1.18
							SE	0.24	0.24	0.14	0.15	0.22	0.11	0.14	0.15	0.22	0.11
							CI Coverage (%)	92.8	94.8	63.9	57.2	91.8	26.1	47.3	67.0	30.1	51.1
							CI Length	0.96	0.96	0.54	0.57	0.86	0.42	0.55	0.58	0.86	0.43
12	1.5	0.41	Low	0.40	0.60	1000	Estimate	0.42	0.37	0.18	0.12	0.37	0.12	0.60	0.42	1.23	0.42
							Bias	0.01	-0.03	-0.22	-0.28	-0.03	-0.28	0.20	0.02	0.83	0.02
							Bias (%)	3.04	-8.23	-54.49	-69.28	-8.34	-69.39	48.93	4.37	203.54	3.97
							SE	0.17	0.17	0.09	0.10	0.16	0.07	0.09	0.10	0.16	0.07
							CI Coverage (%)	94.2	95.6	35.6	7.4	94.0	1.1	43.2	69.8	16.0	57.0
							CI Length	0.62	0.62	0.36	0.38	0.62	0.29	0.37	0.38	0.62	0.29

Table 2.3: Simulation results for the effect of latent class in scenarios with a hazard ratio of 1.5. Estimates are presented on the log scale. MA modal assignment, mPC multiple pseudo class draws, PA partial assignment, PrA proportional assignment, Incl inclusive. One-step models exhibit no or low bias and approximately nominal coverage. Two-step and PA are unbiased with nominal coverage in medium and high entropy scenarios but both exhibit bias when the entropy is low. Incl-mPC and Incl-PrA exhibit no or low bias but poor coverage.

Scenario	HR	log(HR)	Entropy	$\pi_{(1)}$	$\pi_{(2)}$	N	Measure	1step	2step	MA	mPC	PA	PrA	Incl-MA	Incl-mPC	Incl-PA	Incl-PrA
13	2	0.69	High	0.30	0.70	500	Estimate	0.70	0.70	0.56	0.50	0.68	0.50	0.75	0.70	0.94	0.70
							Bias	0.01	0.00	-0.14	-0.19	-0.01	-0.19	0.06	0.01	0.24	0.01
							Bias (%)	1.34	0.39	-19.87	-27.42	-1.20	-27.54	8.08	0.95	34.96	0.75
							SE	0.15	0.15	0.13	0.13	0.15	0.12	0.13	0.13	0.15	0.12
							CI Coverage (%)	95.2	95.5	80.5	72.9	95.5	63.8	85.9	92.2	60.0	87.9
							CI Length	0.59	0.59	0.50	0.51	0.58	0.46	0.50	0.52	0.58	0.46
14	2	0.69	High	0.30	0.70	1000	Estimate	0.69	0.69	0.55	0.50	0.68	0.50	0.76	0.69	0.93	0.69
							Bias	-0.00	-0.00	-0.14	-0.19	-0.01	-0.20	0.06	-0.00	0.24	-0.00
							Bias (%)	-0.07	-0.57	-20.65	-28.06	-1.63	-28.21	9.11	-0.40	34.13	-0.45
							SE	0.11	0.11	0.09	0.09	0.10	0.08	0.09	0.09	0.11	0.08
							CI Coverage (%)	94.8	94.8	63.4	42.8	94.9	32.2	84.0	90.8	41.9	86.4
							CI Length	0.42	0.42	0.35	0.36	0.41	0.32	0.36	0.36	0.41	0.33
15	2	0.69	Medium	0.35	0.65	500	Estimate	0.71	0.68	0.45	0.37	0.66	0.37	0.81	0.70	1.24	0.70
							Bias	0.01	-0.01	-0.24	-0.33	-0.03	-0.33	0.12	0.01	0.54	0.01
							Bias (%)	1.99	-1.57	-34.97	-47.00	-4.07	-47.28	17.31	1.59	78.23	1.21
							SE	0.18	0.18	0.13	0.13	0.17	0.11	0.13	0.14	0.17	0.11
							CI Coverage (%)	95.2	95.1	52.4	25.1	95.0	10.8	69.2	85.2	25.9	76.7
							CI Length	0.70	0.70	0.50	0.53	0.67	0.43	0.51	0.53	0.67	0.43
16	2	0.69	Medium	0.35	0.65	1000	Estimate	0.69	0.68	0.45	0.36	0.67	0.36	0.81	0.69	1.24	0.69
							Bias	0.00	-0.01	-0.24	-0.33	-0.02	-0.33	0.11	-0.00	0.54	-0.00
							Bias (%)	0.24	-1.53	-35.17	-47.68	-3.26	-47.76	16.48	-0.07	78.47	-0.27
							SE	0.13	0.13	0.09	0.09	0.12	0.08	0.09	0.09	0.12	0.08
							CI Coverage (%)	94.7	94.9	23.1	1.5	94.7	0.2	64.3	84.2	9.3	76.1
							CI Length	0.49	0.49	0.35	0.36	0.47	0.30	0.36	0.37	0.48	0.30
17	2	0.69	Low	0.40	0.60	500	Estimate	0.73	0.57	0.30	0.22	0.57	0.22	1.03	0.77	2.13	0.76
							Bias	0.04	-0.12	-0.39	-0.47	-0.13	-0.47	0.34	0.08	1.44	0.06
							Bias (%)	5.83	-17.37	-56.69	-67.66	-18.12	-67.85	48.95	11.23	207.82	9.08
							SE	0.25	0.25	0.14	0.14	0.22	0.11	0.14	0.15	0.22	0.11
							CI Coverage (%)	93.3	94.1	23.4	8.8	90.4	2.4	41.4	66.7	14.1	51.8
							CI Length	0.91	0.91	0.53	0.56	0.85	0.42	0.55	0.57	0.85	0.41
18	2	0.69	Low	0.40	0.60	1000	Estimate	0.71	0.64	0.31	0.21	0.63	0.21	1.00	0.71	1.99	0.71
							Bias	0.02	-0.06	-0.38	-0.48	-0.06	-0.48	0.31	0.02	1.30	0.02
							Bias (%)	2.34	-8.10	-55.07	-69.58	-9.29	-69.78	44.27	2.73	187.78	2.27
							SE	0.17	0.17	0.09	0.10	0.16	0.07	0.09	0.10	0.16	0.07
							CI Coverage (%)	94.4	94.7	2.3	0.1	92.4	0.1	32.8	68.4	2.0	54.1
							CI Length	0.77	0.78	0.36	0.38	0.61	0.28	0.37	0.38	0.61	0.28

Table 2.4: Simulation results for the effect of latent class in scenarios with a hazard ratio of 2. Estimates are presented on the log scale. MA modal assignment, mPC multiple pseudo class draws, PA partial assignment, PrA proportional assignment, Incl inclusive. One-step models exhibit no or low bias and approximately nominal coverage. Two-step and PA are unbiased with nominal coverage in medium and high entropy scenarios but both exhibit bias when the entropy is low. Incl-mPC and Incl-PrA exhibit no or low bias but poor coverage.

Scenario	HR	log(HR)	Entropy	$\pi_{(1)}$	$\pi_{(2)}$	N	Measure	1step	2step	MA	mPC	PA	PrA	Incl-MA	Incl-mPC	Incl-PA	Incl-PrA
19	3	1.1	High	0.30	0.70	500	Estimate	1.12	1.11	0.86	0.77	1.07	0.77	1.17	1.11	1.46	1.10
							Bias	0.02	0.01	-0.24	-0.33	-0.03	-0.33	0.07	0.01	0.36	0.00
							Bias (%)	1.97	0.68	-21.75	-29.59	-2.65	-29.75	6.39	0.62	33.15	0.31
							SE	0.16	0.16	0.13	0.14	0.15	0.12	0.14	0.14	0.16	0.12
							CI Coverage (%)	95.0	95.5	53.6	30.0	94.9	19.0	86.2	92.0	39.9	87.1
							CI Length	0.63	0.63	0.51	0.53	0.60	0.46	0.53	0.55	0.61	0.48
20	3	1.1	High	0.30	0.70	1000	Estimate	1.11	1.11	0.86	0.77	1.07	0.77	1.17	1.10	1.46	1.10
							Bias	0.01	0.01	-0.24	-0.33	-0.03	-0.33	0.07	-0.00	0.36	-0.00
							Bias (%)	1.28	0.68	-21.96	-29.71	-2.47	-29.87	6.37	-0.05	32.98	-0.16
							SE	0.11	0.11	0.09	0.09	0.11	0.08	0.10	0.10	0.11	0.09
							CI Coverage (%)	94.8	95.1	26.3	3.6	94.5	1.6	84.4	91.1	15.7	87.3
							CI Length	0.44	0.44	0.36	0.37	0.42	0.32	0.37	0.38	0.43	0.34
21	3	1.1	Medium	0.35	0.65	500	Estimate	1.13	1.09	0.69	0.56	1.03	0.56	1.28	1.11	1.89	1.10
							Bias	0.03	-0.01	-0.41	-0.54	-0.07	-0.54	0.18	0.01	0.80	0.01
							Bias (%)	3.05	-1.17	-37.30	-49.18	-6.36	-49.43	16.14	1.03	72.39	0.48
							SE	0.19	0.19	0.13	0.14	0.17	0.11	0.14	0.14	0.18	0.11
							CI Coverage (%)	94.4	95.4	12.0	0.4	92.8	0.0	64.3	85.8	8.1	75.8
							CI Length	0.76	0.76	0.51	0.54	0.69	0.43	0.54	0.56	0.70	0.44
22	3	1.1	Medium	0.35	0.65	1000	Estimate	1.12	1.10	0.69	0.55	1.04	0.55	1.27	1.10	1.90	1.10
							Bias	0.02	-0.00	-0.41	-0.54	-0.06	-0.55	0.17	-0.00	0.80	-0.00
							Bias (%)	1.91	-0.15	-37.04	-49.51	-5.09	-49.64	15.32	-0.08	73.20	-0.32
							SE	0.14	0.14	0.09	0.09	0.12	0.08	0.10	0.10	0.13	0.08
							CI Coverage (%)	94.5	95.6	0.5	0.0	93.0	0.0	55.9	85.4	0.6	75.9
							CI Length	0.53	0.54	0.36	0.37	0.49	0.30	0.38	0.39	0.50	0.31
23	3	1.1	Low	0.40	0.60	500	Estimate	1.16	0.89	0.46	0.33	0.86	0.33	1.62	1.24	3.75	1.21
							Bias	0.06	-0.21	-0.64	-0.76	-0.24	-0.77	0.52	0.14	2.65	0.11
							Bias (%)	5.90	-18.68	-58.55	-69.63	-21.80	-69.90	47.51	12.45	241.27	9.86
							SE	0.28	0.27	0.14	0.15	0.22	0.11	0.15	0.15	0.23	0.11
							CI Coverage (%)	94.2	88.9	3.3	0.9	79.5	0.2	35.3	66.9	4.0	51.6
							CI Length	1.31	1.32	0.54	0.57	0.87	0.42	0.58	0.59	0.89	0.41
24	3	1.1	Low	0.40	0.60	1000	Estimate	1.15	1.01	0.47	0.32	0.96	0.32	1.55	1.11	2.93	1.11
							Bias	0.05	-0.08	-0.63	-0.78	-0.14	-0.78	0.45	0.02	1.83	0.01
							Bias (%)	4.87	-7.72	-56.91	-70.97	-12.47	-71.11	40.78	1.46	166.26	1.02
							SE	0.20	0.19	0.09	0.10	0.16	0.07	0.10	0.10	0.16	0.07
							CI Coverage (%)	94.4	92.8	0.0	0.0	84.0	0.0	21.3	66.7	0.0	51.0
							CI Length	0.70	0.66	0.36	0.38	0.62	0.28	0.40	0.40	0.63	0.28

Table 2.5: Simulation results for the effect of latent class in scenarios with a hazard ratio of 3. Estimates are presented on the log scale. MA modal assignment, mPC multiple pseudo class draws, PA partial assignment, PrA proportional assignment, Incl inclusive. One-step models exhibit no or low bias and approximately nominal coverage. Two-step and PA are unbiased with nominal coverage in medium and high entropy scenarios but both exhibit bias when the entropy is low. Incl-mPC and Incl-PrA exhibit no or low bias but poor coverage.

2.5 Discussion

In this chapter, one, two and three-step approaches to latent class modelling with a time-to-event distal outcome were presented and the empirical properties of latent class effect estimates were compared using Monte Carlo simulation. To our knowledge, this is the first study to investigate various approaches to latent class modelling when the distal outcome is a time-to-event variable. Moreover, this is the first study to demonstrate and implement two-step (Bakk and Kuha, 2018) and inclusive bias-correction approaches (Bray et al., 2015) with a time-to-event distal outcome. This study contributes to the emerging body of literature on latent class modelling with a distal outcome variable.

Latent class effect estimates for the one-step model exhibited no or low bias and approximately nominal coverage in most scenarios, although confidence interval coverage was slightly below the nominal value in the low entropy and low sample size scenarios. The lack of bias is consistent with studies with a continuous distal outcome (Clark and Muthén, 2009; Asparouhov and Muthén, 2014; Bakk and Kuha, 2018). Interestingly, standard errors (which determine confidence interval coverage when an estimate is unbiased) in one-step models with a continuous distal outcome have been shown to be both overestimated and underestimated previously when both the entropy and sample size are low (Bakk et al., 2013; Bakk and Kuha, 2018).

The two-step approach resulted in low bias and approximately nominal coverage in the medium and high entropy scenarios. However this approach did exhibit some bias towards the null in the low entropy scenarios, which is consistent with previous research using this approach with continuous distal outcome variables (Bakk and Kuha, 2018). Confidence interval coverage was typically similar to the one-step approach.

Generally, standard three-step approaches resulted in attenuated estimates of the latent class effect with underestimated standard errors, resulting in poor confidence interval coverage. This result is consistent with the research literature in this area (Clark and Muthén, 2009; Bakk et al., 2013; Asparouhov and Muthén, 2014). A surprising result however was that a standard three-step approach using partial assignment produced unbiased estimates and nominal coverage in medium and high entropy scenarios. In the low entropy scenarios, partial assignment exhibited similar levels of bias to the two-step approach, however confidence interval coverage was typically poorer with partial assignment.

The inclusive approach to bias-correction proposed by Bray et al. (2015) was adapted here to include a time-to-event variable as a latent class predictor. As intended, the inclusive approach produced estimates further from the null than their non-inclusive counterpart models and in general improved bias. Proportional and multiple pseudo-class assignment approaches benefited considerably from the inclusive approach with no or low bias in all scenarios. Results for partial assignment (which performed well as a standard approach), however, were worse when combined with the

inclusive approach. Despite the improvements in bias, confidence interval coverage of inclusive approaches was too low, notably producing increased Type I errors compared with standard three-step approaches when simulating under the null hypothesis. An alternative approach to obtaining standard errors with inclusive approaches, such as bootstrapping, may help resolve these issues, as has been suggested previously for a closely related approach (Bakk and Vermunt, 2016).

Despite the superior performance of one-step approaches demonstrated in this study, one-step approaches have a few disadvantages as explicated by Vermunt (2010). The main criticism is that the distal outcome variable can influence latent class composition, possibly affecting the characteristics or even the number of latent classes (Vermunt, 2010; Bakk et al., 2016). Asparouhov and Muthén (2014) give an example of a one-step approach 'failing' where class composition is determined solely by the distal outcome variable, which in that case was simulated from a two component normal mixture distribution. The extent to which a time-to-event submodel could influence latent class composition is not clear and this is a relevant topic for further research. Inclusive three-step methods may also be subject to the same limitation. When fitting one-step latent class models we support the recommendations of Larsen (2004) and Asparouhov and Muthén (2014) in fitting latent class models without the distal outcome variable in model building and/or sensitivity analyses.

In this study, the performance of the two-step (Bakk and Kuha, 2018) and inclusive bias-correction approaches (Bray et al., 2015) were assessed. A number of other correction methods have been proposed (Bolck et al., 2004; Vermunt, 2010; Petersen et al., 2012; Bakk et al., 2013; Lanza et al., 2013), although not all are suitable for modelling with a time-to-event distal outcome. Investigation of the bias-corrected three-step methods described in Vermunt (2010) and Bakk et al. (2013) with a time-to-event distal outcome would be a valuable addition to the research described here. However, these methods are based upon introducing and subsequently correcting classification error and a key advantage of the two-step approach studied here is that this step is avoided (Bakk and Kuha, 2018). Moreover, estimates from the two-step approach were previously found to have better statistical properties than corrected three-step approaches with a continuous distal outcome (Bakk and Kuha, 2018). An interesting additional feature of the two-step approach is that different observations can be used for the latent classification and distal outcome models (Xue and Bandeen-Roche, 2004; Bakk and Kuha, 2018).

In this study, the various models used different hazard functions for the time-to-event outcome variable and, as identified by a reviewer from the Journal of Structural Equation Modeling, this feature warrants special attention. To model the distal outcome in both the standard and inclusive three-step approaches a Cox model was used where the baseline hazard function is not estimated. This is the most common model

used in practice and does not disadvantage the standard or inclusive three-step models in any way, as demonstrated in a supportive analysis in Appendix A. For the inclusive three-step approaches the hazard function used for the distal outcome model should not to be confused with that in latent class prediction in Step 1, see Figure 2.1(c). In this approach, a non-parametric estimate of the unconditional cumulative hazard is used as a latent class predictor, as recommended for multiple imputation, and notably in preference to the observed event time or its logarithm (White and Royston, 2009).

For the one and two-step models, piecewise exponential baseline hazard models were used, see Figure A.1. Setting the partitions of the time-grid for a piecewise exponential model to the observed event times is equivalent to using a non-parametric baseline hazard, as in Larsen (2004). A non-parametric baseline hazard model is, in turn, equivalent to a Cox model (Breslow, 1974). Whilst the piecewise exponential model can offer improved parameter efficiency (it can also result in bias if the time grid is poorly specified), see Han et al. (2014), the main purpose here was to simplify the calculation of standard errors by reducing the number of parameters required to estimate the baseline hazard function. For the one and two-step models we used Louis's method (Louis, 1982) to obtain standard errors, which requires the inversion of the negative Hessian matrix and is not feasible when the number of parameters is large (Larsen, 2004). Bootstrapping has been recommended (Hsieh et al., 2006) but fitting one and two-step models to bootstrap resamples from each simulated data set was found to be overly computationally burdensome.

In this study, time-to-event data were simulated and analysed using a proportional hazards model. In practice the suitability of the proportional hazards assumption should be investigated. Standard residual analyses for time-to-event data (see e.g Collett, 2015) can be used with one-step latent class models by calculating class-specific fitted values and averaging over classes (Proust-Lima et al., 2014). Other possible options are to include a time-dependent latent class effect in the one-step model (Muthén et al., 2009), to estimate separate hazard functions for latent classes (Asparouhov et al., 2006), or by investigating the tenability of the proportional hazards assumption using a pseudo class draw approach to the log-log cumulative hazard plot (Larsen, 2004).

In conclusion, the empirical properties of various latent class effect estimates on a time-to-event distal outcome were compared. One-step models performed very well in general, whilst two-step approaches performed well when classes were well separated. A surprising result was that a standard three-step approach using partial assignment also performed well when classes were well-separated. Although inclusive bias-correction approaches were generally shown to decrease attenuation of the latent class effect estimate, partial assignment was overall the best performing three-step approach. However, when the entropy was low this approach was found to be inferior to one-step approaches and confidence interval coverage was generally worse than the two-step approach.

For the applied researcher, a one-step approach is recommended where possible, although excluding the distal outcome variable in model building and/or accompanying sensitivity analyses is recommended. The suitability of assuming proportional hazards should also be assessed.

Chapter 3

A general joint latent class and time-to-event model

3.1 Introduction

In the previous chapter, a joint model for the simultaneous analysis of latent classes and a time-to-event outcome variable was detailed. When all parameters are estimated simultaneously, the model is referred to as a one-step model, and two and three-step alternatives were also described. A joint model with Bernoulli manifest variables was introduced by Larsen (2004), and an extensive framework for latent class models with a time-to-event outcome was set out by Asparouhov et al. (2006). It is possible to fit joint latent class and time-to-event models using `Mplus` software (Muthén and Muthén, 2011), and an application using `Mplus` software can be found in Muthén et al. (2009). Bakk and Kuha (2018) describe how it is possible to effectively trick `Mplus` in order to fit a two-step model (although appropriate standard errors need to be calculated manually).

Whilst the described models can be fitted using commercial software, to the best of our knowledge no suitable packages or functions for the freely available open source software `R` (R Core Team, 2017) are available. The `poLCA` package (Linzer and Lewis, 2011) can be used to fit LCR models (see Section 2.2.1) with multinomial manifest variables. Other, more basic, latent class `R` packages and functions are described by Linzer and Lewis (2011). The `lcmm()` function in the extensive `lcmm` package (Proust-Lima et al., 2017) can be used to fit latent class linear mixed models with a mixture of continuous and categorical longitudinal outcomes. Also in the `lcmm` package, the `Jointlcmm()` function can be used to fit a joint latent class and time-to-event outcome, with a single continuous longitudinal outcome. As in `Jointlcmm()`, the term 'joint' is

usually used to describe models for the simultaneous analysis of a time-to-event outcome and one or more variables measured longitudinally (see Papageorgiou et al., 2019, for an up-to-date review). Typically, in these longitudinal models, a continuous latent variable is used as a device for joining a linear mixed model and a time-to-event model. Note that continuous latent variables and manifest variables measured longitudinally are not considered in this thesis.

The purpose of this chapter is to: 1) detail a more general joint latent class and time-to-event model than described in the previous chapter, 2) demonstrate how the general model can be fitted using the EM algorithm and how standard errors are obtained, and 3) describe and apply the author-written `LCSM()` function for fitting one and two-step joint latent class and time-to-event models in `R`.

The chapter is structured as follows: In Section 3.2, it is shown how continuous manifest variables can be incorporated into the joint model. In Section 3.3, the time-to-event submodel described in the previous chapter is extended to include class-specific covariate effects (i.e. class-by-covariate interactions) and it is shown how a Weibull submodel can be incorporated. In Section 3.4, the general (conditionally independent) joint model is introduced and it is then shown how various dependencies can be added in Section 3.5. In Section 3.6, a model fitting procedure using the EM algorithm is described, and in Section 3.7 it is shown how standard errors can be obtained. In Section 3.8, the author-written `LCSM()` `R` function is outlined and in Section 3.9 an analysis of the prostate cancer data set is presented. Discussion is given in Section 3.10.

3.2 Mixed manifest variables

In the previous chapter, only categorical (multinomial) manifest variables were considered. Normal, Poisson, binomial, gamma and ordinal categorical are some other possibilities for the conditional distribution of the manifest variables (see Moustaki, 1996; Bartholomew et al., 2011). Only normal and multinomial conditional distributions for the manifest variables are discussed in this thesis.

As defined previously, for a given subject, $\mathbf{Y} = (Y_1, \ldots, Y_M)^\intercal$ denotes a vector of categorical manifest variables, with observed values $\mathbf{y} = (y_1, \ldots, y_M)^\intercal$. Now, let $\mathbf{W} = (W_1, \ldots, W_L)^\intercal$ denote a vector of continuous manifest variables, for a given subject, with observed values $\mathbf{w} = (w_1, \ldots, w_L)^\intercal$. The full vector of categorical and continuous manifest variables is then $(\mathbf{Y}^\intercal, \mathbf{W}^\intercal)^\intercal$, and $(\mathbf{y}^\intercal, \mathbf{w}^\intercal)^\intercal$ is the $(M + L) \times 1$ vector of observed manifest values, for an individual. It is assumed that the conditional distribution for each of the continuous manifest variables is normal,

$$f_{W_l|C}(w_l|j) \sim N(\mu_{lj}, \sigma_l^2), \tag{3.1}$$

$l = 1, \ldots, L$ and $j = 1, \ldots, J$ so that although each conditional mean, μ_{lj}, is class-specific, the variance, σ_l, is assumed to be common across classes for each continuous variable.

3.3 Extending the time-to-event submodel

In the previous chapter, a proportional hazards submodel was presented with a simple latent class effect, and the baseline hazard was assumed to follow a piecewise exponential distribution. One possibility to avoid the assumption of proportional hazards for the latent class effect is to stratify by latent class, so that the hazard function is

$$\alpha(t|\mathbf{Z} = \mathbf{z}, C = j) = \alpha_{0j}(t) \exp(\mathbf{z}^\intercal \boldsymbol{\beta}).$$

However, covariate effects are still assumed to be proportional across strata, and when stratifying by class there is no estimated parameter to test whether the time-to-event prospects differ for the latent classes.

Another possibility to extend the time-to-event submodel is to introduce class-by-covariate interactions. One way of expressing this type of submodel for an individual is

$$\alpha(t|\mathbf{Z} = \mathbf{z}, C = j) = \alpha_0(t) \exp(\mathbf{z}^\intercal \boldsymbol{\beta}_j + \gamma_j),$$

where $\boldsymbol{\beta}_j$ is a vector of class-specific log hazard ratios, and $\gamma_J = 0$ for identifiability. It should not usually be necessary to fit class-specific effects for all covariates in $\mathbf{z}$. β_{pj} denotes the log hazard ratio for the effect of covariate p in class j on the baseline hazard, so that $\beta_{pj} - \beta_{pk}$ is the corresponding class-by-covariate interaction effect.

In the previous chapter, a piecewise exponential distribution was assumed for the baseline hazard function. A fully parametric alternative is the Weibull model, with hazard function

$$\alpha(t|\mathbf{Z} = \mathbf{z}, C = j) = \lambda \phi t^{\phi - 1} \exp(\mathbf{z}^\intercal \boldsymbol{\beta}_j + \gamma_j), \tag{3.2}$$

where λ ($\lambda > 0$) and ϕ ($\phi > 0$) are scale and shape parameters, respectively. Assuming a Weibull distribution for the event times results in a smooth monotonic hazard function (increasing if $\phi > 1$, decreasing if $\phi < 0$ and constant if $\phi = 1$). Whilst the Weibull model is less flexible than the piecewise exponential model, it is has three advantages: 1) the resulting hazard function is smooth, 2) only two parameters are required to estimate the baseline hazard function, and 3) since the Weibull model can be expressed in both PH and AFT form, the acceleration factor for the pth predictor for class j can be calculated as $\alpha_{pj} = \exp(-\beta_{pj}/\phi)$. For a Weibull submodel, the required density

function for an individual is given by

$$f_{T,\Delta|\mathbf{Z},C}(t,\delta|\mathbf{z},j) = \left[\lambda\phi t^{\phi-1}\exp(\mathbf{z}^\intercal\boldsymbol{\beta}_j+\gamma_j)\right]^{\delta}\exp\left[-\lambda t^{\phi}\exp(\mathbf{z}^\intercal\boldsymbol{\beta}_j+\gamma_j)\right].$$

3.4 The general joint model

The joint model described in Section 2.2.1 is now extended to incorporate latent class predictors, mixed manifest variables and the extended time-to-event submodel described above. The model is depicted in Figure 3.1. $\mathbf{Y}$, $\mathbf{W}$ and T are assumed to

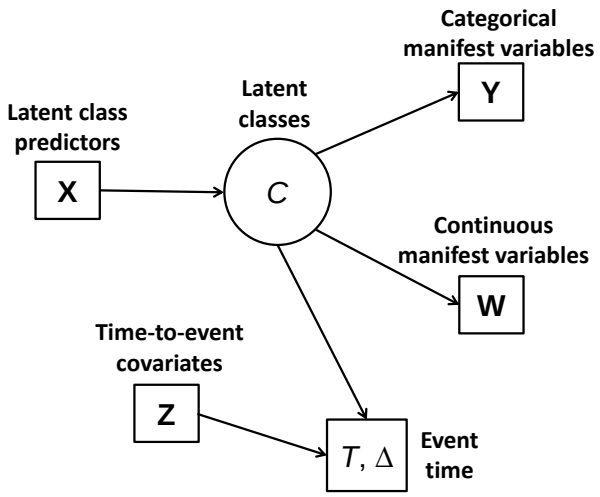

Figure 3.1: Schematic for the conditional independence joint latent class and time-to-event model.

be conditionally independent given C. Additionally, all of the manifest variables in $\mathbf{Y}$ are assumed to be conditionally independent, and all of the manifest variables in $\mathbf{W}$ are assumed to be conditionally independent, given latent class. As a result this model is referred to as the conditional independence joint model. The full joint density function for an individual is

$$f_{\mathbf{Y},\mathbf{W},T,\Delta|\mathbf{X},\mathbf{Z}}(\mathbf{y},\mathbf{w},t,\delta|\mathbf{x},\mathbf{z}) = \sum_{j=1}^{J} P(C=j|\mathbf{X}=\mathbf{x})\prod_{m=1}^{M} P(Y_m=y_m|C=j)\times \prod_{l=1}^{L} f_{W_l|C}(w_l|j)f_{T,\Delta|\mathbf{Z},C}(t,\delta|\mathbf{z},j). \tag{3.3}$$

3.5 Conditional dependency

In this section, it is shown how the conditional independence assumption can be relaxed for different parts of the model. A schematic including some possible dependencies is shown in Figure 3.2. The dependencies are: (i) between continuous manifest variables, (ii) between categorical manifest variables, (iii) between a continuous and categorical manifest variable, (iv) between latent class predictors and continuous manifest variables, (v) between latent class predictors and categorical manifest variables. It is not necessary to add a dependency between the latent class predictors and the time-to-event outcome, as variables in $\mathbf{X}$ and $\mathbf{Z}$ can be common.

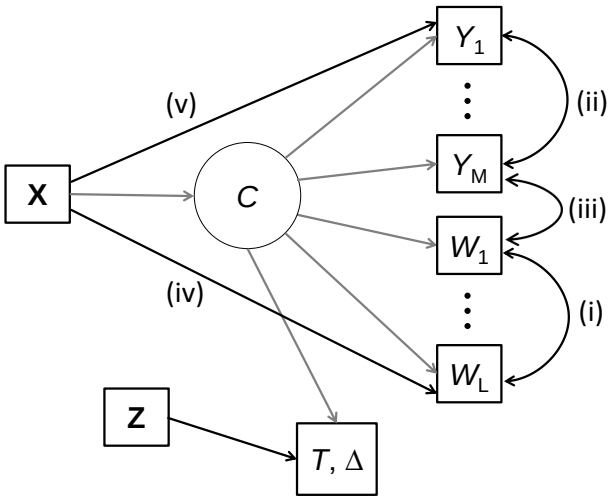

Figure 3.2: Schematic showing some conditional dependencies in the joint latent class and time-to-event model.

3.5.1 (i) Dependency between continuous manifest variables

Hunt and Jorgensen (1999) and McLachlan and Peel (2004) describe how dependencies between continuous manifest variables can be included in latent class models, by introducing covariance parameters. Equation 3.1 implies a conditional multivariate normal distribution (MVN) for the continuous manifest variables,

$$f_{\mathbf{W}|C}(\mathbf{w}|j) \sim \text{MVN}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\mu}_j$ is a vector of class-specific means of length L and $\boldsymbol{\Sigma}$ is an $L \times L$ *diagonal* variance-covariance matrix which is common to latent classes. Dependencies can be introduced by replacing the relevant off-diagonal elements (the zeros) of $\boldsymbol{\Sigma}$ with covariance parameters, $\sigma_{ll'}$. When it is assumed that the variance-covariance matrix is common across latent classes, this is known as the homoscedastic case. It is also possible to introduce class-specific variance-covariance matrices: the heteroscedastic case (see McLachlan and Peel, 2004, page 81).

3.5.2 (ii) Dependency between categorical manifest variables

One way of incorporating dependencies between two categorical manifest variables is to combine them into a single composite variable (see e.g. Hunt and Jorgensen, 1999). Although convenient and simple, Hagenaars (1988) points out that with this approach, additional interaction effects are introduced which may not be desirable, and instead proposes an approach to modelling the relationship between categorical variables by adding latent classes and introducing parameter constraints. Other approaches can be found in Desantis et al. (2012) and Asparouhov and Muthén (2015). For simplicity, the approach of Hunt and Jorgensen (1999) is considered here.

3.5.3 (iii) Dependency between continuous and categorical manifest variables

Hunt and Jorgensen (1999) and McLachlan and Peel (2004) describe a 'location' model with which dependencies are introduced between categorical and continuous variables. In this model, the distribution of the continuous variables can depend on each *category* within each categorical manifest variable. This type of dependency is considered to be overkill for most statistical models and is not considered further.

3.5.4 (iv) and (v) Dependency between latent class predictors and manifest variables

In the conditional independence model (Figure 3.1, equation 3.3) it was assumed that each of the manifest variables were independent of the latent class predictor variables, after conditioning on latent class, e.g.

$$P(Y_m = g|\mathbf{X} = \mathbf{x}, C = j) = P(Y_m = g|C = j).$$

Violations of this assumption are known as 'differential item functioning' (DIF) (see e.g Bandeen-Roche et al., 1997; Larsen, 2004). To invoke DIF, class-specific manifest

category probabilities can be allowed to depend on $\mathbf{x}$, so that for a given subject,

$$P(Y_m = g|\mathbf{X} = \mathbf{x}) = \pi_{mgj}(\mathbf{x}) = \frac{\exp(\iota_{mgj} + \mathbf{x}^\intercal \boldsymbol{\rho}_{mg})}{\sum_{g=1}^{G_m} \exp(\iota_{mgj} + \mathbf{x}^\intercal \boldsymbol{\rho}_{mg})},$$

where $\boldsymbol{\rho}_{mG_m} = \mathbf{0}$ and each $\iota_{mG_mJ} = 0$ for identifiability. Note that previously the first element in $\boldsymbol{x}$ was set to 1 but this is not required here. A similar approach could be applied for continuous variables by allowing each μ_{lj} to vary as a function of $\mathbf{x}$.

3.6 Fitting the joint model using the EM algorithm

Larsen (2004) demonstrated how to fit a joint latent class and time-to-event model with Bernoulli manifest variables and a non-parametric baseline hazard function using the EM algorithm (Dempster et al., 1977). In this section, an adapted version of the estimation routine detailed in Larsen (2004) is presented, which includes some additional features: polytomous categorical manifest variables, continuous (conditionally normal) manifest variables, dependencies between continuous manifest variables, class-specific time-to-event covariate effects and piecewise exponential or Weibull baseline hazard functions. This section involves a considerable amount of detail and borrows heavily from Larsen (2004). For brevity, some of the supportive equations have been placed in Appendix B.

The EM algorithm (Dempster et al., 1977) can be used for maximising the log-likelihood in missing data problems. In latent class models, the class indicator is not observed and therefore fitting a latent class model can be formulated as a missing data problem and the EM algorithm can be used (see McLachlan and Peel, 2004, for an informative introduction). Informally, the idea is to specify the log-likelihood as if class were observed (the *complete data* log-likelihood), then find the expected complete data log-likelihood, which effectively amounts to replacing the latent class indicator with its conditional expectation (the E-step), and then finding the parameter estimates which maximise the log-likelihood (the M-step). The E and M-steps are re-calculated at each iteration until convergence. For some parameters, closed-form maximum likelihood solutions are available in the M-step, whereas for others, maximisation is undertaken iteratively within the M-step.

In order to incorporate summations over subjects, notation is now changed so that, for example, $\mathbf{y}$ represents the observed $N \times M$ matrix of responses for the categorical manifest variables, for $i = 1, \ldots, N$ subjects on the $m = 1, \ldots, M$ categorical manifest variables, and $\mathbf{y}_i$ is the vector of responses for the ith individual. For brevity, random variables have also been omitted from the subscripts for density and probability mass functions.

The full outcome vector of *observed data* for the ith subject is $(\mathbf{x}_i^\intercal, \mathbf{y}_i^\intercal, \mathbf{w}_i^\intercal, \mathbf{z}_i^\intercal, t_i, \delta_i)^\intercal$ and the vector of *complete data* is $(\mathbf{x}_i^\intercal, \mathbf{y}_i^\intercal, \mathbf{w}_i^\intercal, \mathbf{z}_i^\intercal, t_i, \delta_i, j)^\intercal$. Let $\boldsymbol{\theta} = (\boldsymbol{\kappa}^\intercal, \boldsymbol{\pi}^\intercal, \boldsymbol{\mu}^\intercal, \boldsymbol{\sigma}^\intercal, \boldsymbol{\alpha}^\intercal, \boldsymbol{\beta}^\intercal, \boldsymbol{\gamma}^\intercal)^\intercal$. As a reminder, $\boldsymbol{\kappa}$ represents the vector of the effects of the latent class predictors (see Section 2.2.1) which is now included in the full joint model. The *complete data* log-likelihood is then given by

$$\begin{aligned}
\ell_{\text{comp}}(\boldsymbol{\theta}) &= \sum_{i=1}^{N} \log\left[f(\mathbf{y}_i, \mathbf{w}_i, t_i, \delta_i, j_i|\mathbf{x}_i, \mathbf{z}_i)\right], \\
&= \sum_{i=1}^{N} \log\left[P(j_i|\mathbf{x}_i)P(\mathbf{y}_i|j_i)f(\mathbf{w}_i|j_i)f(t_i, \delta_i|\mathbf{z}_i, j_i)\right], \\
&= \sum_{i=1}^{N}\sum_{j=1}^{J} v_{ij} \log\left[P(j_i|\mathbf{x}_i)P(\mathbf{y}_i|j_i)f(\mathbf{w}_i|j_i)f(t_i, \delta_i|\mathbf{z}_i, j_i)\right],
\end{aligned}$$

where v_{ij} is an indicator variable equal to 1 if the ith subject belongs to the jth class and 0 otherwise.

3.6.1 The E-step

The E-Step at the rth iteration is given by

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}) = E_{\boldsymbol{\theta}^{(r)}}[\ell_{\text{comp}}(\boldsymbol{\theta})|\mathbf{x}, \mathbf{y}, \mathbf{w}, \mathbf{t}, \boldsymbol{\delta}]. \tag{3.4}$$

For the initial E-step, $\boldsymbol{\theta}^{(0)}$ represents the starting values for the unknown parameters. The E-step is straightforward as it effectively involves replacing each v_{ij} in equation 3.4 with $\nu_{ij}^{(r)}$, the expected class value or class-specific posterior probability at the rth iteration, given by

$$\nu_{ij}^{(r)} = \frac{P(j_i|\mathbf{x}_i)P(\mathbf{y}_i|j_i)f(\mathbf{w}_i|j_i)f(t_i, \delta_i|j_i, \mathbf{z}_i)}{\sum_{k=1}^{J} P(k_i|\mathbf{x}_i)P(\mathbf{y}_i|k_i)f(\mathbf{w}_i|k_i)f(t_i, \delta_i|k_i, \mathbf{z}_i)},$$

for $i = 1, \ldots, N; j = 1, \ldots, J$.

3.6.2 The M-step

The M-step is more complicated. In the M-Step it is required to find $\boldsymbol{\theta}^{(r+1)}$ which maximises equation 3.4. By taking partial derivatives of equation 3.4, some maximum likelihood estimators can be obtained:

$$\pi_{mgj}^{(r+1)} = \frac{\sum_{i=1}^{N} \nu_{ij}^{(r)} I\{y_{im} = g\}}{\sum_{i=1}^{N} \nu_{ij}^{(r)}},$$

$$\mu_{lj}^{(r+1)} = \frac{\sum_{i=1}^{N} \nu_{ij}^{(r)} w_{li}}{\sum_{i=1}^{N} \nu_{ij}^{(r)}},$$

$$\sigma_{ll'}^{(r+1)} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{J} \nu_{ij}^{(r)} (w_{li} - \mu_{lj})(w_{l'i} - \mu_{l'j})}{\sum_{i=1}^{N} \sum_{j=1}^{J} \nu_{ij}^{(r)}}.$$

where $I\{\cdot\}$ denotes an indicator function. However, $\hat{\boldsymbol{\kappa}}$ does not have a closed-form solution. Bandeen-Roche et al. (1997) and Larsen (2004) suggest updating $\hat{\boldsymbol{\kappa}}$ using a single step of a Newton-Raphson algorithm,

$$\boldsymbol{\kappa}^{(r+1)} = \boldsymbol{\kappa}^{(r)} + I^{-1}_{\boldsymbol{\kappa}^{(r)}} F_{\boldsymbol{\kappa}^{(r)}},$$

where $F_{\boldsymbol{\kappa}^{(r)}}$ is the $P \times 1$ vector of first-order partial derivatives (see Appendix B) and $I_{\boldsymbol{\kappa}^{(r)}}$ is the $P \times P$ matrix of negative second order derivatives evaluated at $\boldsymbol{\kappa}^{(r)}$.

In the time-to-event submodel, the various parameter vectors are dependent on each other. Larsen (2004), describes how one-step of a Newton-Raphson approach can be used when the baseline hazard is non-parametric. The same approach is adopted here for piecewise exponential and Weibull baseline hazard models. Firstly, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are updated using

$$(\boldsymbol{\beta}^{(r+1)\mathsf{T}}, \boldsymbol{\gamma}^{(r+1)\mathsf{T}})^{\mathsf{T}} = (\boldsymbol{\beta}^{(r)\mathsf{T}}, \boldsymbol{\gamma}^{(r)\mathsf{T}})^{\mathsf{T}} + I^{-1}_{\boldsymbol{\beta}^{(r)}, \boldsymbol{\gamma}^{(r)}} \left(F^{\mathsf{T}}_{\boldsymbol{\beta}^{(r)}}, F^{\mathsf{T}}_{\boldsymbol{\gamma}^{(r)}} \right)^{\mathsf{T}}, \tag{3.5}$$

where F denotes the vector of first-order partial derivatives (see Appendix B) and I denotes the corresponding matrix of second-order partial derivatives, which depend on the assumed baseline hazard function.

For a piecewise exponential model, the baseline hazard parameters for the S time intervals, $\boldsymbol{\alpha}_0 = (\alpha_{01} \dots \alpha_{0S})^{\mathsf{T}}$, are updated using

$$\alpha_{0s}^{(r+1)} = \sum_{i=1}^{N} \frac{\psi_{is} \delta_i}{\sum_{i' \in R_s} (t_{i'} - a_{s-1}) \sum_{j=1}^{J} \nu_{ij}^{(r)} \exp(\mathbf{z}_i^{\mathsf{T}} \boldsymbol{\beta}_j + \gamma_j)},$$

where ψ_{is} is an indicator variable which equals 1 if the event occurs in the sth interval for the ith subject, and 0 otherwise. R_s denotes the risk set for the sth time interval and, as defined previously, a_s denotes the upper boundary for the sth interval on the time grid and a_0 equals 0.

For a Weibull model, the shape parameter, ϕ, is included in the Newton-Raphson

update so that equation 3.5 changes to

$$(\boldsymbol{\beta}^{(r+1)\intercal}, \boldsymbol{\gamma}^{(r+1)\intercal}, \boldsymbol{\phi}^{(r+1)\intercal})^\intercal = (\boldsymbol{\beta}^{(r)\intercal}, \boldsymbol{\gamma}^{(r)\intercal}, \boldsymbol{\phi}^{(r)\intercal})^\intercal + I^{-1}_{\boldsymbol{\beta}^{(r)}, \boldsymbol{\gamma}^{(r)}, \boldsymbol{\phi}^{(r)}} \left(F^\intercal_{\boldsymbol{\beta}^{(r)}}, F^\intercal_{\boldsymbol{\gamma}^{(r)}}, F^\intercal_{\boldsymbol{\phi}^{(r)}} \right)^\intercal.$$

The scale parameter, λ, is then updated using

$$\lambda^{(r+1)} = \frac{\sum_{i=1}^N \delta_i}{\sum_{i=1}^N t_i^\phi \sum_{j=1}^J \nu_{ij}^{(r)} \exp(\mathbf{z}_i^\intercal \boldsymbol{\beta}_j + \gamma_j)}.$$

3.7 Standard errors

A difficulty with the EM algorithm is that it does not automatically yield standard errors for the parameter estimates (see e.g. McLachlan and Krishnan, 2007). Larsen (2004) suggested a profile likelihood approach where the baseline hazard parameters are profiled out as nuisance parameters. However, Hsieh et al. (2006) highlighted some theoretical difficulties with this approach and instead advocated non-parametric bootstrapping. Louis's method (Louis, 1982) provides a way of obtaining the observed Fisher information when the EM algorithm is applied, which is suitable if the number of baseline hazard parameters is relatively small.

3.7.1 Louis's method

Temporarily departing from the notation used previously in this chapter for simplicity, let $\mathbf{x}$ represent the complete data (i.e. if classes were known) and $\mathbf{y}$ represent the observed data. Using Louis's method, the observed data information matrix, $I(\boldsymbol{\theta}; \mathbf{y})$, can be expressed as

$$I(\boldsymbol{\theta}; \mathbf{y}) = I_{\text{comp}}(\boldsymbol{\theta}; \mathbf{y}) - \text{Cov}\left[F_{\text{comp}}(\mathbf{x}; \boldsymbol{\theta}) | \mathbf{y}\right], \tag{3.6}$$

where $I_{\text{comp}}(\boldsymbol{\theta}; \mathbf{y})$ is the expected negative second derivative of the complete data log-likelihood (equation 3.4) and $F_{\text{comp}}(\mathbf{x}; \boldsymbol{\theta})$ is the vector of first derivatives (score statistics) of the complete data log likelihood. For the joint model described, using Louis's method is extremely tedious as all first and second-order partial derivatives of the expected complete data log-likelihood, as well as the covariances of all of the first-order partial derivatives, are required. Louis's method was used to find the observed data information matrix for the one and two-step models used in the simulation study in the previous chapter and the requisite calculations are provided in Appendix B.

Evaluating equation 3.6 at the maximum likelihood estimates and taking the inverse,

i.e. $I^{-1}(\hat{\boldsymbol{\theta}}; \mathbf{y})$, gives the required variance-covariance matrix, $\text{Var}(\hat{\boldsymbol{\theta}})$. Standard errors can then be obtained by taking the square-root of the diagonal of $\text{Var}(\hat{\boldsymbol{\theta}})$, and Wald-type confidence intervals can be constructed (see e.g. Pawitan, 2001).

3.7.2 The two-step approach

The two-step approach proposed by Bakk and Kuha (2018) was discussed in Section 2.2.2, and in this approach the parameter vector is partitioned into two, where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\intercal, \boldsymbol{\theta}_2^\intercal)^\intercal$, and both parts are estimated separately. For the piecewise exponential model presented in this chapter, $\boldsymbol{\theta}_1 = (\boldsymbol{\kappa}^\intercal, \boldsymbol{\pi}^\intercal, \boldsymbol{\mu}^\intercal, \boldsymbol{\sigma}^\intercal)^\intercal$ and $\boldsymbol{\theta}_2 = (\boldsymbol{\alpha}^\intercal, \boldsymbol{\beta}^\intercal, \boldsymbol{\gamma}^\intercal)^\intercal$. Clearly, uncertainty in the estimates will be too low if the standard errors from the two separate steps are used. Xue and Bandeen-Roche (2004) and Bakk and Kuha (2018) show how the observed data information matrix can be obtained for two-step estimates. Let

$$I(\boldsymbol{\theta}; \mathbf{y}) = \begin{bmatrix} I_{11} & I_{12} \\ I_{12}^\intercal & I_{22} \end{bmatrix},$$

where the partition corresponds to $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. The variance-covariance matrix of the step two estimates, $\tilde{\boldsymbol{\theta}}_2$, is given by

$$\mathbf{V} = I_{22}^{-1} + I_{22}^{-1} I_{12}^\intercal \Sigma_{11} I_{12} I_{22}^{-1},$$

where Σ_{11} is the variance-covariance matrix of the step one estimates, $\tilde{\boldsymbol{\theta}}_1$. Standard errors can then be obtained as described above.

3.8 The `LCSM()` `R` function

The author-written `LCSM()` `R` is now described. The `R` code can be downloaded from `https://github.com/dlythg/LCSM`.

3.8.1 Description

Estimates joint latent-class and time-to-event models.

3.8.2 Usage

```
LCSM(x = NULL, y = NULL, y.var.vec = NULL, w = NULL, z = NULL, z2 = NULL,
     t = NULL, delta = NULL, nclass = 2, start.list = NULL, hist = FALSE,
     tol = 1e-5, nreps = 1, maxiter = 1000, printiter = TRUE, haz.method =
     "nonpara", part = NULL, ind.mat = NULL, t.mat = NULL, cov.type =
     "unstructured", custom.zeros = NULL, two.step = FALSE)
```

3.8.3 Arguments

`x`	A design matrix with N rows containing the latent class predictors. If not specified then only the class prevalences are estimated (logit scale).
`y`	Optional matrix of categorical variables with N rows. Only 0s and 1s are allowed. For Bernoulli manifest variables (i.e two categories), `x` should have $2 \times M$ columns. For multinomial manifest variables, the number of columns should be equal to the total number of possible categories and `y.var.vec` should be specified. At least one of `y` and `w` must be provided.
`y.var.vec`	A numeric vector with length equal to the total number of columns in `y`, e.g. `y.var.vec = c(1,1,2,2,2)` identifies that the first categorical manifest variable has two categories and the second has three.
`w`	Optional matrix with N rows containing the values for the continuous manifest variables. At least one of `y` and `w` must be provided.
`z`	A design matrix with N rows containing the values for the time-to-event covariates.
`z2`	A design matrix containing the columns of `z` which will be interacted with latent class.
`t`	An N-length vector of event times. If not specified then there is no time-to-event submodel.

`delta`	An N-length vector of censoring indicators (1 = event, 0 = right-censored).
`nclass`	Number of latent classes with default value of 2. If `t` is specified then current functionality only permits a maximum of two classes.
`start.list`	An optional list containing parameter starting values. If not specified then random starting values are generated.
`hist`	Logical for whether the results of each iteration should be stored. Default value is `FALSE`.
`tol`	A tolerance value for determining when the model has converged according to the observed data log-likelihood. Default value is `1e-5`.
`nreps`	Number of times the model is re-fitted with different sets of starting values. Default value is `1`.
`maxiter`	Maximum number of iterations for the estimation process. Default value is `1000`.
`printiter`	Logical for whether the iteration number should be printed. Default value is `TRUE`.
`haz.method`	A character string specifying the type of baseline hazard function. Default value is `"nonpara"` which estimates a parameter for each observed survival time as in Larsen (2004). Other options are `"PEM"` and `"Weibull"` for piecewise exponential and Weibull models, respectively. If `"PEM"` is specified then `part`, `ind.mat` and `t.mat` are required.
`part`	A numeric vector of time partitions for a piecewise exponential model.
`ind.mat`	A matrix with N rows and `(length(part) - 1)` columns specifying each subject's survival status in each time period.
`t.mat`	A matrix with N rows and `(length(part) - 1)` columns specifying the amount of time contributed by each subject within each time period.

