Bayesian and non-Bayesian techniques applied to censored survival data with missing values

Morten Nonboe Andersen

Kongens Lyngby 2007
IMM-PHD-2007-179

Technical University of Denmark
Informatics and Mathematical Modelling
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk

IMM-PHD: ISSN 0909-3192

Summary

This thesis is a comprehensive comparative study of survival analysis methods, in particular the application of the Cox Proportional Hazards (CPH) model to real life data: A data set with 48 right-censored (end of study) patients suffering from multiple myeloma, and the COpenhagen Stroke study (COST) database with 993 right-censored (10 year follow-up) stroke patients.

The most frequently applied method, stepwise selection, is a variable selection technique that fits a single model by searching for significant predictors of the survival time in terms of p-values. However, stepwise selection ignores the between-model uncertainty. This leads to biased and overconfident estimates. We compare stepwise selection to a more advanced approach, Bayesian Model Averaging (BMA), to average over all or a subset of models weighted by their posterior model probabilities. We show how to identify a subset of models using Occam's window subset selection with results comparable to an average over all models.

We show that BMA has several advantages over stepwise selection. Using an average over models, we can evaluate the model uncertainty and obtain more reliable estimates of the risk factor coefficients. BMA also gives probabilistic evaluations of each risk factor, and we can ask questions such as: "What is the probability that this risk factor coefficient is non-zero, i.e. has an effect?" In stepwise selection, risk factors are either significant or not. We also show how to evaluate and compare the predictive power of competing models using the predictive log-score and a novel evaluation score, the predictive $\mathcal{Z}$-score. We show that BMA improves the predictive power of our models.

The CPH model is based on an assumption of proportional hazards. We implement two methods for validating this assumption. One can be used before and the other after a model has been fitted. We also show how to implement time-dependent variables and parameters to give a more general Cox regression (CR) model, and how to apply BMA on this model.

Most real-life data sets have subjects where all values have not been recorded. Standard survival analysis methods cannot handle missing values, and a lot of valuable information is lost. We present three ways to address this problem: Combining BMA and variable selection, we propose a stepwise BMA method, where variables are removed by evaluating the probability of an effect. When we remove variables with missing values, we reduce the number of subjects with missing values, and significantly increase the size of the data set, leading to more accurate parameter estimates and increased predictive power.

Bayesian Networks (BN) have been used in numerous contexts to infer missing values. We show that they are also useful for estimating the missing values in survival data sets. Having estimated the missing values, we apply BMA to the augmented data set with improved evaluation of the risk factors and increased predictive power. We compare several methods for learning the structure and the parameters of a network connecting the risk factors, and show that the best results are obtained using a structural Expectation Maximization (EM) algorithm that is able to handle missing values.

In a final approach, we use a CR model for the failure time distribution, and place fully parametric distributions on the missing data mechanisms and the risk factors. Using an EM algorithm, we iteratively estimate missing values and model parameters. In a simulation, we show how the results of this method depend on the chosen parametric distributions, but that we obtain improved evaluation of risk factors and increased predictive power, when we use the BN structure to propose a distribution on the risk factors. We also propose an improvement to the original EM algorithm by substituting stepwise selection with BMA in the M-step, leading to improved parameter and missing value estimates and increased predictive power.

Results suggest that survival time for stroke patients is lower for male patients, decreases with ageing, severity of stroke, presence of another disabling disease, diabetes, and intermittent claudication, or if the patient has previously experienced a stroke. However, the effect of stroke severity decreases with time. Some results also indicate that survival time decreases with the presence of atrial fibrillation, or if the admission body temperature is $\geq 37.0°$ C. Data showed positive evidence against an effect of hypertension, alcohol consumption, smoking habits, type of stroke, and the presence of an ischemic heart disease, when we adjusted for the possibility of the other risk factors.

Resumé

Denne afhandling er et omfattende komparativt metodestudie med særligt henblik på overlevelsesanalyse, specielt anvendelsen af en Cox Proportional Hazards model (CPH) på virkelige data: Et datasæt bestående af 48 højre-censorerede (afslutning på studiet) patienter med udbredt myelomatose (knoglemarvskræft), og COpenhagen Stroke study (COST) databasen med 993 højre-censorerede (opfølgning efter 10 år) slagtilfælde patienter.

Sædvanligvis anvendes stepwise selection til at evaluere, ved hjælp af p-værdier, hvilken variabel, der vil forbedre modellen mest, hvis den til-/fravælges. Metoden tager hensyn til usikkerheden på parameterestimaterne, men ikke usikkerheden modellerne imellem. Dette medfører biased og overkonfidente estimater. Vi sammenligner metoden med Bayesian Model Averaging (BMA) til at beregne et gennemsnit over alle modeller eller en delmængde heraf. Hver model vægtes med modellens a posteriori sandsynlighed. Vi viser, hvordan man kan identificere en delmængde af modeller ved hjælp af Occam's window subset selection med resultater, der er sammenlignelige med et gennemsnit over alle modeller.

Ved at benytte et gennemsnit over modeller kan vi evaluere modelusikkerheden og opnå mere pålidelige estimater af risikofaktorernes koefficienter. BMA giver også en probabilistisk evaluering af den enkelte risikofaktor, der giver os mulighed for at stille spørgsmål som: "Hvad er sandsynligheden for, at koefficienten for denne risikofaktor er nul, dvs. har en effect?". Vi viser også, at BMA forbedrer modellens prædiktive evne evalueret ved hjælp af den prædiktive log-score og en ny score, den prædiktive $\mathscr{Z}$-score.

Vi beskriver desuden metoder til at validere CPH modellens antagelse om proportionale hazards og implementere tidsafhængige variable og parametre til at opnå en mere generel Cox regressionsmodel (CR).

Endelig kan mange overlevelsesanalysemetoder ikke håndtere manglende værdier, og dermed går store mængder af værdifuld information ofte tabt. Ved at anvende en kombination af BMA og stepwise selection foreslår vi en stepwise BMA metode, hvor variable fjernes på baggrund af sandsynligheden for en effekt. Når vi fjerner variable med manglende værdier, reducerer vi antallet af patienter, der har manglende værdier. Vi viser, at denne metode kan øge størrelsen af datasættet markant og føre til mere præcise parameter estimater og styrket prædiktionsevne.

Vi demonstrerer desuden, hvordan Bayesianske Netværk (BN) kan anvendes til at estimere de manglende værdier, hvorefter vi anvender BMA på det udvidede datasæt og viser forbedringer i evalueringen af risikofaktorerne og øget prædiktionsevne. Vi sammenligner forskellige metoder til at lære strukturen og parametrene i det netværk, der forbinder risikofaktorerne og viser, at de bedste resultater opnås ved at benytte en strukturel Expectation Maximization (EM) algoritme.

Endelig anvendes parametriske fordelinger til at modellere sammenhænge mellem risikofaktorerne og de mekanismer, der resulterer i manglende værdier, mens vi anvender en CR model til at modellere fordelingen af levetiderne. Ved at benytte en EM algoritme kan vi skiftevis estimere parametre og manglende værdier og opnå forbedringer i evalueringen af risikofaktorerne samt øget prædiktionsevne ved at anvende BN strukturen til at modellere sammenhængen mellem risikofaktorerne. Ved at erstatte stepwise selection med BMA i den originale algoritmes M-skridt vises, at vi kan opnå bedre estimater af parametre og manglende værdier samt en styrket prædiktionsevne.

Resultaterne viser, at levetiden for slagtilfældepatienter er kortere for mænd, falder med alderen, slagtilfældets sværhedsgrad, tilstedeværelsen af anden invaliderende sygdom, sukkersyge og forbigående krampe i benene, eller hvis patienten tidligere har haft et slagtilfælde. Effekten af slagtilfældets sværhedsgrad aftager dog med tiden.

Den forventede levetid falder muligvis også med tilstedeværelsen af hjerteflimmer, eller hvis patientens kropstemperatur er $\geq 37.0^\circ$ C ved indlæggelse, men data kunne ikke entydigt påvise disse effekter. Endelig viste analyserne, at der ikke var bevis for en effekt af levetiden ved rygning, indtagelse af alkohol, højt blodtryk, typen af slagtilfælde eller tilstedeværelsen af en iskæmisk hjertesygdom, når vi korrigerede for de øvrige risikofaktorer.

Preface

This thesis was prepared at Informatics and Mathematical Modeling, the Technical University of Denmark in partial fulfillment of the requirements for acquiring the Ph.D. degree in engineering.

The work was funded by the Technical University of Denmark and was supervised by Associate Professor Ole Winther.

The work commenced May 15, 2004 and was completed April 15, 2007.

From January - September 2004, I studied under the supervision of Assistant Professor Nancy E. Reed at the University of Hawai'i.

Lyngby, 15-04-2007

Morten Nonboe Andersen

Publications

Papers included in the thesis

[B] M.N Andersen, K.K. Andersen, L.P. Kammersgaard, and T.S. Olsen. Sex Differences in Stroke Survival: 10-Year Follow-up of the Copenhagen Stroke Study Cohort. *Journal of Stroke and Cerebrovascular Diseases*, 2005: Vol. 14, No. 5 (September-October), pp 215-220.

Other journal papers or conference contributions published during the preparation of the thesis

- M.N. Andersen, K.K. Andersen, L.P. Kammersgaard, and T.S. Olsen. Gender Differences in Stroke Survival: 10-Year Follow-up of the Copenhagen Stroke Study Cohort. *14'th European Stroke Conference. Bologna*, 2005 (May 25-28).

- H.G. Petersen, K.K. Andersen, M.N. Andersen, L.P. Kammersgaard, and T.S. Olsen. Body Mass Index (BMI) and survival after stroke. *Joint World Congress on Stroke. Cape Town, South Africa*, 2006 (October 26-29).

- M.N. Andersen, K.K. Andersen, H.G. Petersen, and T.S. Olsen. Using Bayesian statistics to account for model uncertainty in survival analysis. A study of risk factors in 25 839 patients with acute stroke. *Joint World Congress on Stroke. Cape Town, South Africa*, 2006 (October 26-29).

- K.K. Andersen, T.S. Olsen, H.G. Petersen, and M.N. Andersen. On the importance of a stroke severity score in modelling mortality of stroke patients. *Joint World Congress on Stroke. Cape Town, South Africa*, 2006 (October 26-29).

- K.K. Andersen, H.G. Petersen, M.N. Andersen, L.P. Kammersgaard, and T.S. Olsen. Stroke in diabetics: Frequency, clinical characteristics and survival. A 4-year follow-up study of 24 121 patients with acute stroke. *Joint World Congress on Stroke. Cape Town, South Africa*, 2006 (October 26-29).

- K.K. Andersen, L.P. Kammersgaard, H.G. Petersen, M.N. Andersen, and T.S. Olsen. Intracerebral hematomas versus infarction: Stroke severity and risk factor profile. A Danish nation-wide evaluation of 25 839 patients with acute stroke. *Joint World Congress on Stroke. Cape Town, South Africa*, 2006 (October 26-29).

- T.S. Olsen, K.K. Andersen, M.N. Andersen, and H.G. Petersen. Hemorrhagic strokes in patients with atrial fibrillation: Frequency, clinical characteristics and prognosis. *Joint World Congress on Stroke. Cape Town, South Africa*, 2006 (October 26-29).

- M.N. Andersen, K.K. Andersen, H.G. Petersen, and T.S. Olsen. Women survive stroke better than men. A study of gender-specific differences in survival of 25 839 patients with acute stroke. *Joint World Congress on Stroke. Cape Town, South Africa*, 2006 (October 26-29).

- M.N. Andersen, R.Ø. Andersen, and K. Wheeler. Filtering in Hybrid Dynamic Bayesian Networks. *International Conference on Acoustics, Speech, and Signal Processing. Montral. Canada*, 2004 (May), pp 773-776.

Acknowledgements

I would like to thank my supervisors, Ass. Prof. Ole Winther and Prof. Lars Kai Hansen, IMM, DTU, for their extensive support and valuable discussions.

Prof. Nancy Reed, University of Hawaii, for inviting me to Hawai'i. The time I spent in Holmes 390, and for all the great experiences outside the office (especially on the beach) will always be remembered and appreciated. Mahalo nui loa, a hui hou...

Dr. Tom Skyhøj Olsen and Ass. Prof. Klaus Kaae Andersen for valuable and pleasant collaboration and an unforgettable Joint World Congress on Stroke in Cape Town, South Africa.

All my friends, especially AnneMette, Heidi and Kenneth who had to listen to all my complaints during the writing of this thesis, and for being true friends, when I needed you.

♡ Pølle ♡

However, words and least a "thank you" cannot express my unbounded gratitude to my family. Through good and especially through bad times, you were all there for me. When things looked pretty bleak, and I was on the edge of giving up, you gave me all your love and support, believed in me, and did everything to get me through.

I could not have been impossible without you!

Notation

- X: upper-case letter usually refers to an uncertain variable.
- x: lower-case letter usually refers to a sampled value of an uncertain variable.
- $\boldsymbol{X}$: column vectors or matrices are usually printed in bold type. This also applies to vector or matrix functions.
- $\boldsymbol{X}^{\mathrm{T}}$: a row vector is a transposed column vector indicated by $^{\mathrm{T}}$.
- $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$: vector notation is also used to indicate a list of objects, and in this case we do not distinguish between row and column vectors.
- $p(X)$: probability (distribution) of X.
- $p(x)$: probability of x, corresponds to $P(X = x)$.
- $p(x|y)$: probability of x conditional on y.
- $F(\cdot)$ and $f(\cdot)$ usually denote a function and its density or derivative (also in vector or matrix versions), i.e. $\int \ldots f(t)dt = F(t)$ it F has density f.
- t usually denotes a time, while T is usually a random time (a stopping time, in fact).
- $\mathrm{E}[f(\cdot)]$ or $\mathrm{E}(f(\cdot))$ is the expected value of $f(\cdot)$.
- $\mathrm{V}[f(\cdot)]$ or $\mathrm{V}(f(\cdot))$ is the variance or covariance matrix of $f(\cdot)$.

Abbreviations and Acronyms

Technical Abbreviations and Acronyms

- AFT: Accelerated Failure Time
- BIC: Bayes Information Criterion
- BMA: Bayesian Model Averaging
- BN: Bayesian Network
- BNT: Bayes Net Toolbox
- CC: Complete Case
- CI: Confidence Interval
- COST: COpenhagen Stroke Study
- CPH: Cox Proportional Hazards
- CR: Cox Regression
- CT: Computed Tomography
- DAG: Directed Acyclic Graph
- EM: Expectation Maximization
- GLM: Generalized Linear Model
- IC: Interval Coverage

- i.i.d.: independent, identically distributed
- L&B: Leaps and Bounds
- LL: Log Likelihood
- LPL: Log Partial Likelihood
- LRT: Likelihood Ratio Test
- MAP: Maximum A Posteori
- MAR: Missing At Random
- MCAR: Missing Completely At Random
- MCMC: Markov Chain Monte Carlo
- MH: Metropolis-Hastings
- MI: Multiple Imputation
- ML: Maximum Likelihood
- MLE: Maximum Likelihood Estimate
- MLP: Multi Layer Perceptron
- MMP: Multiple Myeloma Patients
- MPE: Most Probable Explanation
- MPLE: Maximum Partial Likelihood Estimate
- NI: Non-Ignorable
- PL: Partial Likelihood
- PPP: Posterior Parameter Probability
- PPS: Partial Predictive Score
- RSS: Residual Sum of Squares
- SSS: Scandinavian Stroke Scale

Experimental Abbreviations and Acronyms

- af: atrial fibrillation
- age: age of patient in years
- alco: daily alcohol consumption
- apo: previous stroke
- BMI: Body Mass Index
- bun: level of blood urea nitrogen
- ca: serum calcium
- cla: intermittent claudication
- dm: diabetes mellitus
- hb: haemoglobin
- hemo: type of stroke
- hyp: hypertension
- ihd: ischemic heart disease
- odd: other disabling disability
- cells: percentage of plasma cells in the bone marrow
- protein: indicator of whether or not Bence-Jones protein was present in the urine
- sex: sex of patient
- smoke: smoking
- sss: initial stroke severity
- temp: spontaneous body temperature on admission

General Abbreviations and Acronyms

- a.k.a.: also known as
- cf.: confer
- e.g.: exempli gratia (for example)
- et al.: et alii (and others)
- etc.: et cetera (and so forth)
- eq.: equation
- i.e.: id est (that is)
- vs.: versus

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In Denmark, stroke is the third highest cause of death after cardiovascular disorders. Each year, 10,000 Danes experience a stroke, and 1 of 7 Danes will experience a stroke at least once in their lifetime. The disease is very much a lifestyle-related disease, most and foremost caused by arteriosclerosis. All age groups suffer, but the frequency increases strongly with age. Men suffer twice as frequently as women, and blood pressure, diet, exercise habits, smoking, diabetes, and heart diseases are known as predisposing factors. The disease is very important on a community scale, stroke being the most expensive, isolated disease in the Danish hospital service system, and the disease that seizes most bedsides, (www.vfhj.dk, 2007), (Davidsen, 2007).

40% die within a year of stroke onset.

Hence, if we could identify explanatory variables for the survival time of stroke patients, we would be able to guide physicians and patients on how to extend the survival time after a stroke. As the statistics indicate, this would be of value for a very large group of patients worldwide. If the results suggest that, say, smoking has a negative effect on the survival time, physicians could advise the patient to stop smoking, and if, say, the stroke severeness is a significant risk factor, physicians could look for ways to limit the severity, e.g. cool the patient as suggested by Boysen and Christensen (2001), or establish special stroke units as suggested by (Jørgensen et al., 1999a).

In this thesis, which is primarily a comparative method study, we explore possible predictors of the survival time in days from admission to death in a community based, Danish stroke study, the COpenhagen Stroke Study (COST), and possible predictors of the survival time in months from diagnosis of multiple myeloma to death in a smaller study from the Medical Center of University of West Virginia, USA.

The main objective is to identify methods that can make reliable evaluations of how possible risk factors affect the survival time, but the methods should also be able to identify models capable of making reliable predictions of the (expected) survival time for future patients. When a new stroke patient is admitted to the hospital, the doctor should be able to advise the patient on ways to increase the survival time, and what to expect, if he or she does/does not take the advice.

There has been extensive research in survival analysis in general, see e.g. (Collet, 2003), (Lee and Wang, 2003) and (Ibrahim et al., 2005) for many good case studies. Examples of survival analysis on the COST data set includes (Jørgensen et al., 1999a), (Kammersgaard and Olsen, 2006), (Andersen et al., 2006c), (Andersen et al., 2006d), (Andersen et al., 2006b), and (Olsen et al., 2006) just to mention a few.

Although much of the work in this thesis focuses on stroke patients, the explored methods are much more general. We can apply them to any survival analysis study, but in this work we consider the methods in a Cox Proportional Hazards (CPH) and Cox Regression (CR) setting. In survival analysis, we typically have access to various patient information and the time of death or the censoring time. The most common approach is to fit a CPH model, (Cox, 1972), and search for significant, in terms of p-values, independent predictors of the survival time using a stepwise selection method, (Lee and Wang, 2003), (Collet, 2003), which iteratively proposes models that differ from the current one by just one variable, and accepts or rejects the new model based on a significance test statistic.

Queries in Google, Scholar using various search strings gave the following results: “survival analysis stroke patients” (105,000 hits), “survival analysis cox proportional hazards” (50,100), “stroke cox proportional hazards” (8,370), and “copenhagen stroke study cox proportional hazards” (611), indicating the large amount of literature on survival analysis, the application of a CPH model, and stroke data.

Particularly, survival analysis of stroke data using stepwise fitting of a CPH model is very popular, and there are numerous publications based on this approach, (Knuiman et al., 1992), (Tuomilehto et al., 1996), (Anderson et al., 1994), (Gulløv et al., 1998), and (Broderick et al., 1992) just to mention a few. Using the COST database alone, examples are (Andersen et al., 2005b), (Petersen et al., 2006), (Kammersgaard et al., 2002), and (Kammersgaard et al., 2004). These publications all relate to the time of death, but survival analysis, including the CPH model, could easily be used for data sets where the event time is not death. An example of this is (Lee et al., 1999), where the event is the stroke itself.

Stepwise selection is a variable selection method that identifies a *single* model, and makes inference as if the selected model was the true model. However, if we have, say, 20 variables, we have $2^{20} = 1,048,576$ potential models, and by selection just 1(!) model, we subsequently ignore the variables not selected, but we also ignore the uncertainty in the variable selection process. Consequently, the true uncertainty is often limited to the parameter uncertainty. If we ignore the model uncertainty, we underestimate the true uncertainty and get over-confident and biased estimates, (Little and Rubin, 1987), (Ibrahim et al., 2005). This approach will also underestimate the uncertainty about quantities of interest, see e.g. (Madigan and York, 1995), (Raftery, 1994), and (Kass and Raftery, 1994). We will show that the ignored model uncertainty is substantial, even for small data sets with just a few variables.

To address this problem, Raftery et al. (1995), show that accounting for model uncertainty in survival analysis using Bayesian Model Averaging (BMA), improves predictive performance. In BMA we average over all or a subset of models weighted by their posterior model probabilities (PMP), (Hoeting et al., 1999), (Ibrahim et al., 2005). The “significance” of a variable is the sum of posterior model probabilities for models that include the given variable. This gives us an estimate of the importance of each variable that is interpretable, more reliable, and makes a distinction that the p-values cannot. Using p-values we may fail to reject the null hypothesis “no effect” because either a) there is not enough data to detect the effect, or b) the data provide evidence for the null hypothesis. Using posterior probabilities we can make this very valuable distinction. No variables are excluded. If they have no effect, they will not appear in any of the selected models.

However, stepwise selection is still, as of today, the preferred method, although it has been shown in several studies that the method is far from adequate, see e.g. (Miller, 1990) for a thorough review, (Viallefont et al., 2001) for a comparison of BMA and stepwise methods using simulated data, while Wang et al. (2004) compare BMA and stepwise methods in logistic regression. Stepwise selection is very popular, because it is easy to use, is implemented in most statistical software packages, give interpretable models, and is based on familiar terms such as p-values and significance levels. Physicians also find it difficult to adapt to solutions where all variables are included, and are not classified as either significant or not. Furthermore, we do not have a single model, but an average model, and we use unfamiliar terms such as posterior probabilities.

As mentioned, there has been numerous survival studies on fitting a CPH model to COST data. However, all these studies have used stepwise selection. Furthermore, we have not found any study analyzing the survival time of stroke patients using selective model averaging besides (Andersen et al., 2006e), which is also part of this thesis preparation. Volinsky et al. (1997) used BMA and CPH in a study of stroke/non-stroke patients, but assessed the risk of a stroke rather than the survival time of stroke patients. BMA in survival analysis, including the application of a CPH model, has been explored in several publications, e.g. Volinsky and Raftery (2000), Volinsky (1997), Volinsky et al. (1997), and Raftery et al. (1995). However, there are several important aspects that are not addressed in these works, including validation of the proportional hazards assumption, how to include time dependent variables if this assumption is violated, and last, but not least, the problem of missing values was not addressed.

Stepwise selection and other variable selection methods cannot handle cases with missing values, i.e. subjects where one or more of the variables have not been observed. Missing values are very common, especially in data sets that include patient information. In the COST data set, we have 993 subjects and examine 14 potential risk factors. Of the 993 subjects, 441 subjects or 44.4% had missing values in at least one of these risk factors!

These numbers are not unusual for stroke data sets, or other health related data sets for that matter. With such a significant number of subjects with missing values, we would expect that every (survival) analysis of this data set included a missing values analysis. However, such analyses are themselves missing altogether, and it is common practice to simply ignore any subjects with missing values.

Obviously, ignoring such a significant number of subjects can have a great impact on the results, and bias the estimates of the model parameters and the assessment of the uncertainty decisively, (Little and Rubin, 1987). In any case, we significantly reduce the size of the available data and neglect a lot of valuable information stored in the ignored subjects. This is not at all plausible. Another approach is to use fully observed risk factors only. Obviously, it challenges the validity of an analysis to limit the set of possible explanatory variables in the light of such trends. We cannot exclude a variable that we suspect to be a significant predictor of the survival time, simply because it has not been observed for all subjects! A more decent approach to the missing values problem is imputation, the practice of “filling in” missing data with plausible values, e.g. the series mean of the observed values. It is an easy and thus attractive approach to analyze incomplete data, but a naive or unprincipled imputation method may create more problems than it solves, distorting estimates, standard errors and hypothesis tests, as documented by (Little and Rubin, 1987) and others.

The question of how to obtain valid inferences from imputed data was addressed by (Little and Rubin, 1987) using Multiple Imputation (MI). MI is a Monte Carlo technique in which the missing values are replaced by $m > 1$ simulated versions, where m is typically small (e.g. 2-10) to keep the computational challenges at a fair level. Each of the complete data sets is analyzed using standard methods, and the results are combined to produce estimates and confidence intervals that incorporate missing data uncertainty.

Most of the MI techniques presently available, assume that the missing values are Missing At Random (MAR), (Ibrahim et al., 2005), i.e. they assume the missing values carry no information about the probabilities of missingness. This assumption is mathematically convenient, because it allows one to express an explicit probability model for non-response. In most applications, however, ignorability seems artificial or implausible. In the COST database, for example, the stroke severity is estimated using an evaluation of the level of consciousness; eye movement; power in arm, hand, and leg etc. It is easy to associate a missing evaluation with a severe stroke, because the patient will be in such a poor condition that an evaluation is not possible. Most imputing methods have also proven to bias the solution, see e.g. (Little and Rubin, 1987), (Herring and Ibrahim, 2001) and (Ramoni and Sebastiani, 2001).

Instead, we will explore two different approaches to the missing data problem.

Acknowledging the fact that our major concern, in the COST data set at least, are missing *discrete* data values, we suggest the use of Bayesian Networks (BN), (Jensen, 1996), (Heckerman, 1995), (Murphy, 2001a), to identify possible relations between potential risk factors. The method offers a variety of choices as to whether we learn the model structure and parameters separately, using the fully observed cases only, or we include subjects with missing values to learn the structure and parameters interchangeably using a structural Expectation Maximization (EM) algorithm due to Friedman (1998). Furthermore, we can compare different learning algorithms with respect to scoring function and point vs. Bayesian parameter estimates, different inference techniques, and whether we use the most probable assignment of values to the missing variables given the observed evidence, or use the estimated joint distribution to assign a weight to each possible assignment of missing values. Each assignment would then be considered a unique data point with a corresponding weight given by the joint probability of the observed values, and the missing value pattern. Having estimated the missing values using either method, we can use stepwise selection and BMA on the *augmented* data set.

The other is a semi-parametric approach using an Expectation Maximization (EM) algorithm, (Dempster et al., 1977), (Thiesson et al., 1999), to estimate the missing values in the E-step, and update the parameter estimates in the M-step.

To be completely general in terms of the type of missing data, and the type and number of variables subject to missingness, we specify the joint distribution of the missing data mechanism R, the failure time T, and the variable vector $\boldsymbol{Z}$. A general approach would be to specify conditional distributions on $[R|T, \boldsymbol{Z}]$ and $[T|\boldsymbol{Z}]$, and a marginal distribution for $\boldsymbol{Z}$. In this work, we place fully parametric distributions on $[R|T, \boldsymbol{Z}]$ and $\boldsymbol{Z}$, while we use the CPH model for the distribution of $[T|\boldsymbol{Z}]$. However, the original algorithm presented in (Herring et al., 2004) uses stepwise selection to fit a single CPH model. An obvious improvement would be to use BMA on the final, augmented data set.

Using either BN or a semi-parametric approach, we combine the strengths of BMA with an increased data set augmented by estimation of the missing values. This is a very plausible strategy! However, we still apply the techniques separately: First we estimate the missing values, then we apply BMA. Merging BMA and missing values estimation has not yet been addressed, although it seems to be an interesting "all-round" solution applicable to many problems. Using BN to estimate the missing values, there is no obvious way to merge the two, since we do not use the estimates of the parameters in the CPH model to estimate the missing values in the BN. However, the semi-parametric model does. Assuming that we manage the merging, we would then select the subset of models with the highest posterior probabilities, $p(M_i|D)$, given data D for model M_i. To evaluate the posterior probability, we need the marginal likelihood, $p(D|M_i)$ (the Bayesian score), and the model prior, $p(M_i)$. The marginal likelihood is obtained by averaging over the parameters. However, learning models from incomplete data is much harder than learning from complete data as shown by e.g. Volinsky (1997) and (Nielsen, 2003), because the posterior over parameters is no longer a product of independent terms. In other words, there exists a posterior distribution for every completion of the database.

For the same reason, the probability of the data is no longer a product of independent terms. Since the posterior distribution over the parameters of a given model is no longer a product of independent posteriors, we generally cannot represent it in closed form. Instead, we can attempt to approximate it. The simplest approximation is to use the Maximum Likelihood (ML) or the Maximum A Posteori (MAP) parameters. We obtain an approximation to these parameters using either gradient ascent methods, (Bishop, 2006), or an EM algorithm. Since the probability of the data given a model no longer decomposes, we need to approximate it using either stochastic simulation, (Kuo and Smith, 1992), (Liu, 2001), which is extremely expensive in terms of computation, or using large-sample approximations based on Laplace's approximation, e.g. Bayes Information Criterion (BIC), (Weakliem, 1999), (Volinsky and Raftery, 2000).

These approximations require the ML/MAP parameters for each model, before we can score them. Thus, a search in model space requires an expensive evaluation of each candidate. When we are searching in a large space of possible models, this is infeasible. One solution is to use an heuristic search algorithm to quickly remove the majority of candidates. We adopt the approach from (Volinsky, 1997), and use a Leaps and Bounds (L&B) algorithm, (Furnival and Wilson, 2000), (Lawless and Singhal, 1978), to scan the structure × parameter space, and select a (much) smaller set of models using Occam's window subset selection, (Madigan and Raftery, 1994a), small enough to allow a feasible evaluation of each model, but large enough to make fairly sure that we include the important models.

Another aspect of survival analysis using the CPH model is the assumption of proportional hazards. The BMA solutions presented in (Raftery et al., 1995), (Volinsky, 1997), (Volinsky et al., 1997) and other are all based on this assumption. There exist several methods to validate the assumption, before and after the model(s) have been estimated, see e.g. (Collet, 2003) and (Lee and Wang, 2003). Collet (2003) also show how to include time dependent variables, transforming the CPH model into a more general Cox Regression (CR) model. We can use this to implement a more general BMA approach to survival analysis.

All algorithms are implemented using Matlab.

The thesis is divided in two parts, a theoretical part (Chapter 2-4) and an experimental part (Chapter 5-7), and is outlined as follows:

Chapter 2. Survival Analysis. First, we give an introduction to survival analysis and the concept of censoring. We define the survival function, the density function, the hazard function, and their equivalence relationships. We show how to use parametric and non-parametric methods for estimating parameters in survival models, how to include variables, and present various variable selection techniques, including the most widely applied method, stepwise selection. Finally, we introduce the CPH model and the partial likelihood that we will use throughout the thesis.

Chapter 3. The Bayesian View and what it is all about. Then we introduce and discuss the concept of probability, and how it is interpreted by the "frequentists" and their counter-colleagues, the "Bayesians". We discuss various modeling approaches, model learning, model inference, and model selection. The concept of model uncertainty is introduced, and we present methods for comparing and selecting models. We present BMA to average over a subset of models that we select using Occam's razor, and apply it to the CPH. We discuss the concepts of posterior model probability and posterior parameter probability, and compare the latter to the well-known p-value. To round off, we present two methods for evaluating the predictive model performance.

Chapter 4. Missing Values. In this chapter we discuss the importance of missing values in data, and introduce the reader to the MI technique. Furthermore, we show how to use MCMC to estimate the missing values as well as the parameters in the CPH model.

However, MCMC methods can be tricky to implement and computationally expensive. Instead, we present the use of BNs to represent intervariable connections between the risk factors. We outline several methods for learning the graph structure, discuss exact as well as approximate techniques for inferring the missing values, and describe an open source Matlab toolbox to do learning and inference in BNs.

We also present a semi-parametric approach to the missing values problem. The method places fully parametric distributions on the missing data mechanisms and the risk factors, and uses the CPH to model the failure times. We define, and give examples of, different missing data mechanisms, and show how to use an EM algorithm to iteratively estimate the missing values and the parameters. We also show how to include discrete as well as continuous valued variables, and how to improve the algorithm by implementing BMA in the M-step.

Chapter 5. Databases. This is the first chapter in the experimental part of the thesis, and here we outline the two real-life data sets that we apply our methods to. First, we describe the multiple myeloma data set, and then the COST data set.

Chapter 6. Comparison of Stepwise Selection and Bayesian Model Averaging Applied to Real Life Data. In our first experiment, we compare the stepwise selection method to BMA. We apply our methods to the multiple myeloma data set, and then the COST data set. We also outline two methods for validating the proportional hazards assumption, and show how to include time dependent variables.

Chapter 7. Estimating Missing Values in the COST Data Set. In the second experiment, we compare various techniques to estimate the missing values in the COST data set. The chapter is divided in four sections. First, we simple remove variables one by one using the posterior parameter probability to evaluate each risk factor. With fewer variables, there will also be fewer subjects with missing values. Next, we use a BN to estimate the missing values. We compare different algorithms for learning the structure and parameters of the network, and for estimating the missing values. Finally, we estimate the missing values using a semi-parametric approach, and show how to improve the algorithm by incorporating BMA. We also simulate how results are influenced by different distributions. Having increased the size of the data set in either of the three ways, we re-apply stepwise selection and BMA to update the our (model) parameter estimates. In the last section we compare the different approaches with respect to predictive performances, and show examples of survival curves using an estimated model.

Chapter 8. Conclusion and Discussion. We end the thesis with a discussion of the experimental results, and suggest directions for future work.

Chapter 2

Survival Analysis

In *survival analysis*, (Cox and Oakes, 1984), (Collet, 2003), (Lee and Wang, 2003), (Ibrahim et al., 2005), (Andersen et al., 1993) our objective is to model the *survival time*, i.e. the time to the occurrence of a given event. The event could be just about anything. Within the medical field, common examples are the time to development of a disease, response to a treatment, and of course death. The available data often include the survival time, patient characteristics (such as gender, age, and blood pressure), disease information, treatment information, examination data and much more. Often we attempt to predict the probability of survival, response, or mean lifetime given a set of observed variables, compare survival distributions, and identify risk and/or prognostic factors. The aim of this chapter is to give an introduction to survival analysis.

2.1 Censoring

First, however, we need to introduce the concept of censored data. Often, one is tempted to consider survival analysis as the application of a parametric (if the survival times follow a known distribution), or non-parametric method (if the distribution is unknown) to some survival data. However, this is only true if all survival times are exact and known, which is rarely the case. Instead, non-parametric tests based on the *rank ordering* of survival times should be applied. For example, in a clinical study it is very common that some patients are still alive, or not disease-free, at the end of the study. It is also possible to lose patients during the study period, e.g. if the patients are not able to participate because they for some reason no longer qualify, they die by accident, or because they have moved abroad. The survival times for these patients are unknown, and we refer to them as *censored* observations. We have three types of censoring (Lee and Wang, 2003).

Type I censoring Imagine an animal study where the animals are given a lethal dose of some drug at the same time. However, due to limited financial and/or time constraints, the researcher cannot wait for all test subjects to die, and he needs to end the study prematurely. The animals that are still alive at the end of the study are censored, but we do know that their survival time is at least the length of the study period. Also, an animal could be lost, in which case it is censored, but we know that it survived at least to the time is was lost. With no losses, all censored observations are equal to the length of the study period.

Type II censoring If the researcher chooses to end the study when a predetermined number of the animals have died, all censored observations are equal to the largest uncensored observation, if there are no losses.

Type III censoring For many studies involving humans, the study period is time/resource limited, but patients enter the study at different times. The COST data used in this thesis, where patients experience stroke at different times, and the study period ends at a given date, fall within this category. Lost patients that do not want to participate in follow-up, move abroad etc. are censored, and their survival times are at least the time between the stroke and the last contact. Patients that are still alive at the end of the study are censored with survival times at least the time between the stroke and the end of the study. The other data set used in this work, the multiple myeloma data, are also Type III censored.

Figure 2.1 shows examples of these three types of censoring, all *right-censoring* techniques. Type I and II censored observations are called singly censored observations, while Type III censored data are known as progressively or randomly censored data. Left-censoring occurs when all we know is that the event occurred prior to some time t. Interval-censoring is when the event is known to occur between times t_1 and t_2. (Lee and Wang, 2003, chap. 1)

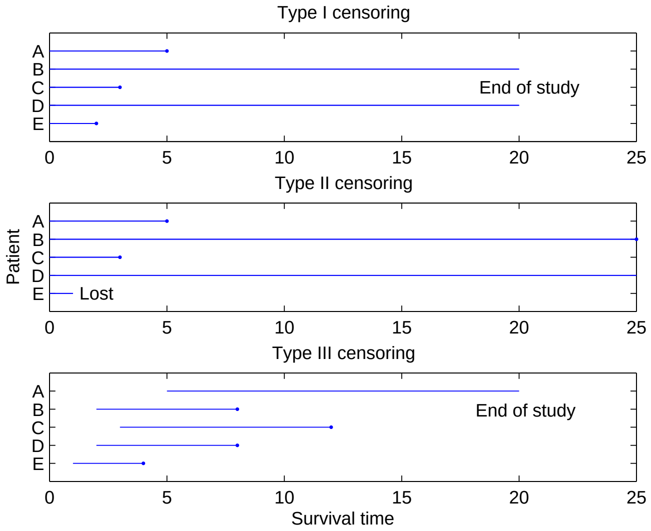

Figure 2.1: Examples of Type I, II, and III censoring.

2.2 Distribution of Survival Times

The distribution of survival times, the time to a given event, is described by three mathematically equivalent functions, i.e. given one of the functions, we can derive the others.

2.2.1 Survival Function

(According to definitions in (Lee and Wang, 2003, chap. 2)). Let T be the survival time for a subject, and let $S(t)$ be the *survival function*, the probability that the subject survives longer than time t, defined as

$$S(t) \equiv p(T > t) \tag{2.1}$$

$S(t)$ is a non-increasing function with

$$S(t) = \begin{cases} 1, & t = 0 \\ 0, & t = \infty \end{cases} \tag{2.2}$$

The *cumulative distribution function*, F(t) of T, is given by

$$F(t) \equiv 1 - S(t) \tag{2.3}$$

and represents the probability that a patient dies before time t.