`cov.type` A character string specifying the type of covariance matrix for the continuous manifest variables. Default value is `"localind"` for conditionally independent continuous manifest variables. Other options are `"unstructured"` and `"custom"`. If `"custom"` then argument `custom.zeros` must be specified. All covariance matrix types are assumed to be homogenous across latent classes.

`custom.zeros` Optional numeric vector with length equal to the number of elements in the lower triangle of the covariance matrix for the continuous manifest variables (excluding the diagonal). If `"custom"` is used then the `custom.zeros` argument is used to identify which covariance parameters should be held at zero.

`two.step` Logical for whether parameter estimates should be obtained using the two-step approach of Bakk and Kuha (2018). Default value is `FALSE`, i.e. parameters are estimated in one step.

3.8.4 Value

`LCSM()` currently returns many different values including all of the input arguments and many of the intermediate calculations. The most important values returned are:

`params` A matrix of parameter estimates.

`post` A $N \times J$ matrix of posterior probabilities.

`loglik` The final observed data log-likelihood.

`AIC` The final AIC.

`BIC` The final BIC.

`iterations` The number of iterations required for the final model.

If `hist == TRUE`, then the return values include results obtained from every iteration of the final chosen model.

3.9 Analysis of the prostate cancer data set

A number of joint models are now fitted to the prostate cancer data set described in Section 1.6.3, as an illustrative example. The data set is from an RCT in which three

different doses of estrogen (diethylstilbestrol) were compared with placebo in patients with advanced (Stage III and IV) prostate cancer (Bailar III et al., 1970; Byar and Corle, 1977).

The purpose of the analysis is to identify underlying subgroups in the data, based on pre-trial covariates, and to establish their effect on survival time. This data set was analysed previously (Jorgensen and Hunt, 1996; Hunt and Jorgensen, 1999) using latent class models with conditional multinomial and normal manifest variables and there was found to be evidence of two underlying subgroups. However, neither survival status nor survival time were incorporated into the statistical modelling. Instead, subjects were modally assigned to classes, and simple tabulations against outcome and survival time ($\leq$ 48 months, $>$ 48 months) were undertaken.

3.9.1 Data handling

The data set contains 506 observations, although only complete cases were considered in this analysis, leaving 474 patients. Survival times were recorded in months, and although the data set includes multiple causes of death, these were consolidated into a single all-cause mortality variable. Around 30% of survival times are right-censored. The pre-trial covariates considered for the analysis are displayed in Table 3.3.

As in the previous analyses of this data set mentioned above, the placebo and 0.2 mg estrogen were combined (Untreated) and the 1.0 and 5.0 mg estrogen treatments (Treated) were combined on account of their survival curves being so similar. The variables acid phosphatase (AP) and tumour size (SZ) were log and square-root transformed respectively to make their distributions more symmetrical, as in Hunt and Jorgensen (1999). Physical performance rating (PF) was consolidated from four categories into two ('Normal', 'Not normal') to avoid sparse cell counts.

The tumour stage variable is worthy of special attention. Stage III patients were defined as those patients with local extension outside the prostate on rectal examination, $AP \leq 1$ (King-Armstrong units - KAU); Stage IV patients were those with either $AP > 1$ and/or distant metastases. The composite variable SG (Index of tumour stage and histologic grade) is a score from 5 to 15 with three components: 1) tumour stage, 2) primary histology pattern and 3) secondary histology pattern (Gleason and Mellinger, 1974). Tumour stages III and IV contribute scores of 3 and 5 respectively, and both histology patterns are scored from 1 to 5 based on increasing histological malignancy. In this analysis, it was deliberately intended to avoid using clinically defined tumour stage as part of one of the manifest variables, so that the latent classes identified in this analysis could be compared with the clinically defined tumour stage. As a result, the stage component was subtracted from SG to create a new score variable, HP, representing the composite primary and secondary histology pattern scores only.

Variable	Abbreviation	Levels
Treatment	Trt	Treated, Untreated
Age (years)	Age	-
Weight (kg)	Wt	-
Physical performance rating ("Confined to bed")	PF	Normal, Not normal
Cardiovascular disease history	HX	No, Yes
Systolic blood pressure (kPa)	SBP	-
Diastolic blood pressure (kPa)	DBP	-
Serum haemoglobin (g/100 ml)	HG	-
Size of primary tumour (mm)	SZ	-
Index of tumour stage and histologic grade	SG	-
Serum prostatic acid phosphatase (KAU)	AP	-
Bone metastases	BM	No, Yes
Histology pattern score	HP	-

Table 3.3: Pre-trial variables in the prostate cancer data set. kPa: Kilopascals, KAU: King-Armstrong Units.

3.9.2 Model selection

The top section in Table 3.4 gives the results of a preliminary analysis of the number of latent classes required for a simple conditional independence latent class model with all pre-trial variables included as manifest variables and ignoring the time-to-event outcome. The AIC and BIC statistics suggest that a two-class model is sufficient.

Type	Classes	AIC	BIC
Starting structure	2	9721.58	9829.77
	3	9774.76	9928.72
	4	9926.80	10126.54
Final structure	2	9621.16	9729.35
	3	9641.66	9791.47
	4	9724.75	9916.16

Table 3.4: Model fit statistics for latent class models with varying numbers of latent classes according to two different model structures.

Given the possible complexity of the full joint model, a pragmatic approach to model selection was undertaken, first building the latent class submodel, then the latent class regression submodel and finally the time-to-event submodel. In each stage of model building, variables and dependencies were sequentially introduced or removed, as required, retaining features which improved Akaike's Information Criterion (AIC). Model selection was undertaken with a non-parametric baseline hazard for the survival submodel, as in Larsen (2004), to minimise assumptions. The model selection process was as follows:

1. Begin with a local independence model with all variables as manifest variables.

2. Introduce dependencies between categorical manifest variables by combining factors and between continuous variables by adding covariance parameters.

3. Remove any manifest variables which do not discriminate between classes.

4. Re-introduce those variables removed in the previous step as latent class predictors.

5. Introduce dependencies between latent class predictors and manifest variables as required.

6. Introduce non-manifest variables and interactions into the survival submodel.

7. Use the estimated non-parametric baseline hazard function from the best model to select candidate time grids for a piecewise exponential baseline hazard model, and fit the final piecewise exponential model.

All confidence intervals (CIs) were constructed at the 95% level from standard errors (SEs) obtained from a non-parametric bootstrap (Carpenter and Bithell, 2000) with 500 bootstrap resamples.

3.9.3 Results

The starting model (Model 1) and final model (Model 14) are depicted in Figures 3.4 and 3.5, respectively. All models fitted as part of the model selection process are detailed in Table 3.5. For a detailed explanation of each of the model selection steps see Appendix B.

In Model 2, where all continuous manifest variables were assumed to be conditionally independent (Figure 3.3), the fit of the curves to *Age* looked to be unsatisfactory and the inclusion of *HP* as a continuous variable appeared to be questionable. *Age* was not included as a manifest variable in the final model.

An estimated non-parametric baseline hazard was used to guide the choice of time grid for a piecewise exponential baseline hazard, and one time interval (i.e. an exponential model) was found to be sufficient (see Figure 3.6 and Table 3.5).

As sensitivity analyses, models excluding the time-to-event submodel were fitted using the final model dependence structure with varying numbers of classes (see Table 3.4, bottom section), and still a two-class solution was found to be suitable. The parameter estimates were almost completely unaffected by the removal of the time-to-event submodel, suggesting that class composition was not influenced by the survival submodel (see Table B.1).

As an additional sensitivity analysis, an exponential model using modal assignment (see Section 2.2.3) from the final joint model was also fitted, and the parameter estimates were the same to two decimal places as the time-to-event submodel from the

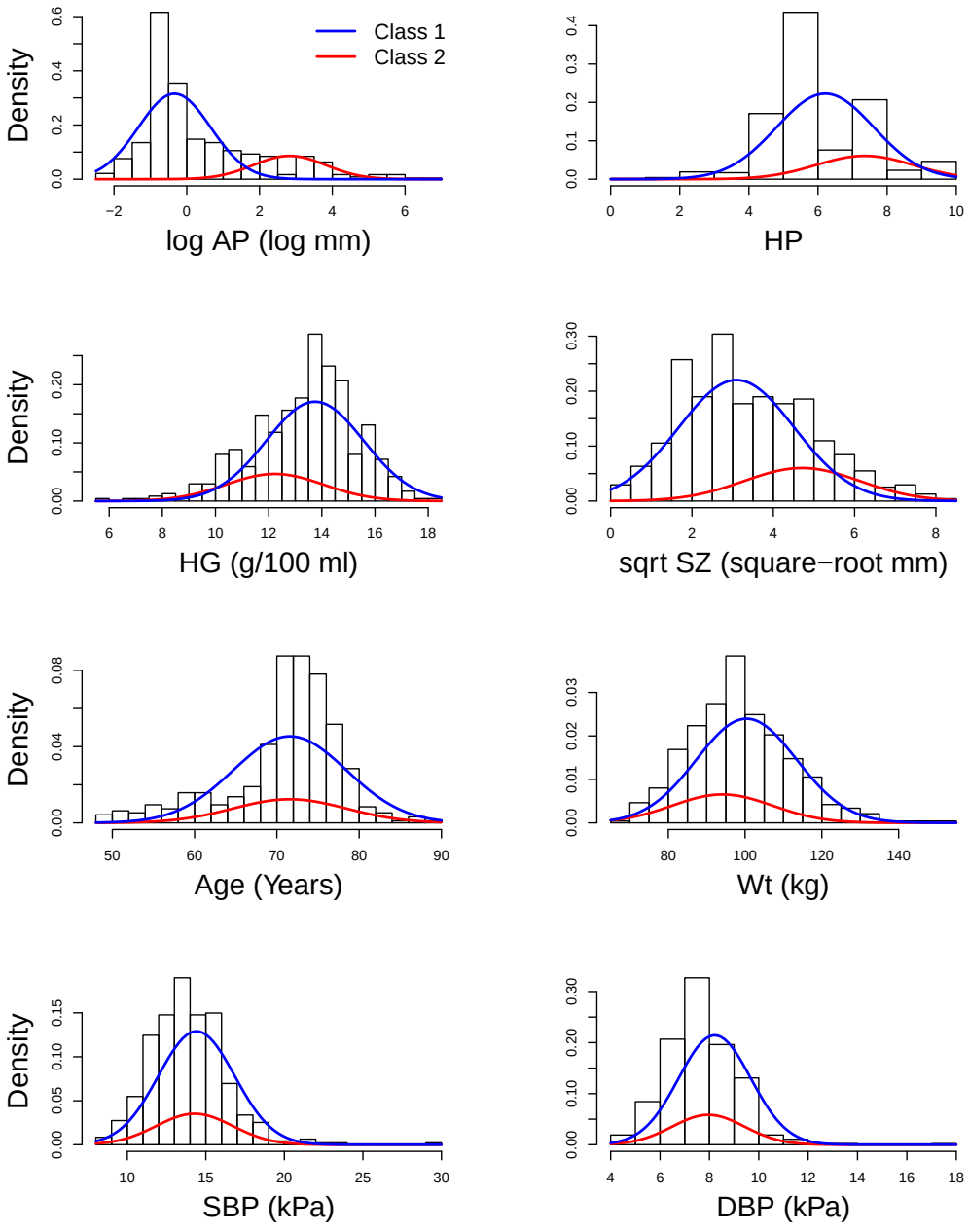

Figure 3.3: Histograms of continuous variables overlaid with fitted curves for latent classes from Model 2 (continuous variables assumed to be conditionally independent).

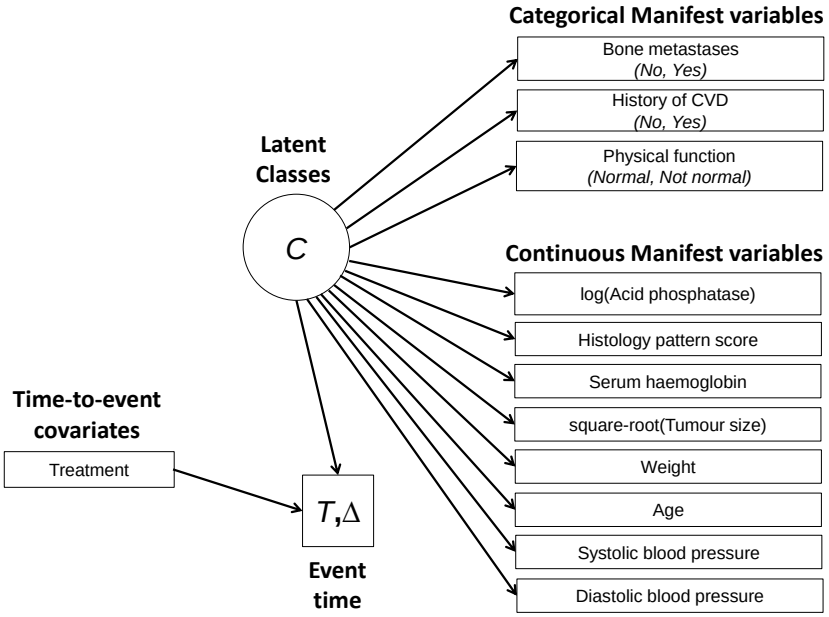

Figure 3.4: Model 1 - Starting model assuming conditional independence between manifest variables and with no latent class predictor variables.

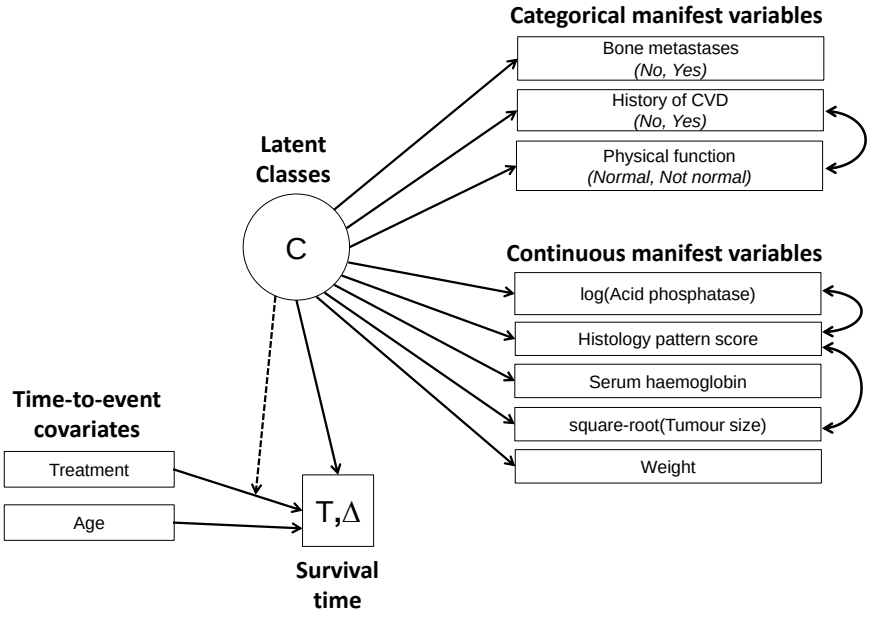

Figure 3.5: Model 14 - Final selected model with Treatment, Age and a Treatment-by-class interaction included in the time-to-event submodel. Conditional dependencies are identified with two-way arrows, and the treatment-by-class interaction is indicated by a dashed arrow.

Model no.	Base haz.	Cat. manifest vars, Y1	Cont. manifest vars, Y2	LC predictors, x	Dependencies	Survival predictors	# time periods	# params	Δ AIC$_{Ref}$	Δ BIC$_{Ref}$
Steps 1 and 2: Start with a Local Independence model and add dependencies between manifest variables										
1 [Ref.]	Non-para.	BM,HX,PF	logAP,HP,HG,sqrtSZ,Age,Wt,SBP,DBP	-	-	Trt	337	509	0	0
2	—"—	—"—	—"—	-	+ HX\PF	—"—	—"—	509	-61.97	-61.97
3	—"—	—"—	—"—	-	+ SBP\DBP	—"—	—"—	510	-299.37	-295.21
4	—"—	—"—	—"—	-	+ HP\sqrtSZ	—"—	—"—	511	-326.31	-317.99
5	—"—	—"—	—"—	-	+ HG\Wt	—"—	—"—	512	-347.97	-335.49
6	—"—	—"—	—"—	-	+ logAP\HP	—"—	—"—	513	-355.75	-339.1
7	—"—	—"—	—"—	-	+ Age\SBP	—"—	—"—	514	-365.57	-344.76
Steps 3, 4 and 5: Remove non-discriminatory manifest variables (Age, SBP and DBP) and add latent class predictors (and dependencies) as required										
8 [Ref.]	Non-para.	BM,HX,PF	logAP,HP,HG,sqrtSZ,Wt	-	HX\PF,logAP\HP,HP\sqrtSZ	Trt	337	502	0	0
9	—"—	—"—	—"—	Age	—"—	—"—	—"—	503	2.05	6.22
10	—"—	—"—	—"—	SBP	—"—	—"—	—"—	503	2.14	6.31
11	—"—	—"—	—"—	DBP	—"—	—"—	—"—	503	-0.06	4.1
Step 6: Add non-manifest variables and interactions to the survival model										
8 [Ref.]	Non-para.	BM,HX,PF	logAP,HP,HG,sqrtSZ,Wt	-	HX\PF,logAP\HP,HP\sqrtSZ	Trt	337	502	0	0
12	—"—	—"—	—"—	-	—"—	+ Age	—"—	503	-11.47	-7.3
13	—"—	—"—	—"—	-	—"—	+ LC*Trt	—"—	504	-15.98	-7.65
Step 7: Compare time grids for PE models										
14 [Ref.]	PE	BM,HX,PF	logAP,HP,HG,sqrtSZ,Wt	-	HX\PF,logAP\HP,HP\sqrtSZ	Trt,Age,LC*Trt	1	31	0	0
15	—"—	—"—	—"—	-	—"—	—"—	2	31	-1.79	2.38
16	—"—	—"—	—"—	-	—"—	—"—	3	31	1.82	10.15

Table 3.5: Model selection process. Final selected model is Model 14. LC: latent class. '\' indicates a dependency between variables after conditioning on latent class.

final joint model. The entropy value is very high at 0.93 and therefore this result is not surprising. It was shown in Chapter 2 that latent class effects on a time-to-event outcome tend to differ most when the entropy is low.

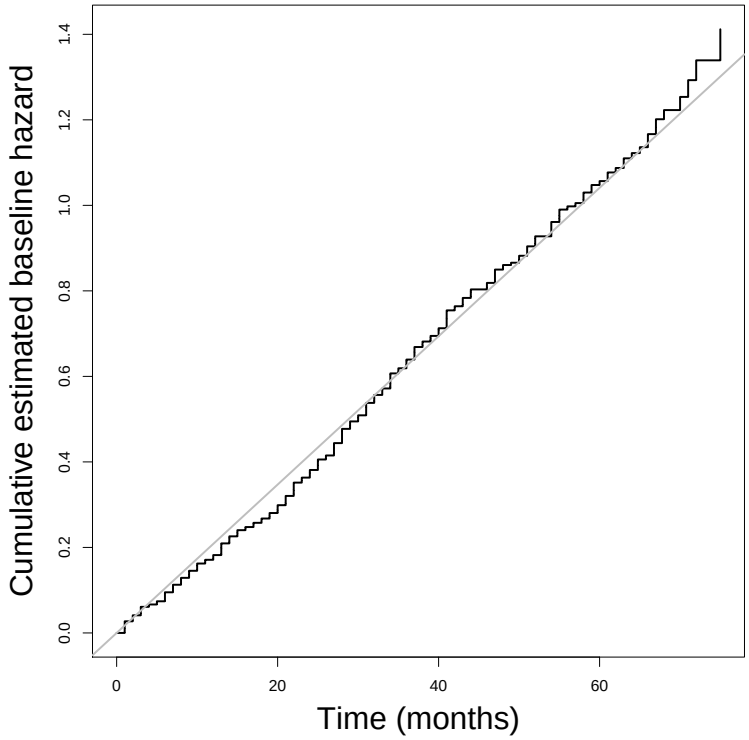

Figure 3.6: Estimated cumulative baseline hazard from a non-parametric baseline hazard model (Model 6). The fitted regression (grey) line supports the use of an exponential hazard model.

3.9.4 Interpretation of the latent classes

The estimates, SEs and CIs for the parameters from the final model are given in Table 3.6. Firstly note that $\hat{\kappa_2}$, following inverse logit transformation, corresponds to a prevalence of 20% (95% CI: 16%, 25%) for Class 2, and hence 80% for Class 1. The posterior probabilities for Class 2 are presented in Figure 3.7 and demonstrate clear class separation with most patients having very high or very low posterior probabilities of belonging to Class 2. This is reflected in the very high entropy value of 0.93 (see Section 2.2.4).

Next, the estimated survival curves are considered, Figure 3.8(b). Since continuous variable *Age* was included in the time-to-event submodel, these curves are predictions based on the mean age of 74 years. Clearly the prognosis for Class 2 patients is considerably poorer than those in Class 1. Whilst treated patients fair better than untreated patients, the treatment effect differs drastically between the two classes. The corresponding hazard ratios (HRs), treated versus untreated, are 0.93 and 0.50 for classes

Parameter	Variable	Estimate	SE	Lower	Upper	Z-value	p-value	Type
$\hat{\kappa}_2$	-	-1.40	0.15	-1.69	-1.11	9.57	<0.0001	logit probability
$\hat{\pi}_{11}$	BM	0.05	0.01	0.02	0.08	3.73	0.0002	probability
$\hat{\pi}_{21}$	CVD\PF	0.02	0.01	0.00	0.03	2.46	0.014	——"——
$\hat{\pi}_{31}$	——"——	0.41	0.03	0.36	0.46	15.9	<0.0001	——"——
$\hat{\pi}_{41}$	——"——	0.05	0.01	0.03	0.08	4.64	<0.0001	——"——
$\hat{\pi}_{12}$	BM	0.62	0.06	0.49	0.74	9.61	<0.0001	——"——
$\hat{\pi}_{22}$	CVD\PF	0.14	0.04	0.06	0.21	3.61	0.0003	——"——
$\hat{\pi}_{32}$	——"——	0.26	0.05	0.16	0.35	5.18	<0.0001	——"——
$\hat{\pi}_{42}$	——"——	0.08	0.03	0.01	0.14	2.43	0.0150	——"——
$\hat{\beta}_1$	Trt	-0.07	0.13	-0.33	0.19	0.53	0.5932	log hazard ratio
$\hat{\beta}_2$	Age	0.23	0.06	0.1	0.35	3.64	0.0003	——"——
$\hat{\gamma}_2$	Class	1.09	0.18	0.74	1.44	6.11	<0.0001	——"——
$\hat{\beta}_3$	Trt*Class	-0.62	0.27	-1.14	-0.1	2.33	0.0199	——"——
$\hat{\alpha}$	-	0.02	0.00	0.01	0.02	11.98	<0.0001	hazard rate
$\hat{\mu}_{11}$	logAP	-0.32	0.05	-0.42	-0.21	5.93	<0.00010	mean
$\hat{\mu}_{21}$	HP	6.26	0.08	6.11	6.42	79.17	<0.0001	——"——
$\hat{\mu}_{31}$	HG	13.71	0.1	13.52	13.9	140.18	<0.0001	——"——
$\hat{\mu}_{41}$	sqrtSZ	3.13	0.08	2.97	3.3	37.72	<0.0001	——"——
$\hat{\mu}_{51}$	Wt	100.26	9.57	81.51	119.02	10.48	<0.0001	——"——
$\hat{\mu}_{12}$	logAP	3.01	0.22	2.58	3.44	13.61	<0.0001	——"——
$\hat{\mu}_{22}$	HP	7.22	0.17	6.88	7.56	41.93	<0.0001	——"——
$\hat{\mu}_{32}$	HG	12.24	0.26	11.74	12.74	47.69	<0.0001	——"——
$\hat{\mu}_{42}$	sqrtSZ	4.69	0.16	4.38	4.99	30.06	<0.0001	——"——
$\hat{\mu}_{52}$	Wt	94.07	17.85	59.1	129.05	5.27	<0.0001	——"——
$\hat{\sigma}_1^2$	logAP	0.91	0.13	0.65	1.17	6.78	<0.0001	variance
$\hat{\sigma}_2^2$	HP	2.06	0.18	1.7	2.41	11.31	<0.0001	——"——
$\hat{\sigma}_3^2$	HG	3.41	0.24	2.94	3.89	14	<0.0001	——"——
$\hat{\sigma}_4^2$	sqrtSZ	2.08	0.15	1.78	2.38	13.49	<0.0001	——"——
$\hat{\sigma}_5^2$	Wt	171.64	0.07	171.49	171.78	2296.76	<0.0001	——"——
$\hat{\sigma}_{12}$	logAP\HP	0.16	0.09	-0.02	0.34	1.69	0.0908	covariance
$\hat{\sigma}_{24}$	HP\sqrtSZ	0.50	0.12	0.26	0.74	4.08	<0.0001	——"——

Table 3.6: Final model - Model 14 parameter estimates and confidence intervals.

1 and 2 respectively, implying 7% and 50% reductions in hazard with treatment. The treatment-by-class interaction effect, $\hat{\beta}_3$, was highly statistically significant (Table 3.6, p = 0.0199).

Table 3.7 gives the class-conditional probabilities of bone metastases and the composite variable of history of CVD and physical function for the two latent classes. Clearly patients in Class 1 represent a healthier subpopulation with low probabilities of bone metastases, more normal physical function and a lower probability of having a history of CVD. In Class 2, around 62% of patients have bone metastases and there is a higher prevalence of abnormal physical function.

Table 3.9 also demonstrates that Class 2 represents a less healthy subpopulation, with higher *logAP* (an indicator of metastasis), worse histology, lower serum haemoglobin levels and larger primary tumours (approximately 10 mm versus 22 mm). Interestingly Class 2 also appear to have lower mean weight by around 6 kg.

In Table 3.8 the lower diagonal of the estimated covariance matrix (transformed

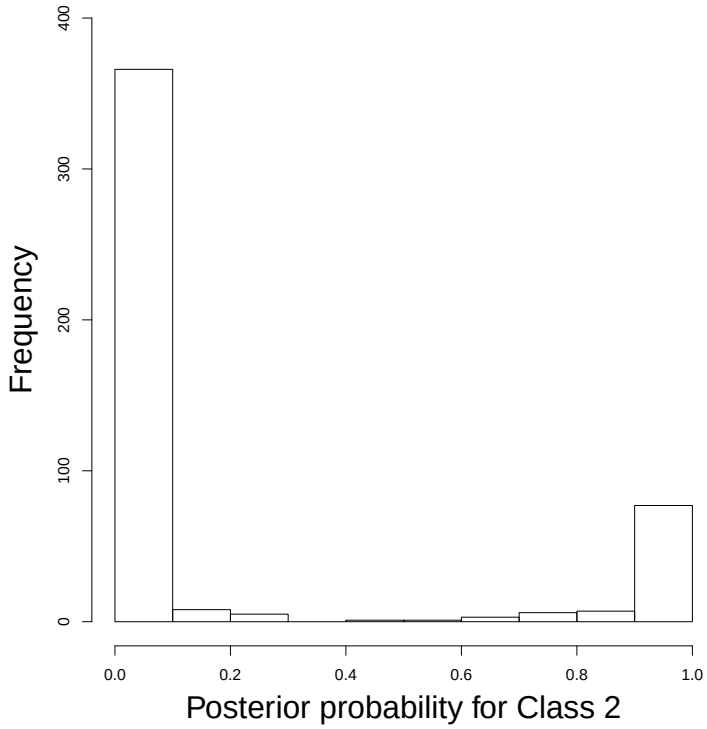

Figure 3.7: Posterior probabilities from the final joint model with exponential baseline hazard (Model 14).

	Bone metastases		History of CVD/Physical function			
	No	Yes	No/Normal	No,Not normal	Yes/Normal	Yes,Not normal
Class 1	0.95	0.05	0.52	0.02	0.41	0.05
Class 2	0.38	0.62	0.53	0.14	0.26	0.08

Table 3.7: Estimated class conditional probabilities for bone metastases (No or Yes) and the combined variable of history of cardiovascular disease (No or Yes) and physical function (Normal or Not normal)

into a correlation matrix) is shown, with estimated class conditional correlations of 0.24 between *sqrtSZ* and *HP*, and 0.11 between *HP* and *logAP*.

3.9.5 Latent class versus tumour stage

It is interesting to compare clinically defined tumour stages with the data (and model) defined latent classes. This data set comprises patients in stages III (42%) and IV (58%) whilst Model 14 indicates an 80-20% split between classes. Figure 3.9 clearly demonstrates that whilst Stage III patients belong to Class 1, most Stage IV patients are also in Class 1. Class 2 therefore represents a subset of Stage IV patients. For a comparison, a Cox model was fitted with *Trt*, *Age*, *Stage* and a *Trt*Stage* interaction. The estimated HRs, treated versus untreated, were 0.94 and 0.76 corresponding to a 6% and 24% reduction in the hazard for treated patients in stages III and IV respectively.

	logAP	HP	HG	sqrtSZ	Wt
logAP	1.00				
HP	0.11	1.00			
HG	0.00	0.00	1.00		
sqrtSZ	0.00	0.24	0.00	1.00	
Wt	0.00	0.00	0.00	0.00	1.00

Table 3.8: Lower diagonal of estimated correlations obtained from the variance-covariance matrix for the continuous manifest variables from Model 14.

	logAP	HP	HG	sqrtSZ	Wt
Class 1	-0.32	6.26	13.71	3.13	100.24
Class2	3.01	7.22	12.24	4.69	94.10

Table 3.9: Estimated class conditional means for the continuous variables

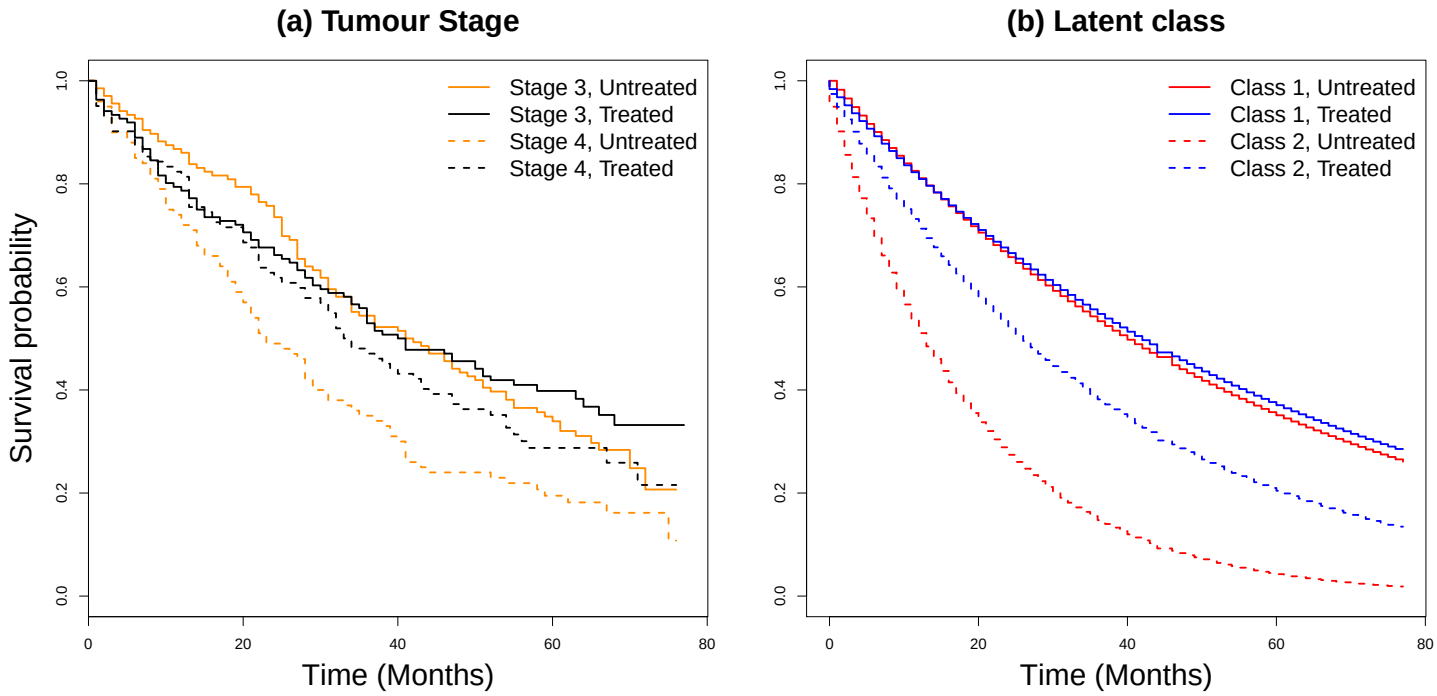

Figure 3.8: From left to right: a) Kaplan-Meier curves for tumour stage and treatment, b) Fitted survival curves from a joint model with a non-parametric baseline hazard (Model 6), c) Fitted survival curves from a joint model with an exponential baseline hazard (Model 14).

Not only are these smaller than those observed for the latent classes, but also the *Trt*Stage* interaction was not statistically significant in this model (p = 0.3172) and inclusion of the interaction term did not improve the AIC. As a result, the conclusion is that there no evidence of a differential treatment effect across tumour stages, however there seemingly is for the latent classes.

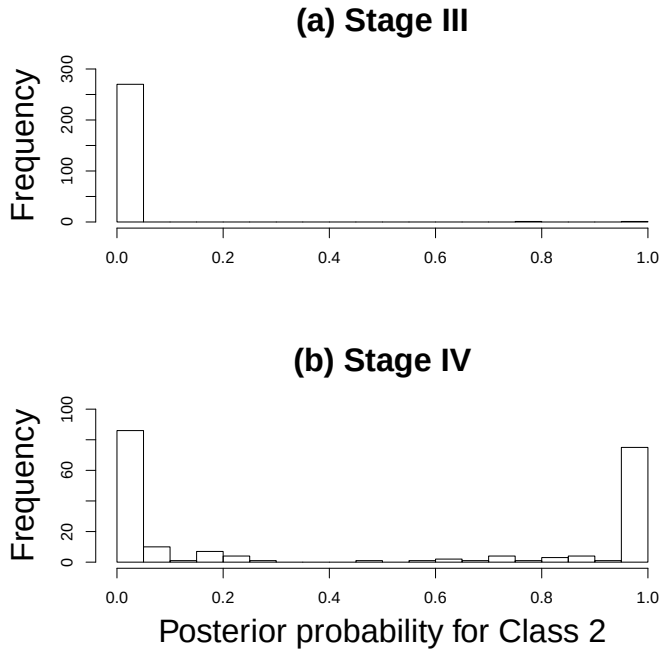

Figure 3.9: Posterior probabilities for Class 2 by tumour stage. The estimated prevalence for Class 2 is 20%. 42% of patients are in Stage III and 58% in Stage IV.

3.9.6 Model fit

Table 3.10 shows that the expected frequencies for the categorical manifest variables are very close to the observed frequencies, and the corresponding Pearson residuals are small, indicating that the model is a good fit to these variables.

Residuals for the continuous manifest variables were obtained from individual and class-specific fitted values averaged over classes (Proust-Lima et al. 2012), i.e. $\hat{y}_{il} = \sum_{j=1}^{J} \hat{\nu}_{ij}\hat{y}_{ijl}$. The Q-Q plots depicted in Figure 3.10 indicated that the model is generally a good fit to these data.

Survival residuals were also obtained using class-average individual fitted values. Figure 3.11(a) depicts the Cox-Snell residuals for the final model, which closely follow the straight line through the origin, indicating that the residuals are exponentially distributed as required and that the model fit is satisfactory (see e.g. Collett, 2015). The deviance residuals, Figure 3.11(b), exhibit a downward trend against the linear predictor indicating that the model may be overestimating the hazard of death at high

values of the linear predictor and underestimating the hazard at low values of the linear predictor. Note that this same residual pattern was observed in a separate analysis using all variables and two-way interactions in a Cox model (not shown). Missing covariates, time-dependent effects or non-linear associations may be responsible, but these were not investigated further.

3.10 Discussion

In this chapter, a general joint latent class and time-to-event model was introduced and the `LCSM()` `R` function to fit the model was described. Although it is possible to fit the general model using commercial software, the `LCSM()` function makes it possible to fit these models using open source software `R` (R Core Team, 2017). An adapted version of the estimation routine detailed in Larsen (2004) was presented, including some additional features: polytomous categorical manifest variables, continuous (conditionally normal) manifest variables, dependencies between continuous manifest variables, class-specific time-to-event covariate effects and piecewise exponential or Weibull baseline hazard functions. Various joint models were fitted to a data set from an RCT, in which two latent classes were identified that exhibited a heterogeneous treatment effect. Interestingly, the identified classes differed from clinically-defined tumour stages, and the use of tumour stage in a time-to-event model would not have led us to believe that there was a heterogenerous treatment effect.

A limitation of the presented model was that conditional variances for continuous manifest variables were constrained to be equal across classes, which can impact the latent classes identified. This constraint was imposed for simplicity. In the clinical example, residual diagnostics for the continuous manifest variables suggested this homogeneity assumption was reasonable in this case. Additionally, currently time-to-event covariate and class effects are limited to be linear and independent of time in the current version of the `LCSM()` function.

In the clinical example, it was very clear *post-hoc* that an exponential time-to-event submodel was suitable and that any *a priori* time grid would be acceptable. However, this is likely to be the exception rather than the rule. A disadvantage of assuming a piecewise exponential baseline hazard is that estimates can be sensitive to the time grid selected, and as a result it has been argued that the time grid should be specified in advance. A discussion of this issue is provided by Han et al. (2014).

In the clinical example, a pragmatic approach was taken to model building based on forward selection. As a result it is possible that alternative versions of the model may provide a superior fit. A single categorical latent variable was assumed. More complex models with additional categorical and/or continuous latent variables, as accommodated in the framework of Asparouhov et al. (2006), could provide improvements on

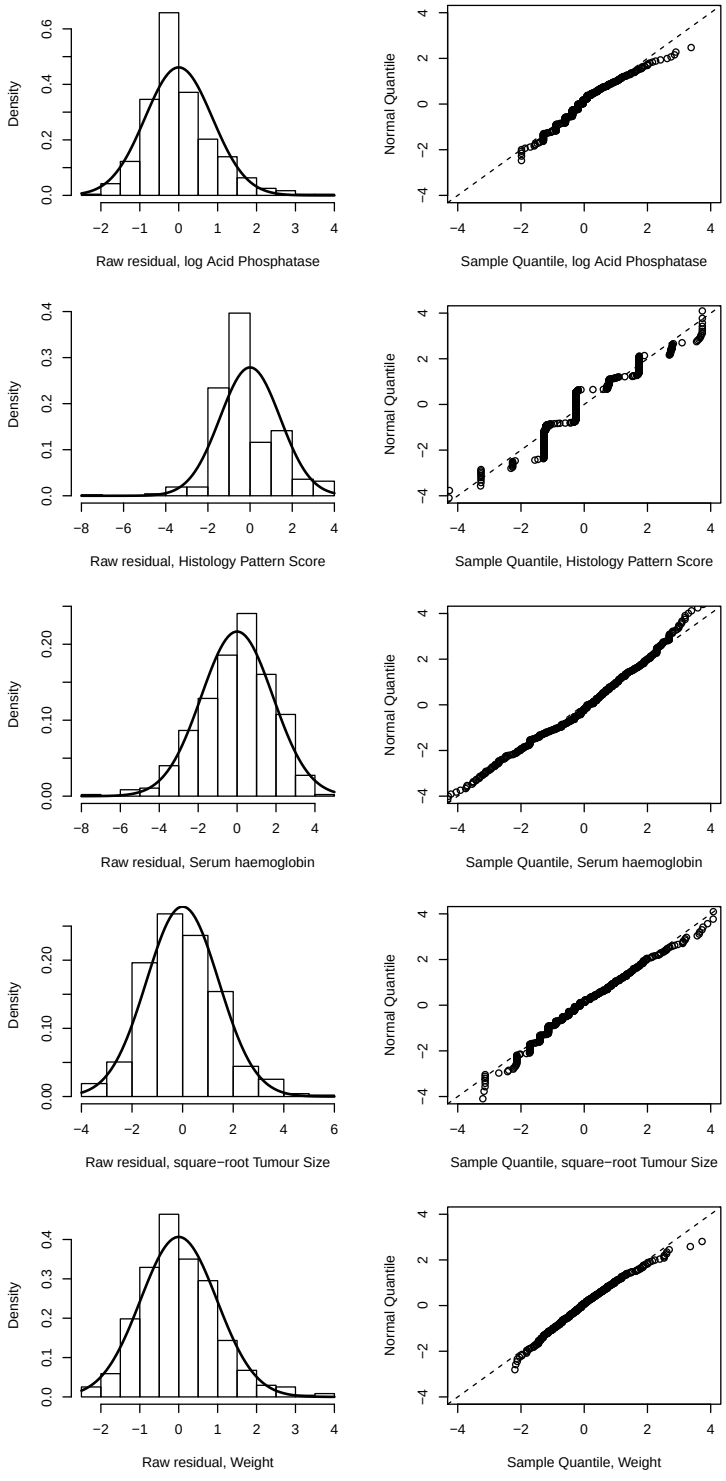

Figure 3.10: Raw residuals for the continuous variables in the final model (Model 14) with overlaid normal density curves (left) and Q-Q plots (right).

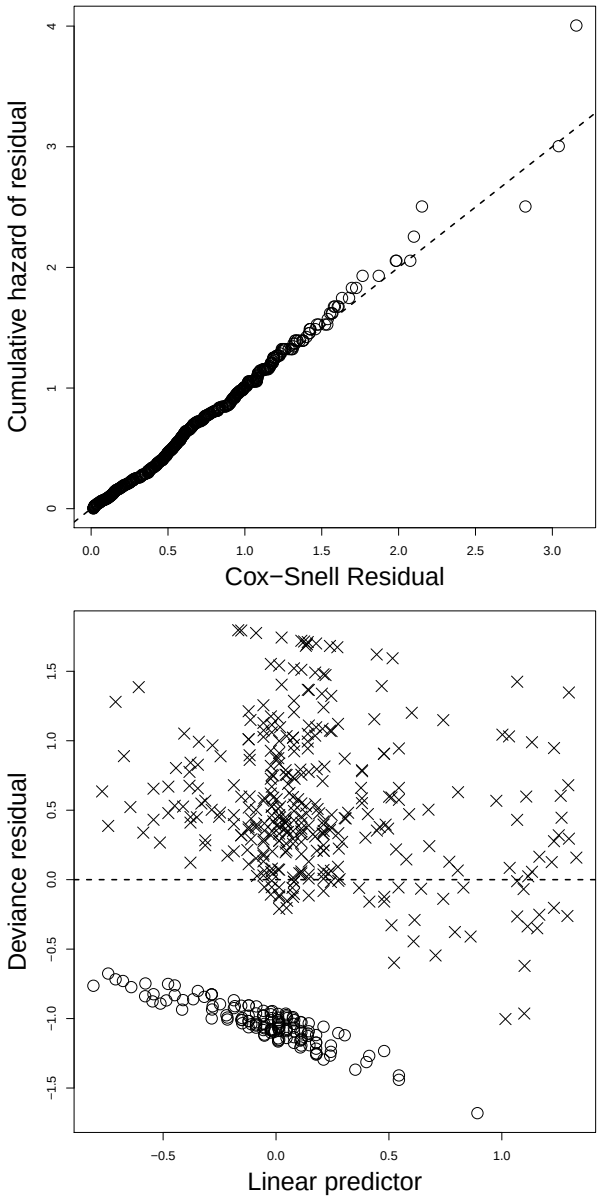

Figure 3.11: Cox-Snell and Deviance residuals for the survival submodel of the final model (Model 14).

Bone metastases		History of CVD/Physical function						
No	Yes	No,Normal	No,Not normal	Yes,Normal	Yes,Not normal	Freq.	Expected	Resid.
1	0	1	0	0	0	210	208.25	0.12
0	1	1	0	0	0	39	40.75	-0.27
1	0	0	1	0	0	8	10.73	-0.83
0	1	0	1	0	0	11	8.27	0.95
1	0	0	0	1	0	160	155.46	0.36
0	1	0	0	1	0	18	22.54	-0.96
1	0	0	0	0	1	19	22.56	-0.75
0	1	0	0	0	1	9	5.44	1.53

Table 3.10: Observed and expected frequencies for the categorical manifest variables from the final model (Model 14).

the final model selected here. However, the model diagnostics suggested that in general the model was a very good fit to the data.

Joint latent class and time-to-event models are a useful exploratory tool which can be used to identify underlying subgroups of observations and estimate their effect on a time-to-event outcome variable. The author-written `R` function `LCSM()` can be used to fit a variety of such models.

Chapter 4

Multidimensional scaling

4.1 Introduction

In the previous chapters, latent class methods were used to assess the relationship between underlying subgroups and a time-to-event outcome variable. The focus of the research is now shifted to a different type of latent variable, the latent dimension, which is associated with MDS. The distinction between latent classes and latent dimensions was given in Chapter 1.

MDS encompasses a broad range of methods for finding a configuration of points in low-dimensional space using 'proximity' data. Informative texts on MDS are Mardia et al. (1979); Cox and Cox (2000); Borg and Groenen (2003); Buja et al. (2008) and Borg et al. (2012). Each point in an MDS configuration represents an observation and the distance between two points, for a good solution, is representative of their proximity, so that 'dissimilar' observations appear far apart and similar observations close together. A plot of an MDS configuration can be an effective visualisation tool for multidimensional data which can lead to unique insights into the data. The use of MDS as a visualisation tool is developed in Chapters 5 and 7. MDS can also be used as a dimension reduction tool, analogous to principal components analysis. MDS is utilised as a dimension reduction tool in Chapter 6. As a result of its utility, MDS has been applied in many scientific disciplines including medicine (e.g. Fuller et al., 2002), genomics (e.g. Freije et al., 2004), marketing (e.g. Carroll and Green, 1997; Cha et al., 2009), psychometrics (e.g. Takane, 2006) and mining (e.g. Jamróz, 2014).

The proximity between two observations can be obtained using one of many possible measures (see e.g. Cox and Cox, 2000). A proximity measure can measure either similarity or dissimilarity, and similarities can be transformed into dissimilarities and vice-versa. Dissimilarities are more intuitive and will be considered in this thesis. Section 4.4 gives an overview of some key dissimilarity measures.

In MDS, a configuration of points is sought such that the inter-point *distances* are representative of the *dissimilarities* between observations. There are two main types of MDS: metric and non-metric. In metric MDS, the actual dissimilarities are used whereas in non-metric scaling, the dissimilarities are replaced by their ranks. Non-metric scaling is effectively an ordinal version of MDS. In this thesis, metric scaling is used almost exclusively and some details are now reviewed.

4.2 Classical scaling

One type of metric scaling is classical scaling, also known as Principal Coordinates Analysis (Gower, 1966). Classical scaling, is a non-iterative method for finding an MDS configuration based on matrix decomposition. Usually explained in terms of a map, if the distances between, say, countries are known, then classical scaling can be used to recover the coordinates of the countries. There is no unique solution, however, since the distances between countries are unaffected by translation, rotation and reflection of the coordinates.

Generally, classical scaling seeks to recover the coordinates, $\mathbf{X}$, of N points in P dimensional space from the distances between points. The full mathematical details of classical scaling can be found in Mardia et al. (1979), Cox and Cox (2000) and Borg and Groenen (2003), and only the main results are presented here. $\mathbf{A}$ is a matrix of squared Euclidean distances, which is doubly centred to produce $\mathbf{B}$. $\mathbf{B}$ is symmetric, positive semi-definite with rank P and hence has P positive eigenvalues and $N - P$ zero eigenvalues. $\mathbf{B}$ can be expressed as

$$\mathbf{B} = \mathbf{X}\mathbf{X}^\intercal, \tag{4.1}$$

and also in terms of its spectral decomposition

$$\mathbf{B} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\intercal,$$

where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$, the diagonal matrix of eigenvalues of $\mathbf{B}$, and $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_N]$, the corresponding matrix of normalised eigenvectors. Conventionally, the eigenvalues are sorted in descending order with their corresponding eigenvectors. Excluding any non-zero eigenvalues,

$$\mathbf{B} = \mathbf{V}_+\mathbf{\Lambda}_+\mathbf{V}_+^\intercal, \tag{4.2}$$

where $\mathbf{\Lambda}_+ = \text{diag}(\lambda_1, \ldots, \lambda_P)$ and $\mathbf{V}_+ = [\mathbf{v}_1, \ldots, \mathbf{v}_P]$. Equation 4.2 can then be rewrit-

ten as

$$\mathbf{B} = \left[\mathbf{V}_+\mathbf{\Lambda}_+^{1/2}\right]\left[\mathbf{V}_+\mathbf{\Lambda}_+^{1/2}\right]^\intercal, \tag{4.3}$$

where $\mathbf{\Lambda}_+^{1/2} = \text{diag}(\lambda_1^{1/2}, \ldots, \lambda_P^{1/2})$. Comparing equations 4.1 and 4.3 it can be seen that

$$\mathbf{X} = \mathbf{V}_+\mathbf{\Lambda}_+^{1/2},$$

so that $\mathbf{X}$ has been recovered. In practice, the goal of classical scaling is not to completely recover $\mathbf{X}$. Instead, $\mathbf{A}$ is based on dissimilarities, not distances, and the aim is to select the S largest eigenvalues in order to find an S-dimensional configuration for which the distances between points approximate these dissimilarities. We refer to this solution as $\mathbf{Z}$. If S is small, say 2 or 3, then $\mathbf{Z}$ can be plotted and interpreted, making MDS an effective visualisation tool.

The classical scaling solution can be shown to minimise the 'Strain' (see Buja et al., 2008),

$$\text{Strain} = \left(\frac{\sum_{i,j}^{N}(\mathbf{B}_{ij} - \langle \mathbf{z}_i, \mathbf{z}_j\rangle)^2}{\sum_{i,j}^{N}\mathbf{B}_{ij}^2}\right)^{1/2}, \tag{4.4}$$

where $\mathbf{z}_i$ is the vector of coordinates for the ith observation, and $\langle\cdot\rangle$ denotes the inner product. A measure of the proportion of variation explained by the chosen solution is given by

$$\frac{\sum_{i=1}^{S}\lambda_i}{\sum_{i=1}^{N-1}\lambda_i}, \tag{4.5}$$

where the summation is only up to $N-1$ because there is always one eigenvalue equal to zero. If the dissimilarities are not calculated using the Euclidean distance, then $\mathbf{B}$ may not be positive semi-definite and some eigenvalues will be negative. In this case, $|\lambda_i|$ can replace λ_i in the denominator of equation 4.5 or the summation in the denominator can be changed to include only the positive eigenvalues (Cox and Cox, 2000).

If the data matrix $\mathbf{X}$ is centred and the Euclidean distance is used to calculate dissimilarities, then the classical scaling and principal component solutions are identical (see e.g. Cox and Cox, 2000, page 43).

4.2.1 Gower's add-a-point method

Gower (1968) showed how the MDS coordinates of a test observation can be determined in classical scaling. The MDS coordinates for the test ($N+1$th) observation are found using

$$\hat{\mathbf{z}}_{N+1} = \frac{1}{2}(\mathbf{Z}^\intercal\mathbf{Z})^{-1}\mathbf{Z}^\intercal(\mathbf{b} - \boldsymbol{\delta}_{N+1}^2),$$

where $\mathbf{b} = (b_{11}, \ldots, b_{NN})^\intercal$, the diagonal elements of $\mathbf{B}$, which represent the squared distance between each point and the centroid of $\mathbf{Z}$, and $\boldsymbol{\delta}_{N+1}^2 = (\delta_{N+1,1}^2, \ldots, \delta_{N+1,N}^2)^\intercal$,

the $N \times 1$ vector of squared dissimilarities between the test observation and the original training observations. This method is used in Chapters 6 and 7 for the purposes of prediction.

4.3 Other types of metric scaling

Other metric MDS methods such as least-squares scaling and Sammon mapping (Sammon, 1969), use iterative techniques to minimise a loss function. 'Stress' is a loss function based on the squared errors between the dissimilarity and the distance between points in the MDS configuration, i.e.

$$\text{Stress} = \sum_{i,j}^{N} (\delta_{ij} - d_{ij})^2.$$

If the matrix of dissimilarities is symmetric, then it is sufficient to half the sum (or sum over only the lower diagonal). For an MDS configuration in Euclidean space, $d_{ij} = \|\mathbf{z}_i - \mathbf{z}_j\|_2$, where $\|\cdot\|_2$ is the Euclidean norm. Many variations of Stress exist (see e.g. Borg and Groenen, 2003, Chapter 11). For example, Sammon (1969) introduced a weighted version of Stress in which smaller dissimilarities contribute a greater weight

$$\text{Stress}_{\text{Sammon}} = \sum_{i,j}^{N} \delta_{ij}^{-1} (\delta_{ij} - d_{ij})^2.$$

Generally, the value of the Stress depends on the scale of the dissimilarities, so that a change in the units of measurement will give a different absolute value. For this reason, a standardised version is sometimes used, and is given by

$$\text{Stress}_{\text{Std}} = \frac{\sum_{i,j} (\delta_{ij} - d_{ij})^2}{\sum_{i,j} d_{ij}^2}.$$

Modern software packages for MDS, such as the `R` package `SMACOF` (de Leeuw and Mair, 2009), use 'iterative majorization' to minimise Stress. Stress is a complicated loss function which may not be differentiable everywhere and may contain multiple local minima. Iterative majorization utilises a simpler surrogate function at each iteration which approximates the Stress function and is easier to optimise. To avoid accepting a locally minimum solution, multiple re-runs with different sets of starting values are required. For comprehensive discussions of iterative majorization for MDS see Borg and Groenen (2003, Chapter 8) and de Leeuw and Mair (2009).

4.4 Dissimilarity

For $i = 1, \ldots, N$ observations, P features or variables are measured, and the dissimilarity between the ith and jth observations is given by δ_{ij}. Two important properties of dissimilarities are: 1) the dissimilarity between any two observations should be greater than or equal to zero, and 2) the dissimilarity between any observation and itself is zero. Different dissimilarity measures exist for different data types and an overview of different measures is given in Cox and Cox (2000). The best known dissimilarity measure is the Euclidean distance, which is applicable to quantitative data, and is given by

$$\delta_{ij} = \left[\sum_{p=1}^{P} (x_{ip} - x_{jp})^2 \right]^{1/2},$$

where x_{ip} is the observed value of the pth variable for the ith observation. To alter the relative influence of a variable, the weighted Euclidean distance can be used, so that

$$\delta_{ij} = \left[\sum_{p=1}^{P} \omega_p (x_{ip} - x_{jp})^2 \right]^{1/2}, \tag{4.6}$$

where ω_p is a variable-specific weight.

In clinical research, variables are usually of mixed type. For example, in the ESPAC3v2 trial (Neoptolemos et al., 2010), baseline data included Age and laboratory biomarker values (quantitative/continuous), Sex (nominal) as well as Tumour grade and WHO Performance Score (ordinal). Gower (Gower, 1971) introduced a general coefficient for calculating the dissimilarity between observations based on multiple variables of mixed type. Using Gower's coefficient, dissimilarities are obtained by

$$\delta_{ij} = 1 - s_{ij} = 1 - \frac{\sum_{p=1}^{P} w_{ijp} s_{ijp}}{\sum_{p=1}^{P} w_{ijp}},$$

where w_{ijp} is a weight for the dissimilarity between subjects i and j on variable p, and s denotes a similarity. For nominal variables the similarity, s_{ijp}, is equal to 1 when the two subjects have the same value for the pth variable and zero otherwise (Table 4.1).

i	j	s_{ijp}	w_{ijp}
+	+	1	1
+	-	0	1
-	+	0	1
-	-	1	1

Table 4.1: Similarities and weights for nominal variables in Gower's coefficient.

For 'asymmetric' binary variables, two subjects are similar if they both possess some attribute ($x_{ip} = x_{jp} = +$). For continuous variables,

$$s_{ijp} = 1 - |x_{ip} - x_{jp}|/R_p,$$

where R_p is the range for the pth variable. As with the weighted Euclidean distance in equation 4.6, variable-specific weights, ω_p, can be incorporated into Gower's coefficient, either via w_{ijp}, or more explicitly by including ω_p, so that

$$\delta_{ij} = 1 - s_{ij} = 1 - \frac{\sum_{p=1}^{P} \omega_p w_{ijp} s_{ijp}}{\sum_{p=1}^{P} \omega_p w_{ijp}}.$$

When either or both subjects have missing data for a particular variable, then the weight, w_{ijp}, can be set to zero. This is a key feature of Gower's coefficient and the implicit assumption is that data are missing completely at random (MCAR, Peugh and Enders, 2004). This feature and some of its implications are discussed in more detail in Section 6.8.

4.5 Summary

In this chapter, MDS and the concept of dissimilarity were introduced. MDS is an effective multivariate tool which can be used for visualisation and dimension reduction. Several alternative MDS methods were described and it was shown how dissimilarities can be calculated when variables are of mixed type (nominal, ordinal, continuous), which is usually the case in clinical data sets. Gower's method for adding test points to an MDS configuration was also detailed and will be used in later chapters to show how MDS can be used for prediction.

Chapter 5

Time-to-event biplot axes

5.1 Introduction

In clinical research, multidimensional data sets are common as usually multiple measurements are taken for each subject. The objective of clinical research is to understand and interpret these data with the aim of improving treatments and patient care. The biplot is a multivariate generalisation of the scatter plot that can be used to visualise the key features of multidimensional data sets in a low-dimensional space. Like the scatter plot, the points on a biplot represent observations and the distance between two points is indicative of how 'dissimilar' they are, in some sense. However, rather than two perpendicular axes, in a biplot, multiple axes can be displayed, which represent the variables. The simultaneous representation of observations and variables is where the 'bi'-plot gets its name. As an exploratory tool, the biplot can reveal multivariate clustering and outliers as well as the correlation structure of the variables. MDS is one method for obtaining a biplot.

Methods for fitting biplot axes for continuous and categorical variables are well-established (see e.g. Gower and Hand, 1995; Greenacre, 2010). However, to the best of our knowledge, biplot axes for time-to-event data have not been considered previously. Time-to-event variables are common in clinical research, for example overall or progression-free survival in oncology. Time-to-event data differs from other data types since it is typically highly-skewed and subject to censoring. In this chapter, the use of AFT models is proposed to fit biplot axes for time-to-event variables. In particular the Weibull AFT model is considered since it can be formulated in both AFT and PH form, the latter representation being far more common in clinical research. Moreover, the possibility of relating an axis for a predictor variable to the time-to-event axis, in order to recover the effect of a clinical variable on the event time, is demonstrated and its utility is evaluated.

Biplots were introduced by Gabriel (1971) and two key texts on the subject are Gower and Hand (1995) and Greenacre (2010). In order to display multivariate data graphically in few dimensions, clearly some dimension reduction is necessary. Classically, the singular value decomposition of a matrix is used to obtain the coordinates of the points on a biplot, but other multivariate methods such as principal components analysis, multiple correspondence analysis, canonical variate analysis and MDS can be used (see e.g. Cox and Cox, 2000, Chapter 7). Whilst the results in this chapter are generalizable to a number of these methods, the focus here is on MDS biplots in which the data are displayed in Euclidean space. One reason for focusing on MDS is that both continuous and categorical variables can be easily incorporated, using Gower's coefficient (Gower, 1971, Section 4.4), and generally in clinical research there is a mix of data types.

The chapter is structured as follows: in Section 5.2, some notation is introduced and in Section 5.3 a simulated data set is described; this data set is used to illustrate various concepts throughout the chapter. In Section 5.4, a general result for projecting a point onto a line or plane is explained. This result is relevant to reading a measurement from a biplot axis. In Section 5.5, it is demonstrated how regression can be used to fit biplot axes in general and specifically it is shown how AFT models can be used to estimate the slope and mark out the scale of a time-to-event biplot axis. In Section 5.6 it is shown how, in principle, the relationship between a variable and the event time can be recovered directly from the biplot by associating biplot axes. In Section 5.7, consideration is given to the precision/uncertainty with which an axis slope is estimated. A coefficient of determination for AFT models and Harrell's concordance index are introduced in Section 5.8. In Section 5.9, the proposed methods are applied to a simulated data set and in Section 5.10 an analysis of the HCC data set is presented. Discussion is given in Section 5.11.

5.2 Notation

To define some general notation used throughout this chapter, observations are indexed as $i = 1, \ldots, N$; $\mathbf{X}$ is an $N \times P$ matrix of predictor variables, with observed values $\mathbf{x}$, and $\mathbf{Z}$ is an $N \times S$ matrix of MDS coordinates, with observed values $\mathbf{z}$, and with columns indexed by $s = 1, \ldots, S$. $\mathbf{Y}$ represents an N-length vector of dependent variables, with observed values $\mathbf{y}$. When the outcome of interest is a non-negative event time, $\mathbf{Y} = \log(\mathbf{T})$, where $\mathbf{T}$ is an N-length vector of event times with observed values $\mathbf{t}$. Let $\boldsymbol{\beta}$ represent a vector of coefficients for the regression of $\mathbf{Y}$ on $\mathbf{X}$ of length P, and let $\boldsymbol{\alpha}$ represent an S-length vector of coefficients for the regression of $\mathbf{Y}$ on $\mathbf{Z}$ (note that intercepts are kept separate from these coefficient vectors).

Let θ_1 denote the slope of a biplot vector or axis on a two-dimensional plot with

intercept θ_0. Bold upper case letters $\mathbf{A}$ and $\mathbf{B}$ are used to denote biplot vectors and axes, e.g. $\mathbf{A} = (\alpha_1, \alpha_2)^\intercal$ is a biplot vector which can be drawn in two-dimensions as an arrow from the origin to point (α_1, α_2). The term 'axis' is used to refer to the line that extends the vector in each direction. Bold lower case letters $\mathbf{a}$, $\mathbf{b}$, $\mathbf{o}$ and $\mathbf{p}$ are used to define specific coordinates, e.g. $\mathbf{a} = (\alpha_1, \alpha_2)^\intercal$ would refer to the point at the tip of vector $\mathbf{A}$, when defined as above.

5.3 A simulated data set

A simple simulated data set is used throughout this chapter to illustrate various concepts. Values for $P = 5$ variables, $X_1, \ldots, X_5$, and $N = 400$ observations were simulated from a multivariate normal distribution for which all variables had a mean of 0 and standard deviation of 1; correlation coefficients were $r_{X_1,X_2} = r_{X_1,X_3} = r_{X_2,X_3} = r_{X_4,X_5} = 0.4$, and otherwise 0, forming a block diagonal correlation matrix. Event times were simulated from an exponential distribution, $T_i \sim \text{Exp}\left[\exp(\boldsymbol{\beta}^\intercal \mathbf{x}_i)\right]$, for $i = 1, \ldots, N$ and with $\beta_1 = \beta_2 = \beta_3 = 0.75$ and $\beta_4 = \beta_5 = 0$, i.e.

$$\log(t_i) = 0.75x_{1i} + 0.75x_{2i} + 0.75x_{3i} + \epsilon_i,$$

where $\epsilon_1, \ldots, \epsilon_N$ are i.i.d. according to a Gumbel distribution. This is an AFT model and these models are discussed in more detail in Section 5.5.2. Additionally, event times were censored by generating censoring times using an exponential distribution with a scale parameter of 1, resulting in 51% of observations being censored.

5.4 Projecting a point onto a line

There are several ways to project a point onto a line. In the context of biplots, an observation can be projected onto an axis in order to read-off a measurement. Figure 5.1 depicts the relationship between a line, $\mathbf{A}$, and a point, $\mathbf{b}$, on a two-dimensional plot. The equation of the line $\mathbf{A}$ is given by:

$$Z_2 = \theta_0 + \theta_1 Z_1, \tag{5.1}$$

where Z_1 and Z_2 are the first and second axes or dimensions of the plot and θ_0 and θ_1 denote the intercept and the slope of $\mathbf{A}$, respectively. The equation of a line perpendicular to $\mathbf{A}$ is

$$Z_2 = \theta_0^* - \frac{1}{\theta_1} Z_1. \tag{5.2}$$

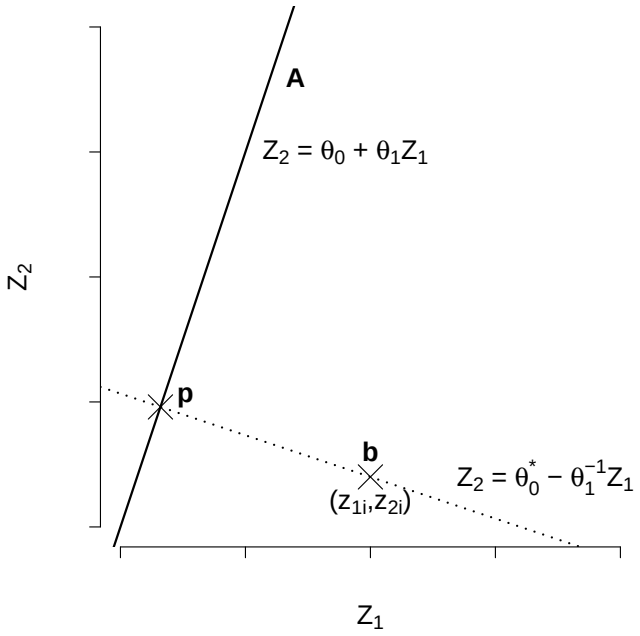

Figure 5.1: Example of the relationship between a line **A** and point $\mathbf{b} = (z_{1i}, z_{2i})$ which must lie on a line perpendicular to **A** (dotted line). Point **p** is the projection of **b** onto **A**.

In order to project point $\mathbf{b} = (z_{1i}, z_{2i})$, onto the line **A**, the perpendicular line which intersects point **b** is required. Substituting the coordinates of **b** into equation 5.2 gives

$$z_{2i} = \theta_0^* - \frac{1}{\theta_1} z_{1i},$$

and rearranging to find θ_0^*:

$$\theta_0^* = z_{2i} + \frac{1}{\theta_1} z_{1i}.$$

Substituting θ_0^* into equation 5.2 gives

$$Z_2 = (z_{2i} + \frac{1}{\theta_1} z_{1i}) - \frac{1}{\theta_1} Z_1.$$

Now the point at which the two lines intersect, **p**, can be found by equating

$$Z_{2i} = (z_{2i} + \frac{1}{\theta_1} z_{1i}) - \frac{1}{\theta_1} Z_{1i},$$

and

$$Z_{2i} = \theta_0 + \theta_1 Z_{1i}. \tag{5.3}$$

It is straightforward to then show that the coordinates of $\mathbf{p}$ are

$$Z_{1i} = \frac{\theta_1 z_{2i} + z_{1i} - \theta_0\theta_1}{\theta_1^2 + 1},$$

and hence

$$Z_{2i} = \theta_0 + \theta_1\left(\frac{\theta_1 z_{2i} + z_{1i} - \theta_0\theta_1}{\theta_1^2 + 1}\right).$$

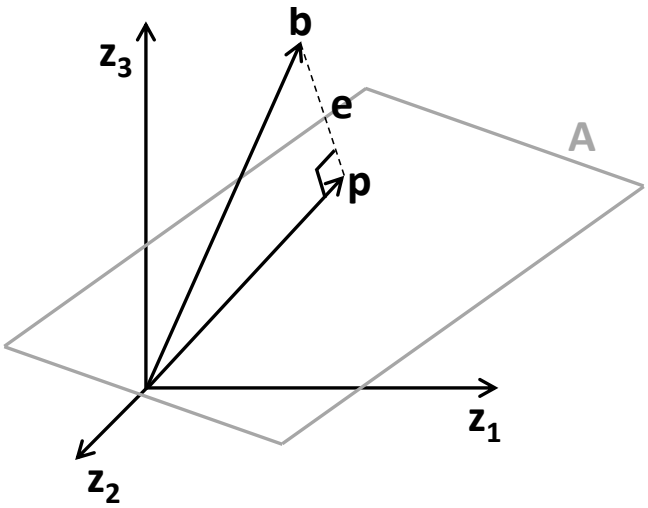

Figure 5.2: Example of a projection of a point, $\mathbf{b}$, onto a three-dimensional plane, $\mathbf{A}$. Point $\mathbf{p}$ is the projection of $\mathbf{b}$ onto $\mathbf{A}$, orthogonal to $\mathbf{A}$, and can be found by solving the linear equation which minimises the sum of squared 'errors', $\mathbf{e}$.

It is similarly straightforward to project a point onto on axis/vector in a higher-dimensional space, or even a plane, as depicted for three dimensions in Figure 5.2. A useful resource in this regard can be found online in "Projections onto subspaces"MIT (2011), on which the following is based. Define a plane $\mathbf{A}$ in $\mathbb{R}^n$ and point $\mathbf{b}$ is to be projected onto $\mathbf{A}$. $\mathbf{A}$ is a matrix of basis vectors, $[\mathbf{a}_1, \ldots, \mathbf{a}_{n-1}]$, and $\mathbf{b}$ is a vector. The point $\mathbf{p}$ is the point closest to $\mathbf{b}$ on the plane $\mathbf{A}$ and is found at the intersection formed by a line $\mathbf{e}$ through $\mathbf{b}$, which is orthogonal to $\mathbf{A}$. Since $\mathbf{p}$ lies on $\mathbf{A}$, it is required to solve $\mathbf{p} = \mathbf{A}\hat{\boldsymbol{\gamma}}$. Now,

$$\begin{aligned}\mathbf{e} &= \mathbf{b} - \mathbf{p} \\ &= \mathbf{b} - \mathbf{A}\hat{\boldsymbol{\gamma}}.\end{aligned}$$

Since $\mathbf{A}$ and $\mathbf{e}$ are orthogonal,

$$\begin{aligned}
\mathbf{A}^\intercal \mathbf{e} &= \mathbf{0}, \\
\mathbf{A}^\intercal(\mathbf{b} - \mathbf{A}\hat{\boldsymbol{\gamma}}) &= \mathbf{0}, \\
\mathbf{A}^\intercal \mathbf{b} &= \mathbf{A}^\intercal \mathbf{A}\hat{\boldsymbol{\gamma}}, \\
(\mathbf{A}^\intercal \mathbf{A})^{-1}\mathbf{A}^\intercal \mathbf{b} &= \hat{\boldsymbol{\gamma}},
\end{aligned}$$

i.e. the least squares estimate, and so

$$\mathbf{p} = \mathbf{A}(\mathbf{A}^\intercal \mathbf{A})^{-1}\mathbf{A}^\intercal \mathbf{b}.$$

If $\mathbf{A}$ is a vector then only a single estimate is returned.

5.5 Fitting the time-to-event axis

5.5.1 Without censoring

The standard approach to fitting a biplot axis for a quantitative variable is to use linear regression. For the ith subject,

$$y_i = \mu + \boldsymbol{\alpha}^\intercal \mathbf{z}_i + \epsilon_i, \tag{5.4}$$

$i = 1, \ldots, N$ and where $\boldsymbol{\alpha}$ is an S-length vector of regression coefficients, $\mathbf{z}_i$ is an S-length vector of MDS coordinates for the ith subject, ϵ_i is an error term and $\epsilon_1, \ldots, \epsilon_N$ are i.i.d. according to a normal distribution, $\epsilon_i \sim N(0, \sigma^2)$. In the case of modelling a non-negative event time, log-transformation may be necessary, so that $y_i = \log(t_i)$. Equation 5.4 can be solved by minimising the sum of squared errors using least squares. The well-known coefficient of determination, R^2, is a measure of how much variation in the response variable is explained by the MDS configuration.

In the case of a two-dimensional biplot, an estimated biplot vector/axis intersects the points $(0, 0)$ and $(\hat{\alpha}_1, \hat{\alpha}_2)$ and therefore z_{1i} and z_{2i} are related through

$$z_{2i} = \frac{\hat{\alpha}_2}{\hat{\alpha}_1} z_{1i}. \tag{5.5}$$

Linking with Section 5.4, the intercept for the biplot axis is typically $\hat{\theta}_0 = 0$ and the slope is $\hat{\theta}_1 = \hat{\alpha}_2/\hat{\alpha}_1$. The intercepts θ_0 and μ should not be confused. For a biplot axis which intersects the origin, it is not necessary to fit the regression model in equation 5.4

without the intercept, μ. Note that the biplot axis can actually be plotted as any line parallel to the fitted axis (it does not necessarily have to intersect the origin). The regression-based approach can be used for any continuous variable and, as an example, biplot axes for the simulated predictor variables are depicted in Figure 5.3(a).

The same idea can be extended to categorical variables using generalised linear models (Greenacre, 2010, Chapter 3). For example, a binary (absence/presence) categorical variable can be assumed to follow a binomial distribution, with the conditional probability of some characteristic being present for the ith subject, $\pi(\mathbf{z}_i)$. The logit or 'log-odds' can be modelled using logistic regression,

$$\text{logit}\left[\pi(\mathbf{z}_i)\right] = \log\left[\frac{\pi(\mathbf{z}_i)}{1-\pi(\mathbf{z}_i)}\right] = \mu + \boldsymbol{\alpha}^\intercal \mathbf{z}_i,$$

for $i = 1, \ldots, N$. Other generalised linear models can be used depending on the nature of the outcome variable (see e.g. Agresti, 2002). Biplot axes for categorical variables are discussed further in Section 5.11.

5.5.2 With censoring

Now we consider the case where the outcome variable is a censored event time. Let C be a censoring indicator with observed value c, where c_i equals 1 if the event is observed for the ith subject and 0 otherwise. If the event times for some subjects are censored then the linear regression model (equation 5.4) is no longer suitable. Instead, an AFT model is proposed. The AFT model can be expressed in log-linear form as

$$y_i = \mu + \boldsymbol{\alpha}^\intercal \mathbf{z}_i + \phi\epsilon_i, \tag{5.6}$$

where ϕ is a scale parameter, the random variable ϵ_i models the deviation of the values of $y_i = \log(t_i)$ from the linear predictor and $\epsilon_1, \ldots, \epsilon_N$ are i.i.d.. For an assumed Weibull, log-normal or log-logistic distribution for the event time, ϵ_i is assumed to follow a Gumbel, normal or logistic distribution, respectively (see e.g. Collett, 2015). When ϵ_i follows a Gumbel distribution and $\phi = 1$, the survival function is exponential, as was introduced in Section 5.3. Leaving the distribution of ϵ_i as unspecified is analogous to not estimating the baseline hazard in Cox proportional hazards regression (Wei, 1992). The Weibull AFT model is particularly useful as it can be expressed both in proportional hazards and acceleration factor form, and is the only time-to-event model that does so. As an example, a biplot axis for the simulated event times, obtained using a a Weibull AFT model, is depicted in Figure 5.3(b). Other features of this plot are discussed in later sections. Observing the likelihood for a parametric AFT model,

it can be seen that censoring is accounted for,

$$L = \prod_{i=1}^{N} \left[f(t_i)\right]^{c_i} \left[S(t_i)\right]^{(1-c_i)},$$

where $f(.)$ and $S(.)$ represent density and survival functions respectively.

5.5.3 Scale of the time-to-event axis

The biplot vector for a time-to-event variable is simply $\mathbf{A} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_S)^\intercal$. We refer to the line that extends the vector in each direction as the axis. In this section, it is shown how markers can be added to the time-to-event axis, which correspond to expected event times, in order to mark out the scale of the axis. This is sometimes referred to as 'calibration' and an example is given in Figure 5.3(b).

Assuming a Weibull distribution for the event times, the AFT model in equation 5.6 implies that $T_i \sim \text{Weibull}\left[\mu\exp(\boldsymbol{\alpha}^\intercal \mathbf{z}_i), 1/\phi\right]$. The expected event time for the ith subject can be shown to be

$$\text{E}(T_i) = \exp(\mu + \boldsymbol{\alpha}^\intercal \mathbf{z}_i)\Gamma(\phi + 1), \tag{5.7}$$

where $\Gamma(.)$ is the Gamma function (see Liu and Lim, 2018, pages 3 and 4). In the special case where $\phi = 1$ (an exponential time-to-event model), $\Gamma(2)$ is just 1. From equation 5.7 it can easily be seen that the origin and tip of the vector for the event time correspond to the expected event time when $\mathbf{z}_i = \mathbf{0}$ and $\mathbf{z}_i = \mathbf{1}$, respectively.

Equation 5.7 can now be rearranged to find coordinates along the time-to-event axis which represent the chosen scale, e.g. 1, 2 and 5 years. Assuming a two-dimensional MDS configuration for ease of explication, the expected event time for the ith subject is

$$\text{E}(T_i) = \exp(\mu + \alpha_1 z_{1i} + \alpha_2 z_{2i})\Gamma(\phi + 1).$$

Substituting for z_{2i} using equation 5.5,

$$\begin{aligned}
\text{E}(T_i) &= \exp\left[\mu + \alpha_1 z_{1i} + \alpha_2 (\frac{\alpha_2}{\alpha_1} z_{1i})\right] \Gamma(\phi + 1), \\
\log\left[\text{E}(T_i)\right] &= \mu + \alpha_1 z_{1i} + \frac{\alpha_2^2}{\alpha_1} z_{1i} + \log\left[\Gamma(\phi + 1)\right], \\
\alpha_1 \log\left[\text{E}(T_i)\right] &= \alpha_1 \mu + \alpha_1^2 z_{1i} + \alpha_2^2 z_{1i} + \alpha_1 \log\left[\Gamma(\phi + 1)\right].
\end{aligned}$$

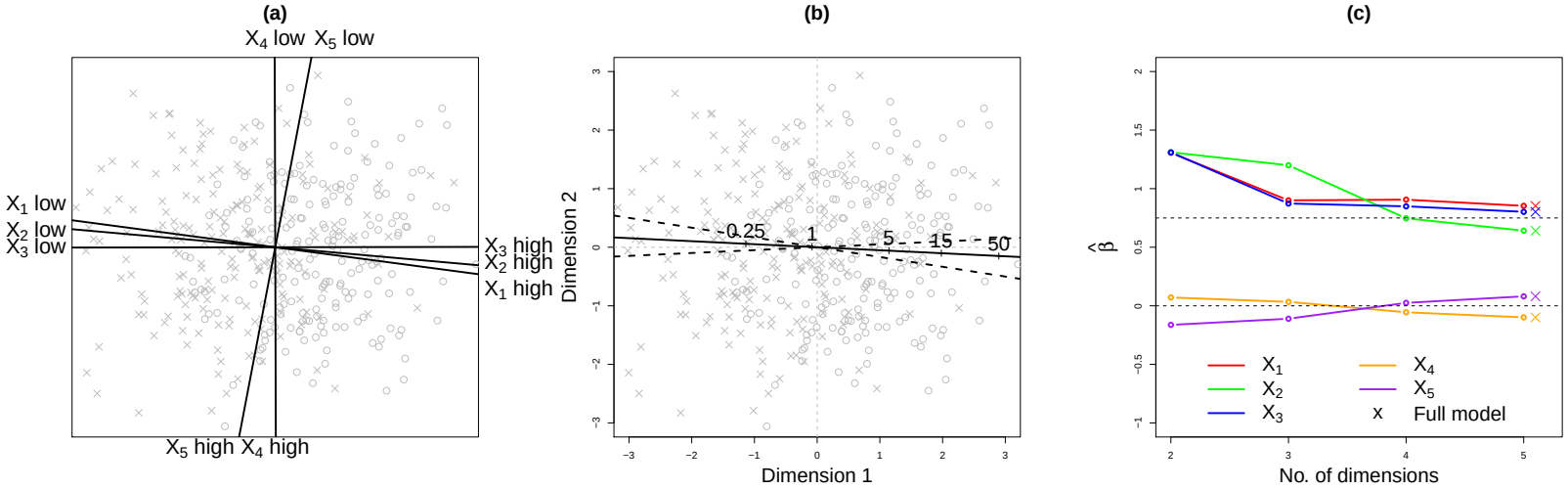

Figure 5.3: Example MDS biplot results obtained using the simulated data set. (a) Biplot and biplot axes for five predictor variables obtained using linear regression, with R^2 values of 0.57, 0.56, 0.67, 0.79 and 0.65 for $\mathbf{x}_1, \ldots, \mathbf{x}_5$, respectively. Observed and censored event times are depicted as crosses and circles, respectively. (b) Time-to-event biplot axis obtained using a Weibull AFT model ($R^2 = 0.64$) with the scale of the axis marked out (arbitrary units). 95% confidence limits for the slope are also shown and were obtained using a non-parametric bootstrap with 1000 resamples. (c) Estimates of the parameters $\hat{\beta}_1, \ldots, \hat{\beta}_5$ obtained by associating the scales in the biplot for $S = 2, \ldots, 5$ dimensional MDS solutions. Dashed lines indicate true parameter values ($\beta_1 = \beta_2 = \beta_3 = 0.75$ and $\beta_4 = \beta_5 = 0$). Coloured crosses indicate the estimates obtained from regressing $\mathbf{y}$ directly on $\mathbf{x}$ using a Weibull AFT model.

Rearranging,

$$\alpha_1 \log\left[\mathrm{E}(T_i)\right] - \alpha_1\mu - \alpha_1 \log\left[\Gamma(\phi+1)\right] = \alpha_1^2 z_{1i} + \alpha_2^2 z_{1i},$$
$$\alpha_1 \Big\{ \log\left[\mathrm{E}(T_i)\right] - \mu - \log\left[\Gamma(\phi+1)\right] \Big\} = z_{1i}\left(\alpha_1^2 + \alpha_2^2\right),$$

and therefore

$$z_{1i} = \frac{\alpha_1}{\alpha_1^2 + \alpha_2^2} \Big\{ \log\left[\mathrm{E}(T_i)\right] - \mu - \log\left[\Gamma(\phi+1)\right] \Big\}. \tag{5.8}$$

z_{2i} is again obtained using equation 5.5. If $S > 2$, the MDS solution can be represented as ${}^{S}\mathrm{C}_2$ two-dimensional plots and equation 5.8 can be easily adapted to find the scale on each plot. For example, for $S = 3$, a Dimension 1 coordinate on the scale for the first two-dimensional plot (Z_1 versus Z_2) would be given by

$$z_{1i} = \frac{\alpha_1}{\alpha_1^2 + \alpha_2^2} \Big\{ \log\left[\mathrm{E}(T_i)\right] - \mu - \alpha_3 z_{3i} - \log\left[\Gamma(\phi+1)\right] \Big\},$$

for a given value, z_{3i}, possibly the observed mean (which will be zero for a centered MDS configuration).

5.6 Associating biplot axes

In the previous section, it was demonstrated how a measurement scale could be added to a time-to-event biplot axis. The same approach can be followed for other variables, replacing equation 5.7 with the appropriate expectation. In principle then, using results for projecting a point onto a line in Section 5.4, the approximate relationship between a variable and the event time, implied by the biplot, can be recovered. The implication for a biplot with a time-to-event axis is that approximate acceleration factors, and hazard ratios if a Weibull distribution is assumed for the event time, may be recovered directly from the biplot.

A two-dimensional configuration is now used to illustrate, as depicted in Figure 5.4. Let $\mathbf{B} = (\hat{\kappa}_1, \hat{\kappa}_2)^{\intercal}$ represent a biplot vector for a centered continuous predictor variable $\mathbf{X}$, with observed values $\mathbf{x}$, from the linear regression model $\hat{x}_i = \hat{\kappa_1} z_{1i} + \hat{\kappa_2} z_{2i}$, for $i = 1, \ldots, N$. Similarly, let $\mathbf{A} = (\hat{\alpha}_1, \hat{\alpha}_2)^{\intercal}$ represent the biplot vector for the event time, modelled using equation 5.6. Let $\mathbf{b}$ represent the point or marker on the axis for vector $\mathbf{B}$ where $\hat{x}_i = 1$. The coordinates of $\mathbf{b}$ can be found from

$$\mathbf{b} = \frac{\mathbf{B}}{\|\mathbf{B}\|},$$

where $\|.\|$ denotes the Euclidean norm and hence vector length. Point $\mathbf{b}$ can then be projected onto the axis for $\mathbf{A}$, as described in Section 5.4, and the projected point is denoted as $\mathbf{p}$. Now, for a Weibull AFT model,

$$E(Y_i) = \mu + \boldsymbol{\alpha}^\intercal \mathbf{z}_i - \phi\xi_i, \tag{5.9}$$

where $-\xi$ is the Euler-Mascheroni constant (≈ 0.57721) (see Liu and Lim, 2018, page 4). Evaluating equation 5.9 at points $\mathbf{p}$ and at the origin, $\mathbf{o}$, the coefficient which represents the effect of a one-point increase, β, in X on Y can be obtained using

$$\begin{aligned}\beta &= E[Y_i|\mathbf{z}_i = \mathbf{p}] - E[Y_i|\mathbf{z}_i = \mathbf{o}] \\ &= \boldsymbol{\alpha}^\intercal \mathbf{p}.\end{aligned}$$

The corresponding AF is $\exp(-\beta)$, and HR is $\exp(-\beta/\phi)$, since $\log(\text{HR}) = \log(\text{AF})/\text{scale}$. As an example, parameter estimates $\hat{\beta}_1, \ldots, \hat{\beta}_5$ for the simulated data set, obtained by associating biplot axes, are shown in Figure 5.3(c) for various numbers of dimensions. The outlined approach for associating biplot axis scales is discussed further in Section 5.11.

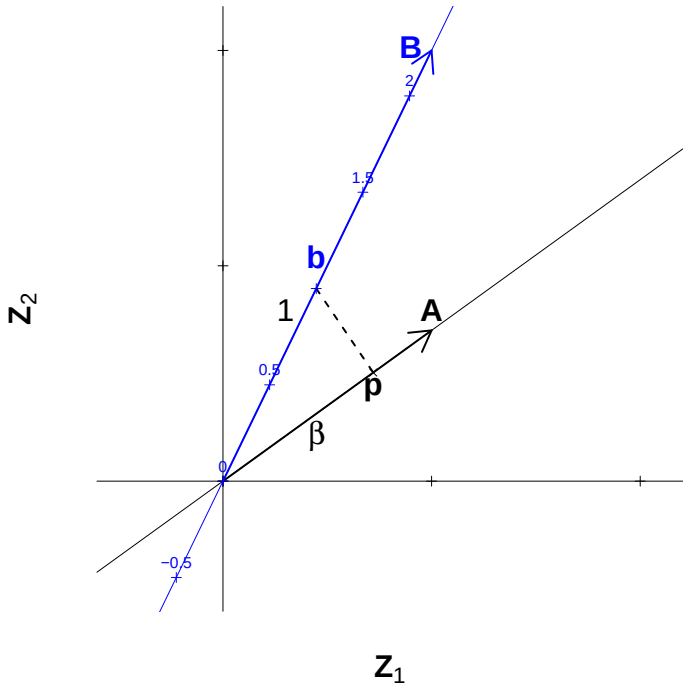

Figure 5.4: Example projection of a point on one axis to another. Axis $\mathbf{B}$ represents a predictor variable X and point $\mathbf{b}$ is located at 1 on axis $\mathbf{B}$. $\mathbf{p}$ is the projection of $\mathbf{b}$ onto axis $\mathbf{A}$, perpendicular to $\mathbf{A}$, which represents an outcome variable, Y.

5.7 Precision of the axis slope

Although not usually presented, the precision or uncertainty with which a biplot axis slope is estimated and can be indicated by adding confidence limits to the biplot. The parameter estimates from the AFT model are assumed to follow a normal distribution and therefore the axis slope, in two dimensions, is a ratio of normal parameters, $\hat{\theta}_1 = \hat{\alpha}_2/\hat{\alpha}_1$. Confidence intervals for ratios are problematic however, since the denominator, $\hat{\alpha}_1$, can take the value of zero (i.e. infinite slope), and both the mean and variance of a ratio of normal random variables are undefined (for a comprehensive discussion see Franz, 2007). Fieller (1940) derived exact confidence limits for a ratio of two normal random variables (see also Franz, 2007; Von Luxburg and Franz, 2009). However, when the denominator is near zero, Fieller's method can lead to a partially or even completely unbounded confidence interval. With MDS biplots, the orientation of the biplot is essentially arbitrary since the MDS solution is invariant to rotation. As a result, it is possible to rotate an MDS solution so that the slope is infinite, arbitrarily producing an unbounded confidence interval.

In practice, we have found that a better solution is to obtain confidence limits based on the angle of the biplot axis. Working in two dimensions, by simple trigonometry, the angle with Dimension 1 is equal to $\tan^{-1}(\hat{\alpha}_2/\hat{\alpha}_1)$, and infinite slopes are not problematic. Nieto et al. (2014) developed an `R` package which uses bootstrapping to obtain a confidence interval for the angle of a biplot axis. To illustrate, we have therefore used a simple non-parametric bootstrap (see e.g. Carpenter and Bithell, 2000) with 1000 resamples to obtain a confidence interval for the angle, and hence slope, of the biplot axis in Figure 5.3(b). In some exploratory work we have found that the width of the confidence interval is unaffected by the orientation of the MDS solution (not shown).

5.8 Measures of predictive ability

5.8.1 Coefficient of determination

A number of attempts have been made to derive a measure of explained variation equivalent to R^2 in linear models for time-to-event models (overviews can be found in Korn and Simon, 1990; Schemper and Stare, 1996). Chan et al. (2018) demonstrated that R^2 is straightforward to obtain for parametric AFT models. Since R^2 represents the proportion of variation in the response variable explained by the model,

$$R^2 = 1 - \frac{\text{Var}(\hat{\phi}\boldsymbol{\epsilon})}{\text{Var}\left[\log(\mathbf{T})\right]} = 1 - \frac{\hat{\phi}^2\text{Var}(\boldsymbol{\epsilon})}{\widehat{\text{Var}}(\mathbf{Z}\boldsymbol{\alpha}) + \hat{\phi}^2\text{Var}(\boldsymbol{\epsilon})}, \tag{5.10}$$

where $\widehat{\text{Var}}(\mathbf{Z}\boldsymbol{\alpha})$ is the sample variance. For parametric AFT models, $\text{Var}(\boldsymbol{\epsilon})$ is determined by the assumed survival distribution: $\pi^2/6$ for Weibull and exponential models, $\pi^2/3$ for logistic and log-logistic models, and 1 for normal and log-normal models. Chan et al. (2018) discuss how this approach can be extended to semi-parametric AFT models. As an example, the R^2 value for the time-to-event axis in Figure 5.3(b), where a Weibull distribution is assumed for the event time, is 0.64.

5.8.2 Concordance

Harrell et al. (1982) defined the concordance or 'c-index', a generalisation of the area under the receiver-operator-characteristic curve applicable to right-censored time-to-event data (see also Newson, 2010; Harrell, 2015). To calculate the concordance, all possible combinations of subject pairs are found, and a pair is determined as concordant if the larger predicted event time corresponds to the larger observed event time or discordant if the larger predicted event time corresponds to the shorter observed time. If both event times are censored, or if only the shorter event time is censored, then the pair is neither concordant nor discordant. The concordance is the probability of concordance between predicted and observed event times, where a concordance of 1 indicates perfect discrimination and a concordance of 0.5 indicates random predictions (analogous to tossing a coin). Usually in time-to-event analysis, the linear predictor rather than the event time is predicted and a higher value for the linear predictor corresponds to a lower expected survival time. A concordant pair is then one for which the higher linear predictor corresponds to the lower observed event time. Therneau and Atkinson (2019) provide useful details on handling ties in Harrell's c-index.

5.9 Analysis of the simulated data set

The example based on simulated data displayed in Figure 5.3 is now considered in more detail. Figure 5.3 depicts the MDS configuration with (a) predictor variable axes and (b) a time-to-event axis. Supporting results are presented in Table 5.1. When interpreting biplots: 1) the distance between observations approximates their dissimilarity, 2) the projection of an observation onto a variable axis approximates its expected value for that variable, and 3) the cosine of the angle between two axes approximates their correlation.

Classical scaling was used to find a two-dimensional MDS configuration, with 65% of the total variation in $\mathbf{x}$ explained. A Weibull AFT model was fitted to the simulated data and, unsurprisingly, event times were found to be approximately exponential, as simulated ($\hat{\phi} = 1.04 \approx 1$). The solid line in Figure 5.3(b) depicts the time-to-event biplot axis with slope $\hat{\theta}_1 = -0.05$. Roughly speaking, the axis is almost horizontal and

suggests that observations to the left of the plot are expected to have shorter survival times than than those to the right of the plot. The point estimate for the acceleration factor for Dimension 1 is $\exp(-\hat{\alpha}_1) = 0.27$, implying that the expected event time is accelerated by around 73% for every one-point decrease in Dimension 1, or conversely that there is an approximately 3.7-fold increase in the expected event time for every one-point increase in Dimension 1. The estimate for Dimension 2 was not statistically significantly different from zero (p = 0.25). The black dashed lines indicate a 95% confidence interval for the axis slope obtained using a non-parametric bootstrap of the angle. The interval is quite narrow suggesting that the slope is a reliable indicator of the direction of the axis. The R^2 value for the time-to-event axis is 0.64, so that 64% of the variation in $\mathbf{y} = \log(\mathbf{t})$ is explained by the MDS configuration, $\mathbf{z}$.

Comparing Figures 5.3(a) and (b) it can be seen that the axes for X_1, X_2 and X_3 are positively correlated with the time-to-event axis, implying that higher values are associated with longer event times, as simulated. The remaining axes are approximately perpendicular to the time-to-event axis, suggesting that they are not correlated with the expected event time, again, as simulated.

In Figure 5.3(c) the parameter estimates for the effects of $X_1, \ldots, X_5$ on Y were estimated by associating biplot scales, as described in Section 5.6, for MDS solutions with $S = 2, \ldots, 5$ dimensions. The parameters are overestimated for $X_1, \ldots, X_3$ when the number of dimensions is low but approach the estimates obtained by regressing $\mathbf{y}$ on $\mathbf{x}$ directly (coloured crosses), as the number of dimensions increases.

	$\hat{\theta}_1$	95% CI	R^2
X_1	-0.14	(-0.23, -0.05)	0.57
X_2	-0.09	(-0.19, -0.01)	0.56
X_3	0.00	(-0.07, 0.07)	0.67
X_4	-362.28	(-773.59, 804.95)	0.79
X_5	5.62	(4.08, 8.88)	0.65
Y	-0.05	(-0.17, 0.05)	0.64

Table 5.1: Supplementary results for the simulated data set. Biplot axis slopes ($\hat{\theta}_1$), 95% confidence intervals and R^2 values for five predictor variables $X_1, \ldots, X_5$ (linear regression models) and the log of the censored survival time regressed on the MDS configuration (Weibull AFT model).

5.10 Analysis of the hepatocellular carcinoma data set

HCC is a common complication of chronic liver disease and, dependent upon tumour-related, liver function and patient-related factors, the prognosis for a patient diagnosed with HCC may be extremely poor. The data from 709 HCC cases from the HCC data set are now analysed (Groups I and IV with patients with invalid survival times

excluded, see Section 1.6.1). Note that this data set is characterised by considerable amounts of missing data.

The purpose of this analysis is to use MDS to obtain a low-dimensional representation of the data based on ten clinical variables and to add a time-to-event axis to illustrate the expected survival prospects of the HCC cases from diagnosis using the methods outlined in previous sections. Moreover, the suitability of the biplot as a representation of the individual variables, and the association between each clinical variable and survival time, are assessed.

5.10.1 Clinical variables

The ten clinical variables included in this analysis are shown in Table 5.2. The variables identified are known to be important measures of HCC, liver function or are patient-related characteristics. Alpha-fetoprotein (AFP), lens culinaris agglutinin (L3) and des-gamma-carboxy prothombin (DCP) are HCC biomarkers and elevated levels are suggestive of HCC. Tumour size is the maximum size of the largest lesion. Albumin and Bilirubin are measures of liver function, with low levels of Albumin and elevated levels of Bilirubin corresponding to poorer liver function.

The Child-Pugh score is a composite measure for assessing the prognosis of cirrhotic patients based on ascites (fluid build-up), encephalopathy (a measure of brain disease), serum levels of Albumin, total Bilirubin and prolongation of clotting time. Patients are classified as A, B or C with respective expected one-year survival probabilities of 100%, 80% and 45%. Cancer Stage is a general grouping based on one of three staging systems, as described in Johnson et al. (2014), with patients categorised as being either Early or Late stage.

From Table 5.2 it can be seen that the extent of missing data across the different variables varies considerably, and is particularly high for L3 and DCP for which less than half of subjects have data. The row labelled as 'Complete' corresponds to the the number of observations with complete data with respect to all ten clinical variables.

5.10.2 Data handling

Values for the continuous variables were log-transformed, centered and scaled and were not categorised in statistical analyses (except for the Kaplan-Meier curves in Figure 5.5). An asterisk (*) is used to indicate when a transformed version of a variable is being referred to specifically.

Variable	n	n with event (%)	Missing (%)
Sex	709	465 (65.6)	0 (0.0)
Age*	709	465 (65.6)	0 (0.0)
Albumin*	708	464 (65.5)	1 (0.1)
Bilirubin*	707	463 (65.5)	2 (0.3)
Child-Pugh Class	700	457 (65.3)	9 (1.3)
AFP*	688	452 (65.7)	21 (3.0)
Tumour stage	686	442 (64.4)	23 (3.2)
Tumour size*	635	410 (64.6)	74 (10.4)
DCP*	317	225 (71.0)	392 (55.3)
L3*	316	224 (70.9)	393 (55.4)
Complete	268	188 (70.1)	441 (62.2)

Table 5.2: Frequencies of patients with complete and missing data for each of ten clinical variables. More than half of patients have missing data for DCP* and L3* and approximately two thirds of patients have missing data for one or more of the ten clinical variables. Note that the denominators used for calculation of percentages in column 3 are in column 2, whilst the denominator used for column 4 is $N = 709$.

	Cox PH HR (CI)	Weibull PH HR (CI)	Weibull AFT AF (CI)	Semi-parametric AFT AF (CI)
Intercept[a]	-	-	-7.03 (-7.35, -6.72)	-6.76 (6.42, 7.11)
Sex (Male vs Female)	1.07 (0.71, 1.60)	1.09 (0.73, 1.63)	1.06 (0.81, 1.39)	1.14 (0.66, 1.17)
Age*	1.15 (0.95, 1.40)	1.17 (0.96, 1.41)	1.11 (0.98, 1.26)	1.06 (0.82, 1.08)
AFP*	1.36 (1.13, 1.62)	1.38 (1.16, 1.64)	1.24 (1.10, 1.39)	1.22 (0.74, 0.90)
L3*	1.32 (1.10, 1.59)	1.33 (1.11, 1.60)	1.21 (1.07, 1.37)	1.18 (0.75, 0.95)
DCP*	1.16 (0.93, 1.44)	1.14 (0.92, 1.42)	1.09 (0.94, 1.26)	1.15 (0.76, 1.00)
Bilirubin*	1.23 (1.00, 1.51)	1.22 (1.00, 1.49)	1.14 (1.00, 1.31)	1.07 (0.80, 1.08)
Albumin*	0.75 (0.61, 0.93)	0.74 (0.60, 0.90)	0.82 (0.71, 0.94)	0.76 (1.14, 1.53)
Child-Pugh (B vs A)	1.40 (0.90, 2.20)	1.43 (0.91, 2.24)	1.27 (0.94, 1.71)	1.27 (0.57, 1.10)
Child-Pugh (C vs A)	1.00 (0.25, 3.97)	0.94 (0.25, 3.59)	0.96 (0.39, 2.35)	1.59 (0.26, 1.55)
Cancer Stage (Late vs Early)	1.81 (1.22, 2.68)	1.84 (1.24, 2.73)	1.50 (1.16, 1.96)	1.57 (0.48, 0.84)
Tumour Size*	1.23 (1.04, 1.46)	1.24 (1.05, 1.47)	1.16 (1.03, 1.30)	1.10 (0.79, 1.04)

Table 5.3: Estimated HRs and AFs for three statistical models with all ten clinical variables included as main effects: 1) Cox model, 2) Weibull AFT model in proportional hazards and AFT form and 3) Semi-parametric AFT model. Estimated HRs for the Weibull model are similar to the Cox model and AFs are generally similar to the semi-parametric AFT model. *Variable is log-transformed, centered and scaled, [a]The intercept for the AFT models is presented on the log-scale.