Given these definitions, it is clear why $S(t)$ is also known as the *cumulative survival rate*, and $F(t)$ as the *cumulative failure rate*. If there are no censored observations, $S(t)$ is estimated by the proportion of patients surviving longer than time t

$$\hat{S}(t) = \frac{\text{\# patients surviving longer than } t}{\text{total \# of patients}} \tag{2.4}$$

However, with censored observations, we may not be able to calculate the numerator of (2.4).

2.2.2 Density Function

(According to definitions in (Lee and Wang, 2003, chap. 2)). The survival time, T, has a *density function*, $f(t)$, defined as the limit of the probability that a patient dies (or more generally fails) in the interval t to $t + \Delta t$ per unit width Δt, i.e. the probability of failure within a small interval per unit time

$$\begin{aligned} f(t) &\equiv \frac{\lim_{\Delta t \to 0} P[\text{patient dies in the interval } (t; t + \Delta t)]}{\Delta t} && (2.5) \\ &= \frac{\lim_{\Delta t \to 0} p(t < T \leq t + \Delta t)}{\Delta t} && (2.6) \end{aligned}$$

The density function, also known as the *unconditional failure rate*, satisfies

$$\begin{aligned} f(t) &\geq 0 \qquad t \geq 0 && (2.7) \\ f(t) &= 0 \qquad t < 0 && (2.8) \end{aligned}$$

and

$$\int_0^\infty f(t) = 1 \tag{2.9}$$

If there are no censored observations, $f(t)$ is estimated as the proportion of patients dying in a short interval per unit width

$$\hat{f}(t) = \frac{\text{\# patients dying in the short interval beginning at time } t}{\text{(total \# of patients)}} \tag{2.10}$$

Again, with censored observations, we may not be able to calculate the numerator of 2.10.

2.2.3 Hazard Function

(According to definitions in (Lee and Wang, 2003, chap. 2)). Finally, let $h(t)$ be the *hazard function*, or the *conditional failure rate*, defined as the limit of the probability of failure during a very small time interval, $t + \Delta t$, given that the patient has survived to time t

$$\begin{aligned} h(t) &\equiv \frac{\lim_{\Delta t \to 0} P[\text{patient dies in } (t; t + \Delta t) | \text{patient survived to time } t]}{\Delta t} \\ &= \frac{\lim_{\Delta t \to 0} p(t < T \leq t + \Delta t | T \geq t)}{\Delta t} \end{aligned} \tag{2.11}$$

or expressed in terms of the density function

$$h(t) = \frac{f(t)}{1 - F(t)} \tag{2.12}$$

In many studies, as is also the case in this thesis, the time t represents the patients age, and thus $h(t)$ expresses the risk of death per unit time, say, days or years, during aging. With no censored observations, $h(t)$ is estimated as the proportion of patients dying in an interval per unit time given survival to the beginning of the interval

$$\hat{h}(t) = \frac{\#\text{ patients dying in the interval } (t; t + \Delta t)}{(\#\text{ patients surviving at time } t) \times (\Delta t)} \tag{2.13}$$

The hazard function can increase (cancer patients that are not treated), decrease (patients undergoing surgery, patients responding to treatment), be constant (the hazard of the patient being struck by lightning), or a combination hereof. Human life, for example, is described by a decreasing hazard rate from the time we are born (high infant mortality), then it remains more or less constant for a period of time, and then it increases as we get older and are more likely to catch diseases, or simply because our body has reached its “use-by”date. We also use the *cumulative hazard function*, $H(t)$, defined as

$$H(t) \equiv \int_0^t h(u) du \tag{2.14}$$

in the interval zero to infinity.

2.2.4 Relationships between Survival Functions

Inserting (2.3) in (2.12), we get the relationship

$$h(t) = \frac{f(t)}{S(t)} \tag{2.15}$$

Furthermore, since the density function is defined as the derivative of the cumulative distribution function, we get

$$f(t) = \frac{d}{dt}[1 - S(t)] = -S'(t) \tag{2.16}$$

Inserting (2.16) in (2.15), we have

$$h(t) = -\frac{S'(t)}{S(t)} = -\frac{d}{dt}\log S(t) \tag{2.17}$$

If we integrate (2.17) from 0 to t using $S(0) = 1$, we get

$$-\int_0^t h(u)du = \log S(t) \tag{2.18}$$

Using (2.14) we get

$$H(t) = -\log S(t) \tag{2.19}$$

or

$$S(t) = \exp[-H(t)] = \exp\left[-\int_0^t h(u)ux\right] \tag{2.20}$$

Inserting (2.20) in (2.15) yields

$$f(t) = h(t)\exp[-H(t)] \tag{2.21}$$

Hence, we have shown that it is possible to derive any of the three functions given the two others are known.

2.3 Estimation of Parameters

Survival analysis in a nutshell is to estimate the three survival (survivorship, density, and hazard) functions defined in the previous chapter. There exist parametric as well as non-parametric methods for this purpose. In case we do not know the exact survival times, estimation of the survival functions becomes much more difficult.

In the most common situation, as with the applications presented in this thesis, we have right-censored data where the patients are followed to death or are censored. Let $t_1, t_2, \ldots, t_k, t_{k+1}^+, \ldots, t_n^+$ be the survival times of n patients with k exact times (patients have died) and $(n-k)$ right-censored times. We assume that the survival times follow a given distribution with density function $f(t, \boldsymbol{\beta})$ and survivorship function $S(t, \boldsymbol{\beta})$ using $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ to denote p unknown parameters in the distribution. As shown later, for the exponential distribution we would have $p = 1$ parameter (λ).

With a discrete survival time, T, say, using the date the patient died, $f(t, \boldsymbol{\beta})$ is the probability of observing the survival time t (an uncensored survival time), and $S(t, \boldsymbol{\beta})$ is the probability that the survival time is greater than t (a right-censored survival time). Hence, the product $\prod_{i=1}^{k} f(t_i, \boldsymbol{\beta})$ represents the joint probability of the k uncensored survival times, and $\prod_{i=k+1}^{n} S(t_i^+, \boldsymbol{\beta})$ represents the joint probability of the remaining right-censored survival times. The product of these two factors represents the joint probability of the complete data set, and is called the likelihood function of the parameter set, $\boldsymbol{\beta}$, (Lee and Wang, 2003), Collet (2003)

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} f(t_i, \boldsymbol{\beta}) \prod_{i=k+1}^{n} S(t_i^+, \boldsymbol{\beta}) \tag{2.22}$$

or the likelihood of observing the data given a set of parameters.

2.3.1 Maximum Likelihood Estimation

In *Maximum Likelihood Estimation (MLE)*, (Lee and Wang, 2003), Collet (2003), (Ibrahim et al., 2005) we estimate the set of parameters that maximizes the likelihood function. For computational ease we use the *Log-Likelihood (LL)* function, $l(\cdot)$, turning products into sums

$$l(\boldsymbol{\beta}) = \log L(\boldsymbol{\beta}) = \sum_{i=1}^{k} \log\left[f(t_i, \boldsymbol{\beta})\right] + \sum_{i=k+1}^{n} \log\left[S(t_i^+, \boldsymbol{\beta})\right] \tag{2.23}$$

The MLE, $\hat{\boldsymbol{\beta}}$, is the parameter set that maximizes $l(\boldsymbol{\beta})$

$$l(\hat{\boldsymbol{\beta}}) = \max\left\{l(\boldsymbol{\beta})\right\} \tag{2.24}$$

corresponding to the solution of

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} = 0 \qquad j = 1, \dots, p \tag{2.25}$$

if

$$\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_j} < 0 \qquad j = 1, \dots, p \tag{2.26}$$

evaluated at $\hat{\boldsymbol{\beta}}$. Otherwise, the solution is a minimum or a saddle-point. In most cases we do not have a closed form solution to (2.25). Instead, we use the numerical Newton-Raphson procedure:, (Bishop, 2006), (Lee and Wang, 2003)

1. Initialize the parameters $(\beta_1, \dots, \beta_p)$ to 0, i.e. let

$$\boldsymbol{\beta}^{(0)} = \mathbf{0} \tag{2.27}$$

2. Take a small step in the likelihood space that increases the LL, that is let the change in $\boldsymbol{\beta}$ be

$$\Delta^{(j)} = \left[-\frac{\partial^2 l(\boldsymbol{\beta}^{j-1})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}}\right]^{-1} \frac{\partial l(\boldsymbol{\beta}^{j-1})}{\partial \boldsymbol{\beta}} \tag{2.28}$$

3. Using 2.28, the updated parameter value at the j'th iteration is

$$\boldsymbol{\beta}^{(j)} = \boldsymbol{\beta}^{(j-1)} + \Delta^{(j)} \qquad j = 1, 2, \dots \tag{2.29}$$

The procedure determines if the step change is below some preselected threshold value, or if the algorithm exceeds a maximum number of iterations. The estimated covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\hat{V}(\hat{\boldsymbol{\beta}}) = \hat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}) = \left[-\frac{\partial^2 l(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}}\right]^{-1} \tag{2.30}$$

2.3.1.1 An Example: The Exponential Distribution

As an example, consider the one-parameter exponential distribution with density function

$$f(t) = \begin{cases} \lambda e^{-\lambda t} & t \geq 0, \lambda > 0 \\ 0 & t < 0 \end{cases} \tag{2.31}$$

survivorship function

$$S(t) = e^{-\lambda t} \qquad t \geq 0 \tag{2.32}$$

and hazard function

$$h(t) = \lambda \qquad t \geq 0 \tag{2.33}$$

Concentrating on right-censored data, let the study begin at time $t = 0$ and terminate at $t = t'$ with n patients entering the study. Let k be the number of patients who die before or at time t', and let $n - k$ be the number of patients who are lost or remain alive at time t'. We order the uncensored data such that

$$t_1 \leq t_2 \leq \ldots \leq t_k, t^+_{k+1}, t^+_{k+2}, \ldots, t^+_n \tag{2.34}$$

Inserting (2.31) and (2.32) in (2.22), we get the likelihood function

$$L(\lambda) = \prod_{i=1}^{k} \lambda e^{-\lambda t_i} \prod_{i=k+1}^{n} e^{-\lambda t^+_i} \tag{2.35}$$

and the LL function

$$l(\lambda) = k \log(\lambda) - \lambda \sum_{i=1}^{k} t_i - \lambda \sum_{i=k+1}^{n} t^+_i \tag{2.36}$$

Differentiating with respect to λ gives

$$\frac{\partial l(\lambda)}{\partial \lambda} = \frac{k}{\lambda} - \sum_{i=1}^{k} t_i - \sum_{i=k+1}^{n} \tag{2.37}$$

and using (2.25) we get the estimator of λ

$$\hat{\lambda} = \frac{k}{\sum_{i=1}^{k} t_i + \sum_{i=k+1}^{n} t^+_i} \tag{2.38}$$

2.3.2 Including Risk Factors

We often need to take into account, especially when working on studies involving humans, that every patient is unique, and that individual differences may influence the survival times significantly. When a physician makes a diagnosis or a prognosis for a patient, he needs to gather a lot of information on the patient, such as personal data, medical history, test results etc. All these patient characteristics, ranging from blood pressure over CT scan results to marital status, are referred to as *prognostic factors* or *risk factors*, see examples of commonly used risk factors for stroke patients in (Andersen et al., 2005b), (Petersen et al., 2006), (Kammersgaard and Olsen, 2006), (Kammersgaard et al., 2002), (Andersen et al., 2006a), (Jørgensen et al., 1999a).

However, a lot of variables also means a lot of information. All information is valuable, but some information is more useful than other depending on the task we are currently working on. The difficult task is to figure out which variables are important to estimate the survival time and how to include them in the model. Usually, an experienced physician is able to eliminate most of the variables that he knows have no influence on the survival time. As an example, the subjects eye color is probably not relevant for the survival time of rats being exposed to a lethal drug, while the gender is important if male rats do not absorb the drug as well as female rats. We need statistical tools to map the relationship between the variables and the survival time, but we often use experienced physicians to limit the domain of variables of potential interest. Otherwise, we would have too many variables, and thus parameters, compared to the amount of available data. Also, we can use the physicians to provide other useful information such as model constraints.

One way to identify the relevant variables is to assume that data are generated by a certain class of models and then fit the parameters of the model. Often, we use regression models for this purpose. To include variables, we assume that we can express the relation between the variables and the survival time explicitly. As with conventional regression methods, we can model the logarithm of the survival time using the *Accelerated Failure Time (AFT) Model*, (Lee and Wang, 2003), (Collet, 2003), (Andersen et al., 1993), as the variables either accelerates or decelerates the time to failure, the survival time. The model assumes a linear relationship between the logarithm of the survival time, T, and the variables, $\boldsymbol{Z}$

$$\log T = \beta_0 + \sum_{j=1}^{p} \beta_j Z_j + \epsilon_i = \mu_i + \epsilon_i \tag{2.39}$$

where Z_j, $j = 1, \ldots, p$ are the variables, β_j, $j = 0, 1, \ldots, p$ their coefficients, and the ϵ_i's are *independent, identically distributed (i.i.d.)* uncertain error terms. Consequently, T is exponentially distributed with the following hazard, density, and survivorship functions

$$h(t, \lambda_i) = \lambda_i = \exp\left[-\left(\beta_0 + \sum_{j=1}^{p} \beta_j z_{ji}\right)\right] = \exp(-\mu_i) \tag{2.40}$$

$$f(t, \lambda_i) = \lambda_i \exp(-\lambda_i t)\epsilon_i \tag{2.41}$$

$$S(t, \lambda_i) = \exp(-\lambda_i t)\epsilon_i \tag{2.42}$$

The model assumes linearity between the variables and the logarithm of the hazard. If we have two patients with hazards $h_i(t, \lambda_i)$ and $h_k(t, \lambda_k)$, the *hazard* *ratio (HR)*, (Lee and Wang, 2003), (Collet, 2003), (Andersen et al., 1993), of these two patients is

$$\frac{h_i(t,\lambda_i)}{h_k(t,\lambda_k)} = \frac{\lambda_i}{\lambda_k} = \exp[-(\mu_i - \mu_k)] = \exp\left[\sum_{j=1}^{p} \beta_j (z_{ji} - z_{jk})\right] \tag{2.43}$$

The ratio is *not* time dependent, but only depends on the difference between the variables. This shows that the exponential regression model is a special case of a broader class of models known as *proportional hazard models*, (Cox, 1972), (Lee and Wang, 2003), (Collet, 2003), (Andersen et al., 1993) where the HR of any two patients is assumed time independent.

2.3.3 Variable Selection

As mentioned, we need to identify the most important variables or risk factors in terms of their effect on the survival time. Hence, we need a method that allows us to compare and choose between models with different subsets of variables. For a known parametric model, e.g. the exponential model, we can use different methods briefly described below. Usually, when researchers publish their results, they compare the outcome of several of these methods and comment on any differences in the selected subsets. This is known as multivariate analysis, as we include more than one variable in our model. We can also apply a univariate analysis, and, for each potential variable, check if the variable is a *significant* risk factor when all other variables are neglected. If it is not, there is no reason to include it in a multivariate analysis. Whether or not a variable is significant is determined by the controversial p-value, see (Collet, 2003), (MacKay, 2003), (Gelman et al., 2004) or any textbook on statistics, e.g. (Johnson, 2005). We say that a variable is *significant* if the p-value is below some preselected threshold, (MacKay, 2003), (Hubbard and Armstrong, 2005).

2.3.3.1 Forward Selection

In the *forward selection* method, (Lee and Wang, 2003), Collet (2003), (Gelman et al., 2004), we begin with an empty variable set, unless we have variables that we force into the model, e.g. in view of expert knowledge. In this case we might include, say, age, even if it is not significant, because we for some reason would like to adjust for age. Otherwise, we add one variable at a time, and once a variable is included, it cannot be removed. The variable that we attempt to add in the next step is the variable with the largest, adjusted χ^2 test statistic, see (Lee and Wang, 2003), Collet (2003), (Andersen et al., 1993) or any textbook on statistics e.g. (Johnson, 2005). If the test statistic is significant at a, say, $\alpha = 0.05$ level, we accept the variable for entrance.

If $\boldsymbol{\beta}_1$ is the parameter vector for the variables already included in the model, we select Z_k to enter the model if the difference between the log-likelihood with and without Z_k is the largest among all Z_j's not in the model, i.e. the coefficient β_k of Z_k satisfies

$$\begin{aligned} X_k &= 2\left[l(\hat{\beta}_k, \hat{\boldsymbol{\beta}}_{1k}(0))\right] && (2.44)\\ &= \max_k \left\{2\left[l(\hat{\beta}_j, \hat{\boldsymbol{\beta}}_1) - l(\boldsymbol{\beta}_{1j}(0))\right], \forall Z_j \text{ not in the model}\right\} && (2.45)\end{aligned}$$

and $X_k > \chi^2_{1,\alpha}$, where β_j is the parameter of Z_j not in the model, $(\hat{\beta}_j, \hat{\boldsymbol{\beta}}_1)$ is the MLE of $(\beta_j, \boldsymbol{\beta}_1)$, $\hat{\boldsymbol{\beta}}_{1j}(0)$ is the MLE of $\boldsymbol{\beta}_1$ given $\beta_j = 0$, and $\chi^2_{1,\alpha}$ is the α-level of the χ^2 distribution with one degree of freedom.

2.3.3.2 Backward Selection

The *backward selection* method, (Lee and Wang, 2003), Collet (2003), (Gelman et al., 2004), is in some sense the opposite of the forward selection method. We begin with the complete set of variables and eliminate them one by one using the Wald test statistic, (Collet, 2003), compared against a χ^2-distribution. In each step, the variable with lowest test statistic that is not above the specified α-level is removed, i.e. we remove Z_k if

$$X_k = \frac{\hat{\beta}_k^2}{\hat{v}_{kk}^2} = \min_j \left\{\frac{\hat{\beta}_j^2}{\hat{v}_{jj}^2}, \forall Z_j \text{ not in the model}\right\} \tag{2.46}$$

and $X_k \leq \chi^2_{1,\alpha}$, where $\hat{\beta}_j$ is the estimated parameter of Z_j, and $\hat{v}_{jj}^2$ is the estimated variance of $\hat{\beta}_j$. Again, if a variable is removed, it cannot re-enter the model.

2.3.3.3 Stepwise Selection

Stepwise selection, (Lee and Wang, 2003), Collet (2003), (Gelman et al., 2004), combines the forward and backward selection procedures. We begin with an empty set of variables and add variables as in forward selection. However, entered variables may now be removed in a later iteration if the variable is no longer significant. The algorithm terminates if there are no significant variables to add, or if the variable just entered is removed and no more variables can be added.

However, many studies show that model selection and in particular stepwise methods have some serious drawbacks, e.g. (Hoeting et al., 1999), Wang et al. (2004), (Volinsky, 1997), (Raftery, 1995), (Viallefont et al., 2001). The most important are:

Credibility of model exaggerated The model appears to have more explanatory power than it really does. Typically we overestimate the goodness-of-fit.

Level of testing procedure unknown We use a complex iterative application of hypothesis tests, and the overall probability of a Type 1 error, (Johnson, 2005), for the family of tests exceeds the specified level for the individual test. The true level for the entire selection procedure is hard to compute.

Criterion level No agreement on the best criterion for addition and deletion of variables. Optimally, we should exclude noise variables and include predictive variables. Suggested significance levels range from 0.01 to 0.50 and affect the results dramatically.

Dichotomization of variables Variables are either "in" or "out". In reality, the variables affect the response on a continuum. It is possible, that non-significant variables, when taken in aggregate, may have important estimation or prediction information.