5.10.3 Exploratory data analysis

Kaplan-Meier curves for the ten clinical variables are presented in Figure 5.5. For display purposes, values for the continuous variables have been categorised. Note that different numbers of patients are depicted in each subplot due to missing data. Visibly, all variables appear to be related to prognosis, except possibly Sex.

Table 5.3 shows the estimates obtained when including patients with complete data for all ten clinical variables in three survival models: a Cox model, a Weibull model in both PH and AFT formulations, and a semi-parametric AFT model. Values greater than one represent worse survival prospects for both HR and AFT estimates. For the transformed continuous variables, estimates represent the relative reduction in expected survival time (/increase in the hazard) for a one standard deviation increase in the transformed variable. Although the estimates are generally very similar between the statistical models, the Weibull PH/AFT model is used for inference. Increased Age*, AFP*, L3*, Bilirubin* and Tumour size* are statistically significantly associated with poorer survival prospects, along with decreased Albumin*. Patients with Late Cancer Stage are also statistically significantly associated with worse prognosis than those with Early Cancer Stage. Although generally the point estimates for the Child-Pugh Class C versus A are not larger than the point estimate for B versus A, as might be expected, the confidence interval is quite wide as a result of there being only 9 patients of 31 in Child-Pugh Class C with an event. Interestingly, this is not the case for the semi-parametric AFT model which is more in line with expectations.

5.10.4 Statistical methods

Gower's coefficient was used to obtain a symmetric matrix of dissimilarities based on the ten clinical variables. Categorical variables (Sex, Cancer Stage and Child-Pugh Class) were down-weighted to half the weight of other variables as categorical variables tend to dominate the MDS configuration with Gower's coefficient. This weighting was found, in a preliminary analysis, to be a reasonable compromise between maintaining distinct clusters as well as some spread of points within the clusters so as not to completely obscure the continuous variables. Child-Pugh Class was treated as a nominal, rather than an ordinal or continuous variable when calculating dissimilarities, and two-level categorical variables were treated as symmetric (see Section 4.4). Metric MDS was used to obtain ten MDS fits, corresponding to $1, 2, \ldots, 10$ dimensional MDS solutions.

A Cox model, a Weibull model in PH and AFT form, as well as a semi-parametric AFT model were fitted including all clinical variables simultaneously as main effects to assess the covariate adjusted relationship between each variable and survival. The same model types were used to model the relationship between overall survival and the MDS dimensions; of these models the Weibull AFT model was used to obtain

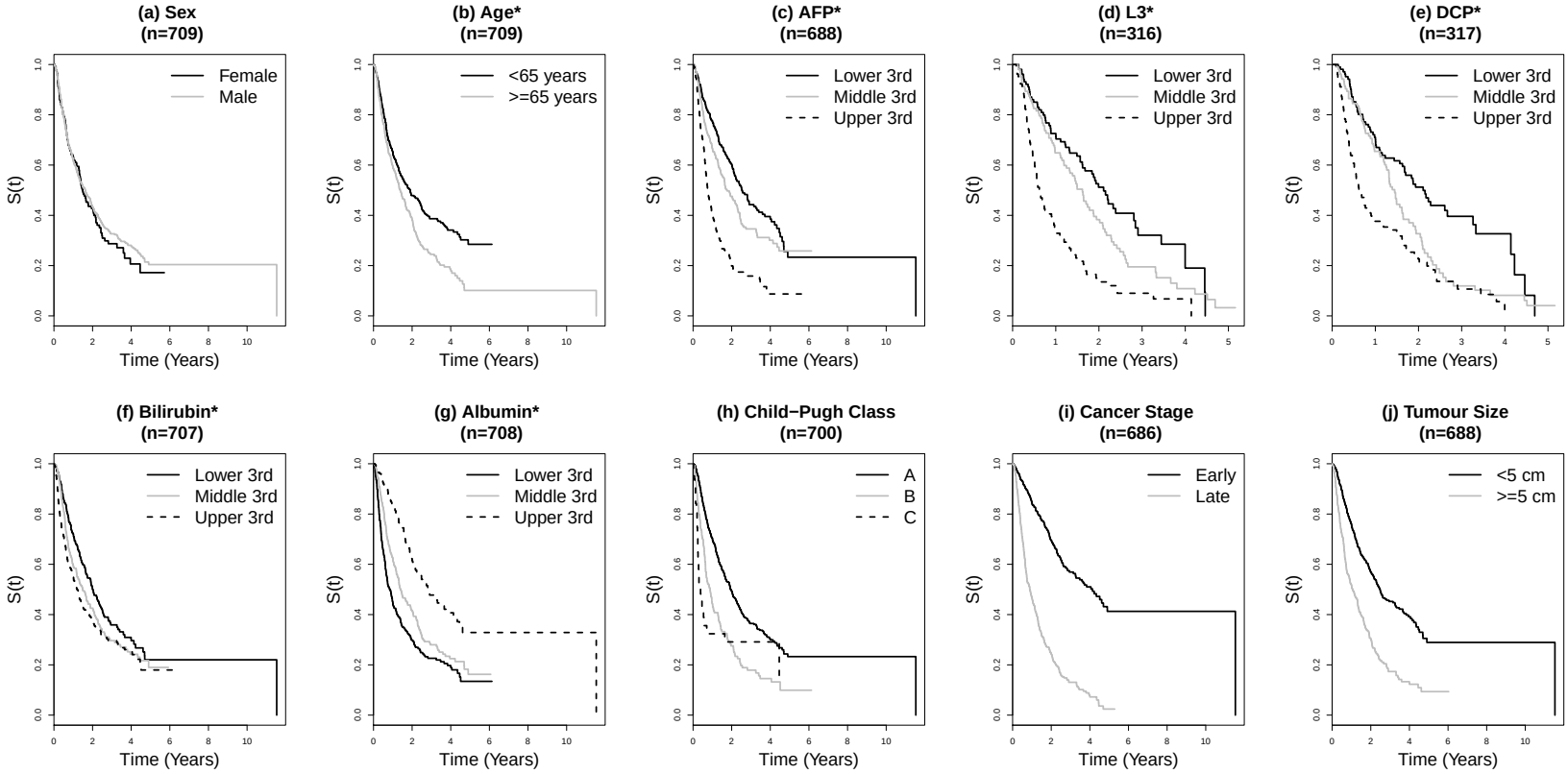

Figure 5.5: Kaplan-Meier plots for the ten clinical variables. Continuous variables have been categorised for the purpose of these plots. The number of observations available for each plot due to missingness is indicated in parentheses. Each plot is suggestive of a prognostic effect, perhaps with the exception of Sex.

a time-to-event biplot axis. Confidence limits for the slope of the time-to-event axis were obtained using a non-parametric bootstrap with 1000 resamples (see Section 5.7). All confidence intervals are 95% confidence intervals and p-values less than 0.05 were considered to be statistically significant.

To obtain biplot vectors/axes for the continuous clinical variables, linear regression was used. For Sex and Cancer Stage, Firth logistic regression (Firth, 1993) was used to account for possible complete separation of the categories. For Child-Pugh class, standard multinomial logistic regression was used. For further discussion on biplot axes for categorical variables, see Section 5.11.

5.10.5 Software

All statistical analyses were conducted using `R` software (version 3.5.2, R Core Team, 2017). The `SMACOF` package (de Leeuw and Mair, 2009), which uses iterative majorization to minimise the stress of the configuration, was used for metric MDS. Dissimilarities were obtained using the `daisy()` function in the `cluster` package (Maechler et al., 2019). Cox and parametric time-to-event models were fitted using the `survival` package (Therneau, 2015), and semi-parametric AFT models were fitted using the generalised estimating equation approach in `aftgee` (Chiou et al., 2014). Firth logistic regression models were fitted using the `logistf` package (Heinze, 2018).

5.10.6 Model fit

Figure 5.6(a) depicts a scree plot of the Stress for MDS solutions with 1 to 10 dimensions. The scree plot suggests that a three or four dimensional solution may suitably represent this data set. However, for illustration purposes the analyses here are restricted to the two-dimensional MDS solution and the suitability of this representation is considered in detail.

Figure 5.6(b) shows the explained variation in overall survival and the concordance of each MDS solution with the survival time. A two-dimensional MDS solution provides a reasonable concordance of 0.73, but the percentage of explained variation is low at 38%. However, neither the concordance nor R^2 improve much as the number of dimensions increases. Note that a Weibull model fitted with all ten clinical variables gave a concordance of 0.75 and an R^2 of 0.40. Although models fitted to the MDS configuration and to the predictor variables cannot be compared directly, as due to missing data they are based on different patient subsets, these statistics demonstrate that the combined predictive power of the variables used in this analysis is not particularly high.

Figures 5.6(c) and (d) show the marginal Kaplan-Meier estimates of the survival and cumulative hazard function respectively. The median survival time from diagnosis for these HCC patients is approximately 1.3 years. The grey fitted curves correspond

to estimates from an unconditional Weibull model. Whilst a Weibull representation of these marginal plots is clearly not perfect, the fitted survival curve is generally consistent with the confidence limits of the Kaplan-Meier estimate. The consistency of the parameter estimates from the Weibull model with both the Cox model and the semi-parametric AFT model (Tables 5.3 and 5.5) provides reassurance that the assumption of a Weibull distributed survival time does not substantially affect the results. Cox-Snell and deviance residuals for the Weibull AFT model indicate that a Weibull distribution is, in general, a reasonably good fit to the data, but that risk may be overestimated for patients with high values of the linear predictor, see Figure 5.7(a) and (c). Results were very similar however even if a Cox model was used and/or if the the values for the clinical variables ($\mathbf{x}$) were used instead of the MDS coordinates ($\mathbf{z}$) (Figure 5.7). The estimated scale parameter from the full Weibull AFT model was $\hat{\phi} = 0.67$ (CI 0.60, 0.75), with the confidence interval excluding 1 implying that a scale parameter is required and that therefore a Weibull AFT model is preferable to an exponential model.

Figure 5.9 illustrates how well the clinical variables are represented by the ten MDS solutions. For all variables the concordance and Nagelkerke's pseudo-R^2 are presented. Additionally, the R^2 is presented for continuous variables. Child-Pugh Class and Cancer Stage are very well represented with only two dimensions, whereas three or four dimensions appear necessary to adequately represent Sex. The concordance for all of the continuous variables is reasonable in two dimensions (>0.6) although clearly four dimensions or more would improve their representation considerably.

The estimated acceleration factors from a Weibull AFT model using the values for ten clinical variables directly (i.e. using $\mathbf{x}$) and those recovered from the two-dimensional biplot (i.e. using $\mathbf{z}$) are shown in Table 5.4. Whilst the signs of the effects from the biplot are correct (log-scale and barring Child-Pugh C vs A), the magnitudes of the acceleration factors are overestimated considerably.

5.10.7 Results

Figure 5.8 depicts the two-dimensional MDS configuration with biplot vectors for the clinical variables and overall survival, (a) and with an overlaid time-to-event biplot axis, (b). The relationship between the clinical variables and time-to-event axis are displayed further in Figure 5.11. In Figure 5.8(b), the time-to-event axis is near to the horizontal, implying that Dimension 1 is strongly related to survival times; the time-to-event biplot axis scale implies that patients to the far right of the plot are expected to survive less than 3 months, whilst those to the far left of the plot are expected to survive beyond 10 years.

In Figure 5.8(a), the relationships between the clinical variables and survival time are shown and are generally in agreement with the estimates in Table 5.3. High values of

the HCC characteristics (HCC biomarkers and Tumour Size) clearly imply an expected decreased survival time. The vectors for Albumin*, Bilirubin* and Child-Pugh Class all lie very close to the same axis which clearly represents liver function: roughly speaking, patients at the bottom of the plot would be expected to have better liver function than those at the top of the plot. Notably, and sensibly, better liver function is correlated with improved survival prospects.

The vector for Age* is approximately perpendicular to the time-to-event axis, implying that Age* is not related to survival time. However, the Kaplan-Meier curve for Age, Figure 5.5(b), and model estimates, Table 5.3, imply that increased Age is associated with poorer prognosis. This discrepancy could be explained by the fact that Age* is not well-represented in this two-dimensional configuration, as shown in Figure 5.9(b).

The categorical clinical variables are displayed in Figure 5.10. Examining the three subplots, it can be seen that there are four distinct clusters that represent combinations of Cancer Stage (Early versus Late) and Child-Pugh Class (A versus B/C, since B and C are not found in distinct clusters in these plots). Patients with missing data for Cancer Stage and/or Child-Pugh Class (grey points) are generally situated between these clusters.

From Figure 5.10(a), it can be seen that Late Cancer Stage is clearly associated with poorer prognosis with near complete separation between these two categories. For Figure 5.10(b), the two biplot vectors (B versus A, C versus A) which represent Child-Pugh Class were found to be almost identical and only one is presented. The vector arrowhead for Child-Pugh class is well outside the plot region; the implication is that a one-point increase in Dimension 1 and Dimension 2 produces an extreme increase in the log-odds of increasing the Child-Pugh Class from A.

In Figure 5.10(c), it can be seen that Sex is poorly represented by the plot, with no clusters corresponding to males and females. Drawing a linear axis through the plot would not discriminate between sexes and this is reflected in the short biplot vector.

5.11 Discussion

In this chapter, it was demonstrated how AFT models can be used to fit time-to-event biplot axes with a measurement scale. To our knowledge, biplot axes for time-to-event data have not been considered previously. It was also shown how, in principle, a biplot axis for a predictor variable can be related to a time-to-event biplot axis in order to recover acceleration factors or hazard ratios directly from a biplot. The utility of MDS biplots with a time-to-event axis was demonstrated using both simulated data and a HCC data set. For the HCC analysis, a two-dimensional MDS solution was found to be an informative representation of the relationship between observations, ten clinical variables of mixed type and overall survival from diagnosis.

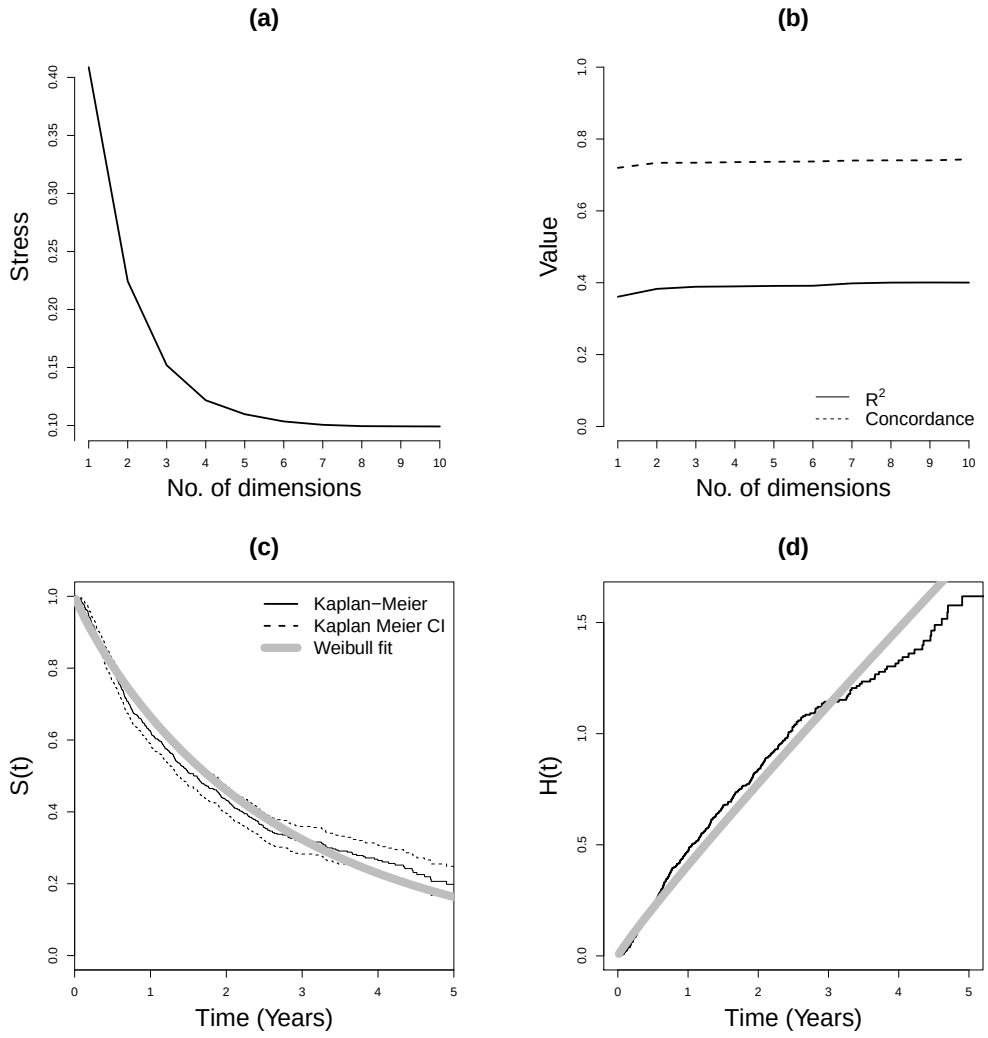

Figure 5.6: (a) Scree plot indicating the Stress of the MDS configuration as a representation of ten clinical variables for $1, 2, \ldots, 10$ dimensional metric MDS solutions. (b) R^2 and concordance for overall survival as a function of the MDS coordinates from Weibull AFT models. (c) Kaplan-Meier estimate and confidence interval for the marginal survival curve for the HCC patients, overlaid with a fitted survival curve from an unconditional Weibull AFT model. (d) Nelson-Aalen estimated marginal cumulative hazard function and overlaid fitted cumulative hazard function from an unconditional Weibull AFT model.

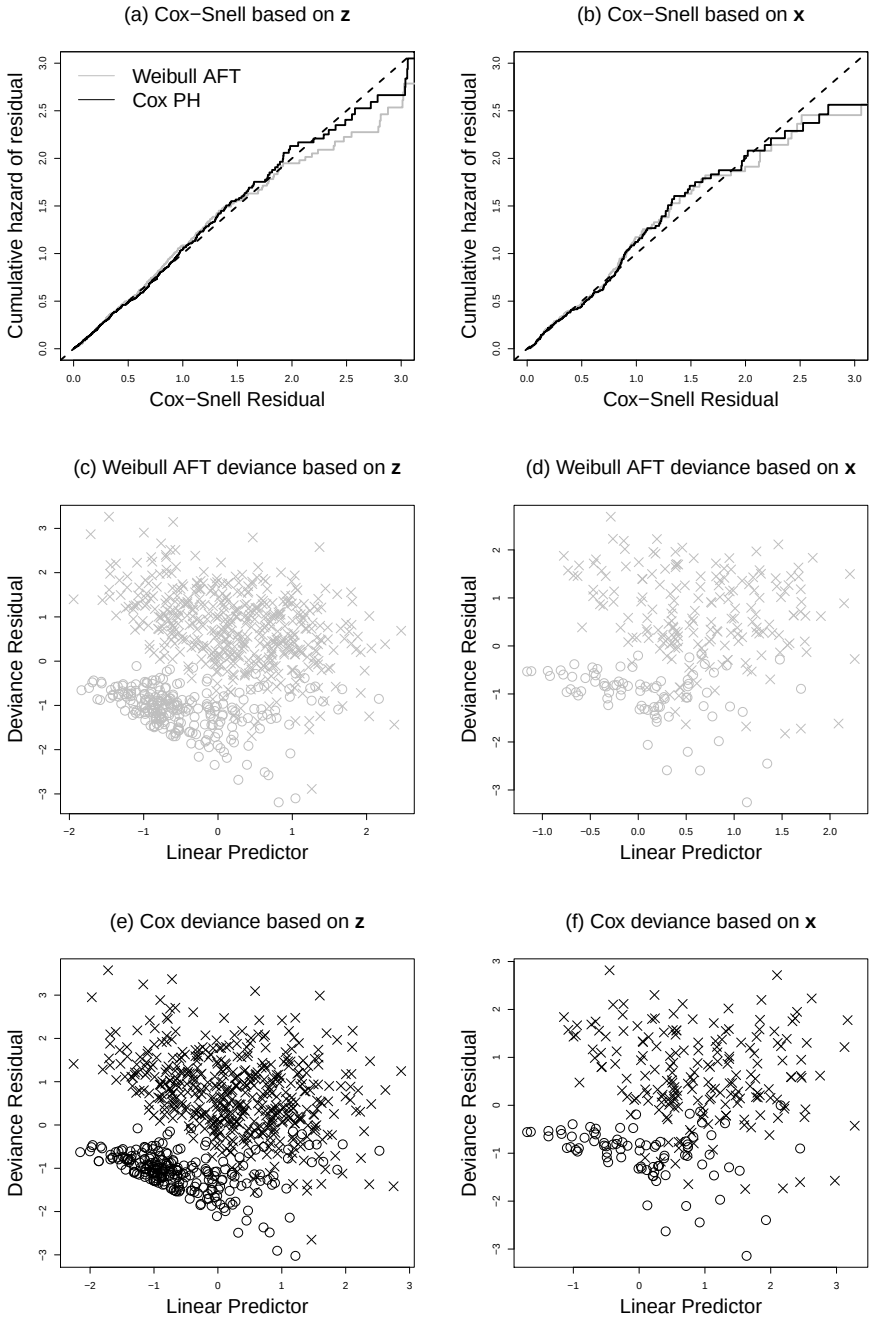

Figure 5.7: Cox-Snell and deviance residuals plots for Weibull AFT and Cox models fitted to the HCC data set. In the left column, MDS coordinates ($\mathbf{z}$) have been used as predictor variables. In the right column, values for the ten clinical variables ($\mathbf{x}$) have been used. Note that 709 observations (465 events) are in the MDS-based analyses compared with only 268 (188 events) in the analyses using the ten clinical variables directly, due to missingness.

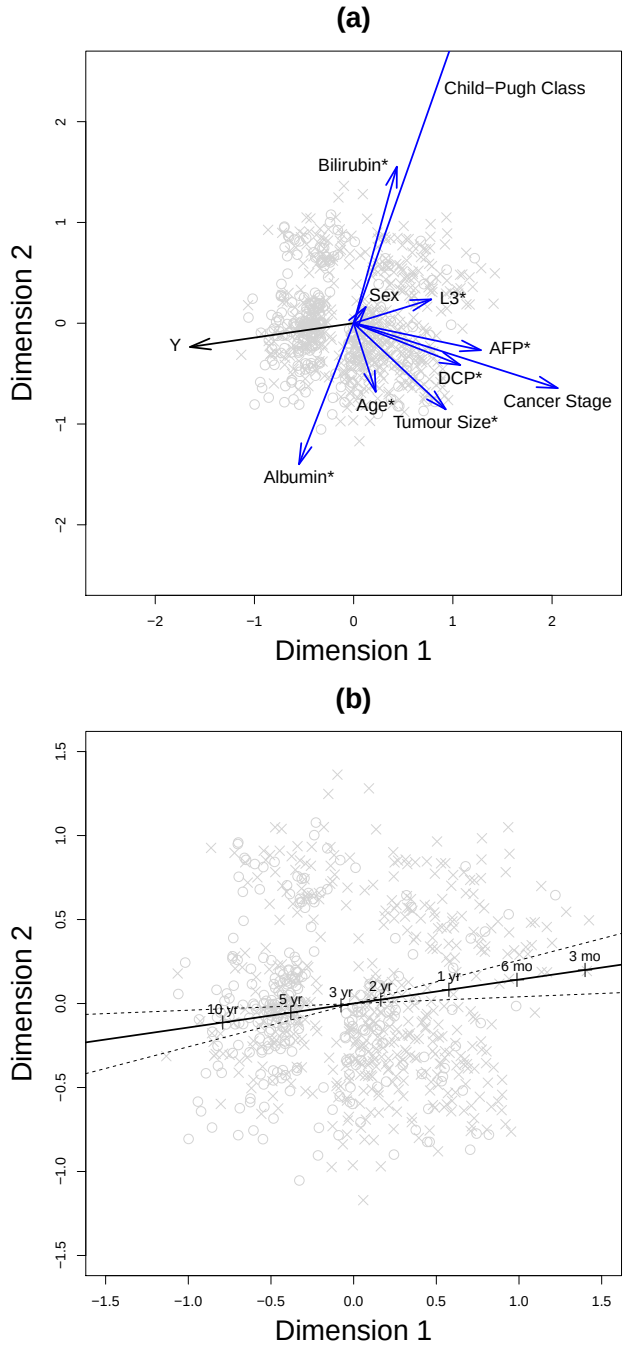

Figure 5.8: The two-dimensional metric MDS solution for the ten clinical variables. Censored and uncensored observations are presented as circles and crosses, respectively. (a) Biplot vectors for overall survival (denoted as Y) and the ten clinical variables. The Bilirubin/Albumin/Child-Pugh-Class 'axis' represents liver function, roughly corresponding to poor liver function at the top of the plot. HCC biomarker vectors point in the opposite direction to the overall survival axis, indicating that large values indicate worse prognosis. (b) Overlaid time-to-event biplot axis and confidence interval obtained using a non-parametric bootstrap with 1000 resamples.

	Weibull AFT	
	AF (using $\mathbf{x}$)	AF (using $\mathbf{z}$)
Sex (Male vs Female)	1.059	3.268
Age*	1.108	1.337
AFP*	1.238	4.793
L3*	1.211	5.193
DCP*	1.092	4.282
Bilirubin*	1.143	1.959
Albumin*	0.815	0.439
Child-Pugh (B vs A)	1.269	2.175
Child-Pugh (C vs A)	0.963	2.118
Cancer Stage (Late vs Early)	1.505	4.500
Tumour Size*	1.156	2.870

Table 5.4: AFs obtained from a Weibull AFT model using the observed values for the ten clinical variables (i.e. using $\mathbf{x}$) compared with those recovered from associating biplot scales in the two-dimensional biplot in Figure 5.8 (i.e. using $\mathbf{z}$). Whilst the signs of the estimates from the biplot are correct on the log-scale, barring Child-Pugh (C vs A), the magnitude of the AFs are overestimated considerably. *Variable is log-transformed, centered and scaled.

	Cox PH Est. (CI)	Weibull PH Est. (CI)	Weibull AFT Est. (CI)	Semi-parametric AFT Est. (CI)
Intercept	-	-	6.75 (6.83, 6.67)	6.29 (6.20, 6.38)
Dimension 1	1.91 (1.69, 2.12)	1.98 (1.79, 2.17)	-1.65 (-1.81, -1.49)	-1.48 (-1.64, -1.33)
Dimension 2	0.31 (0.12, 0.51)	0.28 (0.09, 0.48)	-0.24 (-0.40, -0.07)	-0.27 (-0.46, -0.08)

Table 5.5: Estimates (log-scale) for the association between overall survival and the dimensions from the two-dimensional MDS configuration. The parameter estimates are larger for Dimension 1 than Dimension 2, implying that the overall survival time is more closely associated to Dimension 1. Parameter estimates are similar within the PH and AFT model types.

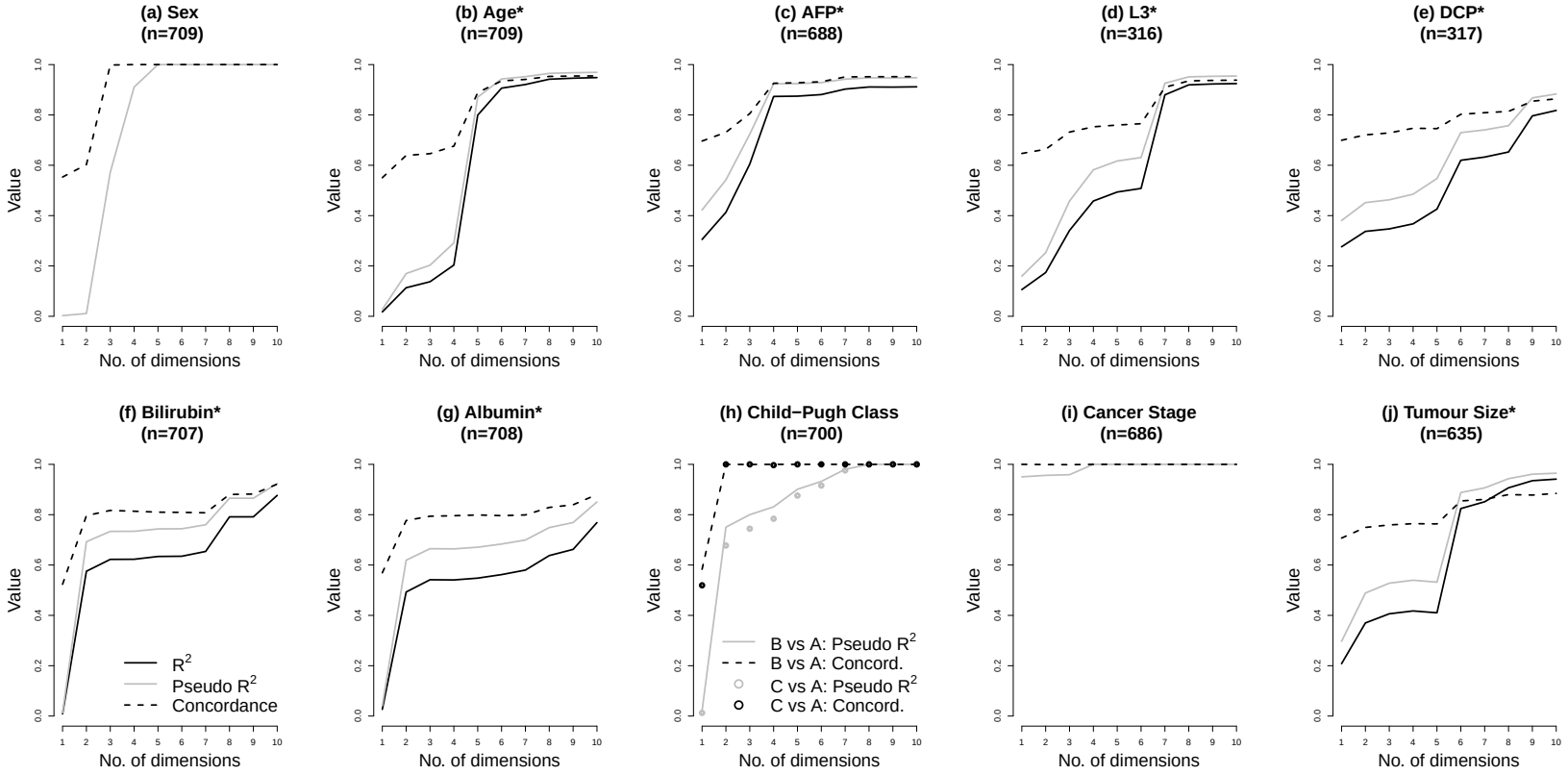

Figure 5.9: Fit statistics for the ten clinical variables as a function of the MDS configurations. More dimensions corresponds to a better representation of the variables. Linear regression was used for continuous variables, Firth logistic regression was used for two-category nominal variables (Sex and Cancer Stage) and multinomial logistic regression was used for Child-Pugh Class.

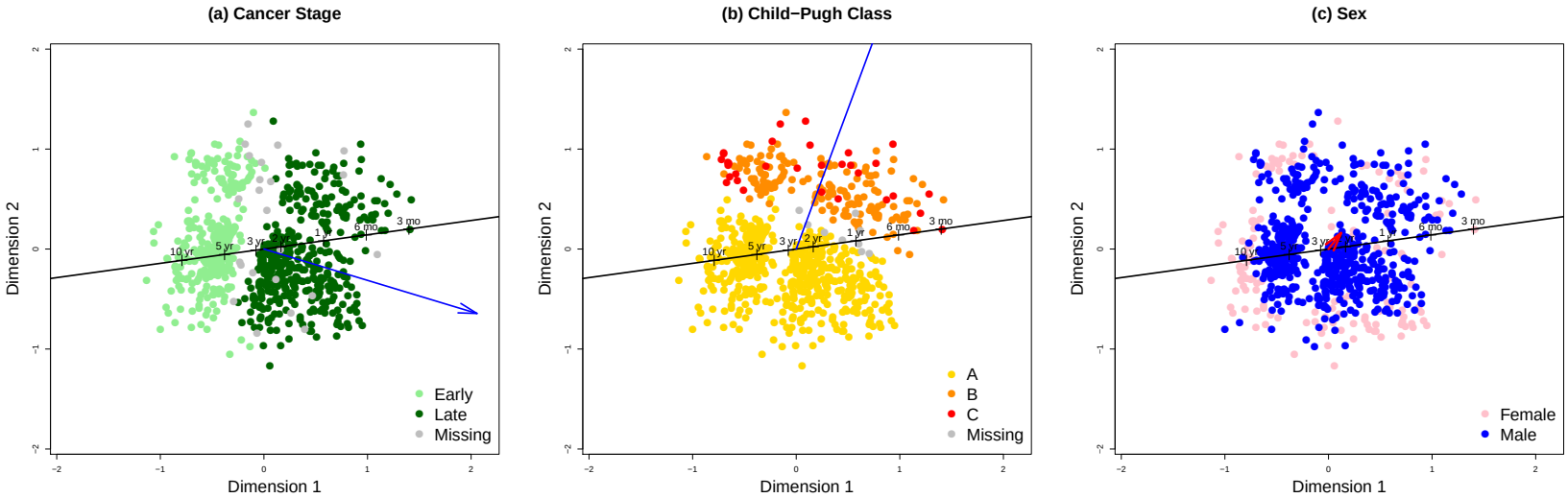

Figure 5.10: Two-dimensional MDS configuration with points coloured according to the categories of the categorical variables, with overlaid time-to-event axis and biplot vector for the clinical variable of interest. (a) Cancer Stage categories form almost completely distinct clusters along the direction of the biplot vector. (b) Child-Pugh Class A cluster is distinct from the B and C cluster in the direction of the biplot vector, but there is considerable overlap between Child-Pugh Class B and C. (c) Sex is not well-represented by the plot and a linear axis does not discriminate between sexes, which is reflected in the lack of clusters corresponding to males and females and the short biplot vector.

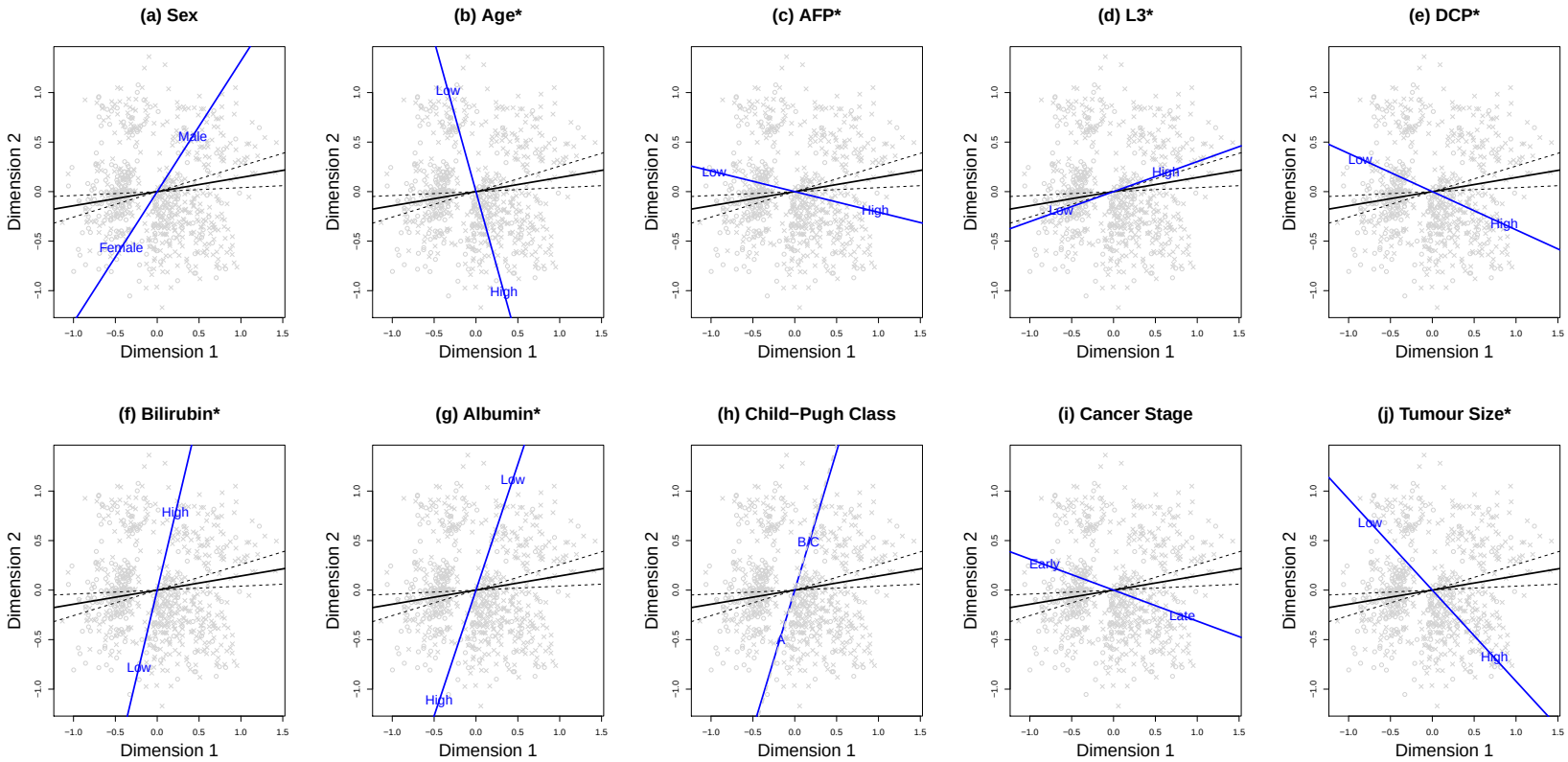

Figure 5.11: Biplots for the two-dimensional MDS solution for the ten clinical variables to illustrate the relationship with the time-to-event axis, its confidence interval, and each of the ten clinical variables. Most variables are correlated with survival, as expected, except Age which is not well-represented by the MDS solution. The biplot axis for Sex is plotted but should be interpreted with caution as Sex was found not to be well-represented by the MDS configuration.

The approach to fitting time-to-event biplot axes proposed in this chapter was based on the use of AFT models, and in particular the Weibull AFT model. If the assumptions of parametric AFT models are overly restrictive then a semi-parametric AFT model can be used. However, when fitting semi-parametric AFT models, it should be noted that the commonly used rank test statistic-based estimation routines do not estimate the intercept (see Wei, 1992; Chiou et al., 2014), which is necessary for marking out the scale of the time-to-event axis. An alternative is the generalised estimating equation approach described by (Chiou et al., 2014).

In this chapter, only linear biplot axes were fitted. Non-linear biplots, in which straight-line axes are replaced by 'trajectories', for displays in both Euclidean and non-Euclidean space, are discussed in Cox and Cox (2000, Chapter 7), and the references therein.

Categorical variable axes in the analysis of the HCC data were fitted using a generalised linear model approach, as described in Section 5.5.1 and in Greenacre (2010, Chapter 3). In the case of a nominal categorical variable, such as Sex, a logistic regression model is convenient. However, complete separation between categories can cause convergence problems and, as we have found, this scenario is likely when using Gower's general coefficient since categorical variables tend to dominate the MDS configuration. Categorical variables can be down-weighted to ameliorate this issue, and/or a penalised approach, such as Firth logistic regression can be used (Firth, 1993). As an alternative to the generalised linear model approach to fitting biplot axes, pseudo-samples can be used to identify centroids representing the different levels of a categorical variable (see Cox and Cox, 2000, Chapter 7).

Another difficulty with categorical variables is that there is no accepted equivalent to the coefficient of determination. Nagelkerke's R^2 was used here as a guide, but it does not have the same interpretation as the R^2 for linear models. Nagelkerke's R^2 is a measure of improvement from the null model rather than of explained variation. An informative introduction to this topic can be found in Royston (2006).

An interesting feature of the presented work is that it was shown how acceleration factors (and hazard ratios if a Weibull distribution is assumed) can be recovered directly from a biplot with a time-to-event axis, by associating biplot axis scales. The usefulness of this feature appears to be limited however. In the HCC example, whilst the signs of the estimates (log-scale) derived from the biplot generally matched those estimated directly from the clinical data, their magnitude was exaggerated considerably. There are several reasons for this. Firstly, two approximate scales are being associated, potentially introducing considerable error. Secondly, the cosine of the angle between axes is only an approximation of the correlation between the predictor variable and time-to-event axis and the angle between axes greatly impacts the projection from one to the other. Finally, as we have found in exploratory analyses (see Appendix C), whilst the approach

can work reasonably well when dissimilarities are obtained using Euclidean distances, estimates can be greatly overestimated when Gower's coefficient is used.

One important feature of MDS is that missing data can be incorporated easily, without the need for imputation or case-wise deletion of observations, by skipping variables with missing data in dissimilarity calculations (see e.g. Buja et al., 2008, see Section 6.8 for further discussion). This was clearly illustrated in the HCC example, where 709 observations (465 events) were available for MDS analysis compared with only 268 (188 events) with complete data. Whilst advantageous, one drawback for interpretation is that different patient subsets are analysed in supportive analyses, such as in fitting the biplot vectors or in multivariable regression using the original clinical variables. This aspect and missing data assumptions when using 'variable skipping' in MDS are discussed in more detail in Section 6.8.

When using Gower's coefficient, categorical variables tend to dominate the MDS solution. In the clinical example, categorical variables were given half the weight of continuous variables in order to ameliorate this issue. However, it may be possible to find a set of optimal variable weights in order to, for example, maximise the average concordance across variables for a fixed number of MDS dimensions.

A drawback of MDS biplots in general is that their usefulness is limited when more than two or three dimensions are necessary for a representative display of the data. In the analysis of the HCC data set it was shown how the biplot can be carefully checked to ensure that the visualisation is not misleading. If a two or three dimensional solution is not feasible, it may be necessary to remove variables.

In conclusion, biplots are a useful generalisation of the scatter plot that can be used to display the key features of a multi-dimensional data set. By adding a time-to-event axis to a biplot, expected event times can be read from the display and the relationships between observations, variables and a time-to-event outcome can be illustrated.

Chapter 6

Supervised distance-based regression

6.1 Introduction

In clinical research, multivariable regression models are used extensively to model and predict clinical outcomes, such as overall survival time, as a function of numerous predictor variables, or 'covariates'. Two key challenges for the statistical analyst when fitting such models are: 1) large numbers of candidate variables relative to the number of subjects/events, 2) missing covariate data. In the previous chapter, the use of MDS for visualisation of multivariate clinical data was considered. In this chapter, MDS is used as a dimension reduction tool for multivariable regression.

If the number of subjects/events is low relative to the number of covariates, parameter estimates can be biased, parameter variances can be over or under-estimated and confidence interval coverage can be poor (see e.g. Peduzzi et al., 1996). In fact, if the number of possible covariates is sufficiently large, a model can become non-estimable. Possible solutions are to reduce the number of covariates, using e.g. stepwise model selection, or reduce the dimensionality of the covariates in some way, using e.g. principal components regression (PCR). When the ratio of subjects/events to parameters is low, stepwise selection is known to result in selection bias, where coefficients for the selected variables are overestimated (see e.g. Steyerberg et al., 1999). Moreover, in clinical research, covariate data is typically of mixed type (categorical, continuous etc.) and for some dimensionality reduction methods, such as PCR, including mixed type data is not straightforward.

When there are missing data, standard statistical software will perform casewise deletion if just one covariate is missing for a subject. Not only is casewise deletion

statistically inefficient, but it is wasteful since data can be costly to collect, particularly in a clinical trial. Whilst the gold standard approach is to use multiple imputation (Rubin, 2004), it requires considered set-up and usually several strong assumptions are made, such as multivariate normality.

Distance-based regression (DBR, Cuadras, 1989) is a simple approach to multivariable regression which offers possible solutions to the aforementioned issues. Put simply, DBR is regression on latent dimensions derived from predictor variables using a dissimilarity measure. The latent dimensions are obtained by classical or metric multidimensional scaling (MDS) and the resulting MDS coordinates are then included as covariates in a regression model. Key advantages of DBR are that both missing data and mixed data types can be easily incorporated using Gower's coefficient (Gower, 1971, Section 4.4).

DBR was introduced by Cuadras (1989) and has been subject to a number of developments (see Cuadras and Arenas, 1990; Cuadras et al., 1996; Boj et al., 2007a; Esteve et al., 2009; Boj et al., 2010, 2012; Melo and Melo, 2013; Melo et al., 2015). Typically, DBR has been used in the context of linear regression, for which the theoretical underpinnings are well established (see Cuadras, 1989; Cuadras and Arenas, 1990), but also in generalised linear models (Boj et al., 2012), mixed models (Melo and Melo, 2013) and beta regression (Melo et al., 2015). A clinical application of conventional DBR in time-to-event analysis can be found in Fuller et al. (2002), in which survival times for patients with brain gliomas were modelled using a Cox model with three MDS dimensions based on gene expression profiles. Interestingly, DBR also appears to have arisen independently in the genomic literature (see e.g. Schaid, 2010).

Typically, the number of latent dimensions included in a DBR model are obtained using cross-validation, and can be numerous. Like PCR, and other 'unsupervised' methods, the main drawback is that the latent dimensions are not necessarily related to the outcome variable. It is plausible then that several latent dimensions may be dominated by variables which do not contribute to the predictive power of the model. When using Gower's coefficient, categorical variables tend to dominate the first few MDS dimensions (Moustaki, 1996), so that a prognostic continuous variable may not be well-represented unless many dimensions are included.

In this chapter, conventional DBR is introduced and an additional supervision step is proposed in order for predictive variables to be better represented in fewer latent dimensions, at the expense of including an additional tuning parameter. Two approaches are proposed. In the first, predictive variables are given a higher weighting in the dissimilarity measure. In the second, candidate variables are screened and only the most predictive variables are included in dissimilarity calculations. The second approach is analogous to that used with PCR previously (Bair and Tibshirani, 2004; Bair et al., 2006). A simple simulated example is used to demonstrate how supervised DBR

can outperform conventional DBR. The supervised approaches and conventional DBR are then compared in a time-to-event analysis using a subgroup from a hepatocellular carcinoma (HCC) data set.

The chapter is organised as follows: in Section 6.2, a linear predictor that is used throughout the chapter is introduced, alongside some corresponding notation. In Section 6.3, the DBR procedure is outlined. In Section 6.4, two different approaches to supervising DBR are proposed, with the intention of obtaining well-fitting DBR models with fewer latent dimensions than conventional DBR. In Section 6.5, K-fold cross-validation is introduced and in Section 6.6 the proposed methods are applied to a simulated data set in which the utility of the supervision step is demonstrated. In Section 6.7, various DBR models are applied to a subgroup from a HCC data set. Discussion is given in Section 6.8.

6.2 Covariate dimension reduced linear predictor

Indexing observations as $i = 1, \ldots, N$, let $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2})$ represent an $N \times P$ matrix of predictor variables. $\mathbf{X_1}$ contains so-called 'protected covariates' ($N \times P_1$) and $\mathbf{X_2}$ contains the 'unprotected covariates' ($N \times P_2$), which will be subjected to dimension reduction. If $P_1 = 0$ then all variables will be subjected to dimension reduction (the usual case in DBR). Let $\mathbf{Z}$ denote an $N \times S$ matrix which is a representation of $\mathbf{X_2}$, obtained using some dimension reduction method, such as MDS. The columns of $\mathbf{Z}$ are indexed by $s = 1, \ldots, S$ and $S \leq P_2$. The observed matrices are denoted as $\mathbf{x}$, $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{z}$. A linear predictor for the ith subject is then given by

$$\eta(\mathbf{x_1}_i, \mathbf{z}_i) = \boldsymbol{\beta}^\intercal \mathbf{x_1}_i + \boldsymbol{\gamma}^\intercal \mathbf{z}_i, \tag{6.1}$$

where $\boldsymbol{\beta}$ ($P_1 \times 1$) and $\boldsymbol{\gamma}$ ($S \times 1$) are vectors of coefficients for $\mathbf{x_1}$ and $\mathbf{z}$, respectively. This linear predictor could be used in any linear model e.g. linear regression, generalised linear model or Cox regression model, depending on the nature of the outcome variable. Bøvelstad et al. (2009) fitted Cox regression models with linear predictors of this form which they referred to as 'clinico-genomic' models, with clinical (low-dimensionality) and genomic (high-dimensionality) covariate vectors included in $\mathbf{x_1}$ and $\mathbf{x_2}$, respectively.

6.3 Distance-based regression

In DBR, a model of the form of equation 6.1 can be fitted using classical or metric MDS to obtain $\mathbf{z}$. The following steps are undertaken:

- Step 1. Using a preferred dissimilarity measure, e.g. Gower's coefficient, calculate

the $N \times N$ dissimilarity matrix, $\boldsymbol{\delta}$, using $\mathbf{x}_2$.

- Step 2. Specify the number of MDS dimensions required, S, and use classical or metric MDS to obtain an $N \times S$ MDS configuration, $\mathbf{z}$. The final choice of S is usually determined by cross-validation (Section 6.5).
- Step 3. Fit a regression model with linear predictor of the form of equation 6.1 to obtain estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$.

For prediction it is necessary to be able to add further observations to the analysis. For the N+1th observation, a predicted value for the outcome variable, $\hat{y}_{N+1}$, can be obtained by continuing the algorithm:

- Step 4. Estimate the position/coordinates of the N+1th observation on the MDS configuration, $\hat{\mathbf{z}}_{N+1}$ (see Section 4.2.1).
- Step 5. Calculate the predicted value, $\hat{y}_{N+1}$, by substituting the parameter and coordinate estimates into equation 6.1, i.e. using

$$\eta(\mathbf{x}_{\mathbf{1},N+1}, \hat{\mathbf{z}}_{N+1}) = \hat{\boldsymbol{\beta}}^{\mathsf{T}} \mathbf{x}_{\mathbf{1},N+1} + \hat{\boldsymbol{\gamma}}^{\mathsf{T}} \hat{\mathbf{z}}_{N+1}.$$

Steps 4 and 5 can then be repeated for all test observations. Cuadras (1989) and Cuadras and Arenas (1990) provide thorough detail on the theoretical properties of DBR, but three pertinent points are: 1) the DBR model fit is invariant to the orientation of the MDS configuration i.e. translation, scaling, rotation and reflection of the MDS configuration will not affect the overall DBR model fit, 2) the number of parameters to be estimated (largely determined by the number of latent dimensions) cannot exceed the number of observations/events, as is usual in (non-penalised) regression, and 3) if Euclidean distances and classical scaling are used then a DBR model can be equivalent to PCR.

6.4 Supervision step

The main shortfall of conventional DBR is that there is no guarantee that those variables in $\mathbf{x}_2$ which are related to the outcome variable will be well-represented, or at least not without possibly including many latent dimensions. Two possible approaches to overcome this issue are now described.

6.4.1 Variable weighting

In the first proposed supervised approach, variables more strongly related to the response are given a higher weighting in dissimilarity calculations. We refer to this

method as variable weighted DBR (WDBR). For example, the weighted Euclidean distance is given by

$$\delta_{ij} = \left[\sum_{p=1}^{P_2} \omega_p (x_{ip} - x_{jp})^2 \right]^{1/2},$$

for $i = 1, \ldots, N$, $j = 1, \ldots, N$ and where ω_p is the weight for the pth variable. The same principle can be applied to other dissimilarity measures, such as Gower's coefficient, as discussed in Section 4.4. There are many possible weighting schemes that could be used for WDBR. We propose to obtain the weight using

$$\omega_p = \frac{a_p^\lambda}{\sum_{q=1}^{P_2} a_q^\lambda} \times P_2, \tag{6.2}$$

for $p = 1 \ldots, P_2$, where λ is a tuning parameter, $\lambda \geq 0$, and each $a_p \geq 0$. The denominator guarantees that $\sum_{p=1}^{P_2} \omega_p = P_2$. One way to obtain each a_p would be to fit $p = 1, \ldots, P_2$ models with linear predictor

$$\eta(\mathbf{x_1}_i, \mathbf{x_2}_{ip}) = \hat{\boldsymbol{\beta}}^\mathsf{T} \mathbf{x_1}_i + \hat{\alpha}_p \mathbf{x_2}_{ip}, \tag{6.3}$$

where $\mathbf{x_2}_{ip}$ denotes the pth variable from the vector $\mathbf{x_2}$, for the ith subject, and let

$$a_p = \left| \frac{\hat{\alpha}_p}{\text{se}(\hat{\alpha}_p)} \right|,$$

where $|.|$ denotes the modulus and se(.) is the standard error. A value of $\lambda = 0$ corresponds to no weighting, i.e. conventional DBR, $\lambda = 1$ weights the variables according to their standardised coefficients, $\lambda > 1$ accentuates the weights according to their standardised coefficients (small standardised coefficients would correspond to weights shrinking towards zero), and $\lambda < 1$ would give weights nearer to 1 than the standardised coefficient. As with the number of latent dimensions, S, a suitable value for λ can be obtained using cross-validation (Section 6.5).

6.4.2 Variable screening

Supervised PCR was introduced by Bair and Tibshirani (2004) and is where variables are screened so that only those most strongly related to the response are included in a PCR model. This approach is also compatible with DBR and we refer to it as variable screened DBR (ScDBR). In this approach, models with a linear predictor of the form of equation 6.3 are fitted, and only those variables for which the absolute standardised

estimates exceed a tuning parameter, θ ($\theta \geq 0$), are retained, i.e. for which

$$\left| \frac{\hat{\alpha}_p}{\text{se}(\hat{\alpha}_p)} \right| \geq \theta. \tag{6.4}$$

A value of $\theta = 0$ corresponds to conventional DBR and larger values for θ correspond to a higher benchmark and fewer variables being included. As with S and λ, θ is a tuning parameter that can be estimated by cross-validation (Section 6.5).

6.5 K-fold cross-validation

Conventional DBR includes a tuning parameter, S, the number of latent dimensions required. The proposed supervised approaches include a second tuning parameter, λ or θ, which, as will be demonstrated, can be used to reduce the total number of dimensions, and hence parameters, required in a DBR model. One way of choosing suitable values for tuning parameters is K-fold cross-validation (Hastie et al., 2003). In K-fold cross-validation, the data set is randomly partitioned into K subsets of equal size, $k = 1, \ldots, K$. The kth subset is held back (the 'validation' subset) and the model is fitted to the remaining $K - 1$ subsets and then applied to the validation subset (Figure 6.1). This process is repeated for each of the $k = 1, \ldots, K$ folds and the prediction error averaged. Commonly, values of 5 or 10 are chosen for K (see Hastie et al., 2003, page 242 for further discussion).

In order to find the optimal values for the tuning parameter(s), the cross-validation process is repeated for each row of a prespecified grid of tuning parameter values. The tuning parameter values with the lowest average prediction error across the validation subsets are selected. There is an important pitfall to be avoided here with supervised DBR. The estimates, $\hat{\alpha}_p$ ($p = 1, \ldots, P_2$), need to be estimated separately for each fold, i.e. observations in a validation subset cannot be used when training the model (see Hastie et al., 2003, page 245).

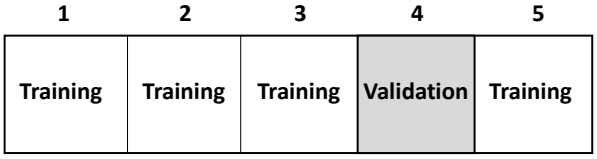

Figure 6.1: Depiction of five-fold cross-validation. A regression model is trained on all folds except the validation fold. The model is then tested on the validation fold and the prediction error calculated. This process is repeated, setting each fold as the validation fold in turn, and prediction errors are averaged over validation folds.

6.6 A simulated example

A simulated example is now used to illustrate the behaviour and performance of the proposed supervised approaches compared with conventional DBR. $P = 200$ variables, for $N = 100$ observations, were simulated from a multivariate normal distribution for which all variables had a true mean of 0 and a standard deviation of 1. A representation of the true correlation matrix for the simulated data is depicted in Figure 6.2; the two blocks correspond to variables $1, \ldots, 10$ and $11, \ldots, 20$, respectively, which are correlated within blocks. A continuous outcome variable was simulated as $\mathbf{y} = \mathbf{x}^\intercal\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is a standard normal variable with a mean of 0 and a standard deviation of 5, and $\boldsymbol{\beta}$ is a vector of true coefficients of length P: $\beta_1, \ldots, \beta_{20} = 3$ and $\beta_{21}, \ldots, \beta_{200} = 0$. An independent test data set was also simulated in the same way. No variables were 'protected' and so $P_2 = P = 200$, and $P_1 = 0$.

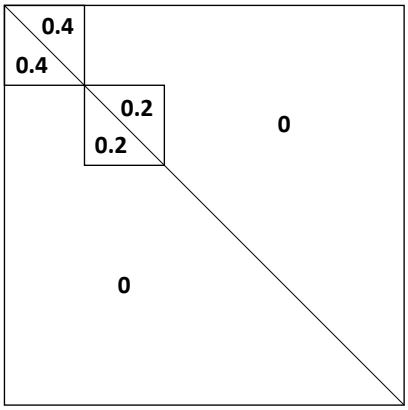

Figure 6.2: A representation of the correlation matrix used to generate multivariate normal data in the simulated example. The two blocks represent variables $1, \ldots, 10$ and $11, \ldots, 20$, respectively. These blocks of variables are correlated within block, with respective correlation coefficients of 0.4 and 0.2. Otherwise, variables were simulated as uncorrelated.

6.6.1 Statistical methods

DBR, WDBR models and ScDBR models were fitted by least squares. Five-fold cross-validation was used to estimate the optimal tuning parameters; the same folds were used for all models. For all models, the number of dimensions permitted was $S = 1, \ldots, 80$. Tuning grid values for λ were 0 to 15 in steps of 0.25, and for θ were 0 to 7.5 in steps of 1.25. Note that WDBR with $\lambda = 0$ and ScDBR with $\theta = 0$ correspond to conventional DBR. Classical scaling was used to find $\mathbf{z}$ and dissimilarities were calculated using the Euclidean (or weighted Euclidean) distance. In addition to the cross-validation process described in Section 6.5, final trained models were also applied to the independent test

data set. Gower's add-a-point method (Section 4.2.1) was used to add test observations. R^2, root mean square error (RMSE) and mean absolute error (MAE) were calculated for the training, validation and test data sets.

6.6.2 Results

The optimal results, i.e. tuning parameters and values that correspond to the lowest mean RMSE across the validation folds, are presented in Table 6.1. DBR selected the most latent dimensions, 63, and was the worst fit to the validation data, by a considerable margin. WDBR provided the best fit, although did so with only slightly fewer latent dimensions than DBR, 56. ScDBR provided a considerably better fit than DBR, and did so using only 14 latent dimensions. For a more complete picture, however, it is necessary to consider the curves in Figures 6.3 and Figures 6.4.

Model	$\hat{S}$	RMSE	R^2	MAE
DBR	63	8.11	0.77	6.23
WDBR ($\hat{\lambda} = 1.25$)	56	6.20	0.86	4.98
ScDBR ($\hat{\theta} = 2.375$)	14	6.39	0.82	5.19

Table 6.1: Comparison of DBR, WDBR and ScDBR with optimal tuning parameter values selected using five-fold cross-validation. RMSE, R^2 and MAE are average values over the validation folds which, by definition, were excluded from model training.

In Figure 6.3, WDBR results are presented for the full sequence of latent dimensions with λ held fixed at: 0 (conventional DBR, black line), the optimal value ($\hat{\lambda} = 1.25$, blue line) and at $\lambda = 10$ (grey line). Note that, strictly speaking, the 'optimal' $\hat{\lambda}$ is only optimal at $\hat{S} = 56$ (Table 6.1). The dashed black lines are the values for the true data-generating model and are a benchmark to identify over and under-fitting. Firstly, notice that as the number of dimensions increases, the fit to the training data improves for all models, as expected, Figure 6.3(a) and (d). All curves exceed the true line at some point, implying that overfitting might be expected if too many latent dimensions are used with WDBR. WDBR with optimal $\hat{\lambda}$ improves upon DBR for any number of latent dimensions. The grey line, WDBR with $\lambda = 10$, demonstrates that a fit nearly as good as optimal WDBR can actually be obtained in very few (approximately 10-15) dimensions, see Figure 6.3(b), (c), (d) and (e). Whilst the fit appears to deteriorate when λ is high and there are too many latent dimensions, in practice a cross-validated S would be low with such a high λ.

In Figure 6.4, ScDBR results are presented for the full sequence of latent dimensions with θ held fixed at: 0 (conventional DBR, black line), the optimal value ($\hat{\theta} = 2.375$, blue line) and at $\theta = 3$ (grey line). Again, strictly speaking, the 'optimal' $\hat{\theta}$ is only optimal at $\hat{S} = 14$ (Table 6.1). Clearly, ScDBR outperforms DBR for any number of latent dimensions and attains its best fit with very few latent dimensions. There is no

improvement in fit beyond approximately 15 dimensions. Overfitting to the training data also appears to be less pronounced than WDBR, Figures 6.3 and 6.4, panels (a) and (d).

A ridge regression model (Hoerl and Kennard, 1970), a linear model in which parameter estimates are penalised and shrunk towards zero was also fitted to these data (Figures 6.3 and 6.4, red dashed line). The fit was similar between ridge regression and optimal WDBR and ScDBR for the validation data, Figures 6.3 and 6.4, panels (b) and (e). Ridge regression performed better than WDBR and ScDBR on the independent test data set, Figures 6.3 and 6.4, panels (c) and (f). Ridge regression was chosen as it was the best performing method in two simulation studies previously (Bøvelstad et al., 2007, 2009).

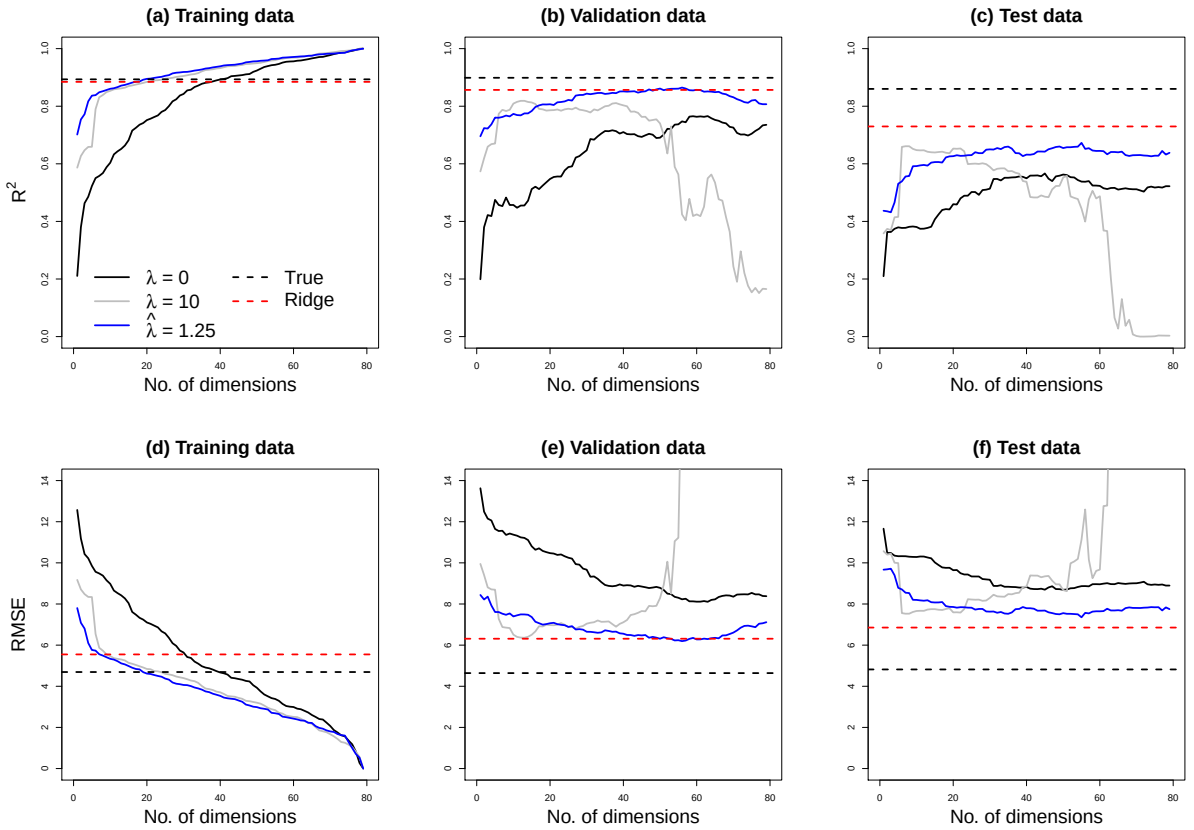

Figure 6.3: Results for WDBR models with increasing numbers of latent dimensions and various values for the tuning parameter, λ, fitted to training, validation and test data sets. $\lambda = 0$ corresponds to conventional DBR (black solid line). $\hat{\lambda}$ is the optimal value selected by five-fold cross-validation (blue solid line). The true data-generating model (black dashed line) and a ridge regression model (red dashed line) are also depicted. WDBR outperforms standard DBR whilst using fewer latent dimensions.

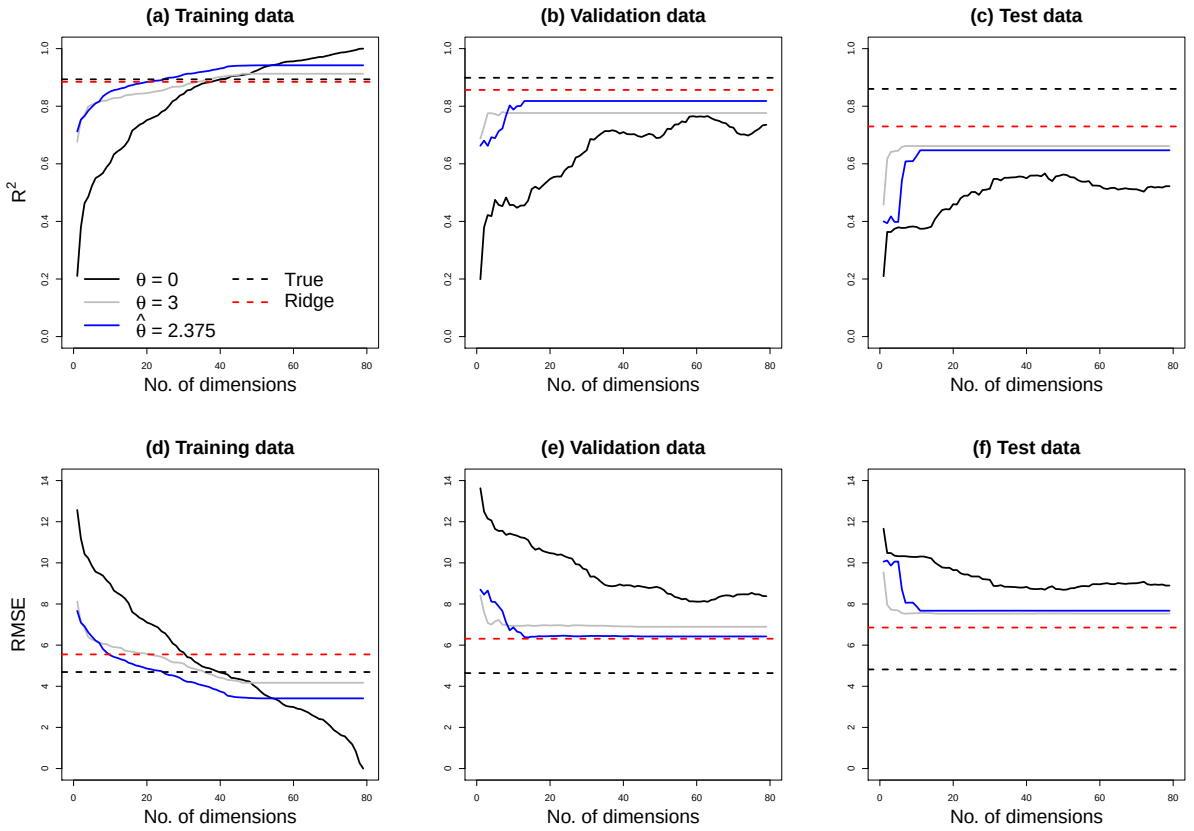

Figure 6.4: Results for ScDBR models with increasing numbers of latent dimensions and various values for the tuning parameter, θ, fitted to training, validation and test data sets. $\theta = 0$ corresponds to conventional DBR (black solid line). $\hat{\theta}$ is the optimal value selected by five-fold cross-validation (blue solid line). The true data-generating model (black dashed line) and a ridge regression model (red dashed line) are also depicted. ScDBR outperforms standard DBR whilst using fewer latent dimensions.

6.7 Subgroup analysis of the hepatocellular carcinoma data set

A subgroup of patients from the HCC data set are now analysed. The subgroup selected are 149 palliative HCC patients with primary aetiology recorded as 'alcoholic' (taken from groups I and IV, see Section 1.6.1). This subgroup was then further split into 120 patients in the training/validation data set and 29 patients in the test data set. The patients in the test data set are part of a cohort which were recruited specifically for internal validation of a diagnostic model in the original study (Johnson et al., 2014). There are 96 (80%) and 16 (55%) recorded deaths ('events') in the training/validation and test data sets respectively

This subgroup, like the full HCC set, contains considerable amounts of missing values (see Section 6.7.2). The purpose of the analysis is to fit and compare distance-based overall survival (Cox) regression models to *all* patients in the training/validation data set, and evaluate their usefulness in prognosis using the test data set.

6.7.1 Clinical variables

Fox et al. (2014) used a version of the HCC data set to model the survival prognosis of HCC patients using three biomarkers: AFP, L3 and DCP, as well as Bilirubin and Albumin. These variables were discussed previously in Sections 1.6.1 and 5.10.1. Other continuous variables in this data set are: Age, Aspartate aminotransferase (AST), Alanine aminotransferase (ALT), International normalised ratio (INR) and Creatinine. AST and ALT are enzymes found in the liver which leak into the bloodstream when liver cells are damaged, INR is a test of how quickly blood clots (higher INR corresponds to slower clotting), and creatinine is a biological waste product with low values being indicative of poor liver function.

A further 30 categorical variables are also available for analysis and are listed, along with their categories/levels, in Table 6.3. Of note are: World Health Organisation (WHO) Performance Status (ranging from normal physical function to completely disabled), Ascites (the accumulation of protein-containing fluid within the abdomen), T-stage (size of the primary tumour), N-stage (lymph node involvement) and M-stage (metastases).

6.7.2 Data handling

Only 37 of 120 patients (31%) in the training/validation data set have complete data, with varying amounts of missing data across variables (Tables 6.2 and 6.4). If a complete-case analysis was undertaken, only 28 events would be available.

Of the variables used by Fox et al. (2014), complete data are only available for Bilirubin* and Albumin* for the training/validation data set (see Table 6.2). These variables were therefore chosen to form the set of 'protected covariates' and are included in $\mathbf{x_1}$, so that $P_1 = 2$. The remaining $P_2 = 38$ variables (8 continuous, 30 categorical) were to be subjected to dimension reduction and included in $\mathbf{x_2}$. Continuous variables were centered and scaled; an asterisk denotes that the transformed version is being referred to.

Observed frequencies for the categorical variables are presented in Table 6.4. Note that some categories are quite sparse. The use of sparse categorical variables in DBR and supervised DBR is considered further in Section 6.8.

	Not missing	Missing
DCP*	67	53
AFP*	119	1
L3*	67	53
Bilirubin*	120	0
Albumin*	120	0
Age*	120	0
AST*	120	0
ALP*	120	0
INR*	116	4
Creatinine*	120	0

Table 6.2: Frequencies of missing values for 10 continuous variables in the HCC training/validation data set. An asterisk indicates that the variable has been centered and scaled.

	Level 1	Level 2	Level 3	Level 4	Level 5
Sex	Female	Male			
Ethnicity	Caucasian	Asian Indian Subcontinent	Asian Oriental	Afro-Caribbean	Other
WHO Performance Status	0	1	2	3	4
Chronic Liver Disease	No	Yes			
Diabetes	No	Yes			
Chronic Hepatitis	No	Yes			
HBV	No	Yes			
HCV	No	Yes			
Autoimmune	No	Yes			
Haemochromatosis	No	Yes			
NAFLD/NASH	No	Yes			
Cirrhosis	No	Yes			
Pain	No	Yes			
Weight Loss	No	Yes			
Malaise	No	Yes			
Nausea/Vomiting	No	Yes			
Diarrhoea	No	Yes			
Decompensated Liver Disease	No	Yes			
Haemoperitoneum	No	Yes			
Symptomatic	No	Yes			
Hepatomegaly	No	Yes			
Stigma of Liver Disease	No	Yes			
Ascites	No	Yes			
Encephalopathy	No	Yes			
Child-Pugh Score	A	B	C		
Tumour Type	Solitary	Multifocal	Metastatic		
Vascular Invasion	None	Vascular Invasion	Minor Vascular Invasion	Hepatic vein	Other
T-Stage	0/1	2	3	4	
N-Stage	No	Yes			
M-Stage	No	Yes			

Table 6.3: Levels for the 30 categorical variables in the HCC data set.

	Level 1	Level 2	Level 3	Level 4	Level 5	Not missing	Missing
Sex	11	109				120	0
Ethnicity	112	5	0	0	3	120	0
WHO Performance Status	26	50	38	3	1	118	2
Chronic Liver Disease	4	112				116	4
Diabetes	74	45				119	1
Chronic Hepatitis	72	27				99	21
HBV	93	5				98	22
HCV	70	28				98	22
Autoimmune	119	1				120	0
Haemochromatosis	114	6				120	0
NAFLD/NASH	110	10				120	0
Cirrhosis	10	97				107	13
Pain	95	18				113	7
Weight Loss	102	9				111	9
Malaise	96	14				110	10
Nausea/Vomiting	108	4				112	8
Diarrhoea	110	3				113	7
Decompensated Liver Disease	89	19				108	12
Haemoperitoneum	112	1				113	7
Symptomatic	47	65				112	8
Hepatomegaly	62	43				105	15
Stigma of Liver Disease	86	17				103	17
Ascites	79	41				120	0
Encephalopathy	103	13				116	4
Child-Pugh Score	78	34	7			119	1
Tumour Type	52	64	0			116	4
Vascular Invasion	80	28	1	3	2	114	6
T-Stage	2	35	26	54		117	3
N-Stage	112	5				117	3
M-Stage	106	10				116	4

Table 6.4: Frequencies of observed categories and missing data for 30 categorical variables in the HCC training/validation data set. The levels for each variable are detailed in Table 6.3.

6.7.3 Statistical methods and software

Cox regression models were used for all survival models, with Efron's methods for ties, and classical scaling for MDS analyses. Gower's coefficient was used to calculate dissimilarities, as described in Section 4.4, with two-level categorical variables treated as symmetric. Gower's 'add-a-point' method (Gower, 1968) was used to add test points to an MDS configuration, as described in Section 4.2.1. DBR, WDBR and ScDBR survival models were fitted, using the procedure described in Section 6.3. Repeated five-fold cross-validation (10 repeats) was used to estimate optimal tuning parameters. The statistic used for cross-validation was the concordance index (see Section 5.8.2). Up to 25 latent dimensions were permitted for each method, and grids of possible values were constructed for both λ and θ from 0 to 2 in steps of 0.25. Some preliminary analyses showed that higher values would not be useful for this data set.

In order to estimate weights for WDBR and select variables for ScDBR, the following Cox models were first fitted for each candidate variable in $\mathbf{x_2}$:

$$h(t|\text{Albumin}^*, \text{Bilirubin}^*, \text{Candidate}_p) = h_0(t)\exp(\hat{\beta}_1\text{Albumin}^* + \hat{\beta}_2\text{Bilirubin}^* + \hat{\alpha}_p\text{Candidate}_p), \quad (6.5)$$

and the standardised coefficient, z-value $= \hat{\alpha}_p/\text{se}(\hat{\alpha}_p)$, recorded. In the case of multiple degree of freedom variables (e.g. WHO Performance Status), the largest absolute z-value was recorded. Note that, due to missing data, modelling in this way means that patients are removed casewise so that different subsets of patients are modelled for each candidate variable. This aspect is discussed in more detail in Section 6.8.

All statistical analyses were conducted using `R` software (R Core Team, 2017, Version 3.5.2). Dissimilarities were calculated using the `daisy()` function from the `cluster` package (Maechler et al., 2019). Cox models were fitted using the `coxph()` function from the `survival` package (Therneau, 2015). Classical scaling was implemented using the `cmdscale()` function. User-written `R` code was used for all DBR models and cross-validation.

6.7.4 Results

Median survival times for the training and test data were 419 and 338 days from diagnosis, respectively, Figure 6.5(a). The eigenvalues for the classical scaling solution using the (unweighted) dissimilarity matrix are depicted in Figure 6.5(b). There are 40 positive eigenvalues, but many are very small, suggesting that even without weighting or screening, dimension reduction is viable.

The z-values for the P_2 candidate variables are depicted in Figure 6.6. The largest

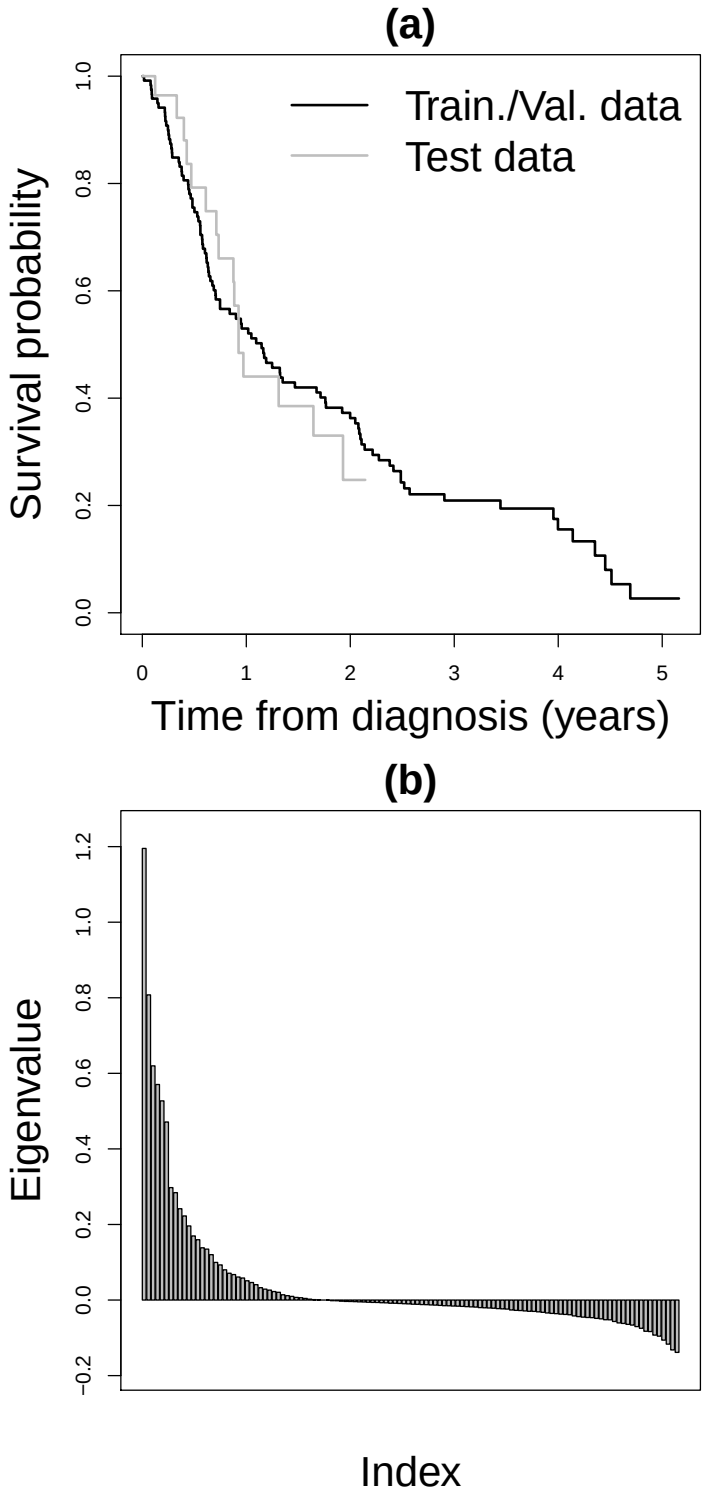

Figure 6.5: (a) Kaplan-Meier curves for the training/validation and test data sets. (b) Eigenvalues for a dissimilarity matrix obtained using Gower's coefficient and using all 38 variables (8 continuous, 30 categorical). There are 40 positive eigenvalues but many are very small.

absolute z-value is for AFP*, implying that this variable would receive the highest weight in WDBR.

Due to the amount of missing data and specifically the lack of complete cases with events, a Cox model including all variables did not converge and stepwise selection was not sensible due to different numbers of observations with complete data for each candidate variable. A Cox model fitted to the combined training/validation data set and including just Albumin* and Bilirubin* had a concordance of 0.70 which fell to 0.48 in the test data. The p-values for Albumin* and Bilirubin* were $p = 0.005$ and $p = 0.051$, respectively.

Results for DBR, WDBR and ScDBR are depicted in Figure 6.7. Note that results for the training and validation data sets are based on repeated cross-validation i.e. they are mean values from 10 repeats with 5 validation folds. Also note that optimal $\hat{\lambda}$ and $\hat{\theta}$ were used for each number of dimensions, Figure 6.7(d) and (e). In Figure 6.7(a), the concordance for the training data improves as the number of dimensions increases, for all models, as expected. WDBR and ScDBR with optimal $\hat{\lambda}$ and $\hat{\theta}$, respectively, have higher concordance than DBR when the number of dimensions is low (less than 5), but otherwise the models perform similarly. The same pattern is seen for the validation data, Figure 6.7(b).

The optimal number of latent dimensions for conventional DBR is $\hat{S} = 8$, as depicted by the dashed line in Figure 6.7(b). This is in fact the best model overall since the concordance for the validation data is highest at 0.77. For WDBR and ScDBR, $\hat{S} = 8$ and $\hat{\lambda} = \hat{\theta} = 0$ i.e. conventional DBR. Therefore, whilst WDBR and ScDBR offer improved concordance when the number of dimensions is very low, conventional DBR is the best performing model for this data set. The DBR model is a considerable improvement over a model including the only the protected variables, Albumin* and Bilirubin*.

In the test data set, Figure 6.7(c), the overall concordance is actually higher than in the training and validation data sets. This is unusual but possible, and could be due to the make-up of the test data set compared with the training/validation data set. The interpretation is much the same as for the training and validation data sets: WDBR and ScDBR outperform DBR when the number of dimensions is very low, but otherwise results are similar or identical. A supportive comparison of the three methods in each of the data sets is shown in Figure 6.8.

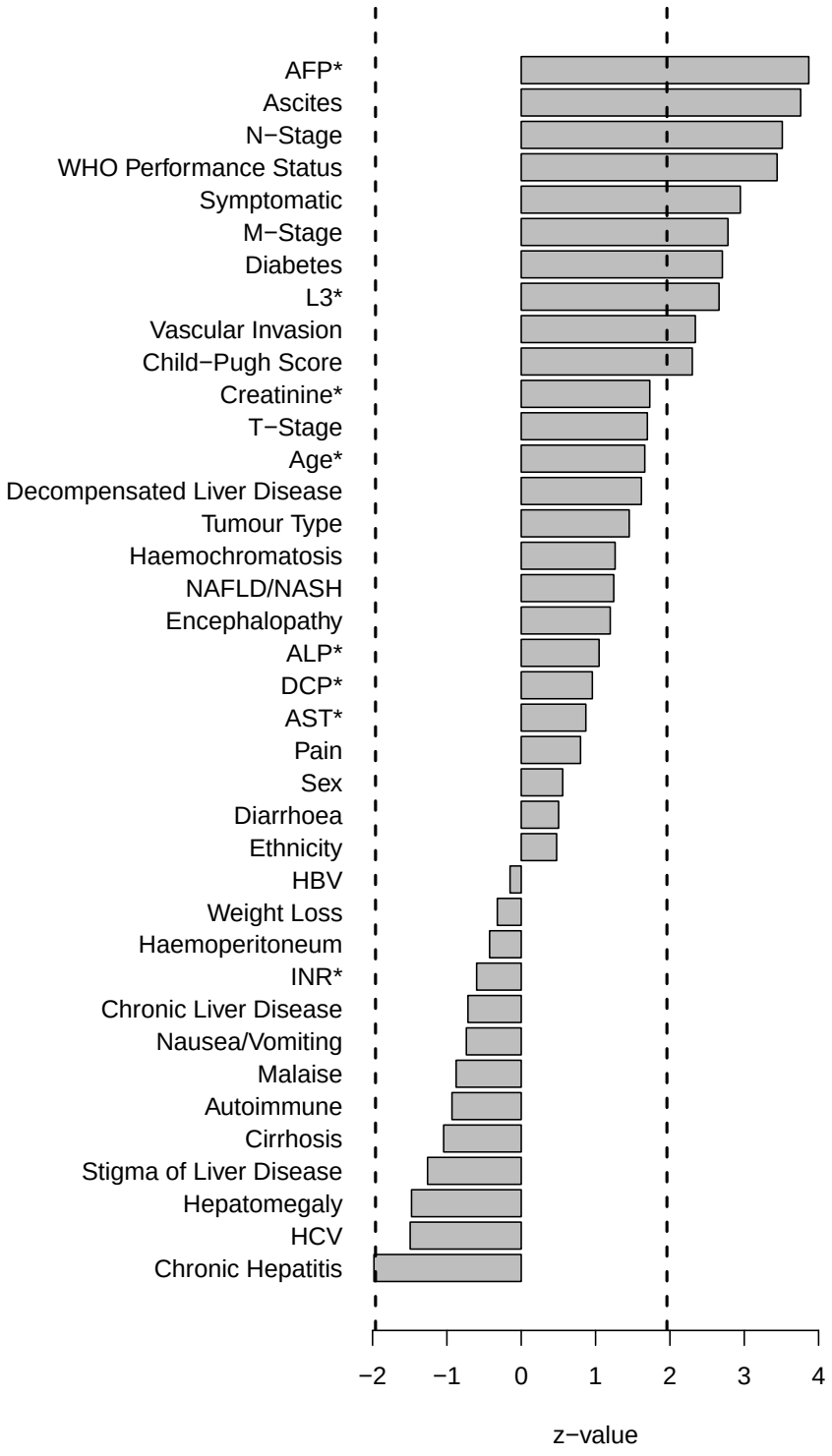

Figure 6.6: z-values for each clinical variable obtained using the models with linear predictors of the form of equation 6.5. AFP* has the largest absolute z-value, implying that it would be weighted highest in WDBR. The overlaid vertical lines represent the quantiles from a standard normal distribution that correspond to probabilities of 2.5% and 97.5%.

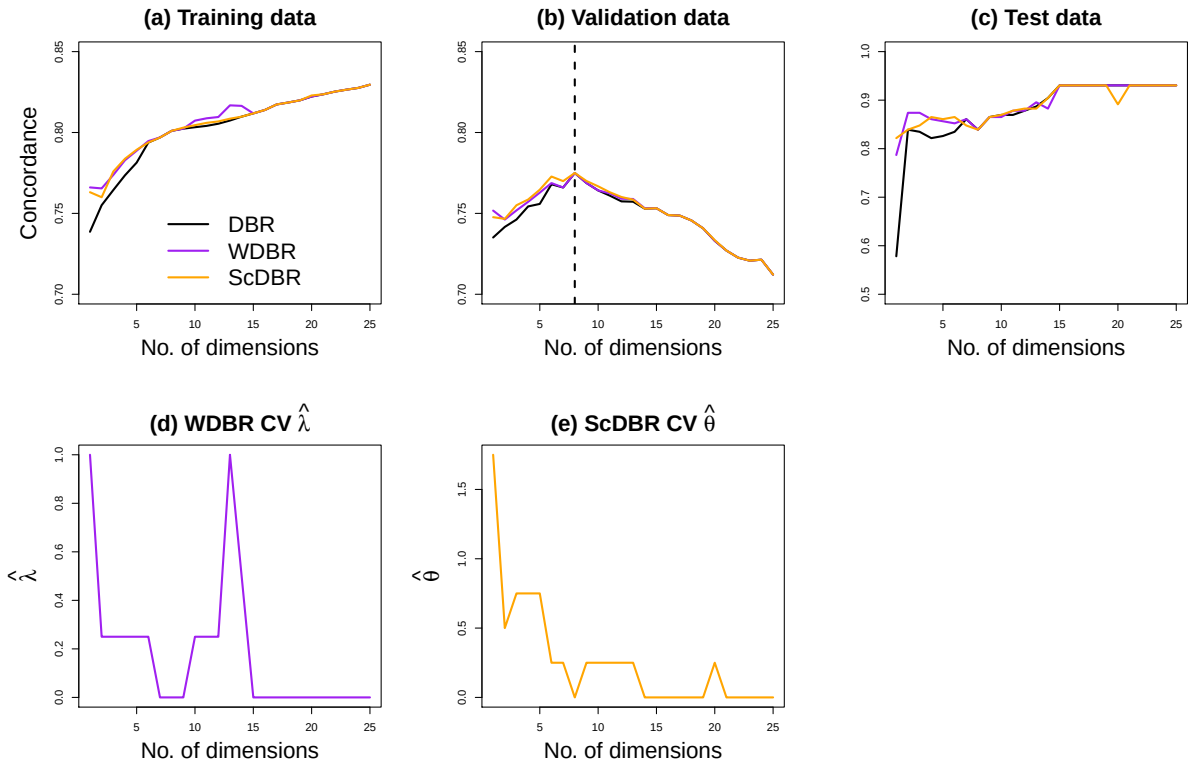

Figure 6.7: Comparison of the concordance for DBR, WDBR and ScDBR models across training, validation and test data sets. Values for the training and test data sets are averages from repeated five-fold cross-validation. (a) WDBR and ScDBR provide the same or better concordance than DBR to the training data, with fewer latent dimensions. (b) For < 8 dimensions, WDBR and ScDBR provide better concordance with the validation data than DBR. However, the best model (dashed black line) is DBR with $\hat{S} = 8$, at which $\hat{\lambda} = \hat{\theta} = 0$, i.e, conventional DBR. (c) The concordance for the test data set is higher than for the training and validation data sets, but the comparison between models is similar. (d, e) Optimal $\hat{\lambda}$ and $\hat{\theta}$ at each number of dimensions. Note that many values are zero, corresponding to conventional DBR.

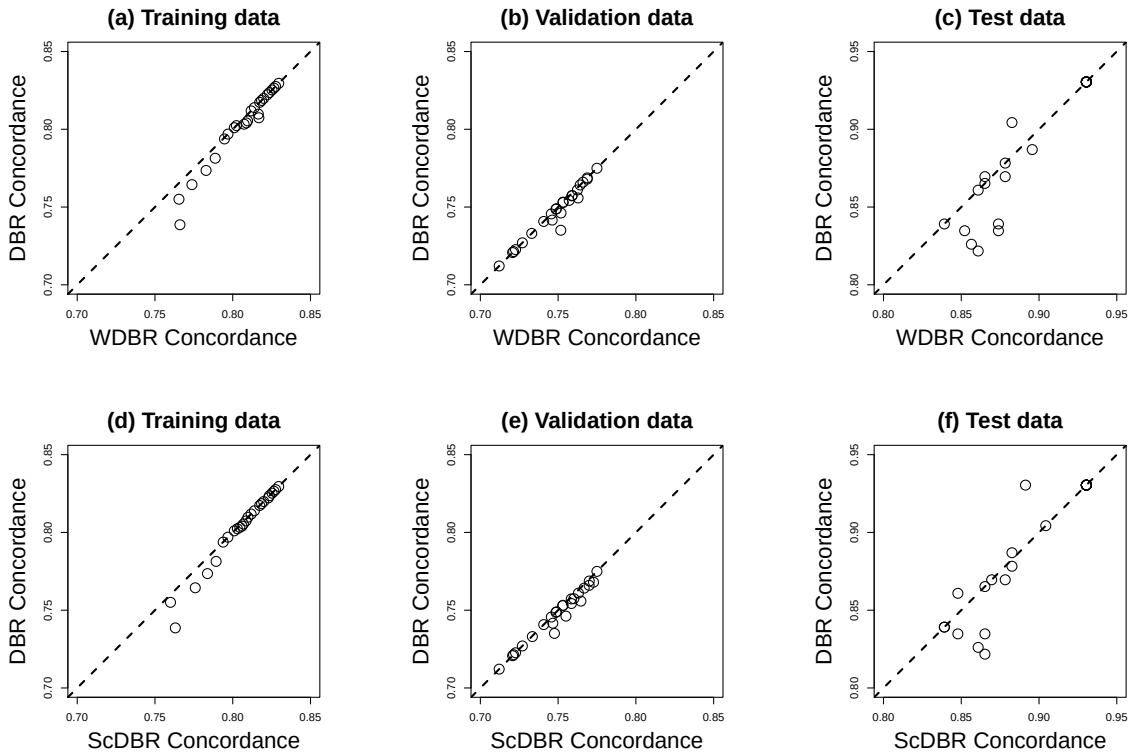

Figure 6.8: Pairwise comparison of the concordance for DBR with WDBR and ScDBR, respectively, across the training, validation and tests data sets. Points in the lower triangle correspond to a better concordance for WDBR/ScDBR at the same number of dimensions as DBR. Values for the training and test data sets are averages from repeated five-fold cross-validation.

6.8 Discussion

In this chapter, DBR was outlined and two approaches to supervised DBR, variable weighted and variable screened DBR, were proposed and explored. The purpose of adding a supervision step to DBR was to obtain well-fitting models with fewer latent dimensions than conventional DBR, at the expense of introducing an additional tuning parameter. Using a simulated data set it was shown that both supervised approaches can improve considerably upon conventional DBR. The utility of supervised and conventional DBR approaches were demonstrated in a subgroup analysis of the HCC data set.

In the simulated example, the proposed supervised DBR approaches considerably improved upon conventional DBR. Whilst variable weighting produced the best fit to validation and test data, variable screening produced only a slightly worse fit than variable weighting, and did so using far fewer latent dimensions. The RMSE was used to assess model fit but a penalised fit statistic, e.g. Adjusted R^2 or AIC, could be used to limit the number of latent dimensions. Moreover, as has been discussed previously (Bøvelstad et al., 2007, 2009), optimal tuning parameter and hence predicted values can be sensitive to the random partitions in K-fold cross-validation. *Repeated* five-fold cross-validation was used in the HCC subgroup analysis to ameliorate this issue, but repeated cross-validation was not computationally feasible for the simulated example due to the dimensions of the tuning grids. An extensive simulation study is required to further assess the properties of variable weighted and variable screened DBR.

The two key benefits of DBR that were focused on in this chapter were: 1) inclusion of observations with missing data without the need for imputation, avoiding casewise deletion of observations, 2) variables of mixed type are easily incorporated, in contrast to some competitor methods such as PCR. A third key benefit not discussed so far is that DBR is straightforward to apply using standard regression methods and software, so that more advanced regression features, e.g. stratification, offsets etc. can be readily applied, using regression procedures familiar to most statistical analysts. Partitioning variables into protected covariates and those for dimension reduction in DBR is also extremely easy.

The approach to handling missing data in Gower's coefficient is worthy of further discussion. This approach has been referred to as 'variable skipping', 'available-case analysis' or less generously as 'pairwise deletion'. Whilst more efficient than complete-case analysis (i.e. casewise deletion), pairwise deletion has been shown to produce biased parameter estimates when data are not MCAR (Little, 1992; Peugh and Enders, 2004). Moreover, whilst pairwise deletion was used here for the final supervised DBR model, in the supervision step proposed in this chapter, observations with missing data are removed casewise, meaning that different subsets of observations are used for different estimates. In general, if there are missing data then supplementary analyses

to DBR that use casewise deletion are not directly comparable. However, if data are MCAR, the supervision step will not introduce bias.

In the simulated example, in which there were no missing data, all DBR methods were outperformed by a ridge regression model. Ridge regression was chosen because it performed best in two previous simulation studies (Bøvelstad et al., 2007, 2009). Whilst a simulation study would be needed to further evaluate the performance of supervised DBR, we would not expect it to outperform ridge regression. One reason is that ridge regression outperformed supervised PCR in the aforementioned simulation studies, which is a special case of variable screened DBR. DBR still holds the advantages over ridge regression of being able to incorporate missing data and being able to fit regression models with advanced features using standard software. Moreover, plotting the dimensions of a DBR model may be informative for model and data interpretation.

In this chapter, a supervision step was proposed in which a simple regression model including the protected covariates and one candidate predictor variable at a time was proposed (see equation 6.3). The principle is that the candidate variable should contribute to the predictive power of the model, over and above the protected variables. This is somewhat simplistic however, as there is no guarantee that, once other variables are included, the variable of interest will still usefully contribute to the model. A more complex multivariable model could be used in the supervision step, e.g. using stepwise selection or model averaging, but the complexity of the model would be limited by the ratio of observations/events to parameters. Boj et al. (2007a) introduced a non-parametric bootstrap approach to establish whether a predictor variable contributes to the predictive power of linear-regression-based DBR model. In contrast, the approach taken here was to test whether the variable may be useful *before* finding latent dimensions, not after. The same univariable/multivariable concerns arise for both approaches.

The model used in the supervision step proposed in this chapter (equation 6.3) was only described for single degree-of-freedom candidate variables. In the HCC subgroup analysis, for multiple degree-of-freedom variables (such as Child-Pugh score with categories A, B and C) the largest standardised parameter estimate was used for weighting and variable screening. In principle, the average standardised parameter estimate or an alternative approach, possibly using the deviance, could be used.

'Sparse' categorical variables, i.e. those for which some categories have low or zero frequencies, in principle, are not a problem for DBR. As an extreme example, if all patients were female and Sex was included as a variable amongst those for dimension reduction, then the dissimilarity for all patients on this variable will just take the value of zero. For supervised DBR, such variables still present an issue since regression models with linear predictors of the form of equation 6.3 are used in the supervision step and resulting estimates can be large or even infinite. Further, with K-fold cross-

validation, such estimates can occur for some folds and not others. In the HCC subgroup example, if this occurred then weights were fixed at zero in variable weighting and the corresponding variable was not included in the variable screened model.

The weighting scheme proposed for variable weighted DBR essentially increases the weight for standardised estimates greater than one and decreases the weight for standardised estimates less than one, so that 'one' is a pivot point. Changing the pivot point is trivial as it can be achieved by simply adding a constant to each a in equation 6.2. Other weighting schemes may be more successful, and this is a useful topic for future work.

In conclusion, DBR is a useful approach to regression modelling when there is a requirement to reduce the dimensionality of the data set and/or include observations with missing data, without the need for multiple imputation. In this chapter, two supervised DBR approaches, variable weighted and variable screened DBR were proposed. It was shown, using simulated and actual data, that an improved model fit can be obtained with these approaches, using fewer latent dimensions than conventional DBR. The cost of supervision is an additional tuning parameter. A simulation study is required to further assess the properties of variable weighted and variable screened DBR.

Chapter 7

Outcome-constrained and outcome-transformed multidimensional scaling

7.1 Introduction

In Chapters 5 and 6, MDS-based methods were used for visualisation and dimension reduction, respectively. In this Chapter, two methods are developed for prediction (using MDS directly as opposed to including MDS dimensions in a regression model) and for visualisation of predictor and outcome data simultaneously.

In MDS, dissimilarities are typically based on measurements of independent or 'predictor' variables. Including dependent or 'outcome' data in an MDS configuration can be useful to obtain effective visualisations of predictor and outcome data simultaneously, and can lead to MDS-based prediction methods. In this chapter, two new approaches for incorporating outcome data into an MDS configuration are proposed for the purposes of prediction and visualisation. Several such methods have been proposed previously, and will be briefly described before the statistical details are presented.