Most of these problems are related to the use of p-values. We discuss this in more details in Section 3.3.1.4, where we also outline a method that does not use p-values.

2.4 Cox Proportional Hazards Model

So far, we have assumed that we know the underlying survival distribution that we can fit with a parametric model. Then we estimate and hypothesis test the parameters using standard asymptotic likelihood techniques. Non-parametric or distribution free methods are more general in the sense that they are more efficient than parametric distributions when survival times do not follow a theoretical distribution or we do not know the theoretical distribution.

The most commonly used survival model is the *Cox Proportional Hazards (CPH)* model, (Cox, 1972), (Cox and Oakes, 1984), (Andersen and Gill, 1982), (Herring and Ibrahim, 2001), (Lee and Wang, 2003), (Collet, 2003), (Andersen et al., 1993). The model is not based on an assumption of a known distribution, and the hazard function can take on any form. The only assumption is that the hazard functions of different patients are proportional and independent of time.

Another justification for the CPH model is that in our main data domain, stroke data, almost any paper published on the subject is based on the application of this model. A query search for “stroke survival analysis cox proportional hazard” in Goggle, Scholar gave 6,150 hits. Dr. Tom Skyhøj Olsen from the stroke unit at Hvidovre Hospital, responsible for collecting the data in the COST database explored in the experimental sections, has published numerous papers on stroke survival analysis, and is co-author on several of the publications written during the preparation of this thesis. According to Dr. Olsen, submitting a research paper for a medical journal or conference *not* based on the CPH model, will most like take flak for analyzing survival data in a way that is not understood or accepted by most readers/reviewers. It is the standard model to analyze survival data in the medical field, physicians understand this model and use it extensively. Indeed, one of the aims of this thesis is to extend the basic use of this model, and show that we can obtain better models using more advanced techniques *based* on the CPH model.

2.4.1 Partial Likelihood

Usually, we calculate the LL function and use it to obtain a point estimate of the parameters in the model. In the CPH model we do not have a theoretical distribution and cannot compute the (log)-likelihood. Instead, we use a quantity known as the *Partial Likelihood (PL)*, (Raftery et al., 1995), (Volinsky, 1997), (Lee and Wang, 2003), (Ibrahim et al., 2005). As we showed for the exponential distribution, all proportional hazard models have the property that the ratio between the hazard functions of any two individuals is constant (independent of time). Hence, we can write the hazard function for any individual as the product of an underlying (common) *baseline* hazard function, and a function of the variables

$$h(t|\boldsymbol{Z}) = h_0(t)g(\boldsymbol{Z}) \tag{2.47}$$

where $g(\boldsymbol{Z})$ is the variable effect and $\boldsymbol{Z} = (Z_1, \dots, Z_p)$. The baseline hazard, $h_0(t)$, expresses the hazard change over time when all variables are ignored (or rather have their baseline/reference values). The CPH model assumes that $g(\boldsymbol{Z})$ is an exponential function of the variables

$$g(\boldsymbol{Z}) = \exp\left(\sum_{j=1}^{p} \beta_j z_j\right) = \exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}) \tag{2.48}$$

which implies

$$h(t|\boldsymbol{Z}) = h_0(t)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}) \tag{2.49}$$

where $\boldsymbol{\beta}$ is the variable coefficient vector. To exemplify the effect of a variable in a treatment study, let there be only one (binary) treatment variable, Z_1, say, gender, where $Z_1 = 0$ for males and $Z_1 = 1$ for females. The HR between female and male patients is

$$\frac{h(t|Z_1 = 1)}{h(t|Z_1 = 0)} = \exp(\beta_1) \tag{2.50}$$

and so the treatment is equally effective if $\beta_1 = 0$, and the treatment introduces lower (higher) risk for females than males if $\beta_1 < 0$ ($\beta_1 > 0$). Using (2.49) and (2.20), we get

$$\begin{aligned}
S(t|\boldsymbol{Z}) &= \exp\left[-\int_0^t h(u)du\right] \\
&= \exp\left[-\int_0^t h_0(u)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z})du\right] \\
&= \exp\left[-\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z})\int_0^t h_0(u)du\right] &&(2.51)\\
&= \exp\left[-\int_0^t h_0(u)du\right]^{\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z})} \\
&= \exp\big[-H_0(t) + H_0(0)\big]^{\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z})} \\
&= \big\{\exp[-H_0(t)]\exp[H_0(0)]\big\}^{\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z})} \\
&= \left[S_0(t)\frac{1}{S_0(0)}\right]^{\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z})} \\
&= [S_0(t)]^{\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z})} &&(2.52)
\end{aligned}$$

i.e. we have incorporated the variables into the survival function. Furthermore, we can rewrite (2.51) using (2.14) to give

$$\begin{aligned} S(t|\boldsymbol{Z}) &= \exp\left[-\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z})\int_0^t h_0(u)du\right] \\ &= \exp\left[-\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z})H_0(t)\right] \end{aligned} \tag{2.53}$$

i.e. we have expressed the survivor function in terms of the cumulative baseline hazard function, $H_0(t)$, which will come in handy in Section 4.4.2.

To identify the subset of the p variables that affects the survival time significantly, we cannot use the likelihood function in (2.22). In the likelihood function we compute the joint probability that each observed failure occurred at the observed times, $f(t)$, and the censored individuals survived at least to the time of censoring, $S(t)$. As we do not have a known (baseline) hazard distribution, we cannot compute the density function nor the survival function. In fact, we do not even assume anything on the form of the baseline hazard function, except that it should always be positive and only defined for $t > 0$. All we require is the hazard functions of any two individuals to be proportional.

However, we can use this assumption to calculate the PL by comparing failing subjects to those not failing at each time, t. Let there be n observations, k distinct and observed failures, and $(n-k)$ right-censored observations. Let $t_1 < t_2 < \ldots < t_k$ be the k ordered failure times and $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_k$ their variable vectors.

Furthermore, let $R(t_i)$ be the risk set at time t_i, i.e. the set of all individuals with survival time at least t_i. The probability that, for failure at time t_i, the failure is on the individual as observed, conditionally on the risk set $R(t_i)$, is the ratio between the hazard function for the observed individual and the sum of hazard functions for the risk set. Using (2.47), we get the PL, L_p, for the variable coefficient vector, $\boldsymbol{\beta}$, at time i

$$\begin{aligned} L_p(\boldsymbol{\beta})_i &= p(\text{subject with } \boldsymbol{z}_i \text{ fails at } t_i|\text{some subject failed at } t_i) \\ &= \frac{p(\text{subject with } \boldsymbol{z}_i \text{ fails at } t_i)}{p(\text{some subject in risk set failed at } t_i)} \\ &= \frac{h(t_i|\boldsymbol{z}_i)}{\sum_{l\in R(t_i)} h(t_i|\boldsymbol{z}_l)} \\ &= \frac{\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}_i)}{\sum_{l\in R(t_i)} \exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}_l)} \end{aligned} \tag{2.54}$$

As the baseline hazard function, $h_0(t_i)$, cancels out, we do not need to estimate it, and implies that the actual times of failure are not important - just the ordering. Censoring times are not important either, as long as we keep track of the risk sets. The key idea is that the PL shifts focus from survival times and the survival distribution to the (relative) hazard of failure.

We get a contribution for each of the k failures, and the joint probability or the PL is

$$L_p(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_i)}{\sum_{l \in R(t_i)} \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_l)} \tag{2.55}$$

giving us the *Log-Partial Likelihood (LPL)*

$$l_p(\boldsymbol{\beta}) = \sum_{i=1}^{k} \left\{ \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_i - \log \left[\sum_{l \in R(t_i)} \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_l) \right] \right\} \tag{2.56}$$

To estimate the parameters we use the Newton-Raphson iterative method described in (2.27)-(2.29) to obtain the *Maximum Partial Likelihood Estimate (MPLE)*, $\hat{\boldsymbol{\beta}}$, (Raftery et al., 1995), (Volinsky, 1997), (Lee and Wang, 2003), (Ibrahim et al., 2005), by solving

$$\frac{\partial l_p(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{k} [z_{ji} - A_{ji}(\boldsymbol{\beta})] = 0 \qquad j = 1, \ldots, p \tag{2.57}$$

where

$$A_{ji}(\boldsymbol{\beta}) = \frac{\sum_{l \in R(t_i)} z_{jl} \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_l)}{\sum_{l \in R(t_i)} \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_l)} \tag{2.58}$$

is the derivative stemming from the second part of the summation in (2.56). The second partial derivatives of the LPL, that we can use to validate that $\hat{\boldsymbol{\beta}}$ is a maximum, are

$$I_{jj'}(\boldsymbol{\beta}) = \frac{\partial^2 l_p(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_j'} = -\sum_{i=1}^{k} C_{jj'i}(\boldsymbol{\beta}) \qquad j, j' = 1, \ldots, p \tag{2.59}$$

where

$$C_{jj'i}(\boldsymbol{\beta}) = \frac{\sum_{l \in R(t_i)} z_{jl} z_{j'l} \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_l)}{\sum_{l \in R(t_i)} \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_l)} - A_{ji}(\boldsymbol{\beta}) A_{j'i}(\boldsymbol{\beta}) \tag{2.60}$$

The covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\hat{\mathrm{V}}(\hat{\boldsymbol{\beta}}) = \hat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}) = \left[-\frac{\partial^2 l_p(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}} \right]^{-1} \tag{2.61}$$

To identify significant risk factors we can use hypothesis testing and selection methods as described earlier for the parametric models. We simple replace the log-likelihood function with the LPL function.

As mentioned, the above derivation assumes that we have k distinct failure times. Now, assume that we have k distinct uncensored times $t_1 < t_2 < ... < t_k$ among n observed survival times. Let m_i be the number of patients who fail at t_i, $R(t_i)$ the risk set at time t_i, and r_i the size of the risk set. From $R(t_i)$ we can randomly select m_i subjects, each selection denoted $\boldsymbol{u}_i$. Let $\boldsymbol{U}_i$ denote the set that contains all the $\boldsymbol{u}_i$'s.

When the survival times are discrete observations, e.g. number of days post stroke, the tied observations are true ties (happen at exact same time). In this case, (Cox, 1972) has proposed a logistic model for the hazard function

$$\frac{h_i(t)dt}{1-h_i(t)dt} = \frac{h_0(t)dt}{1-h_0(t)dt} \exp\left(\sum_{j=1}^{p} \beta_j z_{ji}\right) = \frac{h_0(t)dt}{1-h_0(t)dt} \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_i) \tag{2.62}$$

which corresponds to (2.49) for continuous survival times. Replacing the i'th term in (2.55) with the corresponding term for tied survival times

$$\frac{\exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_{\boldsymbol{u}_i^*})}{\sum_{\boldsymbol{u}_i \in \boldsymbol{U}_i} \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_{\boldsymbol{u}_i})} \tag{2.63}$$

we get the PL function for tied (discrete) survival times, (Lee and Wang, 2003), (Ibrahim et al., 2005)

$$L_p(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_{\boldsymbol{u}_i^*})}{\sum_{\boldsymbol{u}_i \in \boldsymbol{U}_i} \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_{\boldsymbol{u}_i})} \tag{2.64}$$

where $\boldsymbol{z}_{\boldsymbol{u}_i} = \sum_{k \in \boldsymbol{u}_i} \boldsymbol{z}_k = (z_{1\boldsymbol{u}_i}, \ldots, z_{p\boldsymbol{u}_i})$, using $z_{l\boldsymbol{u}_i}$ for the sum of the j'th variable of the m_i subjects in $\boldsymbol{u}_i$. Furthermore, $\boldsymbol{u}_i^*$ is the set of m_i subjects who failed at time t_i and $\boldsymbol{z}_{\boldsymbol{u}_i^*} = \sum_{k \in \boldsymbol{u}_i^*} \boldsymbol{z}_k = (z_{1\boldsymbol{u}_i^*}, \ldots, z_{p\boldsymbol{u}_i^*})$, using $z_{j\boldsymbol{u}_i^*}$ for the sum of the j'th variable of the m_i subjects in $\boldsymbol{u}_i^*$.

2.4.2 Estimation of Survival Function

As we do not know the exact form of the baseline hazard function or the survival function, we cannot estimate the survival function simply be inserting the estimates of the parameters and coefficients. Instead, (Breslow, 1974) assumes that the baseline hazard function is constant between each pair of successive observed failure times and propose the following estimate of the baseline cumulative hazard function

$$\hat{H}_0(t) = \sum_{t_i \le t} \frac{\text{number of failures at time } t_i}{p(\text{some subject in risk set failed at time } t_i)}$$
$$= \sum_{t_i \le t} \frac{m_i}{\sum_{l \in R(t_i)} \exp(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{z}_l)} \tag{2.65}$$

Using (2.20), the estimate of the baseline survival function is

$$\hat{S}_0(t) = \exp\left[-\hat{H}_0(t)\right] = \prod_{t_i \le t} \left\{ \exp\left[\frac{m_i}{\sum_{l \in R(t_i)} \exp(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{z}_l)}\right]\right\} \tag{2.66}$$

Inserting this into (2.52) gives

$$\hat{S}(t, \boldsymbol{Z}) = \left[\hat{S}_0(t)\right]^{\exp(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{Z})} \tag{2.67}$$

that we can use to estimate the probability that a patient survives longer than time t, when the patient has risk factors $\boldsymbol{z}$.

2.4.3 Neural Interpretation

As shown in the previous section, an advantage of the CPH model is that once we have estimated the variable coefficients, we can estimate the parameters of the time dependent baseline hazard function. This sequential estimation can be used to express the CPH model using a *Multi Layer Perceptron (MLP)multi layer perceptron*, Bishop (2006), as shown in (Bakker et al., 2000) and (Bakker et al., 2004).

Recall that the usual likelihood function, the probability of observing the survival data, $\boldsymbol{D}$, given the risk factor coefficients, $\boldsymbol{\beta}$, is

$$L(\boldsymbol{\beta}) = p(\boldsymbol{D}|\boldsymbol{\beta}) = \prod_{t_i \in \text{ uncensored}} f(t_i, \boldsymbol{\beta}) \prod_{t_i^+ \in \text{ censored}} S(t_i^+, \boldsymbol{\beta}) \tag{2.68}$$

as given in (2.22). Note that in (Bakker et al., 2000) and (Bakker et al., 2004), the symbol $\boldsymbol{w}$ is used to denote the parameter vector $\boldsymbol{\beta}$, and F is used to denote the survival function S. The input to the neural network is the risk factor vector $\boldsymbol{Z}$. The weights in the first layer (input to hidden neurons) are $\boldsymbol{w}$. Using exponential transfer functions, Bishop (2006), and a fully connected network, see Figure (1), p. 3 in (Bakker et al., 2000), the output of the hidden units is $h(\boldsymbol{Z}) = \exp(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{Z})$, corresponding to the proportional part of the hazard function. One hidden unit corresponds to the proportional hazards model. With more hidden units we can model non-proportional hazards.

The weights, $\boldsymbol{v}$, in the second layer (hidden to output neurons) are minus the integral up to time t_i of the baseline hazard. Again, using exponential transfer functions, the output of the network at neuron i is $F_i(\boldsymbol{z}) = \exp\left[v_i h(\boldsymbol{z})\right]$ where $v_i = -\int_0^{t_i} h_0(u)du$, i.e. the output corresponds to the survival function for a patient with risk factor vector $\boldsymbol{z}$, i.e. the probability that the patient survives up to time t_i.

We can assume that the baseline hazard has a specific form, or we can assume that the baseline hazard is constant between each pair of successive observed failure times, and use the Breslow estimate in (2.65). The latter approach corresponds to the key idea in the CPH model, where we do not need to model the baseline hazard. The update (estimation) of the parameters (network weights) is done sequentially by feeding the network with cases and update using e.g. back-propagation, Bishop (2006). We need one output node for each time, t_i, making the MLP interpretation a discrete version of survival analysis.

We included this section to show that we do not necessarily need a commercial statistical package to use the CPH model, but can express the CPH model using neural networks normally used in pattern recognition, Bishop (2006).

2.4.4 Time Dependent Variables

A fundamental assumption of the CPH model is the proportional hazards assumption. If the hazards are not proportional, the linear component of the model varies with time. To examine whether this assumption is valid, we use two types of plots for each risk factor, the Schoenfeld plot and the log-cumulative hazard plot. The log-cumulative hazard plot is a straightforward plot that we can use in advance of model fitting to test for non-proportional hazards, while the Schoenfeld plot is based on the residuals of the fitted CPH model.

2.4.4.1 The Log-Cumulative Hazard Plot

According to (2.49), the hazard at time t for the i'th subject is

$$h(t|\boldsymbol{z}_i) = h_0(t)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}_i) \tag{2.69}$$

Integrating over t on both sides gives

$$H(t|\boldsymbol{z}_i) = H_0(t)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}_i) \tag{2.70}$$

where $H(t|\boldsymbol{z}_i)$ and $H_0(t)$ are cumulative hazard functions. Taking the logarithm on both sides yields

$$\log H(t|\boldsymbol{z}_i) - \log H_0(t) = \exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}_i) \tag{2.71}$$

which shows that in the CPH model, the difference in the log-cumulative hazards does not depend on time. Hence, if we plot the log-cumulative hazards for subjects with different risk factor values against time, the lines should be parallel, and for discrete risk factors, we can plot lines for each category. For continuous risk factors we need to transform the values into categorical values, e.g. using quartiles. If we plot the log-cumulative hazards against $\log(t)$ instead of t, the vertical separation will be an estimate of the logarithm of the relative hazard.

An estimate of the cumulative hazard function is obtained from an estimate of the survival function using . The survival function is estimated using the Breslow estimate in (2.67).

2.4.4.2 The Schoenfeld Plot

Another approach is to fit the residuals of the fitted CPH model(s). We can use the residuals to detect whether there is any time dependency for each variable after allowing for the effects of the variable that is expected to be independent of time.

There exist numerous residuals, but the Schoenfeld residual has the advantage that we do not require an estimate of the cumulative hazard function, and that there is not a single value of the residual for each individual, but a set of values, one for each variable included in the fitted CPH model. The i'th Schoenfeld residual for Z_j, the j'th variable in the model, is given by

$$r_{P_{ji}} = \delta_i\{z_{ji} - \hat{a}_{ji}\} \tag{2.72}$$

where

$$\hat{a}_{ji} = \frac{\sum_{l\in R(t_i)} z_{jl}\exp(\hat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{z}_l)}{\sum_{l\in R(t_i)}\exp(\hat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{z}_l)} \tag{2.73}$$

and $R(t_i)$ is the risk set at time t_i. We get non-zero contributions for uncensored observations only. The i'th Schoenfeld residual for Z_j is an estimate of the i'th component of the first derivative of the logarithm of the LPL function with respect to β_j, which, from (2.56), is given by

$$\frac{\partial l(\boldsymbol{\beta})}{\partial\beta_j}\sum_{i=1}^{n}\delta_i\{z_{ji} - a_{ji}\} \tag{2.74}$$

where

$$a_{ji} = \frac{\sum_l z_{jl} \exp(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{z}_l)}{\sum_l \exp(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{z}_l)} \tag{2.75}$$

The i'th term in the summation, evaluated at $\hat{\boldsymbol{\beta}}$, is the Schoenfeld residual for Z_j given in (2.72). Since

$$\left.\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j}\right|_{\hat{\boldsymbol{\beta}}} = 0 \tag{2.76}$$

the Schoenfeld residuals must sum to zero. However, (Grambsch and Therneau, 1994) showed that the scaled Schoenfeld residuals, $r^*_{P_{ij}}$, defined as

$$r^*_{P_{ji}} = k\mathrm{V}(\hat{\boldsymbol{\beta}}) r_{P_{ji}} \tag{2.77}$$

where k is the number of failures among the n individuals, and $\mathrm{V}(\hat{\boldsymbol{\beta}})$ is the covariance matrix of the parameter estimates in the fitted CPH model, are more effective at detecting departures from the assumed model. (Grambsch and Therneau, 1994) also show that the expected value of (2.77) is given by

$$E(r^*_{P_{ji}}) \approx \beta_j(t_i) - \hat{\beta}_j \tag{2.78}$$

where $\beta_j(t_i)$ is the time varying coefficient of Z_j at the i'th failure time t_i, and $\hat{\beta}_j$ is the estimated value of β_j in the fitted CPH model. If we plot the values of $r^*_{P_{ji}} + \hat{\beta}_j$ against the failure times we get information on the form of the time dependent coefficient, $\beta_j(t_i)$. If these values are well fitted by a horizontal line (linear model with slope zero), the coefficient of Z_j does not depend on time (is constant) and the proportional hazards assumption is satisfied.

Chapter 3

The Bayesian View and what it is all about

Probability is relative, in part to our ignorance, in part to our knowledge.

P.S. de Laplace (1840)

Modeling in science remains, partly at least, an art. Some principles exist, however, to guide the modeler. The first is that all models are wrong; some though, are better than others, and we can search for the better ones. At the same time we must recognize that eternal truth is not within our grasp. The second principle (which applies also to artists) is not to fall in love with one model, to the exclusion of alternatives.

McCullagh and Nelder (1983)

The previous chapter showed a common approach to survival analysis, and in particular how to apply the CPH model to survival data. It was also an example of how a “frequentist” or a “classical statistician” would approach the problem. Much has been written about the differences between the “frequentists” and their counter-colleagues, the “Bayesians”, see e.g. (MacKay, 2003) for a thorough (and sometimes quite amusing) discussion. In this chapter we will outline a few important (and interesting) differences, and we will show Bayesian ways to approach survival analysis.

3.1 Bayesians vs. frequentists

A classical probability is a physical property of the world, e.g. the probability that a peanut (more exotic than jelly) butter sandwich will land upside down. A Bayesian probability is a subjective quantity defined as the *degree of belief* that the sandwich will land upside down. In the former case, we can repeat the experiment a number of times (hence the name frequentist) to get an estimate of the probability (and a messy floor), while it is up to the individual Bayesian to assign a degree of belief.

Bayesian probabilities can, by definition, change depending on who assigns them, or how much information is available. Indeed, one name for Bayesian probability is "personal probability". Well, at the end of the day, the sandwich does obey Newton's laws, and in theory, although difficult, we are able to model how the sandwich will behave. The validity of this model and the accuracy of our prediction will certainly depend on how much information is available!

Take an everyday statement from a newspaper weather forecast. Does it make sense to say that there is a 90% chance of rain today? It will either rain or not. A 90% shower does not exist! Should we dump the newspaper? No, because the forecast says that in 90% of previously known weather situations like the one right now (in terms of available information about the atmosphere etc.), it has rained.

How do we use this information? We translate it into advice. If we have an important errand to do, we might get lucky and have a dry day, but chances are that we are better off bringing our umbrella. If we always respond in that way to that advice, on our deathbed, we will have been glad we did about 90% of the time, and we would have brought along a superfluous umbrella only 10% of the time. On average.

So what do we do if the prediction is a 50% chance of rain? As with any other percentage, it is up to you. This is where other factors come in to play. If we live in Waikiki (southern part of the Hawaiian island Oahu with nice little showers) rather than the North Shore (same island, drenching monsoons), we prefer not to carry too much stuff around, but if we are on our way to a date with the Sports Illustrated cover girl, wearing an expensive Armani suit, we cannot afford taking the risk - and we definitely cannot repeat the experiment! The point is that the statistical statement leaves it up to us to decide. It is a report of experience and a projection from experience, but it is also fundamentally a recommendation.

The Bayesian's concede that there is some meaning to objective probability for repeated events like coin (or sandwich) tosses, but not for single events like F.C. Copenhagen qualifying for Champions League. Bayesian's complicate the psychology of decisions like in the Prisoner's Dilemma case[1], and also believe that psychologizing the problem changes the fact that a 75% rain day is 3 times riskier than a 25% rain day for a person who is afraid of umbrellas (umbrellafobia?).

The subjective element is the major point of criticism, as (Bayesian) probabilities at a first glance seem rather arbitrary. Bayesians acknowledge this, but argue that you cannot make inference without making assumptions - as in the examples above. This has also led to the phrase, "Bayes is optimal - when you are!".

Consequently, Bayesians typically refer to X as an *uncertain* variable, because the value of X is uncertain, not *random*, as the classical statisticians claim. The variable is not random - nothing is! Variables are (just) uncertain and the more information we have on a variable, the less uncertain it is. Even random number generators are not random, but governed by deterministic albeit quite complex processes.

Now that we have illustrated the difference between frequentist and Bayesian statistics, let us return to what is all about - modeling. Everything in life can be modeled: traffic, speech, stock prices. The *machine learning* approach to data modeling is to propose or *select* a rather flexible model, and then search for the parameter configuration that best (according to some goodness-of-fit measure) explains the data. We say that we *learn* the model parameters allowing us to *infer* information about one or more of the variables in the model.

The question is, do we need to consider more than one model, and one set of parameters? "Well, usually model G is a good choice for modeling this kind of data. Now, all we need to do is fit the model, and use a goodness-of-fit test to see if the model is significant". However, in many cases a model with a given parameter configuration will be significant, but there may be other models or parameter configurations that are capable of explaining the data just as well or even better. These models should *not* be ignored! How to select a model, or how to use more than one model is the topic of the sections to come.

[1] See e.g. http://plato.stanford.edu/entries/prisoner-dilemma/ for a description.

3.2 Model Selection

Imagine that an evil researcher has persuaded us to use a single model to explain some data. He also tells us that we do not need to average over all possible parameter settings. All we need is to search for the parameters that maximize the probability of the observed data given the parameters, $p(\boldsymbol{D}|\boldsymbol{\theta})$, known as the *likelihood* of $\boldsymbol{\theta}$, and not the likelihood of the data, as it is a function of the parameters rather than the data. The optimal parameter estimate is the *Maximum Likelihood (ML)*, (MacKay, 2003), estimate given by

$$\boldsymbol{\theta}^*_{ML} = \arg\max_{\theta} p(\boldsymbol{D}|\boldsymbol{\theta}) \tag{3.1}$$

If we allow hidden variables $\mathbf{X}$ in the model, we need to integrate (or sum for discrete variables) over all possible values of the hidden variables to get the likelihood

$$p(\boldsymbol{D}|\boldsymbol{\theta}) = \int_{\mathbf{X}} p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{D}|\mathbf{X},\boldsymbol{\theta})d\mathbf{X} \tag{3.2}$$

The integrand in (3.2) is the complete-data likelihood, and (3.2) itself is the incomplete-data likelihood. The observed data is an incomplete account of all factors in the model. For a given data set and parameter setting, we can infer the posterior distribution over the hidden variables using Bayes' rule

$$p(\mathbf{X}|\boldsymbol{D},\boldsymbol{\theta}) = \frac{p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{D}|\mathbf{X},\boldsymbol{\theta})}{p(\boldsymbol{D}|\boldsymbol{\theta})} \tag{3.3}$$

The prior, $p(\mathbf{X}|\boldsymbol{\theta})$, is a subjective quantity that should reflect our *a priori* beliefs about the hidden variables based on all available information, e.g. expert knowledge.

We recognize the denominator in (3.3) as (3.2). With hidden variables, the optimal parameter setting in (3.1) is hard to find, an we need to alternate between estimating the posterior distribution over the hidden variables for a particular setting of the parameters, and re-estimating the optimal parameter setting given that distribution over the hidden variables. We know this scheme as the *Expectation-Maximization (EM)* algorithm (Dempster et al., 1977).

Not satisfied with his results, the evil researcher allows us to use a prior over the parameters, $p(\boldsymbol{\theta})$, to obtain the *Maximum A Posteriori (MAP)* parameter configuration, (MacKay, 2003), (Gelman et al., 2004), (Ibrahim et al., 2005), (Heckerman, 1995)

$$\boldsymbol{\theta}^*_{MAP} = \arg\max_{\theta} \{p(\boldsymbol{\theta})p(\boldsymbol{D}|\boldsymbol{\theta})\} \tag{3.4}$$

However, it is still a point estimate of the true posterior distribution over parameters, and it is possible to find different ML or MAP estimates for different model parameterizations even though the prior and the likelihood are the same. Hence, the MAP estimate is not strictly a Bayesian approach, although often incorrectly claimed to be. The key to a Bayesian approach is not just to include a prior, but to average over all uncertain variables.

We also have to accept that we cannot model all aspects of our data exactly. The aspects of the data that our model cannot account for are referred to as noise. Which aspects are relevant to model and which could be considered noise is hard to know, but we need to find a model that is not too complex nor too simple. More complex models will be better to adapt their shape to fit the data, e.g. a sixth-order polynomial can exactly fit six points, but the additional parameters may not represent anything useful, perhaps those six points are really just distributed about a line. The complexity is generally measured by counting the number of free parameters in the model. If our model is too complex, predictions will be poor, as we have not just fitted the trends in the data, but also the noise (over-fitting), and if the model is too simple, it will not be able to capture all the trends in the data, also giving poor predictions. This is the well-known *bias-variance* trade-off, Bishop (2006). ML or MAP methods do not account for model complexity, and ways to overcome this such as cross-validation are computationally prohibitive with a large number of parameters or large parameter intervals. If we average over all possible configurations, we penalize models with more degrees of freedom and favor simpler models. The phenomena is referred to as *Occam's razor*, (MacKay, 2003), and is illustrated in Figure 3.1, where Dr. Jones has an archaeological digging somewhere in South America.

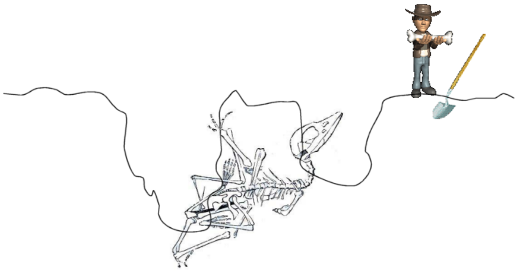

Figure 3.1: Illustration of Occam's razor.

When Dr. Jones start digging, he discovers what he believes to be a skeleton from a rare dinosaur species, R. Unfortunately, some very bad German guys are in Dr. Jones' tail, and he needs to hurry away. Dr. Jones returns to the University with just a single bone and then realizes that the bone could also be a bone from a very common species, C. Dr. Jones looks in his diary to find that the particular bone is believed to have been within the following dimensions (length l cm, diameter d cm):

$$C : \mathrm{l} \,\epsilon\, [50; 200]\,, \mathrm{d} \,\epsilon\, [10; 60] \qquad R : \mathrm{l} \,\epsilon\, [60; 70]\,, \mathrm{d} \,\epsilon\, [15; 20] \tag{3.5}$$

The bone has dimensions $\boldsymbol{D} : \{\mathrm{l} \ = 62, \mathrm{d} \ = 19\}$ which could be any of the species. Dr. Jones is still puzzled! However, he also knows that the common species is believed to have outnumbered the rare species 100:1 that we can express in the ratio of the model priors

$$\frac{p(M_R)}{p(M_C)} = \frac{1}{100} \tag{3.6}$$

Recalling what he learned in “Introduction to Probability Theory”, Dr. Jones knows that

$$p(\boldsymbol{D}|M_R)p(M_R) = p(M_R|\boldsymbol{D})p(\boldsymbol{D}) \tag{3.7}$$

$$p(\boldsymbol{D}|M_C)p(M_C) = p(M_C|\boldsymbol{D})p(\boldsymbol{D}) \tag{3.8}$$

yielding

$$\frac{p(M_R|\boldsymbol{D})}{p(M_C|\boldsymbol{D})} = \frac{p(\boldsymbol{D}|M_R)p(M_R)}{p(\boldsymbol{D}|M_C)p(M_C)} \tag{3.9}$$

To compute the probability of the data, $\boldsymbol{D}$, given M_R, we need to average over the parameters, l and d

$$p(\boldsymbol{D}|M_R) = \int_l \int_d p(\boldsymbol{D}|M_R, l, d)p(l)p(d)\mathrm{d}l \ \mathrm{d}d \tag{3.10}$$

and similar for M_C. For simplicity, Dr. Jones assumes that the parameters can take on only integer values in the given intervals, and that there is a uniform prior over the parameters, i.e. that any length and diameter is equally likely, yielding

$$\frac{p(M_R|\boldsymbol{D})}{p(M_C|\boldsymbol{D})} = \frac{1}{100} \frac{\frac{1}{11}\frac{1}{6}}{\frac{1}{151}\frac{1}{51}} = 1.1668 \tag{3.11}$$

Although C is a much more common species, it is also known to have a large variety of bone length and diameter, and this makes Dr. Jones finding slightly more in favor of the rare species!

Both species models have the same number of free parameters, bone length and diameter, and in this case we can relate the model complexity to the range of data sets (bone findings) they can capture. In the rare species model, the probability is concentrated over a small range of data, making it capable of modeling potentially fewer data sets than the common species model with the ability to model a wide range of data. We also note that the rare species model is in fact a sub-model of the common species model. Since the marginal likelihood as a function of the data, $\boldsymbol{D}$, should integrate to one, the rare species model can assign a higher marginal likelihood to Dr. Jones bone than the common species model. For another quite amusing (homemade) example of Occam's razor, please see Appendix A. Having shown how to use Occam's razor to perform model *selection*, we will move onto model averaging and apply it to survival analysis.

3.3 Bayesian Model Averaging

Historically, statisticians have focused much more on *within* model uncertainty (the parameter uncertainty) than *between* model uncertainty (model uncertainty in short). Once we have selected and fitted a model, the uncertainty reported for values such as predicted values or parameters estimates, is limited to the uncertainty associated with the statistical distributions embedded in the model. If we ignore the uncertainty associated with the model selection procedure, we underestimate the total uncertainty leading to overconfident conclusions.

If our aim is to predict, say, the survival time of a stroke patient, models could have different subset of risk factors as predictors of the survival time. The model with subset A may be able to predict the survival times for female patients above 80, while another with parameter subset B can predict male, diabetes patients and so forth. Using stepwise selection, model A might be the best model, but if the model is only able to explain 70% of the data, we would probably obtain better predictive results if we used a *combination* or an *average* of model A and B. A historical overview of model combination can be found in (Volinsky, 1997), Chapter 1.

The truly Bayesian approach would be to average over all possible models and all possible parameter values as discussed in Section 4.2. However, this can be a quite complicated and computational expensive approach. Instead, we average over all possible models, and for each model we use a point estimate of the parameters. This allows us to use an ensemble of models, and is in some way the happy medium between frequentist statistics and a fully Bayesian approach.

Imagine a data set $\boldsymbol{D} = \{\mathbf{d}_1, \mathbf{d}_2, \ldots \mathbf{d}_N\}$, and let the models have parameters $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots \theta_p\}$. To select a model, Bayesians calculate the posterior distribution over a set of K possible models, $\boldsymbol{M} = \{M_1, M_2, \ldots, M_k\}$, given some a priori knowledge and the observations. The a priori knowledge is expressed in the prior over models, $p(\boldsymbol{M})$, and their parameters, $p(\boldsymbol{\theta}|\boldsymbol{M})$.

For some quantity of interest, Δ, e.g. a future observation, we can use Bayes' rule, (MacKay, 2003), to compute the posterior distribution of Δ given the data, $\boldsymbol{D}$

$$p(\Delta|\boldsymbol{D}) = \sum_{k=1}^{K} p(\Delta|M_k, \boldsymbol{D}) p(M_k|\boldsymbol{D}) \tag{3.12}$$

where $p(\Delta|M_k, \boldsymbol{D})$ is the *predictive distribution* and $p(M_k|\boldsymbol{D})$ is the *Posterior Model Probability (PMP)*, (MacKay, 2003), for model M_k. Hence, we have a weighted *average* of the predictive distributions for each model weighted by their posterior probability. This approach is known as *Bayesian Model Averaging*

(BMA),, (Hoeting et al., 1999), (Gelman et al., 2004), (Ibrahim et al., 2005), (Volinsky, 1997), (Volinsky et al., 1997), and has the obvious advantage that we do not have to select a single model, we do not use (artificial) p-values and significance levels, and variables are not "in" or "out". BMA has been successfully applied to generalized linear models, (Raftery, 1996).

However, even if we define a set of discrete models, it is often practically impossible to average over all models. Instead, we need a way to quickly and efficiently locate a smaller subset of data supported models to average over. We could also use $p(M_k|\boldsymbol{D})$ to locate the model with maximum posterior probability. This would give $K = 1$, and we call it *model selection*. We saw an example of this in the previous section. Madigan et al. (1993) present strategies for graphical model selection, Raftery (1995) use Bayesian model selection in social research, and Shan (2001) use model selection for learning in belief networks with incomplete data.

Note, however, that BMA is *not* model combination as pointed out by (Minka, 2000). BMA is best thought of as "soft model selection" and answers the question: "Given that all of the data so far was generated by exactly one of the hypotheses, what is the probability of observing the new data point, Δ?". The weights (PMPs) in BMA only reflect a statistical inability to distinguish the hypothesis based on limited data. As more data becomes available, it will be easier to compare and value the models, and BMA will always assign its maximum weight to the most probable hypothesis, like a posterior mean of a Gaussian will approximate the sample mean. So the assumption is that just one hypothesis is responsible for the data. If the true hypothesis is *not* within the hypothesis space, BMA will not be able to select the true model even with unlimited data available.

To illustrate the difference between BMA and model combination, Minka (2000) gives an example similar to the following. Let the true class assignments be as illustrated in Figure 3.2. A data point is in class "o" if it is within at least two of the circles. Let the circles be our three hypotheses. The optimal way to combine them is to use the uniform weighting scheme giving 100% accuracy. However, BMA will focus on the top-most circle which is most homogenous and thus most likely to have generated the data (note that the circle placement is not symmetric). With more data available, BMA will assign greater weight to the top-most circle, and in the limit assign it weight 1. As long as the error rates are just slightly different, BMA will assign a larger weight to the hypothesis with the lowest error rate. To do model combination we should not use BMA on the models themselves, rather we should ask the question "Given that all of the data so far was generated by a linear combination of the hypotheses, what is the probability of observing the new data point, Δ?". If we apply BMA to this new hypothesis space, the true hypothesis is now included, and on the circle

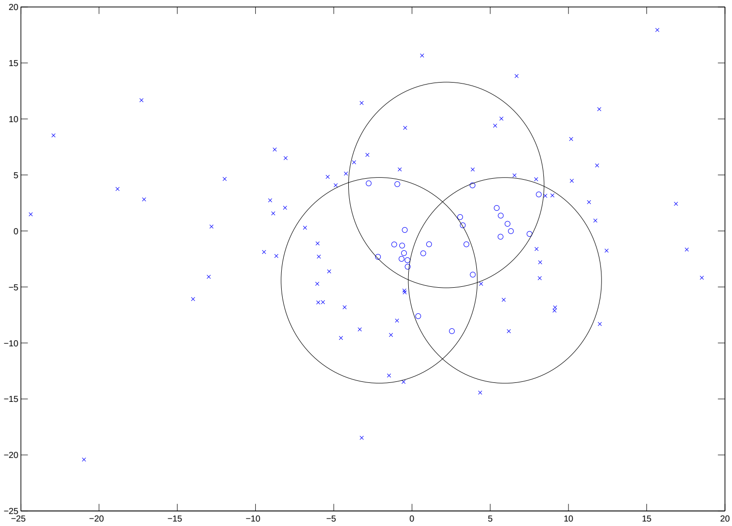

Figure 3.2: Classification problem. Data point is in class 'o' if it is inside at least two of the circles. The optimal solution is a uniform voting between the circles.

problem it will converge to the uniform vote.