Cox and Ferry (1993, CF) incorporated a two-level outcome variable into a non-metric MDS solution with the purpose of classifying new observations. In CF, dissimilarities between observations with different outcome values are accentuated before applying non-metric MDS. A test observation can then be mapped on to the configuration and classified using a discriminant rule. CF was shown to outperform Fisher's discriminant rule as a classification tool, particularly when there were a number of outlying observations.

Witten and Tibshirani (2011) introduced a 'supervised' version of least-squares MDS (SMDS) in which a configuration of points is sought which simultaneously represents the dissimilarities from predictor variable data and separates observations with different outcome measurements. A unique feature of SMDS is that *all* of the MDS dimensions are supervised by the outcome variable. The authors describe how SMDS can be used for classification when the outcome variable is nominal with two or three categories. SMDS was shown to be competitive with a number of well-known classification methods in a simulation study, to be useful for bipartite ranking, and to provide effective data visualisations (Witten and Tibshirani, 2011; Witten, 2013).

In weakly constrained MDS (Borg et al., 2012, WCMDS), external constraints are imposed on an MDS solution that are not strictly enforced (see also Borg and Lingoes, 1980). A classic example of WCMDS is where points representing colours on a colour wheel are constrained to lie near to the circumference of a circle. To our knowledge, WCMDS has not been used for prediction previously.

In this chapter, two simple and flexible extensions of MDS for inclusion of an outcome variable into an MDS configuration are proposed. In the first, Outcome-constrained MDS (OCMDS), a simple adaptation of WCMDS is proposed, whereby a vector of outcome measurements is appended to the predictor variable matrix and then given extra weight when calculating dissimilarities, before carrying out MDS. In the second, Outcome-transformed MDS (OTMDS), the points on an MDS configuration are transformed using simple linear transformations so that they better represent an outcome variable. Transformations are applied to all dimensions simultaneously, so that, like SMDS, all dimensions are supervised. With both of these proposed approaches, either a categorical or continuous outcome can be used, without modification. As will be shown, both OTMDS and OCMDS can be regarded as prediction tools that can also provide effective visualisations of predictor and outcome data simultaneously.

The rest of the chapter is organised as follows: in Section 7.2 the details of existing MDS-based methods and the newly proposed methods are given. In Section 7.3, a simulation study is presented in which the performance of OCMDS and OTMDS as classification tools is assessed. In Section 7.4, the utility of the newly proposed methods for prediction and visualisation is demonstrated using a hepatocellular carcinoma (HCC) data set. Discussion is given in Section 7.5.

7.2 Methods

To introduce some general notation: $\mathbf{X}$ is an $N \times P$ matrix of predictor variables for $i = 1, \ldots, N$ subjects, with observed values $\mathbf{x}$. Let $\mathbf{Z}$ denote an $N \times S$ matrix of MDS coordinates with observed values $\mathbf{z}$ and let $\mathbf{Y}$ represent an $N \times 1$ vector of outcome variables with observed values $\mathbf{y}$. Let $\boldsymbol{\delta}$ represent the $N \times N$ dissimilarity

matrix obtained using a chosen dissimilarity measure with elements δ_{ij} denoting the dissimilarity between subjects i and j. Distances between observations on the MDS configuration are denoted as d_{ij} and are equal to $\|\mathbf{z}_i - \mathbf{z}_j\|_2$, where $\|\cdot\|_2$ is the Euclidean norm.

A simple simulated data set is used to illustrate the effect of applying the various MDS methods discussed in this chapter. Firstly, 100 observations were generated for 10 standard normal variables. Half of the observations were then assigned to outcome group 1 and the the other half to outcome group 2. For group 1, a constant value of 0.05 was subtracted from all values and for group 2, 0.05 was added. Figure 7.1(a) shows a two-dimensional classical scaling solution for these data, where it can be seen that there is considerable overlap between groups.

7.2.1 Cox and Ferry's approach

CF was developed as an MDS-based method for classifying observations and an example CF configuration is shown in Figure 7.1(b). In this approach, a modified dissimilarity matrix is subjected to non-metric MDS to obtain a representation of both predictor and outcome data simultaneously, before using discriminant analysis to classify observations. Suppose that there are two groups labelled as 1 and 2 and let δ_{ij}^{rs} represent the dissimilarity between the ith and jth subject in groups r and s, respectively. In CF, before undertaking non-metric MDS, dissimilarities are accentuated to induce greater between-group separation, so that

$$\delta_{ij}^{*rs} = \gamma \delta_{ij}^{rs},$$

when $r \neq s$ and otherwise $\delta_{ij}^{*rs} = \delta_{ij}^{rs}$, and where $\gamma \geq 1$. Clearly, larger values of the tuning parameter, γ, will result in greater separation between groups, but note that values too large will result in degeneracy (see Cox and Ferry, 1993). In the example in Figure 7.1(b), $\gamma = 3$ was used, which induces complete separation between groups, whereas with standard metric MDS the groups overlap considerably, Figure 7.1(a). Note that separation for CF primarily occurs in a single dimension. Once the MDS configuration is obtained, a discriminant rule is found. A mapping is then estimated using multivariate linear regression, i.e. by regressing the predictor measurements, $\mathbf{x}$, on the the resulting MDS configuration, $\mathbf{z}$. The mapping can be used to add a test observation to the plot, and the test observation is then classified according to the derived discriminant rule. Cross-validation (see Section 6.5) can be used to select a suitable value for γ.

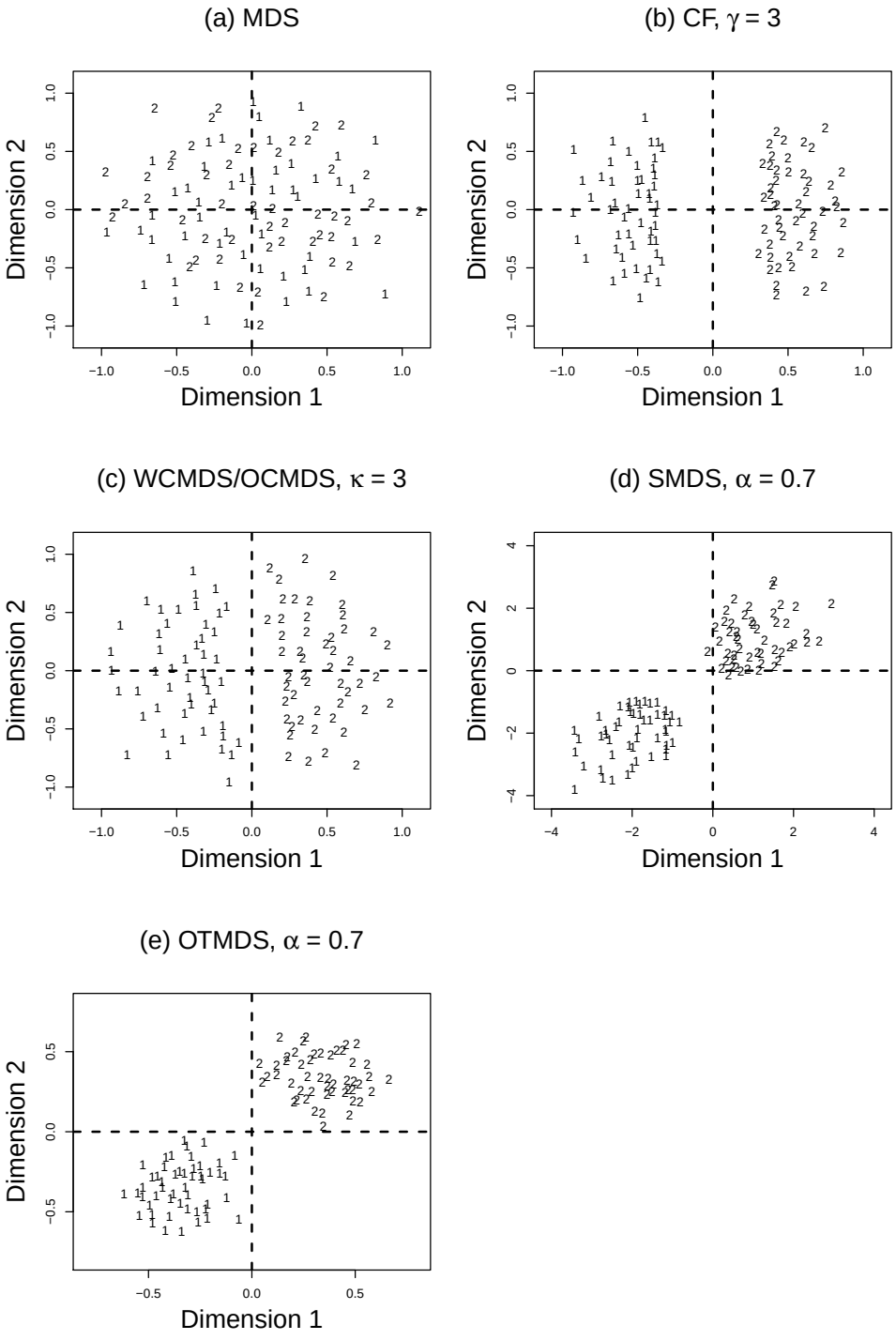

Figure 7.1: Comparison of different methods for incorporating a two-level outcome variable into an MDS configuration. Simulated data consist of 100 observations, 10 (uncorrelated) independent normal random variables with means of -0.05 (group 1) and 0.05 (group 2). Dissimilarities obtained using Euclidean distances: (a) Standard MDS solution, (b) Cox and Ferry's method with tuning parameter $\gamma = 3$, (c) WCMDS (or equivalently OCMDS) with $\kappa = 3$, (d) SMDS with tuning parameter $\alpha = 0.7$, (e) OTMDS with $\alpha = 0.7$

7.2.2 Weakly constrained MDS

In WCMDS (Borg and Groenen, 2003, Chapter 10), an MDS configuration is constrained in some way, but the constraint is not strictly enforced. This is achieved using two dissimilarity matrices which are specified to impose the required constraint. An MDS configuration is then sought which minimises a weighted loss function, as will be shown.

WCMDS is a very general technique in which many different types of constraint can be imposed. However, for our purposes, an MDS configuration is sought for which the distance between points on the MDS configuration is somewhat representative of the dissimilarity between subjects with respect to both predictor and outcome measurements, simultaneously. An example of WCMDS is given in Figure 7.1(c). Let $\delta_{ij}(\mathbf{x})$ and $\delta_{ij}(\mathbf{y})$ denote the dissimilarities between subjects i and j for the predictor measurements, $\mathbf{x}$, and outcome values, $\mathbf{y}$, respectively. A suitable MDS configuration can then be found which minimises a penalised version of the usual Stress:

$$\text{Stress}_{\text{WCMDS}} = \sum_{i,j}^{N} \left[\delta_{ij}(\mathbf{x}) - d_{ij}\right]^2 + \kappa \sum_{i,j}^{N} \left[\delta_{ij}(\mathbf{y}) - d_{ij}\right]^2,$$

where $\kappa \geq 0$. The second term is the penalty term for the fit to the outcome values, with larger values of κ resulting in a greater penalty. As with CF, very large values for the tuning parameter will result in degeneracy. In practice, if the same dissimilarity measure is used for the two dissimilarity matrices, then the vector $\mathbf{y}$ can simply be appended to $\mathbf{x}$, and κ can be set as the weight for $\mathbf{y}$ when calculating the dissimilarities. For example, if the outcome and predictor variables are all continuous and the weighted Euclidean distance is used with WCMDS,

$$\delta_{ij} = \left[\kappa(y_i - y_j)^2 + \sum_{p=1}^{P} \omega_p (x_{ip} - x_{jp})^2\right]^{1/2},$$

where ω_p is the weight for predictor variable p. MDS is then carried out, as usual i.e. there is no requirement for any specialised estimation routine. Moreover, WCMDS can easily include a categorical or continuous outcome variable without modification.

7.2.3 Outcome-constrained MDS

To our knowledge, WCMDS has not been used for prediction previously. Consider a test observation with vector of predictor measurements $\mathbf{x}_{N+1}$, for which the value for the outcome variable, y_{N+1}, is unknown. We propose OCMDS as a simple extension of WCMDS to predict y_{N+1}. In OCMDS, the WCMDS procedure is carried

out as described above, and a value for the outcome variable is specified, $\tilde{y}_{N+1}$. The test observation is then mapped to the MDS configuration using Gower's add-a-point method (Gower, 1968, see Section 4.2.1), with coordinates $\tilde{\mathbf{z}}_{N+1}$, and the value of the Stress-based loss function is calculated as

$$\text{Stress}_{\text{OCMDS}} = \sum_{i,j}^{N+1} \left[\delta_{ij}(\mathbf{x}') - d'_{ij}\right]^2 + \kappa \sum_{i,j}^{N+1} \left[\delta_{ij}(\mathbf{y}') - d'_{ij}\right]^2, \tag{7.1}$$

where $\mathbf{x}'$ and $\mathbf{y}'$ are the observed predictor matrix and outcome vector, respectively, which include the values for the test observation, $\mathbf{x}_{N+1}$ and $\tilde{y}_{N+1}$. Moreover, d'_{ij} is calculated using $\mathbf{z}'$, which is the matrix of MDS coordinates including $\tilde{\mathbf{z}}_{N+1}$. This process is repeated for different values of $\tilde{y}_{N+1}$ and the estimated value, $\hat{y}_{N+1}$, is the value of $\tilde{y}_{N+1}$ which results in the minimum $\text{Stress}_{\text{OCMDS}}$. The only relevant terms in $\text{Stress}_{\text{OCMDS}}$ are actually those involving the test (N+1th) observation and therefore the full summation over all i and j is not required. This approach to classifying observations is similar to that used in SMDS, as will be shown. With OCMDS, κ is now a tuning parameter and a suitable value can be found by cross-validation. Since WCMDS can include a categorical or continuous outcome variable, by extension, so can OCMDS, without any modification to the procedure.

7.2.4 Supervised MDS

For the methods discussed so far, separation between observations with different values for the outcome variable is induced, but it is not guaranteed that separation will occur in all dimensions. In fact, for the the aforementioned methods, the fit to the outcome variable tends to dominate one (usually the first) dimension. In SMDS, all MDS dimensions are supervised, as shown in Figure 7.1(d). SMDS seeks a configuration of points $\mathbf{z}_i, \ldots, \mathbf{z}_N \in \mathbb{R}^S$ for which $\delta_{ij} \approx d_{ij}$, and so that $z_{js} > z_{is}$ tends to occur for all $s = 1, \ldots, S$ dimensions when $y_j > y_i$. Witten and Tibshirani (2011) define a Stress-based loss function for SMDS as

$$\text{Stress}_{\text{SMDS}} = \frac{1}{2}(1-\alpha)\sum_{i,j}^{N}(\delta_{ij} - d_{ij})^2 + \alpha \sum_{i,j:y_j>y_i}^{N} (y_j - y_i) \sum_{s=1}^{S} \left[\frac{\delta_{ij}}{\sqrt{S}} - (z_{js} - z_{is})\right]^2,$$

where $\alpha \in [0, 1]$. Like WCMDS, SMDS includes a penalty term, but a key distinction is that with SMDS the penalty is based on dimension-specific absolute differences i.e. $z_{js} - z_{is}$, and includes a summation over the S dimensions. Iterative majorisation is used to minimise $\text{Stress}_{\text{SMDS}}$. In order to classify a test observation, a similar method as was described for OCMDS is used: $\mathbf{z}_1, \ldots, \mathbf{z}_N$ are held fixed, a value for $\tilde{y}_{N+1}$ is specified, and iterative majorisation is used to find the coordinates, $\tilde{\mathbf{z}}_{N+1}$, which

minimise $\text{Stress}_{\text{SMDS}}$, including the test ($N+1$th) observation in the summations in $\text{Stress}_{\text{SMDS}}$. As with OCMDS, the process is repeated for different values of $\tilde{y}_{N+1}$, and $\hat{y}_{N+1}$ is the value of $\tilde{y}_{N+1}$ which minimises $\text{Stress}_{\text{SMDS}}$.

With SMDS, all dimensions are supervised and an iterative majorisation routine is used to minimise a supervised loss function. The estimation routine presented in Witten and Tibshirani (2011) is specifically for a two-level outcome variable and requires modification for other situations. As can be seen in Figure 7.1(d), for a two-dimensional SMDS configuration, observations with different class labels are forced towards opposing corners of the plot. Another way of forcing observations to the corners of the plot is by using simple linear transformations, and this leads us on to the OTMDS proposal.

7.2.5 Outcome-transformed MDS

OTMDS is now introduced and is where an MDS configuration is found and then subsequently transformed so that the dimensions better correspond to the outcome variable, as shown in Figure 7.1(e). As can be seen, a similar effect to SMDS can be achieved. A direct comparison of OTMDS and SMDS for different tuning parameter values is given in Figure 7.2. The steps involved in the proposed OTMDS approach are now described:

- Step 1. Using a preferred dissimilarity metric, calculate the $N \times N$ dissimilarity matrix, $\boldsymbol{\delta}$, using $\mathbf{x}$.
- Step 2. Use MDS to obtain an $N \times S$ MDS configuration, $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_S)^\intercal$.
- Step 3. Define a chosen orientation for the MDS solution by specifying $S-1$ angles, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{S-1})^\intercal$, for an S-dimensional rotation matrix, $\mathbf{R}$, and rotate the MDS solution accordingly using

$$\dot{\mathbf{z}} = \mathbf{z}\mathbf{R}.$$

For a two-dimensional solution,

$$\mathbf{R} = \begin{bmatrix} \cos\theta & \sin\theta \\ \text{-}\sin\theta & \cos\theta \end{bmatrix},$$

rotates the configuration clockwise by θ radians or $(\theta \times 180)/\pi$ degrees. Choices for $\boldsymbol{\theta}$ are discussed in Section 7.2.5.

- Step 4. Centre and scale $\mathbf{y}$ to match the variance of $\dot{\mathbf{z}}_s$ using

$$\mathbf{y}_s^* = \sqrt{\frac{\text{Var}(\dot{\mathbf{z}}_s)}{\text{Var}(\mathbf{y})}} \left[\mathbf{y} - \text{E}(\mathbf{y})\right], \tag{7.2}$$

  for $s = 1, \ldots, S$, and where $\text{E}(\cdot)$ and $\text{Var}(\cdot)$ denote the expectation and variance operators.

- Step 5. Calculate the linear transformation that maps $\dot{\mathbf{z}}_s$ to $\mathbf{y}_s^*$, i.e. find $\mathbf{a}_s$ such that $\mathbf{y}_s^* = \dot{\mathbf{z}}_s + \mathbf{a}_s$ for $s = 1, \ldots, S$:

$$\mathbf{a}_s = \mathbf{y}_s^* - \dot{\mathbf{z}}_s.$$

- Step 6. Specify a value for the tuning parameter, $\alpha \in [0, 1]$, which controls the extent to which each dimension is transformed according to the outcome variable, and calculate the transformed coordinates, for $s = 1, \ldots, S$, using

$$\mathbf{z}_s^* = \dot{\mathbf{z}}_s + \alpha \mathbf{a}_s. \tag{7.3}$$

The resulting configuration can then be plotted, as shown in Figure 7.1(e). The tuning parameter, α, is effectively the proportion of complete transformation to use, with $\alpha = 1$ corresponding to an exact reproduction of each $\mathbf{y}_s^*$ and $\alpha = 0$ corresponding to no change in the configuration (other than the rotation). The effect of applying different values for α is shown in Figure 7.2. As shown, $\alpha = 0$ corresponds to no change, intermediate values of α cause separation between groups in all dimensions, and $\alpha = 1$ corresponds to each group member taking the exact same coordinates. It is useful to compare the results of $\alpha = 1$ between OTMDS and SMDS; for SMDS there is still some separation between points within groups.

In order to add a test observation, the OTMDS algorithm continues:

- Step 7. Specify a value for $\tilde{y}_{N+1}$ and use Gower's add-a-point formula to obtain $\tilde{\mathbf{z}}_{N+1}$.

- Step 8. Calculate $\tilde{\dot{\mathbf{z}}}_{N+1} = \tilde{\mathbf{z}}_{N+1}^{\intercal} \mathbf{R}$.

- Step 9. Repeat the transformations in Steps 5 and 6 for the test observation only.

In a similar way to OCMDS and SMDS, steps 7 to 9 are repeated for different specified $\tilde{y}_{N+1}$ values. The outcome value which results in the lowest value of the chosen loss function is the predicted value, $\hat{y}_{N+1}$. $\text{Stress}_{\text{SMDS}}$ is one possibility for the loss function.

As Witten and Tibshirani (2011) found for SMDS, empirically better prediction results can be obtained by using the training observations to define a cut-point for

choosing between different values of $\tilde{y}_{N+1}$ (see Witten and Tibshirani, 2011, Section 3.1).

Rotation

The relative distances between objects are preserved for any rotation, reflection, translation or scaling of an MDS configuration (see e.g. Cox and Cox, 2000). The major limitation of linearly transforming the MDS coordinates in OTMDS is that the solution is dependent on the orientation of the configuration. One possibility is to find the optimal angle(s) by minimising the chosen loss function over $\boldsymbol{\theta}$ for a fixed value of α. For a two-dimensional solution, for which most applications of MDS are used, θ can be found using a one-dimensional root finding algorithm, such as those in the `R` functions `optimize` and `uniroot`. Alternatively, a one-dimensional grid search may be feasible, particularly given that OTMDS is extremely computationally inexpensive and an accuracy of one degree should be sufficient for most applications. For visualisation purposes, it may be sufficient for the user to judiciously select the desired orientation.

7.3 Simulation study

A simulation study was conducted to assess the performance of OCMDS and OTMDS as classification tools, compared with logistic regression, SMDS and CF. The simulation conditions are the same as those described in Witten and Tibshirani (2011, Section 2.3), with the exception that a larger sample size scenario ($N = 100$) was added. Briefly, observations were assigned to either group 1 or 2, which were of equal size, and multivariate normal data were generated according to three data generating models which are referred to as 'Two-sided', 'Constant' and 'Linear' (see Appendix D for details). As well as the three data-generating models, the number of predictor variables, $P \in \{5, 15\}$, and sample size, $N \in \{20, 50, 100\}$, were varied, giving $3 \times 2 \times 3 = 18$ simulation scenarios. For each simulation scenario, 100 random samples of size N were generated and these formed the training data sets. For each training data set, an independent test data set containing 100 observations was generated. The models were trained on a training data set and then tested on the corresponding independent test data set. The proportion of classification errors in the test data set was calculated and averaged within each simulation scenario.

To avoid the computational burden of cross-validating tuning parameters, values were specified which were found to work well during some preliminary exploratory analyses. For SMDS and CF, tuning parameter values were fixed at $\alpha = 1$ and $\gamma = 1.5$, respectively. Witten and Tibshirani (2011) showed that generally $\alpha = 1$ worked best for SMDS with these simulation conditions. For OCMDS, $\kappa = 2$ for all scenarios. For

OTMDS, $\alpha = 1$ was used for the Two-sided model and $\alpha = 0.1$ for the Linear and Constant data generating models.

Table 7.1 contains the average proportion of classification errors from the three simulation conditions. For the Two-sided model, SMDS and OTMDS perform well, whilst the remaining models struggle to classify observations at all. In this scenario, the benefit of supervising multiple MDS dimensions is apparent. For the Linear scenarios, OCMDS, SMDS and CF exhibit the lowest error rates and outperform logistic regression. For the Constant scenarios, all the models perform similarly well for $P = 5$, but OCMDS was the best performing model when $P = 15$.

7.4 Analysis of the hepatocellular carcinoma data set

In this section, it is shown how OTMDS and OCMDS can be used for classification (diagnosis) using clinical data. A detailed example of the use of OTMDS for visualisation of prognostic information using clinical data is also given. The HCC data set is used, as described in Section 1.6.1.

7.4.1 Diagnosis

Johnson et al. (2014) developed a diagnostic model 'GALAD' for HCC. GALAD is a predicted score obtained using multivariable logistic regression with three serological biomarkers: AFP, L3 and DCP (see Section 1.6.1), as well as Sex and Age, for diagnosing HCC. GALAD was found to be an extremely effective diagnostic tool and was recommended for use in clinical practice in conjunction with tumour biopsy and radiography.

To compare the diagnostic performance of GALAD with OCMDS, OTMDS and SMDS, a subset of patients from the HCC data set from those used in the original study were used (i.e. Groups I and II, see Section 1.6.1). Moreover, analyses were limited to those patients with complete data for all GALAD variables. Whilst the MDS-based methods do not require data to be complete for an observation, only patients with complete data were used for a fair comparison between these methods and GALAD. The final data set was then split into training and testing subsets according to the original study, giving: training set, $N = 453$ (207 HCC cases, 246 CLD controls) and testing set, $N = 188$ (96 HCC cases, 92 CLD controls).

The number of dimensions, S was fixed at 2 for all MDS-based methods. Age and the biomarkers were log-transformed (base 10) for GALAD, as per the original study, and additionally centred and scaled for the MDS-based methods. In some preliminary exploratory work it was found that, for $S = 2$, it was best to effectively exclude Age by setting the dissimilarity weight equal to zero. The weight for Sex was set to one tenth

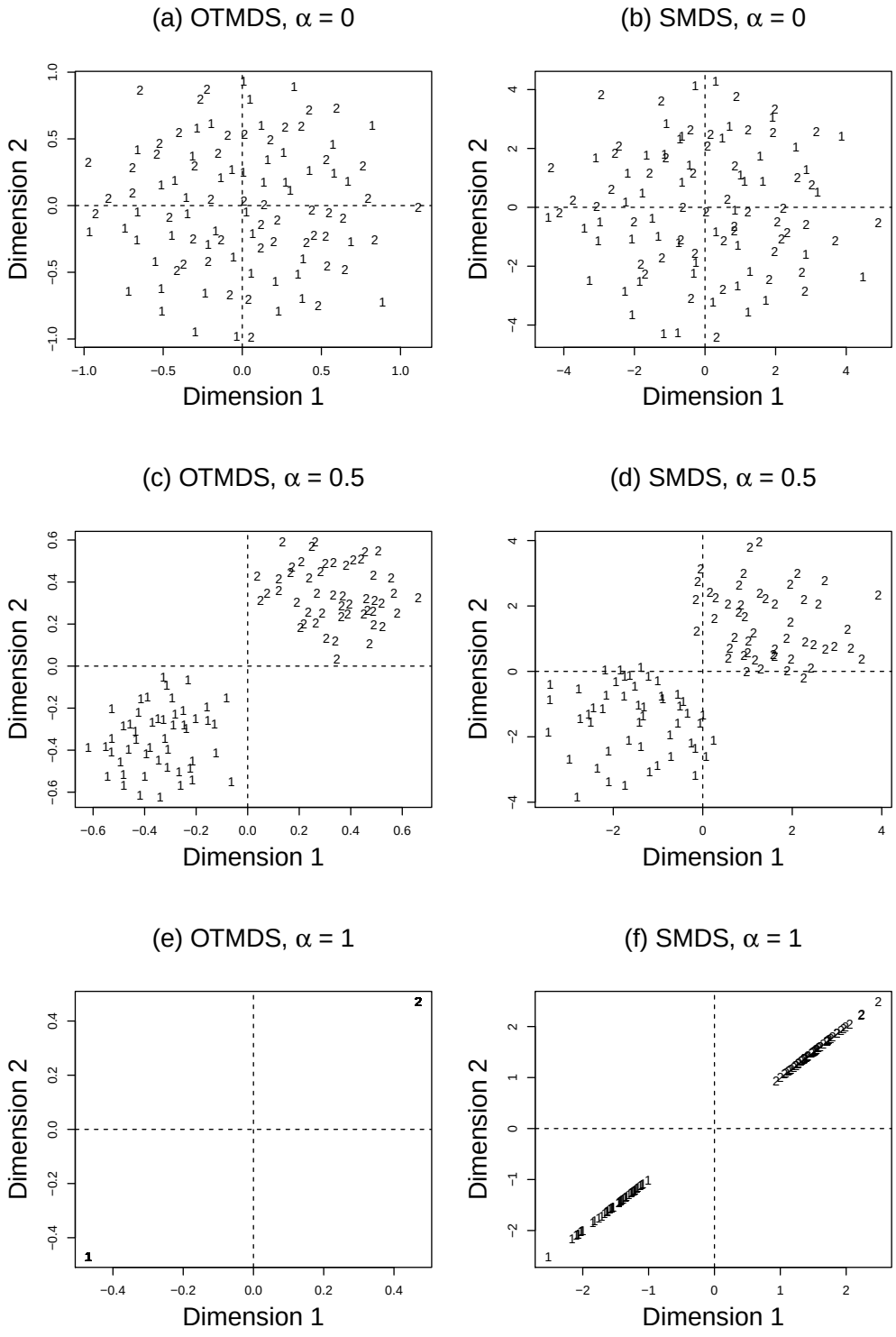

Figure 7.2: Comparison of the effects of tuning parameter α values on OTMDS and SMDS. The clearest difference in the results of the two approaches can be seen in panels (e) and (f) where $\alpha = 1$.

P	N	Data model	Logistic	SMDS	CF	OTMDS	OCMDS
5	20	Two-sided	0.49	0.21	0.47	0.22	0.49
5	50	Two-sided	0.49	0.21	0.46	0.23	0.50
5	100	Two-sided	0.51	0.15	0.47	0.17	0.50
15	20	Two-sided	0.50	0.15	0.50	0.18	0.43
15	50	Two-sided	0.50	0.13	0.51	0.15	0.46
15	100	Two-sided	0.50	0.13	0.49	0.15	0.48
5	20	Linear	0.24	0.13	0.12	0.22	0.14
5	50	Linear	0.26	0.13	0.12	0.21	0.13
5	100	Linear	0.15	0.10	0.10	0.19	0.11
15	20	Linear	0.15	0.10	0.10	0.19	0.07
15	50	Linear	0.12	0.10	0.09	0.18	0.07
15	100	Linear	0.13	0.10	0.10	0.18	0.07
5	20	Constant	0.28	0.28	0.22	0.20	0.24
5	50	Constant	0.29	0.26	0.21	0.20	0.21
5	100	Constant	0.19	0.19	0.15	0.16	0.20
15	20	Constant	0.19	0.19	0.16	0.16	0.11
15	50	Constant	0.16	0.16	0.14	0.14	0.08
15	100	Constant	0.15	0.16	0.14	0.14	0.07

Table 7.1: Simulation results for 18 scenarios in which the numbers of observations (N), predictor variables (P) and the type of data generating model were varied. Numbers in the table correspond to the proportion of classification errors i.e. lower values are better. Models used were logistic regression, SMDS ($\alpha = 1$), CF ($\gamma = 1.5$) and OCMDS ($\kappa = 2$), as well as OTMDS with $\alpha = 1$ for the Two-sided model and $\alpha = 0.1$ for the Linear and Constant data generating models.

of the remaining variables since categorical variables tend to dominate an MDS configuration. Whilst the GALAD logistic regression parameters and cut-point were fixed according to the published model, for the MDS-based methods it was still necessary to estimate the optimal tuning parameter. Five-fold cross-validation (see Section 6.5) was used with five repeats giving: $\hat{\kappa} = 1.5$ (OCMDS), $\hat{\alpha} = 0.25$ (OTMDS) and $\hat{\alpha} = 0.25$ (SMDS).

The results are displayed in Table 7.2. OTMDS correctly classified 87.8% of observations (misclassifying 8 cases and 15 controls), OCMDS 84.0% (misclassifying 3 cases and 27 controls), and SMDS 89.4% (misclassifiying 8 cases and 12 controls). GALAD correctly classified 91.5% of observations (misclassifying 5 cases and 11 controls). A visual comparison of the results can be found in Figures 7.3 and 7.4.

Clearly the aforementioned classification results may be dependent on the particular training/testing data partition (albeit the one used in the original study). As an informal assessment of the sensitivity of the results to the choice of partition, the analysis was repeated five times using five different partitions of the data set previously used for training i.e. using 363 observations for training and 90 observations for testing,

and repeating five times. The results are displayed in Table 7.3. On average, GALAD, OCMDS, OTMDS and SMDS correctly classified 92.4%, 83.3%, 90.7% and 88.2% of subjects, respectively.

7.4.2 Prognosis

In this section, it is demonstrated how OTMDS can be used as a visualisation tool with a continuous external variable. Case and control-specific multivariable Cox regression models (Cox, 1972) were first fitted to the full HCC data set (Groups I to IV, see Section 1.6.1), adjusting for the variables used in the GALAD model, as identified in the previous section. From these models, predicted conditional 3-year survival probabilities were obtained. Note that the estimated 3-year survival probabilities for cases and controls in this data set are approximately 17% and 86% for cases and controls, respectively, emphasising the poor prognosis of patients with HCC. It is now shown how OTMDS can use these predicted probabilities to supervise the MDS solution (note though that a predicted probability is not an outcome variable, as such, but for visualisation purposes this is not important).

Next, Gower's coefficient (Gower, 1971) was used to calculate the dissimilarity between subjects. Sex was weighted as one tenth of the other variables, since categorical variables tend to dominate MDS configurations when using Gower's coefficient. Other weights could be used. A two-dimensional classical scaling solution for these data (equivalent to OTMDS with $\alpha = 0$) is depicted in Figure 7.5(a), and is a good fit, with 73% of the total variation explained in two dimensions. The controls (black triangles) form two distinct and relatively close-knit subgroups corresponding to males and females. The cases (gray circles) are far more disparate, but some of these observations are positioned amongst the controls.

OTMDS was then used with $\alpha \in (0.3, 0.7, 0.95)$, as depicted in Figure 7.5(b) to (d). The optimal angle (287°) was obtained by minimising $\text{Stress}_{\text{SMDS}}$ with $\alpha = 0.5$. Any α could be used, but it is best to use a single angle for all plots, for consistency. As α increases, the observations are translated more closely to their conditional predicted 3-year survival probabilities; clearly the cases and controls diverge as the influence of the predicted probabilities on the configuration increases.

In Figure 7.6, the same data as in Figure 7.5 are displayed, except now the plots have been annotated to demonstrate how OTMDS might be used in more detail. Firstly, the three arrows correspond to increasing values of the three biomarkers (arbitrary magnitude).

Secondly, consider the two shaded regions in each panel of Figure 7.6 which contain all of the cases (solid outline) and controls (dashed outline) with L3 values of 0%. In Figure 7.6(a), the shaded regions are largely overlapping, however as α increases, they soon diverge. The median predicted conditional 3-year survival probability for these

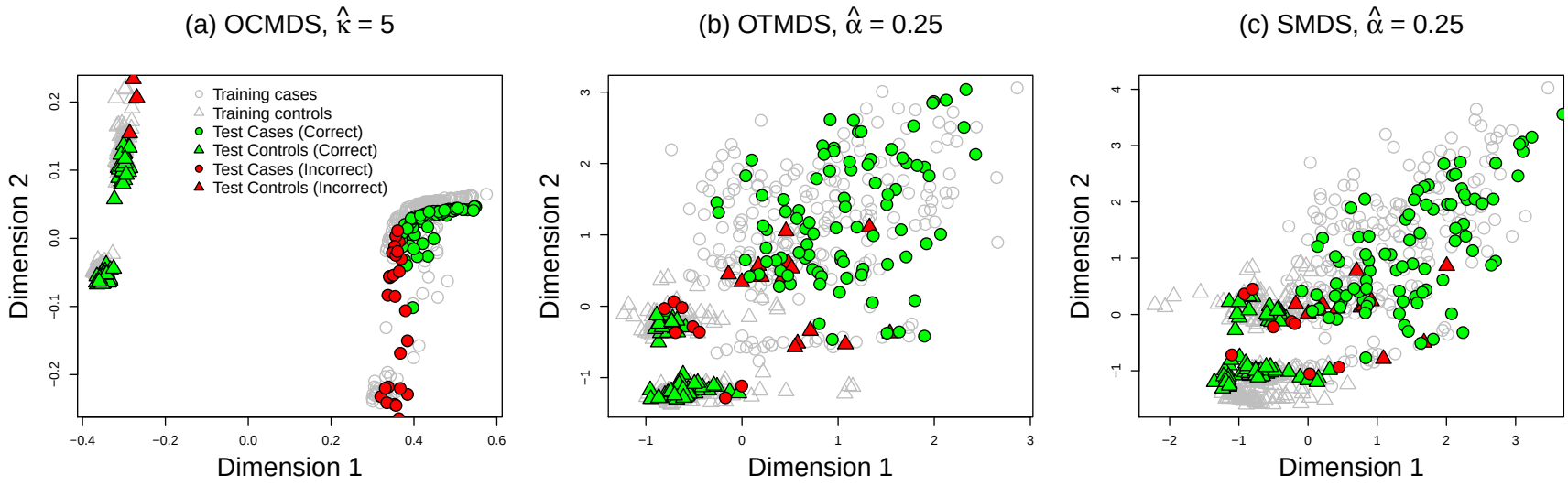

Figure 7.3: Internal validation 1: Comparison of methods for classifying HCC cases (circles) and controls (triangles). Training data (grey symbols), correctly classified test data (green symbols) and incorrectly classified test data (red symbols). (a) OCMDS with $\hat{\kappa} = 5$, (b) OTMDS with $\hat{\alpha} = 0.25$, and (c) SMDS with $\hat{\alpha} = 0.25$

Model	Incorrect			Correct		Sensitivity	Specificity
	Controls	Cases	Total	Total	%		
GALAD	11	5	16	172	91.5	88.0	94.8
OCMDS	27	3	30	158	84.0	76.7	95.8
OTMDS	15	8	23	165	87.8	83.7	91.7
SMDS	12	8	20	168	89.4	87.0	91.7

Table 7.2: Classification results for the prospectively collected test subset of the HCC data set using GALAD, OCMDS ($\hat{\kappa} = 5$), OTMDS ($\hat{\alpha} = 0.25$) and SMDS ($\hat{\alpha} = 0.25$). Tuning parameters were obtained by five-fold cross-validation with five repeats using an independent training data set.

Model	Sample	Incorrect			Correct		Sensitivity	Specificity
		Controls	Cases	Total	Total	%		
GALAD	1	3	4	7	83	92.2	93.9	90.2
	2	3	4	7	83	92.2	93.9	90.2
	3	0	2	2	88	97.8	100.0	95.7
	4	0	4	4	86	95.6	100.0	90.0
	5	4	5	9	81	90.0	91.3	88.6
OCMDS	1	16	2	18	72	80.0	74.6	92.6
	2	12	0	12	78	86.7	80.3	100.0
	3	13	1	14	76	84.4	76.4	97.1
	4	14	1	15	75	83.3	77.8	96.3
	5	14	2	16	74	82.2	75.9	93.8
OTMDS	1	5	9	14	76	84.4	89.8	78.0
	2	2	4	6	84	93.3	95.9	90.2
	3	2	3	5	85	94.4	95.3	93.6
	4	1	7	8	82	91.1	98.0	82.5
	5	2	7	9	81	90.0	95.7	84.1
SMDS	1	7	7	14	76	84.4	85.7	82.9
	2	9	5	14	76	84.4	81.6	87.8
	3	2	2	4	86	95.6	95.3	95.7
	4	2	8	10	80	88.9	96.0	80.0
	5	7	4	11	79	87.8	84.8	90.9

Table 7.3: Classification results for five different samples of the HCC data set using GALAD, OCMDS, OTMDS and SMDS, with tuning parameters obtained by five-fold cross-validation with five repeats.

subjects is around 14% and 91% for cases and controls, respectively. Interestingly, the range for the controls is also very wide at 38-100%, and this is clearly evident in the spread of values in Figure 7.6(c) and (d).

Thirdly, two patients in Figure 7.6(a) who present with contrasting biomarker levels are considered. The first subject (gray inverted triangle) has low L3 and high DCP (<1st and 91st percentiles respectively) whilst the second subject (gray diamond) has the opposite (99th and 40th percentiles respectively). However, as α increases, the two observations converge towards their conditional predicted survival probability, which

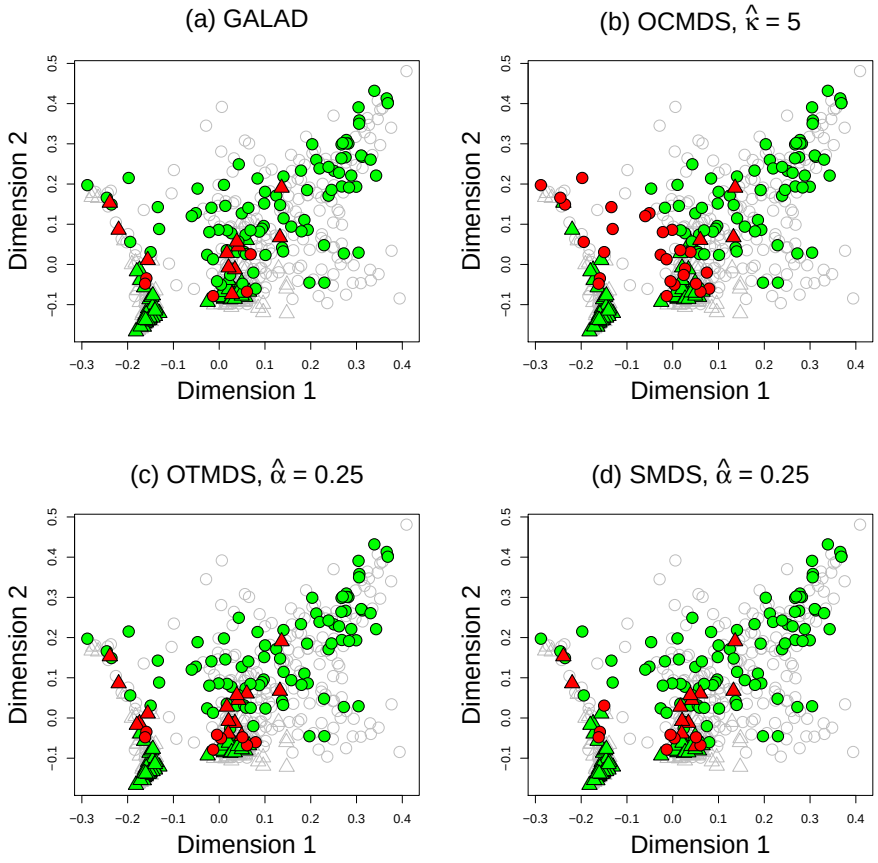

Figure 7.4: Shaded classical scaling representations for a visual comparison of the results displayed in Table 7.2. HCC cases (circles) and controls (triangles). Training data (grey symbols), correctly classified test data (green symbols) and incorrectly classified test data (red symbols). (a) OCMDS with $\hat{\kappa} = 5$, (b) OTMDS with $\hat{\alpha} = 0.25$, and (c) SMDS with $\hat{\alpha} = 0.25$

for both subjects is approximately 16%.

7.5 Discussion

In this chapter, OCMDS and OTMDS were introduced as new prediction and visualisation tools for multivariate data. Both methods were shown to be simple and flexible adaptations of MDS which, without modification, can be used with either a categorical or continuous outcome variable. Using a clinical data set, both methods performed well in classifying (diagnosing) cases and controls, compared with an established clinical prediction tool based on logistic regression as well as with another MDS-based method. A number of potential applications for OTMDS as an effective visualisation tool were also demonstrated using clinical data. In a simulation study, both methods were shown to perform competitively with other well-known classification methods.

All of the MDS-based methods discussed in this chapter rely on having sufficient MDS dimensions to obtain a reasonable fit to the data. Two dimensions were used for diagnosis in the clinical example, but in practice it may be necessary to increase the number of dimensions until a suitable fit is found. Whilst the proposed approaches compared well with the clinical prediction tool, GALAD, effectively only internal validation was conducted and it would be interesting to see how the results might differ on an external validation data set. A detailed comparison of which subjects are correctly classified by the MDS-based methods but incorrectly classified by GALAD could lead to improvements in the GALAD model or to the use of a combination of statistical approaches for HCC diagnosis.

In both OTMDS and SMDS, all dimensions are supervised simultaneously. Contrarily, in OCMDS and CF, supervision tends to occur in a single dimension (see Figure 7.1). The benefit of supervising multiple dimensions is apparent in the results from data simulated from a 'Two-sided' model in the simulation study (Table 7.1) in which CF and OCMDS exhibited approximately 50% misclassification rates compared with $\leq 23\%$ with OTMDS and SMDS.

OTMDS is a heuristic approach. The MDS coordinates are translated simultaneously and by a fixed proportion of the optimal mapping. This has the favourable property that the relative distance between two points with the same value for the outcome variable is maintained. However, this approach limits the possible solutions, in contrast with other MDS-based approaches which seek a globally optimum configuration. Moreover, the performance of OTMDS can depend on the starting configuration (which is effectively arbitrary in MDS). Despite these limitations, the simulation results showed that OTMDS performed competitively with other classification methods when a suitable rotational transformation was found. In exploratory work we have found that a 'good', rather than optimal, orientation is usually sufficient. Moreover, for OT-

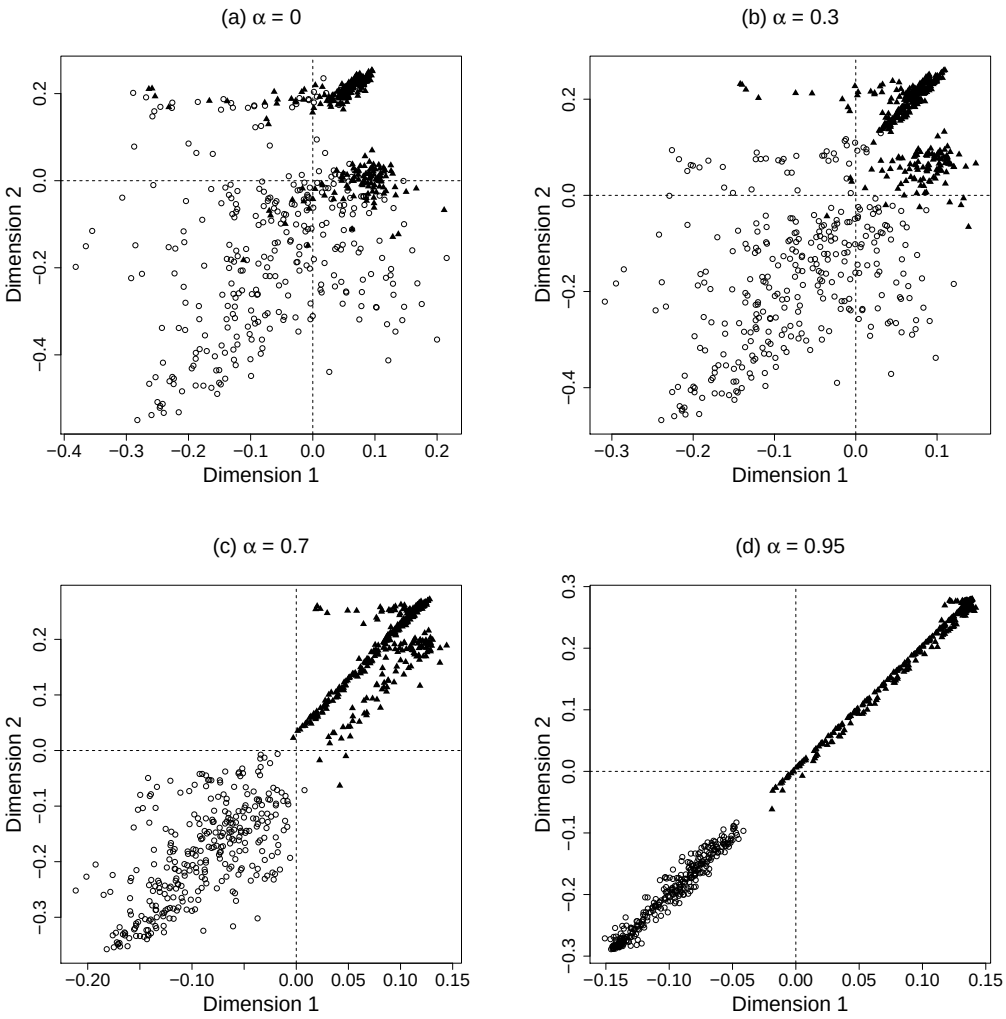

Figure 7.5: TMDS plots of the HCC data set for $\alpha \in (0, 0.3, 0.7, 0.95)$. As α increases, the observations are translated more closely to their conditional predicted 3-year survival probabilities; cases (circles) and controls (triangles) diverge as the influence of the predicted probabilities on the configuration increases.