3.3.1 BMA for Cox Proportional Hazards Models

If we have p potential risk factors, we have $K = 2^p$ potential models (without other constraints). Fortunately, most of the models get very little support from the data. If we are unable to average over all possible models, a good approximation is to average over the subset of "best" models with respect to their posterior probabilities, including only models belonging to the set

$$A = \left\{ M_k : \frac{\max_l\{p(M_l|\boldsymbol{D})\}}{p(M_l|\boldsymbol{D})} \leq C \right\} \tag{3.13}$$

where (Madigan and Raftery, 1994a) have shown that $C = 20$ is a good approximation to an average over the complete model space, i.e. we include models whose posterior probability is at least 1/20 of that of the best model. Of course, the value of C can differ from problem to problem, and is a trade-off between accuracy (the model set should account for most of the PMP mass) and complexity (too many models can make the problem intractable). If we look closer at (3.12), we note that it has three components, each associated with computational difficulties.

3.3.1.1 Predictive Distribution

We get the *predictive distribution* by integrating over the parameters, $\boldsymbol{\theta}_k$, for model k

$$p(\Delta|M_k, \boldsymbol{D}) = \int p(\Delta|\boldsymbol{\theta}_k, M_k, \boldsymbol{D}) p(\boldsymbol{\theta}_k|M_k, \boldsymbol{D}) d\boldsymbol{\theta}_k \tag{3.14}$$

In general, for censored survival models such as the CPH model, the integral cannot be computed analytically, (Volinsky, 1997), (Ibrahim et al., 2005). Instead, we use the ML estimate, $\hat{\boldsymbol{\theta}}_k$, of the model parameters to give

$$p(\Delta|M_k, \boldsymbol{D}) \approx p(\Delta|M_k, \hat{\boldsymbol{\theta}}_k, \boldsymbol{D}) \tag{3.15}$$

However, we are able to compute the exact predictive distribution using *Markov Chain Monte Carlo (MCMC)* methods, see Section 4.2. Unfortunately, these methods are computationally quite expensive, and for the purpose of model averaging, we need to compute the predictive distribution for many models.

3.3.1.2 Posterior Model Probability

The posterior probability for model M_k is proportional to the product of the likelihood and the prior for model M_k

$$p(M_k|\boldsymbol{D}) \propto p(\boldsymbol{D}|M_k) p(M_k) \tag{3.16}$$

where

$$p(\boldsymbol{D}|M_k) = \int p(\boldsymbol{D}|\boldsymbol{\theta}_k, M_k) p(\boldsymbol{\theta}_k|M_k) d\boldsymbol{\theta}_k \tag{3.17}$$

is the integrated likelihood of model M_k, and $p(\boldsymbol{\theta}_k|M_k)$ is the prior density of the model parameters, $\boldsymbol{\theta}_k$, under model M_k. To compare two models, M_1 and M_2, we compute the ratio of the two posterior distributions

$$\frac{p(M_1|\boldsymbol{D})}{p(M_2|\boldsymbol{D})} = \frac{p(M_1)p(\boldsymbol{D}|M_1)}{p(M_2)p(\boldsymbol{D}|M_2)} \tag{3.18}$$

Assuming the two models are equally likely a priori, $p(M_1) = p(M_2) = \frac{1}{2}$, we get the ratio of the marginal likelihoods known as the *Bayes factor*. Approximate

Bayes factors and accounting for model uncertainty in generalized linear models are discussed in (Kass and Raftery, 1994).

Again, however, it is not possible to compute the integrated likelihood analytically. Instead, we use *Bayes Information Criterion (BIC)* approximation, (Ibrahim et al., 2005), (Heckerman and Chickering, 2000), (Volinsky, 1997), and Section 4.3.1.3.

$$\log p(\boldsymbol{D}|M_k) = \log p(\boldsymbol{D}|\hat{\boldsymbol{\theta}}_k, M_k) - \frac{d_k}{2}\log n \tag{3.19}$$

where n is the number of observations, d_k the number of free parameters to be estimated in model M_k, and $\hat{\boldsymbol{\theta}}_k$ the ML parameter estimate. We see that the first term increases with the model complexity (number of free parameters). A more complex model (that includes the simpler model) will always fit the data as well or better, but the second term also increases with d_k and penalizes more complex models, making BIC a way to balance gain and penalty. Given any two estimated models, the model with the lower value of BIC is the one to be preferred.

However, Volinsky (1997) and Volinsky and Raftery (2000) argue that for censored survival models, setting n to be the number of uncensored individuals rather than the total number of individuals, gives better predictive performance and corresponds to more appropriate priors on the parameters. See also Weakliem (1999) for a critique of the BIC for model selection.

When we apply BMA to the CPH model, the parameter vector is $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \boldsymbol{h}\}$, where $\boldsymbol{h} = \{h_0(t) : t \in \mathbb{R}_+\}$ is the baseline hazard and $\boldsymbol{\beta}$ are the risk factor coefficients. Using the PL in (2.55) as the likelihood for $\boldsymbol{\theta}$ with $\boldsymbol{h}$ integrated out, we get

$$p(\boldsymbol{D}|M_k) = \int PL(\boldsymbol{\theta}_k|M_k)p(\boldsymbol{\theta}_k|M_k)d\boldsymbol{\theta}_k \tag{3.20}$$

as an approximation to (3.17).

The justification for this approximation is that if we discard the time of failure (death), the PL becomes a full likelihood for the reduced data composed of the order in which individuals die, and the risk set, R_i, for each failure. The actual time of failure does not, as also mentioned in Section 2.4.1, contain much information about the risk factor coefficients, $\boldsymbol{\beta}$, the primary arguing point for competing models.

The second part of (3.16), the model prior, can be expressed as

$$p(M_k) = \prod_{j=1}^{p} \pi_j^{\delta_{kj}}(1-\pi_j)^{1-\delta_{kj}} \tag{3.21}$$

where $\pi_j \in [0;1]$ is the prior probability that $\beta_j \neq 0$, and δ_{kj} is an indicator of whether or not variable j is included in model M_k. Using $\pi_j = 0.5$ for all j corresponds to a uniform prior, while $\pi_j < 0.5$ for all j implies a penalty for complex models. Finally, $\pi_j = 1$ ensures that variable j is included in all models. In this thesis all models are assumed equally likely a priori, as we do not want to rule out any models. Instead, we let the data decide which models too choose, but as shown, it is easy to specify explicit prior model knowledge.

3.3.1.3 Subset Selection

As mentioned in Section 3.3, we need a way to quickly search through our model space, and select a subset of models to average over, when the hypotheses space is too large to include all models.

Let the full model have parameter vector $\boldsymbol{\theta}$, and sub-model M_k parameter vector $\boldsymbol{\theta}_k$. Then we can rewrite $\boldsymbol{\theta}_k$ as $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ so that model M_k corresponds to $\boldsymbol{\theta}_2 = \mathbf{0}$ of length q. The standard way to test a sub-model versus the full model is to use either (i) the *Likelihood Ratio Test (LRT)* statistic, (Collet, 2003), (Lee and Wang, 2003), (Volinsky, 1997)

$$\Lambda = -2\left[l(\tilde{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}})\right] \tag{3.22}$$

where $l(\cdot)$ is the LL, $\hat{\boldsymbol{\theta}}$ the MLE of $\boldsymbol{\theta}$ under the full model, and $\tilde{\boldsymbol{\theta}}$ the MLE with the restriction that $\boldsymbol{\theta}_2 = \mathbf{0}$, or (ii) the asymptotic normal distribution to the distribution of $\hat{\boldsymbol{\theta}}$.

In (i) we assume that Λ is approximately χ^2_q distributed under the hypothesized sub-model, with large Λ supporting evidence against the sub-model. In (ii) we use

$$\Lambda' = \hat{\boldsymbol{\theta}}_2^{\mathrm{T}} \boldsymbol{C}_{22}^{-1} \hat{\boldsymbol{\theta}}_2 \tag{3.23}$$

where $\boldsymbol{C} = \boldsymbol{I}^{-1} = \begin{pmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} \\ \boldsymbol{C}_{12}^{\mathrm{T}} & \boldsymbol{C}_{22} \end{pmatrix}$ is the inverse of the observed information matrix $\boldsymbol{I}$ with entries $I(\boldsymbol{\theta}_{uv}) = -\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_u \theta_v}$ for the full model so that $\boldsymbol{C}_{22}$ has size $q \times q$.

Recall that $\boldsymbol{\theta}_2$ is the vector of parameters that are 0 in the sub-model. We can consider $\boldsymbol{\theta}_2$ as the "error" source that we introduce in the sub-model versus the full model. Hence, under the hypothesized sub-model, Λ' is asymptotically χ^2_q distributed, again with large Λ' supporting evidence against the sub-model.

The "best" models with p variables are the models with the smallest Λ or Λ' values, where p is the length of $\boldsymbol{\theta}_1$. If we use Λ we need to (iteratively using

Newton-Raphson) compute ML estimates for each model considered. If we use Λ', we need to compute the inverse $\boldsymbol{C}_{22}^{-1}$ matrices for each model considered. This can be very expensive, but we can avoid this when the number of possible models, K, is not very large. Consider the normal linear model

$$\boldsymbol{Y}_{1\times n} = \boldsymbol{\theta}^{\mathrm{T}}_{1\times p}\boldsymbol{X}_{p\times n} + \boldsymbol{\epsilon} \tag{3.24}$$

where $\epsilon_i \sim N(0, \sigma_i^2), i = 1, \ldots, n$. In (Furnival and Wilson, 2000), the authors present algorithms for sequentially generating and evaluating, in terms of the *Residual Sum of Squares (RSS)*[2], all 2^p possible sub-models of (3.24), or the m best models of each size. All information in the data is contained in

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} & \boldsymbol{X}\boldsymbol{Y}^{\mathrm{T}} \\ \boldsymbol{Y}\boldsymbol{X}^{\mathrm{T}} & \boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}} \end{bmatrix} \tag{3.25}$$

of size $(p+1) \times (p+1)$. (Furnival and Wilson, 2000) then use sweep operations on this matrix to search for and evaluate models.

The principle is adopted in (Lawless and Singhal, 1978), where the authors present algorithms to search for sub-models in the non-linear regression model domain, and show that for non-linear models, we can substitute $\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ with the information matrix, $\boldsymbol{I}$, to give

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{I}\hat{\boldsymbol{\theta}} \\ \hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{I} & \hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{I}\hat{\boldsymbol{\theta}} \end{bmatrix} \tag{3.26}$$

onto which we can apply the same sweep operations and "fit" each of the 2^p models. Note that $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$ under the full model.

However, each model is not truly fitted, as the estimate $\tilde{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ for each sub-model are not the ML estimates of $\boldsymbol{\theta}$, but only approximations to these based on the asymptotic normal approximation to $l(\boldsymbol{\theta})$. Analogous to the RSS and covariance matrix for $\tilde{\boldsymbol{\theta}}$ in the normal model, we get the approximate LRT statistic, Λ', of (3.23), and the asymptotic covariance matrix $\boldsymbol{C}_{11}^{-1}$ for $\tilde{\boldsymbol{\theta}}_1$.

Using the algorithms of (Furnival and Wilson, 2000), we get (for each sub-model)

- the first order approximation to the MLE $\tilde{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ under the sub-model.
- the asymptotic covariance matrix $\boldsymbol{C}_{11}^{-1}$ for $\tilde{\boldsymbol{\theta}}_1$.
- the approximate LRT statistic, Λ', of (3.23).

[2] In RSS, the error for each case in the data set is squared, then added together and the square root is taken.

As our objective is model screening, we do not have to compute the exact MLE, $\tilde{\boldsymbol{\theta}}$, of $\boldsymbol{\theta}$ under each sub-model, and the exact LRT statistic, Λ, of (3.22), as the approximate values are adequate. To identify a subset of models to average over, we do not want to fit all 2^p models, just the $m \geq 1$ best models of size $1, \ldots, p$.

In (Furnival and Wilson, 2000), the authors develop a *Leaps and Bounds (L&B)* algorithm based on the statistical fact that for two linear models, A and B, if $A \subset B$ then $RSS(A) > RSS(B)$, because model B includes model A, and therefore model B is able to explain at least the data that model A is able to explain. By representing all models using a tree structure with the full model as the root node, the algorithm can compare RSS values and select subsets for further investigation.

The principle is best explained using the example in Figure 3.3. We have 4 potential variables A, B, C, and D and the full model $ABCD$ is the root node with RSS value 5 (shown in parenthesis). As all other models are sub-models of this model, the full model has the lowest RSS. In the next level we have all 3 parameter models and their respective RSS values. Model ABC has the lowest RSS value, so we calculate all 2 parameter models in this subset. Then we realize that model AB has a lower RSS value (12) than the three parameter model BCD (22). As the RSS can only increase when we remove variables, it implies that any of the two parameter models in model BCD have a higher RSS value than model AB. Same argument applies for model ACD (15), but not necessarily for model ABD (11). If we are looking for the best 2 parameter model, we do not have to consider the sub-models of model ACD and BCD. In

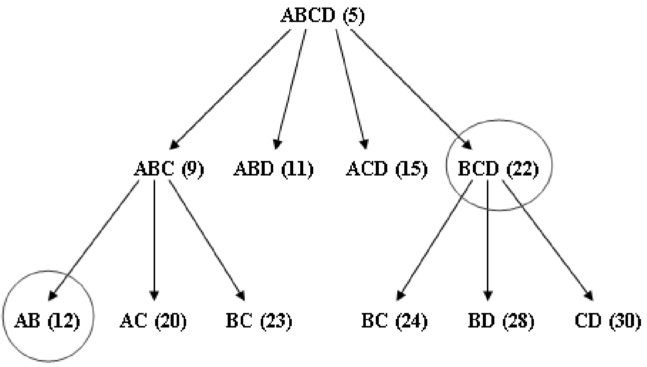

Figure 3.3: Demonstration of the L&B algorithm. Full model is $ABCD$. Two-parameter model AB has lower RSS than three-parameter model BCD.

(Furnival and Wilson, 2000) it is noted that the computational cost associated with the location of, say, the 10 best models of a given size is not much more than the cost of finding the single best model of the given size.

When we have non-linear models, we use the fact that $LL(A) > LL(B)$, if $B \subset A$, and when we apply the modified L&B algorithm to our hypotheses space, we get the previously listed output plus the m best models of each size. Optimally, we would like the models in (3.13), but as long as m is large enough, we get all the models included in (3.13) plus many not included, (Madigan and Raftery, 1994b). We can use the approximate LRT statistic to identify the models most likely to be in (3.13) by keeping only those models whose (approximate) PMP is at least $1/C'$ the PMP of the best model, where C' is greater than C from (3.13). In this work we use $C' = C^2$ as in (Madigan and Raftery, 1994b).

Practically, we use the logarithm of the PMP given by (3.16), where the LL is given by the BIC approximation in (3.19) with the LRT statistic inserted. We refer to this subset selection as the *soft Occam window* subset selection. Next, we loop all these models and make the true ML fit for each model. In turn this gives us the exact LRT statistic, Λ, and the exact BIC value for each model. Finally, we use the exact PMP values for the *hard Occam window* subset selection, accepting only those models that are in (3.13), (Hoeting et al., 1999), (Ibrahim et al., 2005), (Madigan and Raftery, 1994a), (Volinsky, 1997), (Madigan et al., 1993).

After normalizing the PMP over the model set, we can use them as weights in our model averaging. We can also use them as a measure of the model uncertainty. If we have many models with non-negligible posterior probabilities, we have a substantial amount of model uncertainty. On the other hand, if the the majority of the posterior probability mass is assigned to one model, the model is a reasonable stand-alone fit to the data. Whereas the standard methods such as stepwise selection identifies a single, “optimal” model, BMA selects an ensemble of models capable of fitting the data better or at least as good as the single, “best” model. In the experimental sections we will explore whether stepwise selections and BMA agree on which model is the best, and if these models are capable of fitting the data by themselves. Otherwise, we need more models to explain the data.

Because the L&B algorithm selects the m best models of *each* size, we can choose whether models with more likely sub-models are eliminated. Otherwise, the algorithm returns all models whose posterior model probability is within a factor of $1/C$ of that of the best model. In this work we keep the sub-models as we feel that this will give a more correct evaluation of the risk factors.

3.3.1.4 Evidence of Effect - those p-values again...

For a given model parameter, β_j, representing the j'th variable or risk factor, the point mass at zero represents the probability that this parameter equals zero and should not be included in the model. The sum of the PMPs for models that include this variable tells us how likely it is that this variable has an effect in the *average model* given the selected sub-domain of models. It is known as the *Posterior Probability of the Parameter (PPP)* or $p(\beta_j \neq 0)$, (Kass and Raftery, 1994), (Volinsky, 1997), (Volinsky et al., 1997). In (Kass and Raftery, 1994), the authors give standard rules of thump for interpreting this value. These are shown in Table 3.1. If we compare the PPP with the infamous p-value for measuring the significance of a variable, we will see that they are indeed very different. (Hubbard and Armstrong, 2005) provide a historical background on the widespread confusion of the p-value. Using p-values, we may fail to reject the null hypothesis "no effect" because either a) there is not enough data to detect the effect, or b) the data provide evidence for the null hypothesis. Using the PPP we can make this distinction.

PPP	Interpretation
$< 50\%$	positive evidence against an effect
$50 - 75\%$	weak evidence for an effect
$75 - 95\%$	positive evidence for an effect
$95 - 99\%$	strong evidence for an effect
$> 99\%$	very strong evidence for an effect

Table 3.1: PPP levels and their interpretation.

There are also several common misunderstandings about p-values, see e.g. (Sterne, 2001)

- The p-value is **not** the probability that the null hypothesis is true, and it is not a "rule" that p-values close to zero are significant. Frequentist statistics does not, and cannot, assign probabilities to hypotheses. Comparison of Bayesian and classical approaches shows that a p-value can be very close to zero, while the posterior probability of the null hypotheses is very close to unity. This is the Jeffreys-Lindley paradox, (Lindley, 1957).

- The p-value is **not** the probability that a finding is just a fluke. As the calculation of a p-value is based on the assumption that a finding is the result of chance alone, it cannot, at the same time, be used to measure the probability of that assumption being true.

- The p-value is **not** the probability of falsely rejecting the null hypothesis.
- The p-value is **not** the probability that a replicating experiment would not yield the same conclusion.
- $[1 - (p\text{-value})]$ is **not** the probability of the alternative hypothesis being true.
- The significance level of the test is **not** determined by the p-value. The significance level of a test is a value that should be decided a priori by the researcher, and is compared to the p-value or any other statistic calculated *after* the test has been performed.