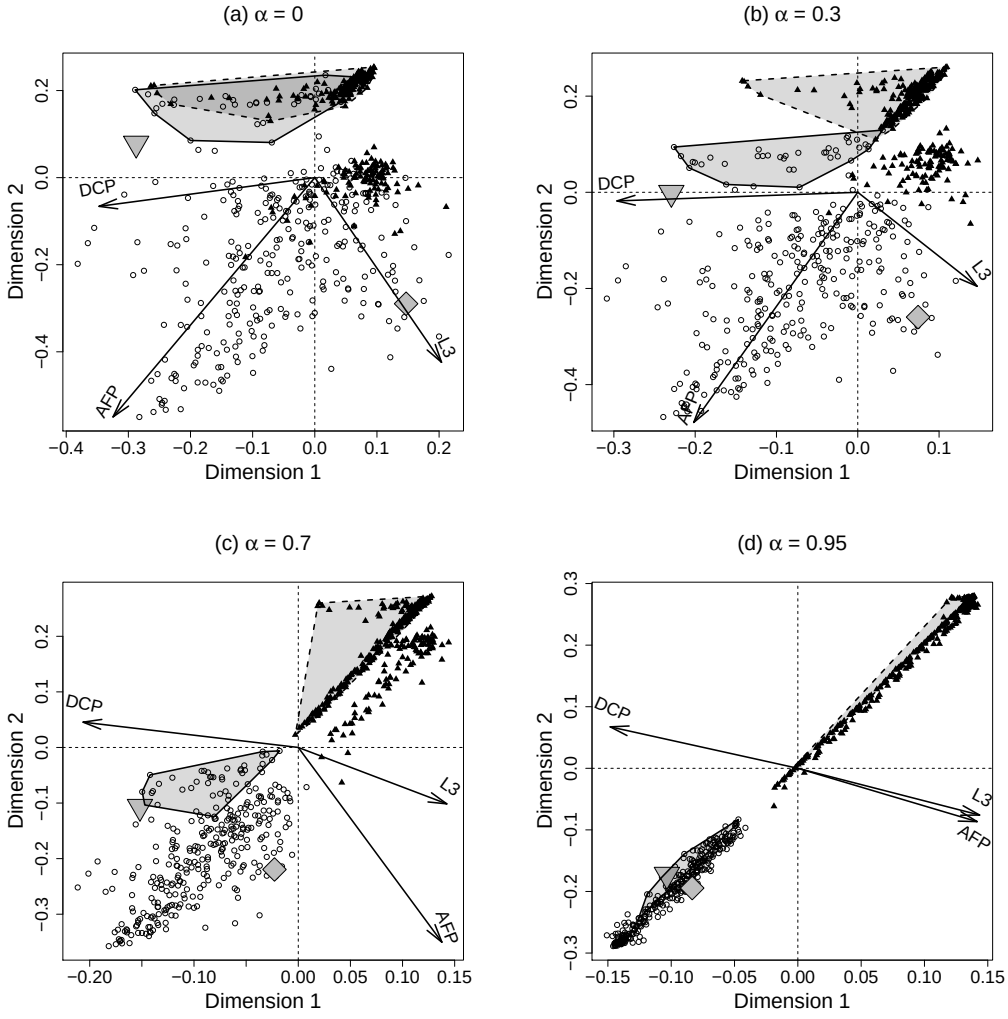

Figure 7.6: Annotated TMDS plots of the HCC data set for $\alpha \in (0, 0.3, 0.7, 0.95)$. Arrows represent the axes of the three serological biomarkers with arbitrary magnitude. The shaded regions ('convex hulls') represent cases (solid border) and controls (dashed border) that present with L3 values of zero. For further description, see the text in Section 7.4.2.

MDS, the Stress-based criterion introduced by Witten and Tibshirani (2011) was used as a goodness-of-fit measure. If classical scaling is used then the Strain rather than the Stress is minimised and a Strain-based loss function could be used instead.

OCMDS and OTMDS can be used with a continuous outcome variable without modification and for OTMDS it was shown how effective visualisations of clinical data can be obtained with a continuous variable. Whilst, in principle, OCMDS and OTMDS can also be used for prediction with a continuous outcome variable, we have found that generally they perform poorly. Some exploratory work is presented in Appendix D. Performance was particularly poor when the relationship between predictors and outcome was highly non-linear and non-monotonic. Polynomial, spline, or other non-linear regression models should be preferred in these situations.

Conceptually, OTMDS and OCMDS work by distorting the (unsupervised) MDS configuration and then using a supervised loss function to classify new observations. Valid questions are: is it necessary to distort the MDS configuration in the first place? And if so, do we need to use the same tuning parameter for the distortion of the MDS configuration as is used in the loss function when classifying observations? In some exploratory work, we have found that it is, a) not always essential to distort the MDS configuration, but that usually improved results are obtained when doing so, b) it is critical to supervise the loss function to effectively classify observations, c) generally, better predictive results are obtained when the same tuning parameter is used for both parts, but sometimes results equally as good can be obtained when the configuration is distorted to a lesser extent. If the tuning parameter value is very high, then the plot can be obscured, see e.g. Figure 7.3(a), so it is reassuring that good predictive results can be obtained without having to obscure the plot so much. On the other hand, finding optimal values amounts to optimising two tuning parameters, plus the number of MDS dimensions if varied, giving three in total. It would be useful to see if these findings apply to SMDS, particularly since the best tuning parameter value for SMDS is usually 1, which makes the SMDS plot difficult to interpret.

A limitation of this work is the down-weighting of categorical variables in the clinical examples. Sex was set to one tenth of the weight of the other variables as categorical variables tend to dominate an MDS configuration using Gower's coefficient. Other weights could have been used. In principle, the weight could be optimised according to some criterion, and this is an interesting subject for further work.

OCMDS and OTMDS are simple and flexible MDS-based prediction and visualisation tools. In this chapter, their utility was demonstrated using a clinical data set in which they performed competitively with other classification methods and were shown to be an effective visualisation tools for multivariate clinical data.

Chapter 8

Discussion

8.1 Topics covered

The focus of this thesis was on developing and applying latent variable models in clinical research. In particular, there was an emphasis on the use of latent variables in time-to-event modelling. This thesis contains five pieces of original research based on developments of either latent class methods (Chapters 2 and 3) or MDS (Chapters 5 to 7), for the statistical analysis of clinical data.

In Chapter 2, one, two and three-step approaches to latent class modelling with a time-to-event distal outcome were introduced and the empirical properties of the corresponding latent class effect estimates were compared using Monte Carlo simulation. To our knowledge, this was the first study to investigate various approaches to latent class modelling when the distal outcome is a time-to-event variable, and contributes to the emerging body of literature on latent class modelling with a distal outcome variable. Additionally, a potential solution to the label switching problem in latent class simulation studies was proposed.

In Chapter 3, a general joint latent class and time-to-event model was presented and the user-written `LCSM R` function was described. Various joint models were applied to a prostate cancer data set in which the identified latent classes were found to differ from the clinically defined tumour stage, and there was evidence of a differential treatment effect between latent classes on survival.

In Chapter 5, it was demonstrated how AFT models could be used to fit time-to-event biplot axes with a measurement scale. With a time-to-event axis, a biplot can be used to display the key features of a multidimensional data set and it's association with a time-to-event variable. To our knowledge, biplot axes for time-to-event data have not been considered previously. The utility of MDS biplots with a time-to-event axis was demonstrated using both simulated and clinical data sets.

In Chapter 6, two approaches to supervised DBR were proposed and explored: variable weighted and variable screened DBR. The purpose of adding a supervision step to DBR was to obtain well-fitting models with fewer latent dimensions than conventional DBR, at the expense of introducing an additional tuning parameter. Using simulated data, it was shown that with both of the proposed supervised approaches it is possible to improve considerably upon conventional DBR.

In Chapter 7, two novel approaches to incorporating an outcome variable into an MDS configuration were proposed: outcome-constrained and outcome-transformed MDS. It was shown how these methods could be used as prediction and visualisation tools for multivariate clinical data. Both were shown to be simple and flexible adaptations of MDS. Using a clinical data set, these methods performed well in classifying (diagnosing) cases and controls, compared with an established clinical prediction tool based on logistic regression. In a simulation study, the proposed methods were shown to perform competitively with other well-known classification methods.

8.2 Limitations and further work

In the simulation study in Chapter 2, one-step methods generally outperformed the alternatives. However, the main criticism of one-step (and inclusive) methods is that the distal outcome variable can influence latent class composition. A pragmatic solution is to fit the model with and without the distal outcome variable, but a useful area of further research would be to investigate the extent to which a time-to-event variable could influence latent class composition, particularly if a non-parametric or piecewise exponential baseline hazard function is assumed. It would also be interesting to see how these methods perform under model misspecification e.g. if proportional hazards are assumed but not met, what might be the impact on latent class composition? Some inclusive three-step methods performed well in terms of bias, but not in terms of confidence interval coverage. An alternative approach to obtaining confidence intervals, such as bootstrapping, would be worthy of further investigation for inclusive methods.

For the author-written `R` function for joint latent class and time-to-event models presented in Chapter 3, further work is required to extend the functionality and develop a full `R` package.

The latent class models presented in Chapters 2 and 3 are examples of finite mixture models. An inherent difficulty with mixture models is that the statistical analyst can never be sure whether the latent classes identified correspond to genuine underlying subgroups or whether they are a byproduct of the model finding that a mixture of parametric distributions improve the fit to the data. Collaboration with clinical colleagues is important to verify that the subgroups are clinically plausible and, where possible, the results should be supported by similar findings in other data sets. Sensitivity anal-

yses in which manifest variables are added/removed may also be useful to assess the influence of different variables on class composition.

For the time-to-event biplot axes in Chapter 5, it was shown how, in principle, a biplot axis for a predictor variable could be related to a time-to-event biplot axis in order to recover acceleration factors or hazard ratios directly from a biplot, by associating biplot axis scales. The usefulness of this feature appears to be limited however, as it seems that too much error is introduced for the approximation to be adequate. A simulation study could be used to assess the accuracy of this approach further under different conditions.

For the proposed supervised DBR approaches in Chapter 6, there are numerous other statistical models based on distances/dissimilarities that share some of the advantages of DBR. Faraway (2014) discusses the use of dissimilarity matrices in regression in general, and even instances where an outcome matrix is a dissimilarity matrix. 'Kernel' methods cover a broad class of distance-based models commonly used in the genomic literature, including Support Vector Machines and its relatives (Schaid, 2010). More directly related to the work in Chapter 6 is the method of dissimilarity-based Partial Least Squares (DB-PLS; Martin et al., 1995; Boj et al., 2007b), which is an alternative to the supervised DBR approaches proposed here. In future work, it would be useful to compare DB-PLS with variable-weighted and variable screened DBR, and even evaluate other possible weighting schemes for variable-weighted DBR.

The MDS-based methods for visualisation and prediction discussed Chapter 7, rely on having sufficient MDS dimensions to obtain a reasonable fit to the data. In practice it may be necessary to increase the number of dimensions until a suitable fit is found, although this would limit the usefulness of these methods as visualisation tools. Further investigation of the use of the same or different tuning parameters in the transformation of the plot, as well as in the Stress function is also an interesting area of further research that also applies to the supervised MDS approach introduced by Witten and Tibshirani (2011).

In Chapters 5 and 7, categorical variables were down-weighted, since categorical variables tend to dominate an MDS configuration using Gower's general coefficient. In principle, it should be possible to find optimal weights according to some fit criterion and this is an interesting topic for future research.

Appendices

Appendix A

Supplementary material to accompany Chapter 2

A.1 Simulation of time-to-event data

Survival times were simulated in a similar way to that described by Bender et al. (2005). Two classes, 1 and 2, and two treatment groups, A and B, were simulated in each scenario. The reference survival curve used was based on the Kaplan-Meier estimate of the gemcitabine arm in the ESPAC3v2 trial 1.6.2, Figure A.1.

Let S_0 represent the reference survival curve for subjects belonging to *true* latent class 2 and treatment group B (i.e. $c_i = 0$ and $z_i = 0$), so that survival probabilities corresponding to proportional hazard effects can be obtained using

$$S_i = S_0^{\exp(\beta z_i + \gamma_j c_i)},$$

in this case for z_i and $c_i \in \{0, 1\}$. As described in Section 2.3.3, the hazard ratio for the effect of Treatment A relative to Treatment B, $\exp(\beta)$, was fixed at 0.75 and the hazard ratio for the effect of latent class 1 relative to latent class 2 was varied, $\exp(\gamma_1) \in \{1, 1.5, 2, 3\}$. 'True' survival probabilities were obtained for each of the four permutations of class and treatment over a sequence of 0 to 60 months in steps of 0.1 months. High-dimensional spline fits were used to approximate these survival curves, as shown for the reference survival curve in Figure A.1(a). The splines were fitted separately to each of the four survival curves by regressing the time sequence on polynomials of the survival probabilities. A survival probability was then simulated for each subject from Uniform$(0, 1)$ and a corresponding survival time obtained from the relevant spline fit. Administrative censoring was applied at 60 months and uniform

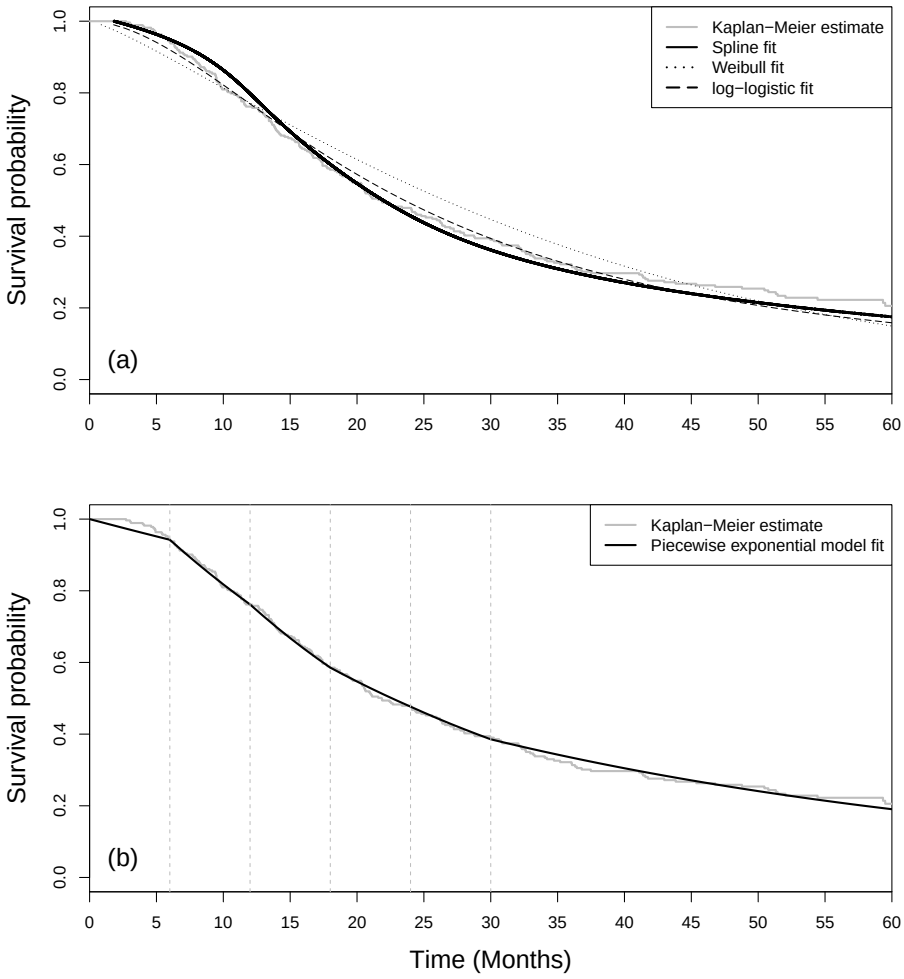

Figure A.1: Kaplan-Meier estimate of overall survival for the gemcitabine arm from the ESPAC3v2 study and overlaid fitted models. (a) Fitted polynomial spline, Weibull and log-logistic (parametric) models. (b) A piecewise exponential survival model with five partitions approximates the Kaplan-Meier estimate well.

censoring was added by generating censoring times from an exponential distribution such that overall approximately 50% of survival times were right-censored.

A.2 Comparison of the use of different hazard functions in three-step models

In this study, one and two-step models used a piecewise exponential baseline hazard function whilst the three-step models used a Cox model in which the baseline hazard is left unspecified. The choice of a piecewise exponential model for the one and two-step models was primarily motivated by the fact that standard errors are easier to obtain when there are few baseline hazard parameters (see Section 2.5). To illustrate that the three-step methods are not disadvantaged by the choice of hazard function the table below contains results from a small simulation study for Scenario 17 (Low entropy, $N = 500$, HR=2). The results demonstrate that the results are practically unaffected by the choice of baseline hazard function.

Model	Estimate	Bias	CI Coverage (%)
MA using Cox model	-0.30	0.39	23.4
MA using PE model	-0.30	0.39	23.8
PA using Cox model	-0.57	0.13	90.4
PA using PE model	-0.57	0.13	90.7

Table A.1: Comparison of simulation results for modal assignment and partial assignment when using unspecified (Cox) an piecewise exponential baseline hazard functions. MA Modal assignment, PA Partial assignment. The results demonstrate that the statistical properties of the latent class effect estimates are practically unaffected by the different hazard functions compared here.

Appendix B

Supplementary material to accompany Chapter 3

B.1 First-order derivatives used in Newton-Raphson steps

In this section, first-order derivatives required for the Newton-Raphson steps in the estimation routine are given. For a piecewise exponential baseline hazard submodel:

$$F_{\kappa_{pj}} = \sum_{i=1}^{N} x_{pi}\nu_{ij}^{(r)}\left[1 - \frac{\exp(\mathbf{x}_i^\intercal\boldsymbol{\kappa}_j)}{\sum_{k=1}^{J}\exp(\mathbf{x}_i^\intercal\boldsymbol{\kappa}_k)}\right],$$

$$F_{\beta_{qj}} = \sum_{i=1}^{N}\sum_{s=1}^{S} z_{qi}\Bigg\{\psi_{is}\delta_i - \psi_{is}\left[\alpha_s(u_i - a_{s-1}) + \sum_{g=1}^{s-1}\alpha_g(a_g - a_{g-1})\right]\times \sum_{j=1}^{J}\nu_{ij}^{(r)}\exp(\mathbf{z}_i^\intercal\boldsymbol{\beta}_j + \gamma_j)\Bigg\},$$

$$F_{\gamma_j} = \sum_{i=1}^{N}\sum_{s=1}^{S}\nu_{ij}^{(r)}\psi_{is}\Bigg\{\delta_i - \left[\alpha_s(u_i - a_{s-1}) + \sum_{g=1}^{s-1}\alpha_g(a_g - a_{g-1})\right]\exp(\mathbf{z}_i^\intercal\boldsymbol{\beta}_j + \gamma_j).$$

For a Weibull submodel:

$$F_\phi = \sum_{i=1}^{N}\frac{\delta_i}{\phi} + \delta_i\log(t_i) - \lambda\log(t_i)t_i^\phi\sum_{j=1}^{J}\nu_{ij}^{(r)}\exp(\mathbf{z}_i^\intercal\boldsymbol{\beta}_j + \gamma_j),$$

$$F_{\beta_{qj}} = \sum_{i=1}^{N} z_{qi} \left[\delta_i - \lambda t_i^{\phi} \sum_{j=1}^{J} \nu_{ij}^{(r)} \exp(\mathbf{z}_i^{\intercal} \boldsymbol{\beta}_j + \gamma_j) \right],$$

$$F_{\gamma_j} = \sum_{i=1}^{N} \nu_{ij}^{(r)} \left[\delta_i - \lambda t_i^{\phi} \exp(\mathbf{z}_i^{\intercal} \boldsymbol{\beta}_j + \gamma_j) \right].$$

B.2 Louis's method

In this section, the requisite calculations for Louis's method (Louis, 1982) to obtain standard errors for the joint model described in Chapter 2 are given. For this purpose, Louis's method is extremely tedious as all first and second-order partial derivatives of the expected complete data log-likelihood, as well as the covariances of all of the first-order partial derivatives, are required.

For a given subject, the density function for the joint latent class and survival model is

$$f_{\mathbf{Y},T,\Delta|\mathbf{Z}}(\mathbf{y}, t, \delta|\mathbf{z}) = \sum_{j=1}^{J} \eta_j \prod_{m=1}^{M} f_{Y_m|C}(y_m|j) f_{T,\Delta|\mathbf{Z},C}(t, \delta|\mathbf{z}, j),$$

where $\sum_{j=1}^{J} \eta_j = 1$. Assuming only binary manifest variables ($y_m = 0, 1$ for $m = 1, \ldots, M$),

$$P(Y_m = 1|C = j) = \pi_{mj}^{y_m} (1 - \pi_{mj})^{(1-y_m)},$$

and assuming a piecewise exponential model for the survival time

$$f_{T,\Delta|\mathbf{Z},C}(t, \delta|\mathbf{z}, j) = \prod_{s=1}^{S} \left[\alpha_{0s} \exp(\mathbf{z}^{\intercal} \boldsymbol{\beta} + \gamma_j) \right]^{\delta \psi_s} \times$$
$$\exp \left\{ - \psi_s \left[\alpha_{0s}(t - a_{s-1}) + \sum_{h=1}^{s-1} \alpha_{0h}(a_h - a_{h-1}) \right] \exp(\mathbf{z}^{\intercal} \boldsymbol{\beta} + \gamma_j) \right\}.$$

The posterior probability of a given subject belonging to class j is obtained by

$$P(C = j|\mathbf{Y} = \mathbf{y}, \mathbf{Z} = \mathbf{z}, T = t, \Delta = \delta) = \nu_j = \frac{\eta_j f_{\mathbf{Y}|C}(\mathbf{y}|j) f_{T,\Delta|\mathbf{Z},C}(t, \delta|\mathbf{z}, j)}{\sum_{k=1}^{J} \eta_k f_{\mathbf{Y}|C}(\mathbf{y}|k) f_{T,\Delta|\mathbf{Z},C}(t, \delta|\mathbf{z}, j)},$$

and $\sum_{j=1}^{J} \nu_j = 1$ for each subject. The *observed data* likelihood for N subjects is

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} \sum_{j=1}^{J} \eta_j \prod_{m=1}^{M} f_{Y_{im}|C_i}(y_{im}|j) f_{T_i,\Delta_i|\mathbf{Z}_i,C_i}(t_i, \delta_i|\mathbf{z}_i, j).$$

If class was known for each subject then the *complete data* likelihood would be

$$L_{\text{comp}}(\boldsymbol{\theta}) = \prod_{i=1}^{N}\prod_{j=1}^{J}\Big[\eta_j \prod_{m=1}^{M} f_{Y_{im}|C_i}(y_{im}|j) f_{T_i,\Delta_i|\mathbf{Z}_i,C_i}(t_i,\delta_i|\mathbf{z}_i,j)\Big]^{v_{ij}},$$

where v_{ij} is an indicator variable and each subject can only belong to one class. Assume only two latent classes ($J=2$) so that: $v_{i2} = 1 - v_{i1}$, $\eta_1 = 1 - \eta_2$ but η_2 is the parameter to be estimated, and that $\gamma_1 = 0$ for identifiability. The log-likelihood of the complete data is given by

$$\begin{aligned}
\ell_{\text{comp}}(\boldsymbol{\theta}) &= \sum_{i=1}^{N}\sum_{j=1}^{J} v_{ij} \log\Big[\eta_j \prod_{m=1}^{M} f_{Y_{im}|C_i}(y_{im}|j) f_{T_i,\Delta_i|\mathbf{Z}_i,C_i}(t_i,\delta_i|\mathbf{z}_i,j)\Big], \\
&= \sum_{i=1}^{N}\sum_{j=1}^{J} v_{ij}\log(\eta_j) \quad + \\
&\quad v_{ij}\Big[\sum_{m=1}^{M} y_{im}\log\pi_{mj} + (1-y_{im})\log(1-\pi_{mj})\Big] \quad + \\
&\quad v_{ij}\Bigg(\sum_{s=1}^{S}\delta_i\psi_{is}\log\alpha_{0s} + \delta_i\psi_{is}(\mathbf{z}_i^{\intercal}\boldsymbol{\beta}+\gamma_j) \quad - \\
&\qquad \psi_{is}\left[\alpha_{0s}(t_i - a_{s-1}) + \sum_{h=1}^{s-1}\alpha_{0h}(a_h - a_{h-1})\right]\exp(\mathbf{z}_i^{\intercal}\boldsymbol{\beta}+\gamma_j)\Bigg).
\end{aligned}$$

The Expected complete data log-likelihood, $E[\ell_{\text{comp}}(\boldsymbol{\theta})]$, is just obtained by replacing v_{ij} with ν_{ij}, for $i = 1,\ldots,N$ and $j = 1,\ldots,J$.

B.2.1 First derivatives of the complete data log-likelihood

For convenience let the cumulative baseline hazard $A_{0s} = \Big[\alpha_{0s}(t_i - a_{s-1}) + \sum_{h=1}^{s-1} \alpha_{0h}(a_h - a_{h-1})\Big]$. The partial first derivatives are then given by:

$$\frac{\partial \ell_{\text{comp}}(\boldsymbol{\theta})}{\partial \eta_2} = \frac{\partial}{\partial \eta_2}\left\{ \sum_{i=1}^{N} v_{i2} \log \eta_2 + (1 - v_{i2}) \log(1 - \eta_2) \right\} = \sum_{i=1}^{N} \frac{v_{i2}}{\eta_2} - \frac{(1 - v_{i2})}{(1 - \eta_2)},$$

$$\frac{\partial \ell_{\text{comp}}(\boldsymbol{\theta})}{\partial \pi_{mj}} = \frac{\partial}{\partial \pi_{mj}}\left\{ \sum_{i=1}^{N} v_{ij}[y_{im} \log \pi_{mj} + (1 - y_{im}) \log(1 - \pi_{mj})] \right\} = \sum_{i=1}^{N} v_{ij} \left[\frac{y_{im}}{\pi_{mj}} - \frac{(1 - y_{im})}{(1 - \pi_{mj})} \right],$$

$$\begin{aligned}
\frac{\partial \ell_{\text{comp}}(\boldsymbol{\theta})}{\partial \beta_q} &= \frac{\partial}{\partial \beta_q}\left\{ \sum_{i=1}^{N} \sum_{j=1}^{J} v_{ij} \left(\sum_{s=1}^{S} \delta_i \psi_{is} (\mathbf{z}_i^\intercal \boldsymbol{\beta} + \gamma_j) - \psi_{is} A_{0s} \exp(\mathbf{z}_i^\intercal \boldsymbol{\beta} + \gamma_j) \right) \right\}, \\
&= \sum_{i=1}^{N} \sum_{j=1}^{J} v_{ij} z_{iq} \sum_{s=1}^{S} \psi_{is} \Big(\delta_i - A_{0s} \exp(\mathbf{z}_i^\intercal \boldsymbol{\beta} + \gamma_j) \Big),
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \ell_{\text{comp}}(\boldsymbol{\theta})}{\partial \gamma_2} &= \frac{\partial}{\partial \gamma_2}\left\{ \sum_{i=1}^{N} v_{i2} \left(\sum_{s=1}^{S} \delta_i \psi_{is} (\mathbf{z}_i^\intercal \boldsymbol{\beta} + \gamma_j) - \psi_{is} A_{0s} \exp(\mathbf{z}_i^\intercal \boldsymbol{\beta} + \gamma_2) \right) \right\}, \\
&= \sum_{i=1}^{N} v_{i2} \sum_{s=1}^{S} \psi_{is} \Big(\delta_i - A_{0s} \exp(\mathbf{z}_i^\intercal \boldsymbol{\beta} + \gamma_2) \Big),
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \ell_{\text{comp}}(\boldsymbol{\theta})}{\partial \alpha_{0s}} &= \frac{\partial}{\partial \alpha_{0s}}\left\{ \sum_{i=1}^{N} \sum_{j=1}^{J} v_{ij} \left(\sum_{s'=1}^{S} \delta_i \psi_{is'} \log \alpha_{0s'} + \delta_i \psi_{is'} (\mathbf{z}_i^\intercal \boldsymbol{\beta} + \gamma_j) - \psi_{is'} \left[\alpha_{0s'}(t_i - a_{s'-1}) + \sum_{h=1}^{v-1} \alpha_{0h}(a_h - a_{h-1}) \right] \exp(\mathbf{z}_i^\intercal \boldsymbol{\beta} + \gamma_j) \right) \right\}, \\
&= \sum_{i=1}^{N} \sum_{j=1}^{J} v_{ij} \left(\frac{\delta_i \psi_{is}}{\alpha_{0s}} - t_{is} \exp(\mathbf{z}_i^\intercal \boldsymbol{\beta} + \gamma_j) \right),
\end{aligned}$$

where t_{is} is the amount of time that the ith subject contributes to the sth interval. The summations can be taken two steps further:

$$
\begin{aligned}
&= \sum_{i=1}^{N} \frac{\delta_i \psi_{is}}{\alpha_{0s}} - t_{is} \exp(\mathbf{z}_i^{\intercal}\boldsymbol{\beta}) \sum_{j=1}^{J} v_{ij} \exp(\gamma_j), \\
&= \sum_{i=1}^{N} \left[\frac{\delta_i \psi_{is}}{\alpha_{0s}} \right] - \sum_{i' \in R_s} \left[t_{i's} \exp(\mathbf{z}_{i'}^{\intercal}\boldsymbol{\beta}) \sum_{j=1}^{J} v_{i'j} \exp(\gamma_j) \right],
\end{aligned}
$$

where the summation of all $i' \in R_s$ refers to all subjects at risk during the sth time period.

B.2.2 Covariance of the first derivatives of the complete data log-likelihood

Some useful covariance results:

$$\mathrm{Cov}(X, Y) = E[(X - E(X)) \times (Y - E(Y))],$$

$$E(v_{ij}) = \nu_{ij},$$

$$E(v_{ij}^2) = \nu_{ij},$$

$$E(v_{i1} v_{i2}) = 0.$$

The following covariances can then be obtained:

$$\mathrm{Cov}\left\{ \frac{\partial \ell_{\mathrm{comp}}(\boldsymbol{\theta})}{\partial \eta_2}, \frac{\partial \ell_{\mathrm{comp}}(\boldsymbol{\theta})}{\partial \eta_2} \right\} = \mathrm{Var}\left\{ \frac{\partial \ell_{\mathrm{comp}}(\boldsymbol{\theta})}{\partial \eta_2} \right\} = \sum_{i=1}^{N} \nu_{i1} \nu_{i2} \left[\frac{1}{\eta_2 (1 - \eta_2)} \right]^2,$$

$$\text{Cov}\left\{\frac{\partial \ell_{\text{comp}}(\boldsymbol{\theta})}{\partial \eta_2}, \frac{\partial \ell_{\text{comp}}(\boldsymbol{\theta})}{\partial \pi_{mj}}\right\} = \sum_{i=1}^{N} \nu_{i1}\nu_{i2} \frac{1}{\eta_2(1-\eta_2)}\left[\frac{y_{im}}{\pi_{mj}} - \frac{(1-y_{im})}{(1-\pi_{mj})}\right] \text{if } j = 2,$$
$$= \sum_{i=1}^{N} -\nu_{i1}\nu_{i2} \frac{1}{\eta_2(1-\eta_2)}\left[\frac{y_{im}}{\pi_{mj}} - \frac{(1-y_{im})}{(1-\pi_{mj})}\right] \text{if } j = 1,$$

$$\text{Cov}\left\{\frac{\partial \ell_{\text{comp}}(\boldsymbol{\theta})}{\partial \eta_2}, \frac{\partial \ell_{\text{comp}}(\boldsymbol{\theta})}{\partial \beta_q}\right\} = \sum_{i=1}^{N} \nu_{i1}\nu_{i2} \frac{1}{\eta_2(1-\eta_2)} z_{iq} \sum_{s=1}^{S} \psi_{is}\left(\delta_i - A_{0s}\exp(\mathbf{z}_i^\intercal \boldsymbol{\beta})\right)[\exp(\gamma_2) - \exp(\gamma_1)],$$
$$= \sum_{i=1}^{N} \nu_{i1}\nu_{i2} \frac{1}{\eta_2(1-\eta_2)} z_{iq} \sum_{s=1}^{S} \psi_{is}\left(\delta_i - A_{0s}\exp(\mathbf{z}_i^\intercal \boldsymbol{\beta})\right)[\exp(\gamma_2) - 1],$$

$$\text{Cov}\left\{\frac{\partial \ell_{\text{comp}}(\boldsymbol{\theta})}{\partial \eta_2}, \frac{\partial \ell_{\text{comp}}(\boldsymbol{\theta})}{\partial \gamma_2}\right\} = \sum_{i=1}^{N} \nu_{i1}\nu_{i2} \frac{1}{\eta_2(1-\eta_2)} \sum_{s=1}^{S} \psi_{is}[\delta_i - A_{0s}\exp(\mathbf{z}_i^\intercal \boldsymbol{\beta} + \gamma_2)],$$

$$\text{Cov}\left\{\frac{\partial \ell_{\text{comp}}(\boldsymbol{\theta})}{\partial \eta_2}, \frac{\partial \ell_{\text{comp}}(\boldsymbol{\theta})}{\partial \alpha_{0s}}\right\} = \sum_{i=1}^{N} \nu_{i1}\nu_{i2} \frac{1}{\eta_2(1-\eta_2)}\Big[- t_{is}\exp(\mathbf{z}_i^\intercal \boldsymbol{\beta})[\exp(\gamma_2) - 1]\Big],$$

$$\text{Cov}\left\{\frac{\partial \ell_{\text{comp}}(\boldsymbol{\theta})}{\partial \pi_{mj}}, \frac{\partial \ell_{\text{comp}}(\boldsymbol{\theta})}{\partial \pi_{lk}}\right\} = \sum_{i=1}^{N} -\nu_{i1}\nu_{i2}\left[\frac{y_{im}}{\pi_{mj}} - \frac{(1-y_{im})}{(1-\pi_{mj})}\right]\left[\frac{y_{il}}{\pi_{lk}} - \frac{(1-y_{il})}{(1-\pi_{lk})}\right] \text{ if } j \neq k,$$
$$= \sum_{i=1}^{N} \nu_{i1}\nu_{i2}\left[\frac{y_{im}}{\pi_{mj}} - \frac{(1-y_{im})}{(1-\pi_{mj})}\right]\left[\frac{y_{il}}{\pi_{lk}} - \frac{(1-y_{il})}{(1-\pi_{lk})}\right] \text{if } j = k,$$

$$\text{Cov}\left\{\frac{\partial \ell_{\text{comp}}(\boldsymbol{\theta})}{\partial \pi_{mj}}, \frac{\partial \ell_{\text{comp}}(\boldsymbol{\theta})}{\partial \beta_q}\right\} = \sum_{i=1}^{N} \nu_{i1}\nu_{i2}\left[\frac{y_{im}}{\pi_{mj}} - \frac{(1-y_{im})}{(1-\pi_{mj})}\right]\left[z_{iq}\sum_{s=1}^{S}\psi_{is}\left(\delta_i - A_{0s}\exp(\mathbf{z}_i^\intercal\boldsymbol{\beta})\right)[\exp(\gamma_2)-1]\right] \text{if } j=2,$$
$$= \sum_{i=1}^{N} \nu_{i1}\nu_{i2}\left[\frac{y_{im}}{\pi_{mj}} - \frac{(1-y_{im})}{(1-\pi_{mj})}\right]\left[z_{iq}\sum_{s=1}^{S}\psi_{is}\left(\delta_i - A_{0s}\exp(\mathbf{z}_i^\intercal\boldsymbol{\beta})\right)[1-\exp(\gamma_2)]\right] \text{if } j=1,$$

$$\text{Cov}\left\{\frac{\partial \ell_{\text{comp}}(\boldsymbol{\theta})}{\partial \pi_{mj}}, \frac{\partial \ell_{\text{comp}}(\boldsymbol{\theta})}{\partial \gamma_2}\right\} = \sum_{i=1}^{N} \nu_{i1}\nu_{i2}\left[\frac{y_{im}}{\pi_{mj}} - \frac{(1-y_{im})}{(1-\pi_{mj})}\right]\sum_{s=1}^{S}\psi_{is}[\delta_i - A_{0s}\exp(\mathbf{z}_i^\intercal\boldsymbol{\beta}+\gamma_2)] \text{if } j=2,$$
$$= \sum_{i=1}^{N} -\nu_{i1}\nu_{i2}\left[\frac{y_{im}}{\pi_{mj}} - \frac{(1-y_{im})}{(1-\pi_{mj})}\right]\sum_{s=1}^{S}\psi_{is}[\delta_i - A_{0s}\exp(\mathbf{z}_i^\intercal\boldsymbol{\beta}+\gamma_2)] \text{if } j=1,$$

$$\text{Cov}\left\{\frac{\partial \ell_{\text{comp}}(\boldsymbol{\theta})}{\partial \pi_{mj}}, \frac{\partial \ell_{\text{comp}}(\boldsymbol{\theta})}{\partial \alpha_{0s}}\right\} = \sum_{i=1}^{N} \nu_{i1}\nu_{i2}\left[\frac{y_{im}}{\pi_{mj}} - \frac{(1-y_{im})}{(1-\pi_{mj})}\right]\left[-t_{is}\exp(\mathbf{z}_i^\intercal\boldsymbol{\beta})[\exp(\gamma_2)-1]\right] \text{if } j=2,$$
$$= \sum_{i=1}^{N} \nu_{i1}\nu_{i2}\left[\frac{y_{im}}{\pi_{mj}} - \frac{(1-y_{im})}{(1-\pi_{mj})}\right]\left[-t_{is}\exp(\mathbf{z}_i^\intercal\boldsymbol{\beta})[1-\exp(\gamma_2)]\right] \text{if } j=1,$$

$$\text{Cov}\left\{\frac{\partial \ell_{\text{comp}}(\boldsymbol{\theta})}{\partial \beta_q}, \frac{\partial \ell_{\text{comp}}(\boldsymbol{\theta})}{\partial \beta_r}\right\} = \sum_{i=1}^{N} \nu_{i1}\nu_{i2}\left[z_{iq}\sum_{s=1}^{S}\psi_{is}[\delta_i - A_{0s}\exp(\mathbf{z}_i^\intercal\boldsymbol{\beta})[1-\exp(\gamma_2)]\right]\left[z_{ir}\sum_{s=1}^{S}\psi_{is}[\delta_i - A_{0s}\exp(\mathbf{z}_i^\intercal\boldsymbol{\beta})[1-\exp(\gamma_2)]\right],$$
$$= \sum_{i=1}^{N} \nu_{i1}\nu_{i2}z_{iq}z_{ir}\left[\sum_{s=1}^{S}\psi_{is}[\delta_i - A_{0s}\exp(\mathbf{z}_i^\intercal\boldsymbol{\beta})[1-\exp(\gamma_2)]\right]^2,$$

$$\text{Cov}\left\{\frac{\partial \ell_{\text{comp}}(\boldsymbol{\theta})}{\partial \beta_q}, \frac{\partial \ell_{\text{comp}}(\boldsymbol{\theta})}{\partial \gamma_2}\right\} = \sum_{i=1}^{N} \nu_{i1}\nu_{i2}\left[\sum_{s=1}^{S}\psi_{is}[\delta_i - A_{0s}\exp(\mathbf{z}_i^\intercal\boldsymbol{\beta}+\gamma_2)]\right]\left[z_{iq}\sum_{s=1}^{S}\psi_{is}[\delta_i - A_{0s}\exp(\mathbf{z}_i^\intercal\boldsymbol{\beta})[\exp(\gamma_2)-1]\right],$$

$$\text{Cov}\left\{\frac{\partial \ell_{\text{comp}}(\boldsymbol{\theta})}{\partial \beta_q}, \frac{\partial \ell_{\text{comp}}(\boldsymbol{\theta})}{\partial \alpha_{0s}}\right\} = \sum_{i=1}^{N} \nu_{i1}\nu_{i2}\left[z_{iq}\sum_{s'=1}^{S}\psi_{is'}[\delta_i - A_{0s'}\exp(\mathbf{z}_i^\intercal\boldsymbol{\beta})][1-\exp(\gamma_2)]\right]\left[-t_{is}\exp(\mathbf{z}_i^\intercal\boldsymbol{\beta})[1-\exp(\gamma_2)]\right],$$

$$\text{Cov}\left\{\frac{\partial \ell_{\text{comp}}(\boldsymbol{\theta})}{\partial \gamma_2}, \frac{\partial \ell_{\text{comp}}(\boldsymbol{\theta})}{\partial \gamma_2}\right\} = \text{Var}\left\{\frac{\partial \ell_{\text{comp}}(\boldsymbol{\theta})}{\partial \gamma_2}\right\} = \sum_{i=1}^{N}\nu_{i1}\nu_{i2}\left[\sum_{s=1}^{S}\psi_{is}[\delta_i - A_{0s}\exp(\mathbf{z}_i^\intercal\boldsymbol{\beta}+\gamma_2)]\right]^2,$$

$$\text{Cov}\left\{\frac{\partial \ell_{\text{comp}}(\boldsymbol{\theta})}{\partial \gamma_2}, \frac{\partial \ell_{\text{comp}}(\boldsymbol{\theta})}{\partial \alpha_{0s}}\right\} = \sum_{i=1}^{N}\nu_{i1}\nu_{i2}\left[\sum_{s'=1}^{S}\psi_{is'}[\delta_i - A_{0s'}\exp(\mathbf{z}_i^\intercal\boldsymbol{\beta}+\gamma_2)]\right]\left[-t_{is}\exp(\mathbf{z}_i^\intercal\boldsymbol{\beta})[\exp(\gamma_2)-1]\right],$$

$$\begin{aligned}\text{Cov}\left\{\frac{\partial \ell_{\text{comp}}(\boldsymbol{\theta})}{\partial \alpha_{0s}}, \frac{\partial \ell_{\text{comp}}(\boldsymbol{\theta})}{\partial \alpha_{0s'}}\right\} &= \sum_{i=1}^{N}\nu_{i1}\nu_{i2}\left[-t_{is}\exp(\mathbf{z}_i^\intercal\boldsymbol{\beta})[1-\exp(\gamma_2)]\right]\left[-t_{is'}\exp(\mathbf{z}_i^\intercal\boldsymbol{\beta})[1-\exp(\gamma_2)]\right], \\ &= \sum_{i=1}^{N}\nu_{i1}\nu_{i2}t_{is}t_{is'}\left[\exp(\mathbf{z}_i^\intercal\boldsymbol{\beta})[1-\exp(\gamma_2)]\right]^2.\end{aligned}$$

B.2.3 Second derivatives of the complete data log-likelihood

$$\frac{\partial^2 \ell_{\text{comp}}(\boldsymbol{\theta})}{\partial \eta_2^2} = \frac{\partial}{\partial \eta_2}\left\{\sum_{i=1}^{N}\frac{v_{i2}}{\eta_2} - \frac{(1-v_{i2})}{(1-\eta_2)}\right\} = \sum_{i=1}^{N} -\frac{v_{i2}}{\eta_2^2} - \frac{(1-v_{i2})}{(1-\eta_2)^2},$$

$$\frac{\partial^2 \ell_{\text{comp}}(\boldsymbol{\theta})}{\partial \pi_{mj}^2} = \frac{\partial}{\partial \pi_{mj}}\left\{\sum_{i=1}^{N} v_{ij}\left[\frac{y_{im}}{\pi_{mj}} - \frac{(1-y_{im})}{(1-\pi_{mj})}\right]\right\} = \sum_{i=1}^{N} v_{ij}\left[-\frac{y_{im}}{\pi_{mj}^2} - \frac{(1-y_{im})}{(1-\pi_{mj})^2}\right].$$

All off-diagonal elements containing η_2 and π_{mj} are equal to zero.

$$\frac{\partial^2 \ell_{\text{comp}}(\boldsymbol{\theta})}{\partial \beta_q \partial \beta_r} = \frac{\partial}{\partial \beta_r}\left\{ \sum_{i=1}^{N}\sum_{j=1}^{J} v_{ij} z_{iq} \sum_{s=1}^{S} \psi_{is}\left(\delta_i - A_{0s}\exp(\mathbf{z}_i^\intercal \boldsymbol{\beta} + \gamma_j)\right)\right\},$$
$$= \sum_{i=1}^{N}\sum_{j=1}^{J} v_{ij} z_{iq} z_{ir} \sum_{s=1}^{S} -\psi_{is} A_{0s}\exp(\mathbf{z}_i^\intercal \boldsymbol{\beta} + \gamma_j),$$

$$\frac{\partial^2 \ell_{\text{comp}}(\boldsymbol{\theta})}{\partial \beta_q \partial \gamma_2} = \frac{\partial}{\partial \gamma_2}\left\{ \sum_{i=1}^{N}\sum_{j=1}^{J} v_{ij} z_{iq} \sum_{s=1}^{S} \psi_{is}\left(\delta_i - A_{0s}\exp(\mathbf{z}_i^\intercal \boldsymbol{\beta} + \gamma_j)\right)\right\},$$
$$= \sum_{i=1}^{N} v_{i2} z_{iq} \sum_{s=1}^{S} -\psi_{is} A_{0s}\exp(\mathbf{z}_i^\intercal \boldsymbol{\beta} + \gamma_2),$$

$$\frac{\partial^2 \ell_{\text{comp}}(\boldsymbol{\theta})}{\partial \beta_q \partial \alpha_{0s}} = \frac{\partial}{\partial \alpha_{0s}}\left\{ \sum_{i=1}^{N}\sum_{j=1}^{J} v_{ij} z_{iq} \sum_{s'=1}^{S} \psi_{is'}\left(\delta_i - \left[\alpha_{0s'}(t_i - a_{s'-1}) + \sum_{h=1}^{s'-1}\alpha_{0h}(a_h - a_{h-1})\right]\exp(\mathbf{z}_i^\intercal \boldsymbol{\beta} + \gamma_j)\right)\right\},$$
$$= \sum_{i=1}^{N}\sum_{j=1}^{J} -v_{ij} z_{iq} t_{is}\exp(\mathbf{z}_i^\intercal \boldsymbol{\beta} + \gamma_j),$$

$$\frac{\partial^2 \ell_{\text{comp}}(\boldsymbol{\theta})}{\partial \gamma_2^2} = \frac{\partial}{\partial \gamma_2}\left\{ \sum_{i=1}^{N} v_{i2} \sum_{s=1}^{S} \psi_{is}\left(\delta_i - A_{0s}\exp(\mathbf{z}_i^\intercal \boldsymbol{\beta} + \gamma_2)\right)\right\},$$
$$= \sum_{i=1}^{N} v_{i2} \sum_{s=1}^{S} -\psi_{is} A_{0s}\exp(\mathbf{z}_i^\intercal \boldsymbol{\beta} + \gamma_2),$$

$$\frac{\partial^2 \ell_{\text{comp}}(\boldsymbol{\theta})}{\partial \gamma_2 \partial \alpha_{0s}} = \frac{\partial}{\partial \alpha_{0s}} \left\{ \sum_{i=1}^{N} v_{i2} \sum_{s'=1}^{S} \psi_{is'} \left(\delta_i - \left[\alpha_{0s'}(t_i - a_{s'-1}) + \sum_{h=1}^{s'-1} \alpha_{0h}(a_h - a_{h-1}) \right] \exp(\mathbf{z}_i^\intercal \boldsymbol{\beta} + \gamma_2) \right) \right\},$$

$$= \sum_{i=1}^{N} -v_{i2} t_{is} \exp(\mathbf{z}_i^\intercal \boldsymbol{\beta} + \gamma_2),$$

$$\frac{\partial^2 \ell_{\text{comp}}(\boldsymbol{\theta})}{d \alpha_{0s}^2} = \frac{\partial}{d \alpha_{0s}} \left\{ \sum_{i=1}^{N} \sum_{j=1}^{J} v_{ij} \left(\frac{\delta_i \psi_{is}}{\alpha_{0s}} - t_{is} \exp(\mathbf{z}_i^\intercal \boldsymbol{\beta} + \gamma_j) \right) \right\},$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{J} -\frac{v_{ij} \delta_i \psi_{is}}{\alpha_{0s}^2} = \sum_{i=1}^{N} -\frac{\delta_i \psi_{is}}{\alpha_{0s}^2},$$

$$\frac{\partial^2 \ell_{\text{comp}}(\boldsymbol{\theta})}{\partial \alpha_{0s} \partial \alpha_{0s'}} = 0. \tag{B.1}$$

B.3 Analysis of the prostate cancer data set

B.3.1 Model selection

In this section, a detailed description of the model selection process shown in Table 3.5 is given.

In models 1 to 7, dependencies were added sequentially to the model. The combination of *HX* and *PF* into a single four-level categorical variable improved discrepancies between observed and expected frequencies considerably (not shown) as well as the AIC. After introducing a covariance parameter between *SBP* and *DBP*, the corresponding estimated correlation was quite high at 0.63 (Model 3), strongly indicating that *SBP* and *DBP* were not independent after conditioning on class. For the remaining continuous variables, introduction of further covariance parameters produced marked improvements in AIC (models 4 to 7), but the resulting correlations were small (<0.25).

After establishing dependencies between manifest variables the expected means and frequencies were examined across classes to identify variables that did not discriminate between classes. Figure 3.3 depicts the densities and overlaid continuous mixtures from Model 2, in which all continuous manifest variables were assumed to be conditional independent. Clearly, the means for the two classes are very close for *Age*, *SBP* and *DBP* indicating that these variables do not discriminate between classes. This was not changed by the incorporation of dependencies in later models, and from Model 8 onwards *Age*, *SBP* and *DBP* were removed as manifest variables since the corresponding estimated mean differences between classes were very small at 0.03 years, 0.04 kPa (0.30 mmHg) and 0.25 kPa (1.88 mmHg) respectively.

Inclusion of *Age*, *SBP* and *DBP* in turn as latent class predictors (models 9, 10 and 11) did not improve the AIC, and as a result the LCR submodel was left as an intercept only model, implying class prevalences were unaffected by covariates.

With treatment and a latent class effect already included in the time-to-event submodel, *Age* and a class-by-treatment effect were added and both were found to considerably improve the AIC (models 12 and 13). The inclusion of other variables and two-way interactions did not improve the AIC. Note that variables retained as manifest variables were not considered as candidates for the time-to-event submodel.

The estimated non-parametric baseline hazard from Model 13 (Figure 3.6) was used to guide the choice of time grid for a piecewise exponential baseline hazard. From Figure 3.6 a constant hazard appeared reasonable, suggesting an exponential model would be appropriate (Model 14). Adding partitions at 30 days (Model 15) and 15 days (Model 16), to capture the apparent deviation from the fitted straight line in Figure 3.6, did not improve the AIC and as a result Model 14 was selected as the final model.

B.3.2 Sensitivity of results to inclusion/exclusion of time-to-event submodel

Table B.1 gives the parameter estimates from the final selected model, compared with the same model excluding a time-to-event submodel. Parameters estimates are almost completely unaffected by the time-to-event submodel.

Parameter	Model 14	LCR
$\hat{\kappa}_2$	-1.41	-1.40
$\hat{\pi}_{11}$	0.05	0.05
$\hat{\pi}_{21}$	0.02	0.02
$\hat{\pi}_{31}$	0.40	0.41
$\hat{\pi}_{41}$	0.06	0.05
$\hat{\pi}_{12}$	0.61	0.62
$\hat{\pi}_{22}$	0.14	0.14
$\hat{\pi}_{32}$	0.26	0.26
$\hat{\pi}_{42}$	0.07	0.08
$\hat{\mu}_{11}$	-0.32	-0.32
$\hat{\mu}_{21}$	6.27	6.26
$\hat{\mu}_{31}$	13.7	13.71
$\hat{\mu}_{41}$	3.14	3.13
$\hat{\mu}_{51}$	0.09	0.09
$\hat{\mu}_{12}$	3.04	3.01
$\hat{\mu}_{22}$	7.19	7.22
$\hat{\mu}_{32}$	12.25	12.24
$\hat{\mu}_{42}$	4.69	4.69
$\hat{\mu}_{52}$	-0.37	-0.37
$\hat{\sigma}_1^2$	0.89	0.91
$\hat{\sigma}_2^2$	2.07	2.06
$\hat{\sigma}_3^2$	3.42	3.41
$\hat{\sigma}_4^2$	2.08	2.08
$\hat{\sigma}_5^2$	0.96	0.96
$\hat{\sigma}_{12}$	0.17	0.16
$\hat{\sigma}_{24}$	0.51	0.50

Table B.1: The Model 14 non-time-to-event parameter estimates compared with a latent class model with no time-to-event submodel. Parameter estimates are almost completely unaffected by the inclusion/exclusion of a time-to-event submodel.

Appendix C

Supplementary material to accompany Chapter 5

C.1 Associating biplot axes: some exploratory results

A simulated data set, similar to that used in Section 5.3, is now used to demonstrate the influence of various factors on how well parameter estimates can be recovered from a biplot, by associating biplot axes, using the methods outlined in Section 5.6. The factors examined are: MDS type (metric, classical), the dissimilarity metric (Euclidean, Gower) and the weights used in the dissimilarity metric (held at 1 for continuous variables; 0, 0.5 or 1 for categorical variables). The number of fitted dimensions was varied from 2 to 6.

Values for $P = 6$ variables, $X_1, \ldots, X_6$, and $N = 400$ observations were simulated from a multivariate Normal distribution for which all variables had a mean of 0 and standard deviation of 1; correlation coefficients were $r_{X_1,X_2} = r_{X_3,X_4} = r_{X_5,X_6} = 0.4$, and otherwise 0. Observations for variables X_5 and X_6 were subsequently dichotomised at the mean, to give four continuous variables and two categorical variables in total.

Event times were simulated from an exponential distribution, $T_i \sim \text{Exp}\left[\exp(\boldsymbol{\beta}^\intercal \mathbf{x}_i)\right]$, for $i = 1, \ldots, N$ and with $\beta_1 = \beta_2 = 0.75$ and $\beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$, i.e.

$$\log(t_i) = 0.75x_{1i} + 0.75x_{2i} + \epsilon_i,$$

where $\epsilon_1, \ldots, \epsilon_N$ are i.i.d. according to a Gumbel distribution. Event times were additionally censored by generating censoring times using an exponential distribution with a scale parameter of 1, resulting in 52% of observations being censored.

The results are shown in Figure C.1 and Figure C.2. w in the figure titles refers to

the weight of the categorical variables. Figure C.1(a), (c) and (e) show the estimated parameters recovered from the MDS biplots using the Euclidean distance, classical scaling and varying the weight of the categorical variables. Whilst β_1 and β_2 are overestimated in low dimensions, the results are sensible as they approach the correct values as the number of dimensions increases. This is what was seen in the simulated example in Section 5.9. When Gower's coefficient is used however, Figure C.1(b), (d) and (f), the estimates for β_1 and β_2 were poor, as was seen in Section 5.10. In Figure C.2, where metric instead of classical MDS was used, estimates when using Euclidean distances were overestimated, (a), (c) and (e), to a greater extent than with classical scaling. Estimates for Gower's coefficient, Figure C.2(b), (d) and (f) improved compared with classical scaling, but were still poor.

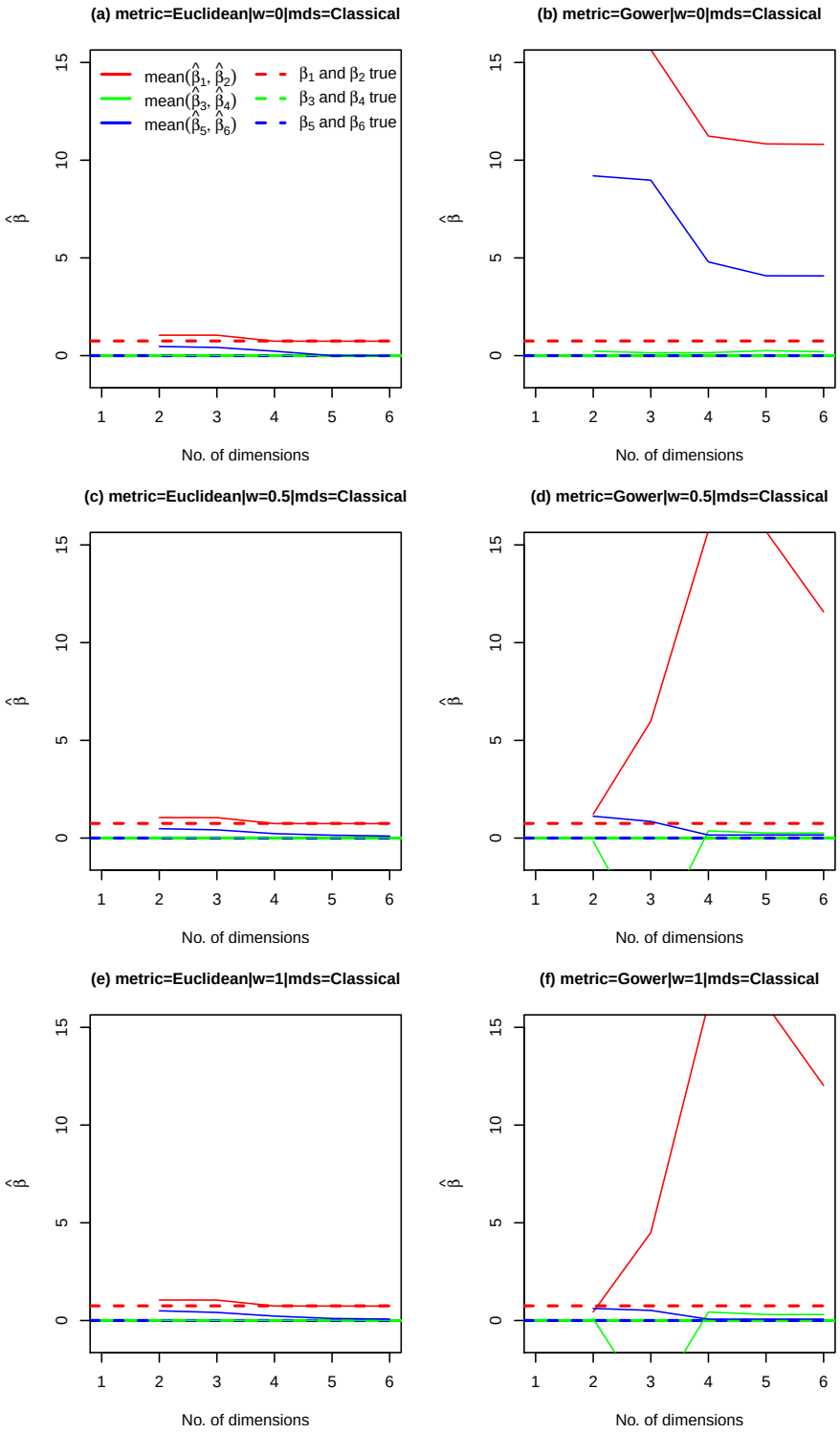

Figure C.1: Classical scaling results for associating biplot axis scales to estimate each $\hat{\beta}$. Euclidean distances (left) and Gower's coefficient (right) were used to obtain dissimilarities and weights for categorical variables (X_5 and X_6) were varied: 0 (top row), 0.5 (middle row) and 1 (bottom row). Results have been averaged over relevant variable pairs for simplicity.

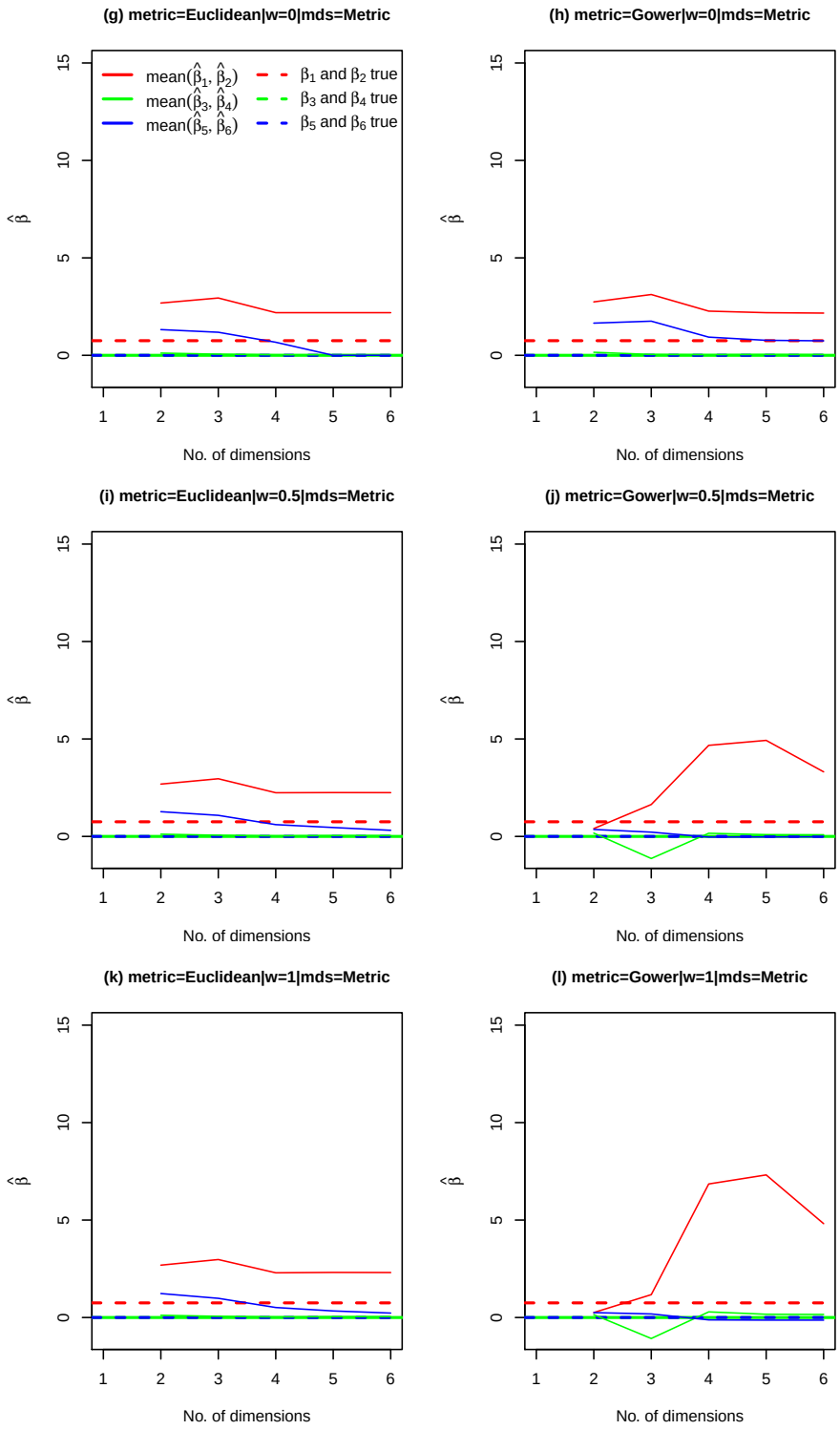

Figure C.2: Metric scaling results for associating biplot axis scales to estimate each $\hat{\beta}$. Euclidean distances (left) and Gower's coefficient (right) were used to obtain dissimilarities and weights for categorical variables (X_5 and X_6) were varied: 0 (top row), 0.5 (middle row) and 1 (bottom row). Results have been averaged over relevant variable pairs for simplicity.

Appendix D

Supplementary material to accompany Chapter 7

D.1 Data generating models for the simulation study

The following data generating models from Witten and Tibshirani (2011) were used for the simulation study in Section 7.3.

D.1.1 Two-sided model

For the two-sided data-generating model, observations in group 2 fall into two distinct groups:

$$\mathbf{x}_i \sim \begin{cases} N(\mathbf{0}, \mathbf{I}_P) & \text{if } y_i = 1, \\ N(\mathbf{1}, \mathbf{I}_P) \text{ or } N(-\mathbf{1}, \mathbf{I}_P) & \text{with equal probability if } y_i = 2, \end{cases}$$

where $\mathbf{0}$ and $\mathbf{1}$ P-length vectors with each element equal to 0 or 1, respectively, and $\mathbf{I}_P$ is the $P \times P$ identity matrix.

D.1.2 Linear model

For the linear data-generating model, data are generated with a linear trend as a function of the observation index, i):

$$\mathbf{x}_i \sim N\left(\frac{\mathbf{3i}}{\mathbf{N}}, \mathbf{I}_P\right),$$

where $\frac{\mathbf{3i}}{\mathbf{N}}$ is a P-length vector with elements $\frac{3i}{N}$. For observations $1, \ldots, N/2$, $y_i = 1$ and otherwise $y_i = 2$.

D.1.3 Constant model

For the constant mean data-generating model, there is a constant mean for each group:

$$\mathbf{x}_i \sim \begin{cases} N(-\mathbf{0.4}, \mathbf{I}_P) & \text{if } y_i = 1, \\ N(\mathbf{0.4}, \mathbf{I}_P) & \text{if } y_i = 2, \end{cases}$$

where $\mathbf{0.4}$ is a P-length vector which each element equal to 0.4.

D.2 A simulated non-linear continuous outcome variable

A simulated data set is now used to demonstrate that OTMDS can, in principle, be used for prediction of a continuous outcome variable. For $N = 130$ observations, an $N \times 5$ matrix of uncorrelated predictor variables, $\mathbf{x}$, were simulated from a standard normal distribution and two scenarios were considered: 1) $\mathbf{y}_1 = \text{sine}[\frac{1}{2}(\mathbf{x}_1 + \mathbf{x}_2)]$ and 2) $\mathbf{y}_2 = \text{sine}[2(\mathbf{x}_1 + \mathbf{x}_2)]$, as depicted in Figure D.1(a) and (b). These scenarios were chosen to demonstrate OTMDS when the outcome variable is non-linearly related to the predictors, but also the scenarios respectively represent low and high levels of monotonicity. The data set was then randomly partitioned into independent training and test sets, with 100 and 30 observations respectively, and OTMDS applied with $\alpha = 0.8$, Figure D.1(c) and (d).

Scenario 1 is depicted in Figure D.2, and plots (a) and (b) show that the predicted values from both the training and test data sets, $\hat{y}_1$, are strongly linearly related to the observed value, y_1 (Pearson's correlation coefficient, $r \geq 0.80$). However, the standardised residual plots, Figure D.2 (c) and (d) show that TMDS underestimates and overestimates the true values at the tails i.e. where the relationship is most non-linear and non-monotonic. Scenario 2, where the non-linearity and non-monotonicity is more pronounced demonstrates OTMDS performing poorly.

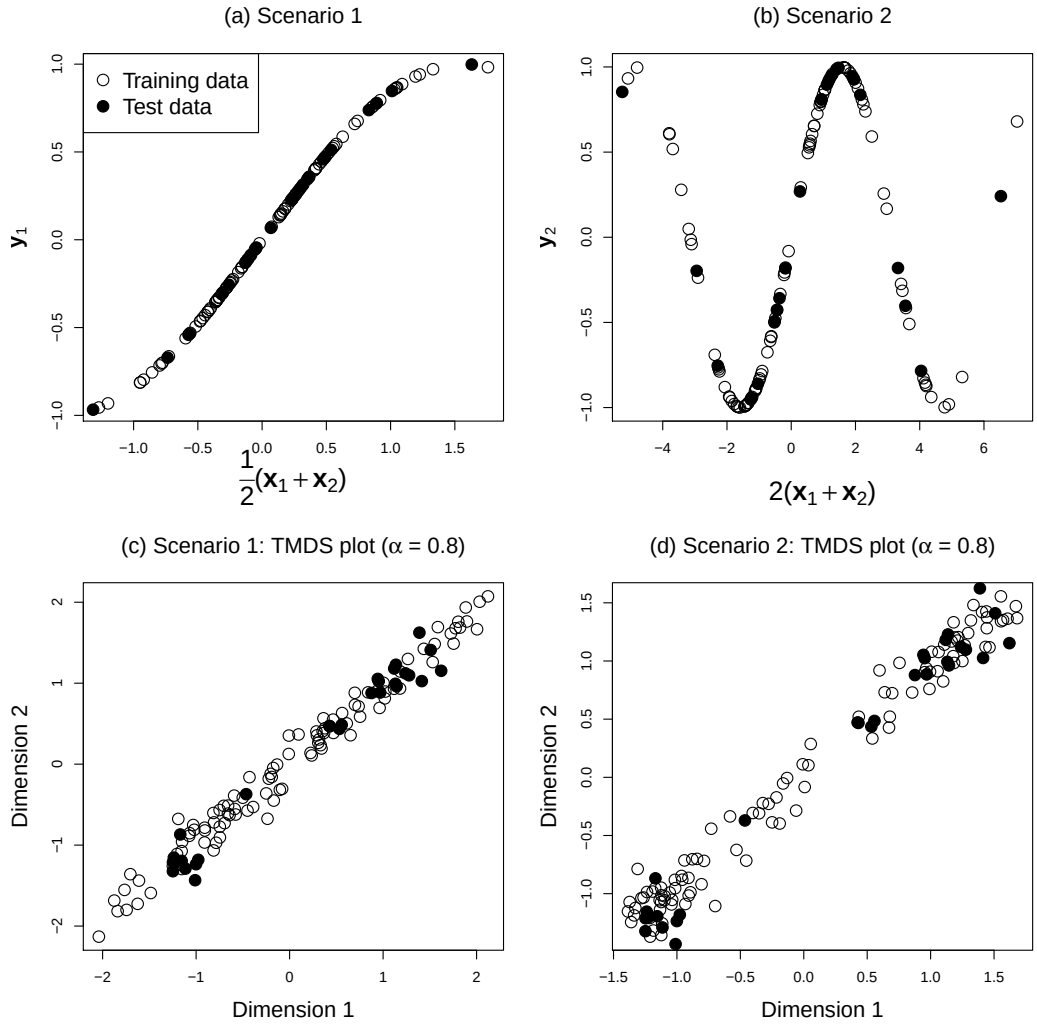

Figure D.1: Top: Plots of the relationship between $\mathbf{x}$ and $\mathbf{y}$ for both the training and test data in two simulated scenarios. Bottom: OTMDS plots for the two scenarios with $\alpha = 0.8$.

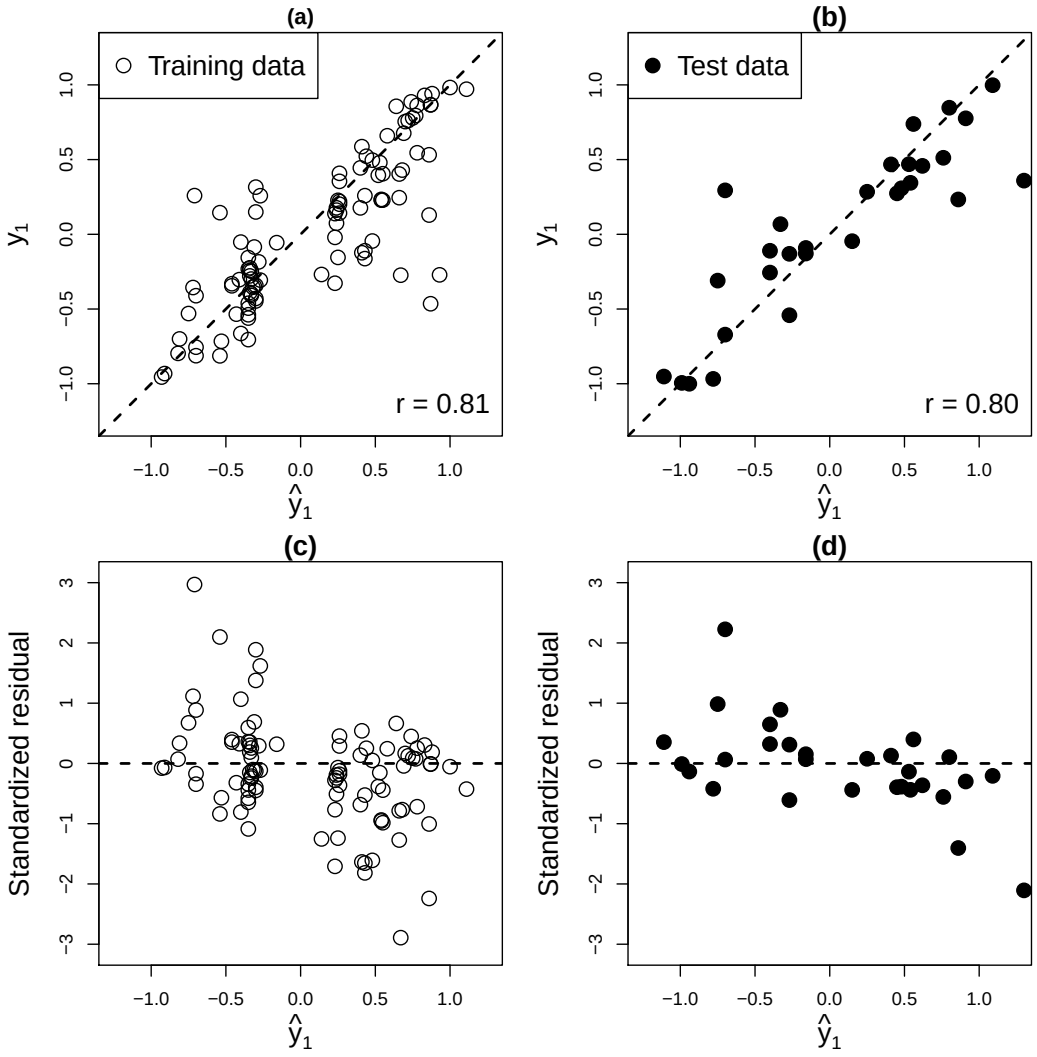

Figure D.2: Top: Plots of predicted versus actual outcome variable values in Scenario 1, where $\mathbf{y}_1 = \text{sine}[\frac{1}{2}(\mathbf{x}_1 + \mathbf{x}_2)]$ show high linear correlation. Diagonal dashed lines are lines of inequality. Bottom: Predicted values versus standardised residual plots illustrate a non-random pattern, with large residuals tending to occur in the tails.

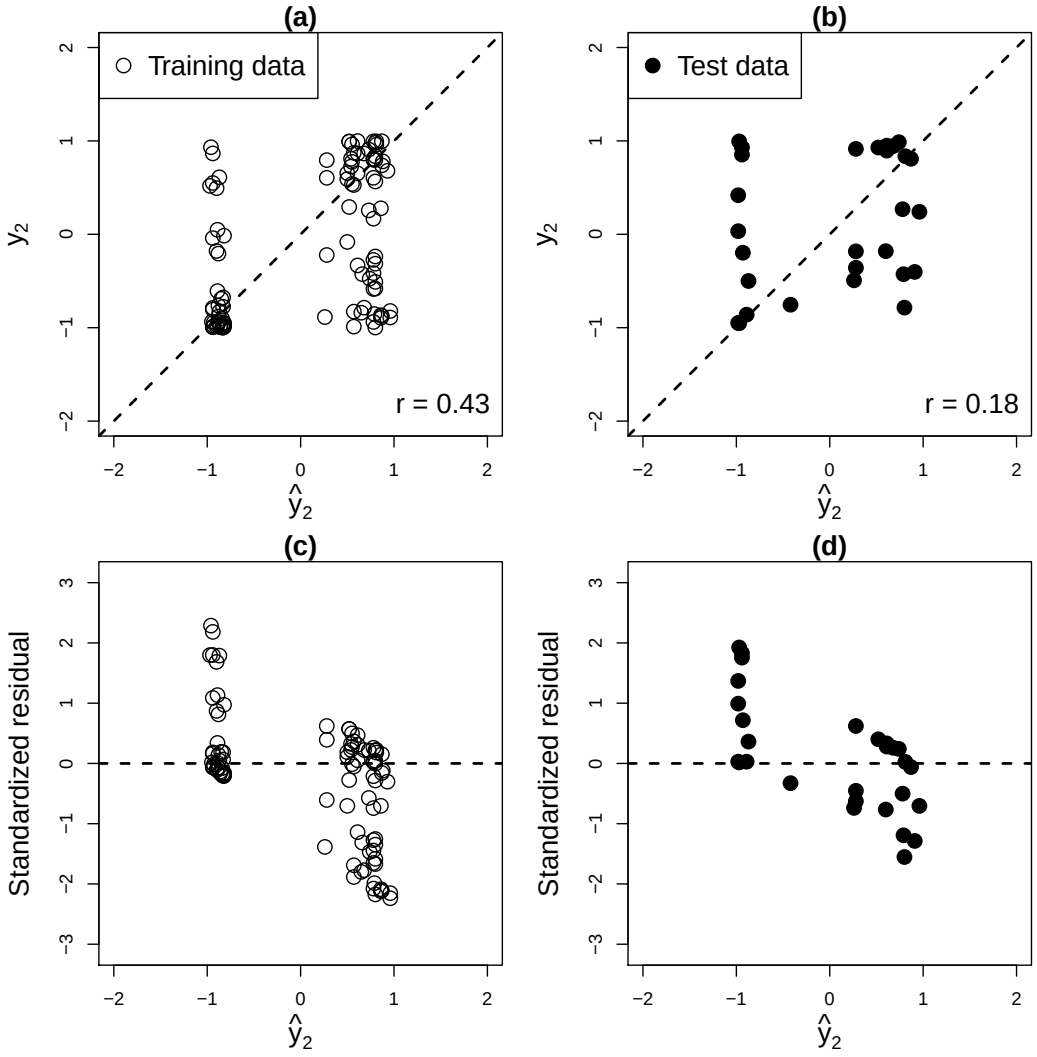

Figure D.3: Top: Plots of predicted versus actual outcome variable values in Scenario 1, where $\mathbf{y}_1 = \text{sine}[2(\mathbf{x}_1 + \mathbf{x}_2)]$ show low linear correlation. Diagonal dashed lines are lines of inequality. Bottom: Predicted values versus standardised residual plots illustrate a non-random pattern, with large residuals tending to occur in the tails.

Bibliography

Aaronson, N. K., Ahmedzai, S., Bergman, B., Bullinger, M., Cull, A., Duez, N. J., Filiberti, A., Flechtner, H., Fleishman, S. B., de Haes, J. C., et al. (1993). The European organization for research and treatment of cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. *JNCI: Journal of the National Cancer Institute*, 85(5):365–376.

Agresti, A. (2002). *Categorical data analysis*. Springer, Berlin Heidelberg.

Asparouhov, T., Masyn, K., and Muthén, B. (2006). Continuous time survival in latent variable models. In *Proceedings of the Joint Statistical Meeting in Seattle*, pages 180–187.

Asparouhov, T. and Muthén, B. (2014). Auxiliary variables in mixture modeling: Three-step approaches using Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(3):329–341.

Asparouhov, T. and Muthén, B. (2015). Residual associations in latent class and latent transition analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 22(2):169–177.

Auer, M. F. F., Hickendorff, M., Putten, C. M. V., Béguin, A. A., and Heiser, W. J. (2016). Multilevel latent class analysis for large-scale educational assessment data: Exploring the relation between the curriculum and students' mathematical strategies. *Applied Measurement in Education*, 29(2):144–159.

Azuero, A. (2016). A note on the magnitude of hazard ratios. *Cancer*, 122(8):1298–1299.

Bailar III, J. C., Byar, D. P., and Group, V. A. C. U. R. (1970). Estrogen treatment for cancer of the prostate. Early results with 3 doses of diethylstilbestrol and placebo. *Cancer*, 26(2):257–261.

Bair, E., Hastie, T., Paul, D., and Tibshirani, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association*, 101(473):119–137.

Bair, E. and Tibshirani, R. (2004). Semi-supervised methods to predict patient survival from gene expression data. *PLoS biology*, 2(4):0511–0522.

Bakk, Z. and Kuha, J. (2018). Two-step estimation of models between latent classes and external variables. *Psychometrika*, 83(4):871–892.

Bakk, Z., Oberski, D. L., and Vermunt, J. K. (2016). Relating latent class membership to continuous distal outcomes: Improving the LTB approach and a modified three-step implementation. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(2):278–289.

Bakk, Z., Tekle, F. B., and Vermunt, J. K. (2013). Estimating the association between latent class membership and external variables using bias-adjusted three-step approaches. *Sociological Methodology*, 43(1):272–311.

Bakk, Z. and Vermunt, J. K. (2016). Robustness of stepwise latent class modeling with continuous distal outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(1):20–31.

Bandeen-Roche, K., Miglioretti, D. L., Zeger, S. L., and Rathouz, P. J. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association*, 92(440):1375–1386.

Bartholomew, D. J., Knott, M., and Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*, volume 904. John Wiley & Sons, Hoboken, New Jersey.

Bender, R., Augustin, T., and Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, 24(11):1713–1723.

Boj, E., Claramunt Bielsa, M. M., and Fortiana, J. (2007a). Selection of predictors in distance-based regression. *Communications in Statistics—Simulation and Computation*, 36(1):87–98.

Boj, E., Delicado, P., and Fortiana, J. (2010). Distance-based local linear regression for functional predictors. *Computational Statistics & Data Analysis*, 54(2):429–437.

Boj, E., Delicado, P., Fortiana Gregori, J., Esteve, A., and Caballé, A. (2012). Local distance-based generalized linear models using the dbstats package for R. https://ssrn.com/abstract=2063822 [Online; accessed May-2019].

Boj, E., Grané, A., Fortiana, J., and Claramunt, M. M. (2007b). Implementing PLS for distance-based regression: computational issues. *Computational Statistics*, 22(2):237–248.

Bolck, A., Croon, M., and Hagenaars, J. (2004). Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political Analysis*, 12(1):3–27.

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology*, 53(1):605–634. PMID: 11752498.

Borg, I. and Groenen, P. (2003). Modern multidimensional scaling: Theory and applications. *Journal of Educational Measurement*, 40(3):277–280.

Borg, I., Groenen, P. J., and Mair, P. (2012). *Applied multidimensional scaling*. Springer Science & Business Media.

Borg, I. and Lingoes, J. C. (1980). A model and algorithm for multidimensional scaling with external constraints on the distances. *Psychometrika*, 45(1):25–38.

Bøvelstad, H. M., Nygård, S., and Borgan, Ø. (2009). Survival prediction from clinico-genomic models - a comparative study. *BMC Bioinformatics*, 10(1):413.

Bøvelstad, H. M., Nygård, S., Størvold, H. L., Aldrin, M., Borgan, Ø., Frigessi, A., and Lingjærde, O. C. (2007). Predicting survival from microarray data — a comparative study. *Bioinformatics*, 23(16):2080–2087.

Bray, B. C., Lanza, S. T., and Tan, X. (2015). Eliminating bias in classify-analyse approaches for latent class analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 22(1):1–11.

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics*, 30(1):89–99.

Buja, A., Swayne, D. F., Littman, M. L., Dean, N., Hofmann, H., and Chen, L. (2008). Data visualization with multidimensional scaling. *Journal of Computational and Graphical Statistics*, 17(2):444–472.

Burton, A., Altman, D. G., Royston, P., and Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, 25(24):4279–4292.

Byar, D. P. and Corle, D. K. (1977). Selecting optimal treatment in clinical trials using covariate information. *Journal of Chronic Diseases*, 30(7):445–459.

Cancer Research UK (2017). Common cancers. https://www.cancerresearchuk.org/health-professional/cancer-statistics/mortality/common-cancers-comparedheading-Zero [Online; accessed October-2019].

Cancer Research UK (2018). Liver cancer. https://www.cancerresearchuk.org/about-cancer/liver-cancer/survival [Online; accessed October-2019].

Carpenter, J. and Bithell, J. (2000). Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine*, 19(9):1141–1164.

Carroll, J. D. and Green, P. E. (1997). Psychometric methods in marketing research: Part II, multidimensional scaling. *Journal of Marketing Research*, 34(2):193–204.

Celeux, G., Hurn, M., and Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95(451):957–970.

Cha, J.-E., Kim, S., and Lee, Y. (2009). Application of multidimensional scaling for marketing-mix modification: A case study on mobile phone category. *Expert Systems with Applications*, 36(3):4884–4890.

Chan, P. H., Xu, R., and Chambers, C. D. (2018). A study of R^2 measure under the accelerated failure time models. *Communications in Statistics - Simulation and Computation*, 47(2):380–391.

Chiou, S. H., Kang, S., and Yan, J. (2014). Fitting accelerated failure time models in routine survival analysis with R package aftgee. *Journal of Statistical Software*, 61(11):1–23.

Chung, H., Flaherty, B. P., and Schafer, J. L. (2006). Latent class logistic regression: Application to marijuana use and attitudes among high school seniors. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(4):723–743.

Clark, S. L. and Muthén, B. (2009). Relating latent class analysis results to variables not included in the analysis. https://www.statmodel.com/download/relatinglca.pdf [Online; accessed March-2017].

Clogg, C. (2013). Latent class models. In Arminger, G., Clogg, C. C., and Sobel, M. E., editors, *Handbook of statistical modeling for the social and behavioral sciences*, chapter 6. Springer Science & Business Media.

Collett, D. (2015). *Modelling survival data in medical research (2nd ed.)*. CRC press, Boca Raton, Florida.

Collier, Z. K. and Leite, W. L. (2017). A comparison of three-step approaches for auxiliary variables in latent class and latent profile analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 24(6):819–830.

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

Cox, T. F. and Cox, M. A. (2000). *Multidimensional scaling*. Chapman and hall/CRC, Boca Raton, Florida.

Cox, T. F. and Ferry, G. (1993). Discriminant analysis using non-metric multidimensional scaling. *Pattern Recognition*, 26(1):145–153.

Cuadras, C., Areans, C., and Fortiana, J. (1996). Some computational aspects of a distance—based model for prediction. *Communications in Statistics-Simulation and Computation*, 25(3):593–609.

Cuadras, C. and Arenas, C. (1990). A distance based regression model for prediction with mixed data. *Communications in Statistics-Theory and Methods*, 19(6):2261–2279.

Cuadras, C. M. (1989). Distance analysis in discrimination and classification using both continuous and categorical variables. In *Statistical data analysis and inference*, pages 459–473. Elsevier Science, North Holland, Amsterdam.

Dayton, C. M. and Macready, G. B. (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association*, 83(401):173–178.

de Leeuw, J. and Mair, P. (2009). Multidimensional scaling using majorization: SMACOF in R. *Journal of Statistical Software*, 31(3):1–30.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Denson, N. and Ing, M. (2014). Latent class analysis in higher education: An illustrative example of pluralistic orientation. *Research in Higher Education*, 55(5):508–526.

Desantis, S. M., Andrés Houseman, E., Coull, B. A., Nutt, C. L., and Betensky, R. A. (2012). Supervised Bayesian latent class models for high-dimensional data. *Statistics in Medicine*, 31(13):1342–1360.

Downing, A., Harrison, W. J., West, R. M., Forman, D., and Gilthorpe, M. S. (2010). Latent class modelling of the association between socioeconomic background and breast cancer survival status at 5 years incorporating stage of disease. *Journal of Epidemiology & Community Health*, 64(9):772–776.

Dziak, J. J., Lanza, S. T., and Tan, X. (2014). Effect size, statistical power, and sample size requirements for the bootstrap likelihood ratio test in latent class analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(4):534–552.

Esteve, A., Boj, E., and Fortiana, J. (2009). Interaction terms in distance-based regression. *Communications in Statistics—Theory and Methods*, 38(19):3498–3509.

Faraway, J. J. (2014). Regression for non-Euclidean data using distance matrices. *Journal of Applied Statistics*, 41(11):2342–2357.

Fieller, E. C. (1940). The biological standardization of insulin. *Supplement to the Journal of the Royal Statistical Society*, 7(1):1–64.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38.

Formann, A. K. (1992). Linear logistic latent class analysis for polytomous data. *Journal of the American Statistical Association*, 87(418):476–486.

Fox, R., Berhane, S., Teng, M., Cox, T., Tada, T., Toyoda, H., Kumada, T., Kagebayashi, C., Satomura, S., and Johnson, P. (2014). Biomarker-based prognosis in hepatocellular carcinoma: validation and extension of the BALAD model. *British Journal of Cancer*, 110(8):2090.

Franz, V. H. (2007). Ratios: A short guide to confidence limits and proper use. arXiv preprint arXiv:0710.2024, https://arxiv.org/pdf/0710.2024.pdf [Online; accessed April-2019].

Freije, W. A., Castro-Vargas, F. E., Fang, Z., Horvath, S., Cloughesy, T., Liau, L. M., Mischel, P. S., and Nelson, S. F. (2004). Gene expression profiling of gliomas strongly predicts survival. *Cancer Research*, 64(18):6503–6510.

Friedman, M. (1982). Piecewise exponential models for survival data with covariates. *The Annals of Statistics*, 10(1):101–113.

Fuller, G. N., Hess, K. R., Rhee, C. H., Yung, W. A., Sawaya, R. A., Bruner, J. M., and Zhang, W. (2002). Molecular classification of human diffuse gliomas by multidimensional scaling analysis of gene expression profiles parallels morphology-based classification, correlates with survival, and reveals clinically-relevant novel glioma subsets. *Brain pathology*, 12(1):108–116.

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

Gleason, D. F. and Mellinger, G. T. (1974). Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging. *The Journal of urology*, 111(1):58–64.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3-4):325–338.

Gower, J. C. (1968). Adding a point to vector diagrams in multivariate analysis. *Biometrika*, 55(3):582–585.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 27(4):857–871.

Gower, J. C. and Hand, D. J. (1995). *Biplots*, volume 54. CRC Press, Boca Raton, Florida.

Greenacre, M. J. (2010). *Biplots in practice*. Fundacion BBVA.

Grün, B. and Leisch, F. (2009). Dealing with label switching in mixture models under genuine multimodality. *J. Multivar. Anal.*, 100(5):851–861.

Hagenaars, J. A. (1988). Latent structure models with direct effects between indicators: Local dependence models. *Sociological Methods & Research*, 16(3):379–405.

Han, G., Schell, M. J., and Kim, J. (2014). Improved survival modeling in cancer research using a reduced piecewise exponential approach. *Statistics in Medicine*, 33(1):59–73.

Harrell, F. E. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer, New York.

Harrell, F. E., J., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. (1982). Evaluating the Yield of Medical Tests. *JAMA*, 247(18):2543–2546.

Hastie, T., Tibshirani, R., and Friedman, J. (2003). *Elements of statistical learning: data mining, inference, and prediction*. Springer, New York.

Heinze, G. (2018). *Firth's bias-reduced logistic regression*. version 1.23.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Hougaard, P. (1999). Fundamentals of survival data. *Biometrics*, 55(1):13–22.

Hsieh, F., Tseng, Y. K., and Wang, J. L. (2006). Joint modeling of survival and longitudinal data: Likelihood approach revisited. *Biometrics*, 62(4):1037–1043.

Huang, G. H. and Bandeen-Roche, K. (2004). Building an identifiable latent class model with covariate effects on underlying and measured variables. *Psychometrika*, 69(1):5–32.

Hunt, L. and Jorgensen, M. (1999). Theory & methods: Mixture model clustering using the MULTIMIX program. *Australian & New Zealand Journal of Statistics*, 41(2):154–171.

Jamróz, D. (2014). Application of multidimensional scaling to classification of various types of coal. *Archives of Mining Sciences*, 59(2):413–425.

Johnson, P. J., Pirrie, S. J., Cox, T. F., Berhane, S., Teng, M., Palmer, D., Morse, J., Hull, D., Patman, G., Kagebayashi, C., et al. (2014). The detection of hepatocellular carcinoma using a prospectively developed and validated model based on serological biomarkers. *Cancer Epidemiology and Prevention Biomarkers*, 23(1):144–153.

Jorgensen, M. and Hunt, L. (1996). Mixture model clustering of data sets with categorical and continuous variables. In *Proceedings of the Conference ISIS*, volume 96, pages 375–384. World Scientific.

Kalbfleisch, J. D. and Prentice, R. L. (2011). *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, Hoboken, New Jersey.

Korn, E. L. and Simon, R. (1990). Measures of explained variation for survival data. *Statistics in Medicine*, 9(5):487–503.

Lanza, S. T., Tan, X., and Bray, B. C. (2013). Latent class analysis with distal outcomes: A flexible model-based approach. *Structural Equation Modeling: A Multidisciplinary Journal*, 20(1):1–26.

Larsen, K. (2004). Joint analysis of time-to-event and multiple binary indicators of latent classes. *Biometrics*, 60(1):85–92.

Lazarsfeld, P. F. (1959). *Latent structure analysis. In S. Koch (Ed.) Psychology: A study of a science. Conceptual and systematic.*, volume 3. McGrawHill, New York.

Leigh, L., Hudson, I. L., and Byles, J. E. (2015). Sleeping difficulty, disease and mortality in older women: A latent class analysis and distal survival analysis. *Journal of sleep research*, 24(6):648–657.

Linzer, D. and Lewis, J. (2011). poLCA: An R package for polytomous variable latent class analysis. *Journal of Statistical Software, Articles*, 42(10):1–29.

Little, R. J. (1992). Regression with missing x's: a review. *Journal of the American Statistical Association*, 87(420):1227–1237.

Liu, E. and Lim, K. (2018). Using the Weibull accelerated failure time regression model to predict time to health events. *BioRxiv*, page 362186.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):226–233.

Lythgoe, D. T., Garcia-Fiñana, M., and Cox, T. F. (2019). Latent class modeling with a time-to-event distal outcome: A comparison of one, two and three-step approaches. *Structural Equation Modeling: A Multidisciplinary Journal*, 26(1):51–65.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2019). *cluster: cluster analysis basics and extensions*. R package version 2.1.0.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate analysis*. Academic press, London.

Martin, Y. C., Lin, C. T., Hetti, C., and DeLazzer, J. (1995). PLS analysis of distance matrixes to detect nonlinear relationships between biological potency and molecular properties. *Journal of Medicinal Chemistry*, 38(16):3009–3015.

McLachlan, G. and Krishnan, T. (2007). *The EM algorithm and extensions*, volume 382. John Wiley & Sons, Hoboken, New Jersey.

McLachlan, G. and Peel, D. (2004). *Finite mixture models*. John Wiley & Sons, Hoboken, New Jersey.

Melo, O. O., Melo, C. E., and Mateu, J. (2015). Distance-based beta regression for prediction of mutual funds. *Advances in Statistical Analysis*, 99(1):83–106.

Melo, S. E. and Melo, O. O. (2013). Distance-based approach in univariate longitudinal data analysis. *Journal of Applied Statistics*, 40(3):674–692.

MIT (2011). Projections onto subspaces. https://ocw.mit.edu/courses/mathematics/18-06sc-linear-algebra-fall-2011/least-squares-determinants-and-eigenvalues/projections-onto-subspaces/MIT18_06SCF11_Ses2.2sum.pdf [Online; accessed June-2019].

Moustaki, I. (1996). A latent trait and a latent class model for mixed observed variables. *British Journal of Mathematical and Statistical Psychology*, 49(2):313–334.

Muthén, B., Asparouhov, T., Boye, M., Hackshaw, M., and Naegeli, A. (2009). Applications of continuous-time survival in latent variable models for the analysis of oncology randomized clinical trial data using Mplus. http://www.statmodel2.com/download/lilyFinalReportV6.pdf [Online; accessed February-2018].

Muthén, L. K. and Muthén, B. O. (2004). *Mplus Technical Appendices.* Los Angeles, California.

Muthén, L. K. and Muthén, B. O. (2011). *Mplus User's Guide. Sixth Edition.* Los Angeles, California.

Neoptolemos, J. P., Stocken, D. D., Bassi, C., Ghaneh, P., Cunningham, D., Goldstein, D., Padbury, R., Moore, M. J., Gallinger, S., Mariette, C., et al. (2010). Adjuvant chemotherapy with fluorouracil plus folinic acid vs gemcitabine following pancreatic cancer resection: A randomized controlled trial. *JAMA*, 304(10):1073–1081.

Newson, R. B. (2010). Comparing the predictive powers of survival models using Harrell's C or Somers' D. *The Stata Journal*, 10(3):339–358.

Nieto, A. B., Galindo, M. P., Leiva, V., and Vicente-Galindo, P. (2014). A methodology for biplots based on bootstrapping with R. *Revista colombiana de estadística*, 37(2):367–397.

Papageorgiou, G., Mauff, K., Tomer, A., and Rizopoulos, D. (2019). An overview of joint modeling of time-to-event and longitudinal outcomes. *Annual review of statistics and its application*.

Pawitan, Y. (2001). *In all likelihood: statistical modelling and inference using likelihood*. Oxford University Press, Oxford.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., and Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, 49(12):1373–1379.

Petersen, J., Bandeen-Roche, K., Budtz-Jørgensen, E., and Larsen, K. G. (2012). Predicting latent class scores for subsequent analysis. *Psychometrika*, 77(2):244–262.

Peugh, J. L. and Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, 74(4):525–556.

Proust-Lima, C., Philipps, V., and Liquet, B. (2017). Estimation of extended mixed models using latent classes and latent processes: The R package lcmm. *Journal of Statistical Software, Articles*, 78(2):1–56.

Proust-Lima, C., Séne, M., Taylor, J. M., and Jacqmin-Gadda, H. (2014). Joint latent class models for longitudinal and time-to-event data: A review. *Statistical Methods in Medical Research*, 23(1):74–90. PMID: 22517270.

R Core Team (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rahbar, M. H., Ning, J., Choi, S., Piao, J., Hong, C., Huang, H., Del Junco, D. J., Fox, E. E., Rahbar, E., and Holcomb, J. B. (2015). A joint latent class model for classifying severely hemorrhaging trauma patients. *BMC research notes*, 8(1):602.

Ramaswamy, V., Desarbo, W. S., Reibstein, D. J., and Robinson, W. T. (1993). An empirical pooling approach for estimating marketing mix elasticities with PIMS data. *Marketing Science*, 12(1):103–124.

Royston, P. (2006). Explained variation for survival models. *The Stata Journal*, 6(1):83–96.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons, Hoboken, New Jersey.

Sammon, J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on computers*, 100(5):401–409.

Schaid, D. J. (2010). Genomic similarity and kernel methods I: advancements by building on mathematical and statistical foundations. *Human Heredity*, 70(2):109–131.

Schemper, M. and Stare, J. (1996). Explained variation in survival analysis. *Statistics in Medicine*, 15(19):1999–2012.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Chapman and Hall/CRC, Boca Raton, Florida.

Smith, I. C. (2017). The Advanced Research Computing Condor service. http://condor.liv.ac.uk/index.html [Online; accessed November-2017].

Snuderl, M., Chi, S. N., DeSantis, S., Stemmer-Rachamimov, A., Betensky, R. A., DeGirolami, U., and Kieran, M. W. (2008). Prognostic value of tumor microinvasion and metalloproteinases expression in intracranial pediatric ependymomas. *The FASEB Journal*, 22(1 Supplement):706.8.

Sperrin, M., Jaki, T., and Wit, E. (2010). Probabilistic relabelling strategies for the label switching problem in Bayesian mixture models. *Statistics and Computing*, 20(3):357–366.

Stapinski, L. A., Edwards, A. C., Hickman, M., Araya, R., Teesson, M., Newton, N. C., Kendler, K. S., and Heron, J. (2016). Drinking to cope: a latent class analysis of coping motives for alcohol use in a large cohort of adolescents. *Prevention Science*, 17(5):584–594.

Steyerberg, E. W., Eijkemans, M. J., and Habbema, J. D. F. (1999). Stepwise selection in small data sets: A simulation study of bias in logistic regression analysis. *Journal of Clinical Epidemiology*, 52(10):935–942.

Takane, Y. (2006). 11 Applications of multidimensional scaling in psychometrics. *Handbook of statistics*, 26:359–400.

Tan, P. S. and Huang, Q. D. (2018). Hepatocellular carcinoma. https://bestpractice.bmj.com/topics/en-gb/369 [Online; accessed October-2019].

Therneau, T. and Atkinson, E. (2019). Concordance. https://cran.r-project.org/web/packages/survival/vignettes/concordance.pdf [Online; accessed August-2019].

Therneau, T. M. (2015). *A package for survival analysis in S.* version 2.38.

Tueller, S. J., Drotar, S., and Lubke, G. H. (2011). Addressing the problem of switched class labels in latent variable mixture model simulation studies. *Structural Equation Modeling: A Multidisciplinary Journal*, 18(1):110–131.

van der Heijden, P. G. M., Dessens, J., and Bockenholt, U. (1996). Estimating the concomitant-variable latent-class model with the EM algorithm. *Journal of Educational and Behavioral Statistics*, 21(3):215–229.

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, 18(4):450–469.

Von Luxburg, U. and Franz, V. H. (2009). A geometric approach to confidence sets for ratios: Fieller's theorem, generalizations and bootstrap. *Statistica Sinica*, pages 1095–1117.

Wang, C. P., Brown, C. H., and Bandeen-Roche, K. (2005). Residual diagnostics for growth mixture models. *Journal of the American Statistical Association*, 100(471):1054–1076.

Wei, L. J. (1992). The accelerated failure time model: A useful alternative to the Cox regression model in survival analysis. *Statistics in Medicine*, 11(14-15):1871–1879.

White, I. R. and Royston, P. (2009). Imputing missing covariate values for the Cox model. *Statistics in Medicine*, 28(15):1982–1998.

Witten, D. M. (2013). Package 'superMDS'. https://cran.r-project.org/web/packages/superMDS/superMDS.pdf [Online; accessed March-2019].

Witten, D. M. and Tibshirani, R. (2011). Supervised multidimensional scaling for visualization, classification, and bipartite ranking. *Computational Statistics & Data Analysis*, 55(1):789–801.

Xue, Q. and Bandeen-Roche, K. (2004). Combining complete multivariate outcomes with incomplete covariate information: A latent class approach. *Biometrics*, 58(1):110–120.

Yao, W. (2015). Label switching and its solutions for frequentist mixture models. *Journal of Statistical Computation and Simulation*, 85(5):1000–1012.

Zhang, J. J. and Wang, M. (2010). Latent class joint model of ovarian function suppression and DFS for premenopausal breast cancer patients. *Statistics in Medicine*, 29(22):2310–2324.

```
rm(list=ls())
```