The standard level of significance used to justify a claim of a statistically significant effect is 0.05 (see (Dallal, 2007) and (Bross, 1971) for a historical background to the origins of p-values and the choice of 0.05 as the cut-off for significance), and a p-value of 0.05 is often interpreted in the sense that the variable has a 5% chance of being zero, or that the null hypothesis ($H_0 : \beta_j = 0$) has a 5% chance of being true. The true interpretation is that, "If the null hypothesis is true, the probability of collecting data as extreme as or more extreme than the observed is 5%, *assuming* the observed data were the result of chance alone." See any textbook in statistics or (Johnson, 2005), (MacKay, 2003), (Bross, 1971). On the contrary, Bayesians are more interested in relevant questions to which you can assign a probability, such as "What is the probability that the model is true?", or, "What is the probability that the coefficient is non-zero?".

Many people believe that the p-value is a Bayesian probability (for example the posterior probability of the null hypothesis), but as shown in several examples in (MacKay, 2003) it is not, and in cases where we have a p-value below the "magical" 0.05 value, data can actually be in favor of the null-hypothesis (in the Bayesian sense)!

As mentioned, we can compute the posterior probability that a given risk factor has an effect, $p(\beta_j \neq 0)$, simply by summing the posterior probabilities of the models that include this variable. How large is the effect, given that there is one? The answer is given by the posterior distribution over β_j, namely

$$p(\beta_j|\boldsymbol{D}) = \sum_{\mathcal{T}} p(\beta_j|M_k, \boldsymbol{D}) p(M_k|\boldsymbol{D}) \tag{3.27}$$

where $\mathcal{T} = \{M_k : \beta_j \neq 0\}$. A possible objection to this solution is that β_j has different meanings in different models depending on what risk factors are included, and so it does not make sense to combine inferences from different models. However, (3.27) can be viewed in two ways: the first as a mixture across different models. This is the one that is hard to interpret. It can also, however, be viewed as the posterior distribution from the single model with all risk factors, but with a prior distribution that assigns probability 1/2 to each coefficient being equal to zero. When viewed this way, we see no problem with the interpretation of $p(\beta_j|\boldsymbol{D})$ as the posterior distribution distribution of β_j controlling for all the risk factors, but allowing for the possibility that they have no effect. An appropriate terminology would be “the effect of the treatment after adjustment for the *possibility* of (all) the risk factors”.

Furthermore, we can compute the posterior mean of the coefficient vector by

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{BMA} = E_M(\hat{\boldsymbol{\beta}}) &= \sum_{k=1}^{K} \hat{\boldsymbol{\beta}}_k p(M_k|\boldsymbol{D}) && (3.28)\\
&= \frac{\sum_{k=1}^{K} \hat{\boldsymbol{\beta}}_k p(M_k|\boldsymbol{D})}{\sum_{k:\boldsymbol{\beta}_k \in M_k}^{K} p(M_k|\boldsymbol{D})} \times \sum_{k:\boldsymbol{\beta}_k \in M_k}^{K} p(M_k|\boldsymbol{D}) && (3.29)\\
&= E(\hat{\boldsymbol{\beta}}|\boldsymbol{\beta}_k \in M_k) \times p(\boldsymbol{\beta} \neq 0) && (3.30)
\end{aligned}$$

corresponding to the conditional posterior mean of $\boldsymbol{\beta}$ multiplied by its posterior probability. We will use $\exp(\hat{\boldsymbol{\beta}}_{BMA})$ as the posterior estimate of the vector of hazard ratios using the ensemble of models. Each element of this vector expresses how each variable changes the hazard.

Let $p_k = p(M_k|\boldsymbol{D})$ and $V_k = V(\hat{\boldsymbol{\beta}}|M_k, \boldsymbol{D})$, then we can compute the variance of the regression coefficient vector as

$$\begin{aligned}
V(\hat{\boldsymbol{\beta}}) &= E(\hat{\boldsymbol{\beta}}^2) - \left(\sum_{k=1}^{K} p_k \hat{\boldsymbol{\beta}}_k\right)^2\\
&= \sum_{k=1}^{K} p_k (V_k + \hat{\boldsymbol{\beta}}_k^2) - \left(\sum_{k=1}^{K} p_k \hat{\boldsymbol{\beta}}_k\right)^2\\
&= \sum_{k=1}^{K} p_k V_k + \sum_{k=1}^{K} p_k \hat{\boldsymbol{\beta}}_k^2 - \left(\sum_{k=1}^{K} p_k \hat{\boldsymbol{\beta}}_k\right)^2\\
&= \sum_{k=1}^{K} p_k V_k + \sum_{k=1}^{K} p_k \left(\hat{\boldsymbol{\beta}}_k - \sum_{k=1}^{K} p_k \hat{\boldsymbol{\beta}}_k\right)^2 && (3.31)
\end{aligned}$$

The first term is the weighted variance over models. The second term expresses the variance of the parameter estimates across models. The more the parameter estimates differ over models, the higher the posterior variance. This implies that the regression coefficient variance includes the model uncertainty.

3.3.2 Predictive Performance

Reading medical journals and conference papers, we noticed that physicians are mostly interested in which risk factors are significant (they all use p-values) predictors of the survival time. However, we (and others) also find it important to investigate and compare the predictive performance of competing models to assess which model(s) and modeling strategy to prefer. If we split our data set into a training set, $\boldsymbol{D}_{train}$, and a test set, $\boldsymbol{D}_{test}$, we can use the training set to select models, estimate the parameters, and compute the PMP, PPP etc. If we use all available data for training we can compare models using Bayes Factors, but if we leave some data for the test set, we can compare models in terms of generalization error or test error, i.e. how well the models fit data they have not seen before.

3.3.2.1 Partial Predictive Log Score

The first score we use to measure the predictive performance (test error) is a predictive log score. The log score for model M_k is based on the observed ordinate of the predictive density for the subjects in the test set (using log to transform the product of predictive densities in the test set into a sum)

$$\sum_{i=1}^{N_{test}} \log p(\boldsymbol{d}_i|M_k, \boldsymbol{D}_{train}) \tag{3.32}$$

In stepwise selection we have just one model, but In BMA we need to average over all models considered

$$\sum_{i=1}^{N_{test}} \log \left\{ \sum_{k=1}^{K} p(\boldsymbol{d}_i|M_k, \boldsymbol{D}_{train}) p(M_k|\boldsymbol{D}_{train}) \right\} \tag{3.33}$$

Unfortunately, as we estimate the *cumulative* hazard in the CPH model rather than the hazard itself, we do not have a predictive density, but an estimated predictive cumulative distribution function, a step function like the cumulative hazard, (Breslow, 1974), that we cannot differentiate into a density. Instead, inspired by the PL, (Volinsky, 1997) have proposed the alternative predictive density

$$p(\boldsymbol{d}_i|M_k, \boldsymbol{D}_{train}) = \left(\frac{\exp(\hat{\boldsymbol{\beta}}_k^{\mathrm{T}} \boldsymbol{z}_i)}{\sum_{l \in R_i} \exp(\hat{\boldsymbol{\beta}}_k^{\mathrm{T}} \boldsymbol{z}_l)} \right)^{\delta_i} \tag{3.34}$$

If we insert (3.34) into (3.32) and (3.33), we get the *Partial Predictive log Score (PPS)*, which is greater (less negative) for the method that best fits the data in the test set, i.e. it is the inverse test error. We can use this score to compare BMA to any single model. Note that we only get a non-zero contribution for an event (failure).

3.3.2.2 Predictive $\mathcal{Z}$-score

In a subsequent discussion of (Raftery et al., 1995) (included in (Raftery et al., 1995)), Draper claims that the PPS has the drawback that it shifts from survival times to a "less relevant domain". We still believe that the PPS is a valid and very useful scoring method, since we consider each subject in the test set and calculate the (log) probability of observing the given survival time for a subject with a given set of risk factor values. Even though we transform the survival time into a probability, PPS still expresses how well each method predicts the data points in the test set.

However, Draper proposes that it would be interesting to pretend at random that some of the uncensored survival times were censored (this will be our test set), compute predictive distributions for these subjects, and calculate predictive $\mathcal{Z}$-scores[3]

$$\mathcal{Z}_i = \frac{\log(t_i) - \bar{t}_i}{\sigma_i} \tag{3.35}$$

where t_i is the true survival time, $\bar{t}_i$ the predictive mean or median of the log survival time, and σ_i the predictive standard deviation of the log survival time for the i'th subject in the test set. A "better" method should give predictive distributions with lower standard deviations (large predictive standard deviations implies uncertain predictions) that we are and say, 95% *Confidence Intervals (CI)* (see any textbook in statistics, e.g. (Johnson, 2005)) that more often contain the true survival times.

The estimated mean survival time is estimated as the area under the estimated survival curve, $\hat{S}(t)$. We get an estimate of the survival function using (2.67). The estimator is based upon the entire range of data. (DW and Lemeshow, 1999) point out that it will bias the estimate of the mean downwards, if we use only the data up to the last observed event, and they recommend that the entire range of data is used.

However, instead of the mean (log) survival time we use the median (log) survival time. Samples of survival times are often highly skewed and the median is generally a better measure of central location than the mean, (DW and Lemeshow,

[3] Z is normally used to denote a test statistic variable, but to avoid confusion with the variable vector, we use $\mathcal{Z}$.

1999). The median survival time is calculated as the smallest survival time for which the survival function is less than or equal to 0.5.

Using Greenwood's formula, (Collet, 2003), we can also calculate the estimated variance of the survival function

$$\mathrm{V}\big[\hat{S}(t)\big] \approx [\hat{S}(t)]^2 \sum_{i=1}^{p} \frac{d_i}{n_i(n_i - d_i)} \tag{3.36}$$

for $t_k \leq t < t_{k+1}$ where n_i is the number of individuals at risk at time (interval) i, and d_i is the number of individuals who fail (die) at time (interval) i. Because we have discrete times, we sum over time (intervals) instead of integrating. We can use this formula to find the predictive variance and in turn the standard error by evaluating (3.36) at the estimated mean or median survival time. Once the predictive standard error has been calculated, we can calculate corresponding 95% CIs for the estimated mean or median survival time by assuming that the estimated value of the survival function at t is normally distributed with mean $S(t)$ and estimated variance given by (3.36).

Chapter 4

Missing Values

So far, we have assumed that the available data are fully observed, that is we do not have any missing risk factors or failure times. This is the *Complete Case (CC)* scenario. However, it is more a rule than an exception to have missing values: Unavailable risk factors because the patient died or was too weak to collect information such as the results of CT scans, motor skill tests, survey non-response, patients failing to report for evaluations, patients unable to or refuse to answer questionnaires, lost data, lack of time or finances to perform tests and evaluations etc.

Most methods for analyzing survival data including the stepwise selection method cannot handle missing values. The standard solution is to discard all subjects with missing values and perform a CC analysis. The shortcomings of various case-deletion strategies have been well documented, (Little and Rubin, 1987), (Ramoni and Sebastiani, 2001), (Herring and Ibrahim, 2001), (Herring et al., 2004). If the discarded cases form a representative and relatively small portion of the entire data set, case deletion may be a reasonable approach, but only when missing data are missing completely at random (see Section 4.4.1) in the sense that the probability of response does not depend on any data values, observed or missing. For example, if the CT scan information for stroke patients is only available for those patients that do not die within the first week of admittance, the missingness of the CT variable is related to the failure time, and our CC analysis would then be based on patients with less severe strokes, leading to biased results. When we discard data, the efficiency of the analysis decreases as the fraction of missing data increases, whether or not bias is involved.

In other words, case deletion implicitly assumes that the discarded cases are like a random sub-sample. When the discarded cases differ systematically from the rest, estimates may be seriously biased. Moreover, in multivariate problems, case deletion often results in a large portion of the data being discarded and an unacceptable loss of power. After conducting a CC, it is common to mention the fraction of missing data and to add some assessment about whether these missing data are likely to bias the results.

Hence, there is substantial reason to look for ways to incorporate subjects with missing values. In the COST data set for example, 441 (44.1%) of the 993 patients have one or more missing values. If we could include these subjects, we would almost double the amount of available data! In this chapter we will present several techniques for estimating the missing values. We will assume that the response (failure time or censoring) is always observed and focus on missing risk factor data.

4.1 Imputation

Imputation, the practice of "filling in" missing data with plausible values, is an attractive approach to analyze incomplete data, (Little and Rubin, 1987). If the proportion of missing values is small, then *single imputation* in which we replace a missing value with a single estimate may be reasonable. Standard statistical procedures for complete data analysis can then be applied on the filled-in data set. For example, each missing value can be imputed using the variable mean of the complete cases, or the mean conditional on observed values of other variables. This approach treats missing values as if they were known in the complete-data analysis. Single imputation does not reflect the uncertainty about the predictions of the unknown missing values, and the resulting estimated variances of the parameter estimates will be biased towards zero. Without special corrective measures, single-imputation inference tends to overstate the precision because it omits the between-imputation component of variability, (Little and Rubin, 1987).

Multiple Imputation (MI) is a Monte Carlo technique in which the missing values are replaced by $m > 1$ simulated versions, where m is typically small (e.g. 2-10). In Rubin's, (Little and Rubin, 1987), method for "repeated imputation" inference, each of the simulated complete data sets are analyzed using standard methods, and the results are combined to produce estimates and confidence intervals that incorporate missing data uncertainty. With the advent of new computational methods and software for creating MI, the technique has become increasingly attractive for researchers in the biomedical, behavioral, and social sciences whose investigations are hindered by missing data.

In order to generate imputations for the missing values, we must impose a probability model on the complete data (observed and missing values). Except in trivial settings, the probability distributions that we must draw from to produce proper MI tend to be complicated and intractable. For this purpose, MCMC methods have spawned a small revolution. See e.g. (Leong et al., 2001) for an MCMC EM analysis of uncomplete variables in the Cox model with applications to biological marker data. In this work, however, we will focus on two other methods to estimate the missing values, but we incorporate MCMC methods as part of the proposed solution in Section 4.4.

4.2 The Truly Bayesian Approach

The paradigmatic setting for missing data imputation is regression, where we are interested in the model $p(\boldsymbol{t}|\boldsymbol{z}, \boldsymbol{\theta})$, but have missing values in the risk factor matrix $\boldsymbol{z}$. The parameter matrix $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{h_0})$ includes the risk factor coefficients as well as the baseline hazards for all possible models.

To be Bayesian at heart we would then average over all models, their parameters, and the missing value distributions. One approach would be to apply an MCMC, (Liu, 2001), (Gilks et al., 1996), (MacKay, 2003), (Gelman et al., 2004), (Neal, 1993), method to sample from both the missing variables and the parameters of the CPH model, and then sum over the models. We could also sample the models, but assuming we have a finite number of models, we are better off just summing over all possible models.

The samples are used to approximate the joint posterior density, $p(\boldsymbol{\theta}, \boldsymbol{z}_{\text{miss}}|\boldsymbol{D})$, of the model parameters, $\boldsymbol{\theta}$, and the missing risk factors, $\boldsymbol{z}_{missing}$, given the data, $\boldsymbol{D} = (\boldsymbol{z}_{\text{obs}}, \boldsymbol{t})$, i.e. the observed risk factor matrix and the survival times.

We would then compute the expectation of this distribution to get the expected risk factor coefficients, and in turn the HRs. In general, the expectation value of some function, $f(\boldsymbol{X})$, e.g. $f(\boldsymbol{X}) = \boldsymbol{X}$, is

$$\mathrm{E}[f(\boldsymbol{X})] = \int f(\boldsymbol{X}) p(\boldsymbol{X}|\boldsymbol{D}) d\boldsymbol{X} = \int g(\boldsymbol{X}) d\boldsymbol{X} \tag{4.1}$$

In straight Monte Carlo integration we pick n points uniformly distributed in a multi-dimensional volume V that covers all regions where $g(\boldsymbol{X})$ contributes (significantly) to the integral. Then we get

$$\mathrm{E}[f(\boldsymbol{X})] = \int_V g(\boldsymbol{X}) d\boldsymbol{X} \approx V \times \mathrm{E}[g(\boldsymbol{X})] \pm V \times \sqrt{\frac{\mathrm{E}[g^2(\boldsymbol{X})] - \mathrm{E}[g(\boldsymbol{X})]^2}{n}} \tag{4.2}$$

where

$$\mathrm{E}[g(\boldsymbol{X})] = \frac{1}{n} \sum_{i=1}^{n} g(X_i) \qquad \mathrm{E}[g^2(\boldsymbol{X})] = \frac{1}{n} \sum_{i=1}^{n} g^2(X_i) \tag{4.3}$$

and n is the number of random samples of $g(\boldsymbol{X})$. Although the law of large numbers guarantees that the approximation can be made arbitrarily accurate when the samples are independent, we waste too much time sampling in regions of low probability in terms of $p(\boldsymbol{X}|\boldsymbol{D})$. However, the samples do not have to be independent as long as we generate samples from the target distribution, $p(\boldsymbol{X}|\boldsymbol{D})$, in the correct proportions.

MCMC methods are algorithms for sampling from any probability distribution by constructing a Markov chain that has the desired distribution as its stationary distribution. First, we take a large number of steps/samples, forget them and then use the present state of the chain as a sample from the desired distribution. It is difficult to determine how many steps are needed to converge to the stationary distribution within an acceptable error margin, so we would like to have a chain with rapid mixing, i.e. the stationary distribution is reached quickly independent of the starting position.

Random walk methods[1], (Liu, 2001), (Gilks et al., 1996), where an ensemble of "walkers" move around randomly is an example of a Monte Carlo method, but whereas the random samples used in a conventional Monte Carlo algorithm are statistically independent, samples in MCMC are correlated. MCMC methods move around the equilibrium distribution in steps that are relatively small and not necessarily in the same direction. These methods are fairly easy to implement and analyze, but may not converge as it takes a long time to explore the entire sample space, and they tend to move around in space already explored.

4.2.1 Metropolis-Hastings

In general, we generate a random walk using a Markov chain such that the probability of being in a region is proportional to the posterior density for that region. The new sample, $\boldsymbol{x}_t$, depends only on the previous state, $\boldsymbol{x}_{t-1}$, through the transition probability, $p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)$, assumed to be time-independent.

Optimally, we would like to sample from our *target density*, $p(\boldsymbol{X})$, the density that we would like our Markov chain to have as its stationary distribution. Unfortunately, this distribution is often hard to sample from. Instead, we use a *proposal density*, $Q(\boldsymbol{X}'; \boldsymbol{x}_t)$, that depends on the current state $\boldsymbol{x}_t$ to propose a new state/sample $\boldsymbol{x}'$, e.g. a Gaussian proposal density centered on the current state $\boldsymbol{x}_t$

$$Q(\boldsymbol{X}'; \boldsymbol{x}_t) \sim N(\boldsymbol{x}_t, \sigma^2)\boldsymbol{I} \tag{4.4}$$

This proposal density would generate samples centered around the current state, $\boldsymbol{x}_t$, with variance $\sigma^2\boldsymbol{I}$. Having drawn a new proposal state, $\boldsymbol{x}'$, with probability $Q(\boldsymbol{X}'; \boldsymbol{x}_t)$, we compute the likelihood ratio between the proposed state, $\boldsymbol{x}'$, and the previous state, $\boldsymbol{x}_t$, multiplied with the ratio of the proposal density in two directions (from $\boldsymbol{x}_t$ to $\boldsymbol{x}'$ and vice versa)

$$a = \frac{p(\boldsymbol{x}')}{p(\boldsymbol{x}_t)} \frac{Q(\boldsymbol{x}_t; \boldsymbol{x}')}{Q(\boldsymbol{x}'; \boldsymbol{x}_t)} \tag{4.5}$$

[1] Ever heard of *uncertain walk methods*?

We use a to decide whether or not to keep the proposed state based on the rule

$$\boldsymbol{x}_{t+1} = \begin{cases} \boldsymbol{x}' & \text{if } a > 1 \\ \boldsymbol{x}' \text{ with probability } a & \text{if } a < 1 \end{cases} \tag{4.6}$$

The original algorithm proposed by (N. Metropolis and Teller, 1953) considered only symmetric proposal densities where the density ratio is equal to 1, and the new state is accepted with probability 1 if the likelihood increases, and with probability $\frac{p(\boldsymbol{x}')}{p(\boldsymbol{x}_t)}$ otherwise.

The generalization of this algorithm as outlined above is referred to as the *Metropolis-Hastings (MH)* algorithm, (Liu, 2001), (Gilks et al., 1996), (Neal, 1993). We can use this algorithm to draw samples from any probability distribution, $p(\boldsymbol{X})$, as long as we can to evaluate the density at $\boldsymbol{x}$. The Gibbs sampling algorithm is a special case of the MH algorithm. A simple explanation of Gibbs sampling can be found in (MacKay, 2003). To initialize the algorithm we choose the first state at random[2]. As we do not expect the first state and the states visited shortly here-after to be representative of the target distribution, we let the algorithm run for, say, a few hundred iterations so that this initial state is “forgotten”. This phase is known as the “burn-in” phase and the samples are simply discarded.

Obviously, we get the best samples if the proposal density matches the shape of the target distribution, $Q(\boldsymbol{X}'; \boldsymbol{x}_t) \approx p(\boldsymbol{X}')$, but as $p(\boldsymbol{X}')$ can be hard to sample from, $Q(\boldsymbol{X}'; \boldsymbol{x}_t)$ is chosen as a distribution that is easier to sample from. If $p(\boldsymbol{X}')$ is unknown, we cannot evaluate $p(\boldsymbol{x}')$ exactly, and we have to choose a $p(\boldsymbol{X}')$ that we believe is close to the true target density. When the algorithm runs, we can survey the acceptance rate, i.e. the fraction of samples that are accepted within the last N samples according to the above rule, and use it to optimize our proposal density during the burn-in period. For example, for the Gaussian proposal distribution, we can tune the variance parameter σ^2. If the proposal steps are too small, the acceptance rate will be high, but the chain will mix slowly, i.e. it will move around the space and converge slowly to $p(\boldsymbol{X})$. If the proposal steps are too large, the acceptance rate will be very low, because the proposals are likely to occur in regions of much lower probability density causing $\frac{p(\boldsymbol{x}')}{p(\boldsymbol{x}_t)}$ to be very small, Liu (2001).

In this work we do not implement a method that averages over all possible models, all possible parameter settings and the missing values. We acknowledge that this is, in theory at least, an ideal solution, but we leave it to others. However, we use the MH algorithm to sample and estimate the missing continuous values in the solution presented in Section 4.4.

[2] Bayesian would probably say, “without any prior preference” instead of random.

4.3 Graphical Models

Our first approach to the missing data problem is to use *graphical models*, (Murphy, 2001a), (Jensen, 1996), (Jensen, 1996), (Heckerman, 1995). to represent inter-dependence between risk factors, and use them to infer or estimate the values of the missing variables.

To represent our knowledge or rather our beliefs about a domain using probabilistic relations is known as a *probabilistic model*, (MacKay, 2003). Statistical modeling problems often involve a large number of interacting uncertain variables, and it is often convenient to express the dependencies between these variables graphically. The key idea in these *graphical models* is modularity, i.e. that we - graphically - represent a complex system of uncertain variables, and relations between these variables, by a set of simpler subsystems and use probability theory to ensure consistency. In particular, graphical models are used as a visual tool aiding humans in capturing the *conditional independency*, (Murphy, 2001a), relationships between variables. The variable a is said to be conditionally independent of b, given c (written $a \perp\!\!\!\perp b|c$) if and only if $p(a, b|c) = p(a|c)p(b|c)$. If $a \perp\!\!\!\perp b|c$ then b gives us no new information about a once we know c.

In a (probabilistic) graphical model nodes represent variables, and arcs represent conditional dependencies, and missing arcs represent conditional independence assumptions. More formally, we can use the graphical model to represent a set of variables $\boldsymbol{Z} = \{\boldsymbol{X}, \boldsymbol{Y}\}$ by the joint distribution $p(\boldsymbol{X}, \boldsymbol{Y})$. Here, $\boldsymbol{X}$ represents the set of hidden (never observed) variables, and $\boldsymbol{Y}$ the set of observable variables. Our data set $\boldsymbol{D}$ would then be instances of $\boldsymbol{Y}$, possibly with missing values.

To represent the joint distribution of N variables, we would normally require $\mathcal{O}(2^N)$ parameters. In a graphical model we can take advantage of the independence assumptions using (possibly) exponentially fewer parameters, (Murphy, 2001a). This is a major advantage for learning and inference problems as described in Section 3.2 and 4.3.1.

We distinguish between two kinds of graphical models: Undirected graphical models (a.k.a. *Markov networks* or *Markov random fields*), and directed graphical models (a.k.a. *Bayesian networks*, *belief networks*, *generative models*, *causal models*, etc.), where the arcs have directionality only in the latter case, (Murphy, 2001a), (Bishop, 2006). A model with both directed and undirected arcs is known as a *chain graph.* In this work we will focus on directed graphical models, and we will refer to them as *Bayesian Networks (BN).* It is important to understand that Bayesian networks do not imply the use of Bayesian methods. The Bayesian reference is to Bayes' rule, used for inference as in (3.3).

A BN for a set of variables $\boldsymbol{Z} = \{Z_1, Z_2, \ldots, Z_N\}$ is represented by a *Directed Acyclic Graph (DAG)*, (Murphy, 2001a), or structure $\mathcal{S}$, where each variable corresponds to a node in the graph, and directed arcs between nodes implies variable dependencies. A DAG is a graphical model in which there exist no directed path including the same variable more than once. In combination with a set of local probability distributions $\boldsymbol{\mathcal{P}}$ for each variable, the joint distribution for $\boldsymbol{Z}$ is defined.

This often leads to the interpretation that the directed arcs in the graph imply causality. Hence, in the case $A \rightarrow B$ we would say that A *causes* B, which is why BNs are also referred to as *causal models*.

The joint distribution for a graphical structure, $\mathcal{S}$, is given by

$$p(\boldsymbol{Z}) = \prod_{i=1}^{N} p(\boldsymbol{Z}_i | \mathrm{pa}(\boldsymbol{Z}_i)) \tag{4.7}$$

where $\mathrm{pa}(\boldsymbol{Z}_i)$ denotes the *parent* nodes of the i'th node in the network, i.e. the nodes that have directed arcs going into the i'th node. In the same terminology, $\boldsymbol{Z}_i$ is referred to as the *child* of $\mathrm{pa}(\boldsymbol{Z}_i)$. Hence, the local probability distributions, $\boldsymbol{\mathcal{P}}$, are the conditional probability distributions of (4.7). Note that there is no need for a normalization constant in (4.7), because, by the definition of the conditional probabilities, it is equal to one. The *descendants* of a node includes its children and its children's descendants, and the *ancestors* of a node are its parents and the parents' ancestors.

As a simple example, consider the network depicted in Figure 4.1. In this strictly boolean network, we have a node, *Sleep* (S), representing whether or not we easily fall asleep on a given night. The value of this node depends on its two parent nodes, $Coffee$ (C) and scary $Movie$ (M), whether or not we had coffee and/or watched a scary movie just before turning in. Furthermore, the value of the M node is conditioned on whether or not we are home *Alone* (A).

Using the chain rule of probability, the joint distribution is given by

$$p(A, M, C, S) = p(A) \times p(M|A) \times p(C|A, M) \times p(S|A, M, C) \tag{4.8}$$

Taking advantage of the conditional independence assumptions we get

$$p(A, M, C, S) = p(A) \times p(M|A) \times p(C) \times p(S|M, C) \tag{4.9}$$

using the fact that C has no parents and that $S \perp\!\!\!\perp A | M$. The conditional independence assumptions give a more compact representation of the joint distribution, and fewer parameters makes learning easier. In general, N binary nodes requires $\mathcal{O}(2^N)$ parameters to represent the joint distribution, while the

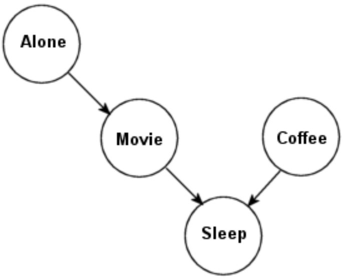

Figure 4.1: Simple Bayesian network example where drinking $Coffee$ (C) and/or watching a scary $Movie$ (M) affect the probability of falling a $Sleep$ (S). Value of M depends on whether or not we are home $Alone$ (A).

factored form requires $\mathcal{O}(N2^q)$ parameters, where q is the maximum fan-in of a node (one plus the number of parents).

Using Bayesian networks, we can represent many statistical models, e.g. Principal Component Analysis, Independent Component Analysis, Hidden Markov Models etc. For a nice overview, please see (Murphy, 2001a).

4.3.1 Inference in Graphical Models

In a nutshell, inference in a Bayesian network is to *infer* information about an unobserved (set) of variables given the observed variables (the data). When we observe the leaves of a network and try to infer the values of the hidden causes, we call it *diagnosis*, or bottom-up reasoning. When we observe the roots, and try to predict the effects, we call it *prediction*, or top-down reasoning.

To infer information about the unobserved variable, X, based on some data set, $\boldsymbol{D}$, we can use (3.3) and integrate over the uncertain parameters

$$p(X|\boldsymbol{D}) = \frac{\int p(\boldsymbol{D}|X)p(X|\boldsymbol{\theta})d\boldsymbol{\theta}}{\int p(\boldsymbol{D}|\boldsymbol{\theta})d\boldsymbol{\theta}} \tag{4.10}$$

We could be interested in the distribution itself or moments hereof. In most cases we are not interested in the distribution over all unobserved or hidden variables in the model, but a subset hereof (perhaps just a single variable). The hidden nodes can represent physical quantities that we for some reason cannot observe, or hidden nodes introduced to obtain a certain network structure that do not necessarily have a physical interpretation. These variables also need to be integrated or marginalized out, and often leaves us with very complex integrals that are not analytically tractable. To overcome this we can use approximate schemes.

4.3.1.1 Exact Inference

Consider the network in Figure 4.1 where we observe the following probabilities (using $0 \equiv \mathit{false}$ and $1 \equiv \mathit{true}$ for clarity).

Prior on *Alone*

p(A=0)	p(A=1)
0.9	0.1

Prior on *Coffee*

p(C=0)	p(C=1)
0.4	0.6

Movie conditioned on *Coffee*

A	p(M=0|A)	p(M=1|A)
0	0.9	0.1
1	0.7	0.3

Sleep conditioned on *Movie* and *Alone*

M	C	p(S=0|M,C)	p(S=1|M,C)
0	0	0.01	0.99
0	1	0.3	0.7
1	0	0.4	0.6
1	1	0.9	0.1

If we cannot easily fall asleep, $S = 0$, we can explain it by a late cup of coffee or the memories of a late night scary movie. To infer which explanation is more likely, we use Bayes' rule to compute the posterior probability of the two variables

$$\begin{aligned} p(M = 1|S = 0) &= \frac{M = 1, S = 0}{p(S = 0)} \\ &= \frac{\sum_{A,C} p(A = a, M = 1, C = c, S = 0)}{p(S = 0)} \\ &= \frac{0.0264}{0.3593} = 0.0735 \end{aligned}$$

and

$$\begin{aligned} p(C = 1|S = 0) &= \frac{C = 1, S = 0}{p(S = 0)} \\ &= \frac{\sum_{A,M} p(A = a, M = m, C = 1, S = 0)}{p(S = 0)} \\ &= \frac{0.2232}{0.3593} = 0.6212 \end{aligned}$$

with

$$\begin{aligned} p(S = 0) &= \sum_{A,M,C} p(A = a, M = m, C = c, S = 0) \\ &= 0.3593 \end{aligned}$$

as the normalizing constant. We have used the fact that both the numerator and the denominator of (3.3) correspond to marginalized versions of the joint distribution (4.7) with evidence $\boldsymbol{d}_1 = \{S = 0, M = 1\}$ and $\boldsymbol{d}_2 = \{S = 0, C = 1\}$.

Comparing the two explanations shows that it is more than 8 times more likely to have been our weakness for hot coffee that caused an unpleasant night's sleep

$$\frac{p(C=1|S=0)}{p(M=1|S=0)} = \frac{0.2232}{0.0264} = 8.4545 \tag{4.11}$$

As indicated by this simple example with just a few binary nodes, it quickly gets very complicated to compute posterior estimates using Bayes' rule, e.g. the normalizing constant, in general, involves a sum over an exponential number of terms. For continuous variables, the sum is replaced by integrals that are analytically intractable (except for special cases like Gaussian's).

4.3.1.2 Variable Elimination

However, we can take advantage of the conditional independence assumptions encoded in the graph and can compute the normalizing constant using the factored representation of the joint distribution

$$\begin{aligned} p(S=s) &= \sum_a \sum_m \sum_c p(A=a, M=m, C=c, S=s) && (4.12)\\ &= \sum_a \sum_m \sum_c p(A=a) \times p(M=m|A=a) && (4.13)\\ &\times p(C=c) \times p(S=s|M=m, C=c) \end{aligned}$$

The trick in variable elimination is to make as few summations as possible using the distributivity law of × over +. In the example, only the *Sleep* node and the *Movie* have parents, and there is no need to sum $p(A=a)$ or $p(C=c)$ for all instances of *Movie* yielding

$$\begin{aligned} p(S=s) = \sum_a p(A=a) \sum_c p(C=c) \sum_m p(M=m|A=a) && (4.14)\\ \times p(S=s|M=m, C=c) \end{aligned}$$

where we first sum the variables with no parents, and push the conditional $p(M=m|A=a)$ as far possible. If we substitute the inner-most sum with the term

$$L_1(A, C, S) = \sum_m p(M=m|A=a) \times p(S=s|M=m, C=c) \tag{4.15}$$

that does not depend on the summed variable M, we get

$$p(S=s) = \sum_a p(A=a) \sum_c p(C=c) \times L_1 \tag{4.16}$$

Repeating this we get

$$L_2(A, S) = \sum_c p(C = c) \times L_1(A, C, S) \tag{4.17}$$

and obtain

$$p(S = s) = \sum_a p(A = a) \times L_2(A, S) \tag{4.18}$$

This principle is the foundation of many algorithms, such as the Baum-Welch algorithm, the Fast Fourier Transform, Viterbi's algorithm and Pearl's Belief Propagation algorithm, see e.g. (Aji and McEliece, 2000), (Lauritzen and Spiegelhalter, 1988), and (Spiegelhalter and Lauritzen, 1990).

The algorithm's complexity is bounded by the size of the largest term. Choosing a summation (elimination) ordering to minimize this is NP-hard, (Arnborg et al., 1987), although greedy algorithms work well in practice (Kjaerulff, 1990), (Huang and Darwiche, 1994).

Usually we are not interested in computing just one marginal, but several marginals at a time by repeating the variable elimination algorithm for each marginal, leading to a large number of redundant computations. If the corresponding *undirected* graph is also acyclic, i.e. a tree, we can apply a local message passing algorithm due to Pearl, (Pearl, 1988). The algorithm is a generalization of the forwards-backward algorithm for Hidden Markov models, (Rabiner and Juang, 1986). If the BN has (undirected) loops, a local message passing algorithms may double count evidence and not converge. For example, if we had connected the *Alone* and *Coffee* nodes as shown in Figure 4.2, the information from M and C passed on to S would no longer be independent, because it came from a common parent, A. To solve this, we convert the graphical model into

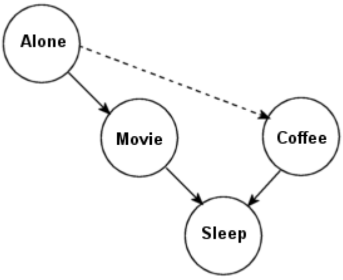

Figure 4.2: Simple Bayesian network example (loopy version).

a tree by clustering nodes together, and then run a local message passing algorithm. The network in Figure 4.1 is already a tree, but the network in Figure 4.2 is not. In this case we would cluster the nodes as illustrated in Figure 4.3. The most common algorithm is the *Junction Tree* algorithm due to Pearl (1988), where the original graph is converted into a Junction Tree where probabilistic information can be locally distributed and collected.

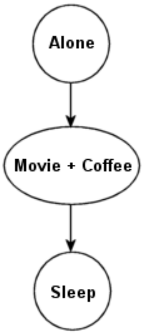

Figure 4.3: Simple Bayesian network example (clustered version).

The complexity of the Junction Tree algorithm is exponential in the size of the largest clique in the moralized, triangulated graph, assuming all hidden nodes are discrete. Although the largest clique size may be much smaller than the total number of nodes for a sparsely connected graph, exact inference using the Junction Tree algorithm is still intractable in most cases. There exist more efficient algorithms for graphical models with special structures, see e.g. (Guo and Hsu, 2002), but there is still a demand for approximate alternatives. Especially, when we have continuous valued nodes where the corresponding integrals in Bayes' rule cannot be evaluated in closed form. For more on exact inference, see e.g. (Agresti and Hitchcock, 2005), (Huang and Darwiche, 1994), (Heckerman, 1989), (Beinlich et al., 1989), (Morris, 2002), and (El-Hay, 2001).

4.3.1.3 Approximate Inference

When the integrals (or summations) in (4.10) are analytically intractable, we need to use approximate techniques. As we saw in Chapter 3.2, we also face complicated integrals when we compute the posterior distribution over models or parameters. To compute the posterior distribution, we need the marginal likelihood which averages over parameters. These integrals quickly become analytically intractable. Hence, there is an over-all need for numerical integration techniques, and we therefore treat the problem in general, and then apply it to the inference problem and for the purpose of computing the marginal likelihood.

Roughly speaking, we have two alternatives: *deterministic* or *non-deterministic* (*Monte Carlo*) methods. The latter approach was discussed in Section 4.2. Instead, we will focus on deterministic or analytical approximations. We will review the Laplace method (the Gaussian approximation), (de Bruijn, 1981), and Bayes Information Criterion, (Schwarz, 1978). Both methods are analytical and in different ways try to account for the probability mass around the MAP parameter configuration. The MAP value is usually easy to find and makes these approximations attractive.

The Gaussian approximation is based on the idea that for large samples, the true integrand can often be approximated by a multi-variate Gaussian distribution. Let

$$h(\boldsymbol{X}) \equiv \log\left[f(\boldsymbol{X})\right] \tag{4.19}$$

If we make a second order Taylor expansion of $h(\boldsymbol{X})$ around its mode $\hat{\boldsymbol{X}}$, we get

$$h(\boldsymbol{X}) \approx h(\hat{\boldsymbol{X}}) + \boldsymbol{g}^{\mathrm{T}}(\boldsymbol{X} - \hat{\boldsymbol{X}}) + \frac{1}{2}(\boldsymbol{X} - \hat{\boldsymbol{X}})\boldsymbol{H}(\boldsymbol{X} - \hat{\boldsymbol{X}})^{\mathrm{T}} \tag{4.20}$$

$$\boldsymbol{g} = \left.\frac{\partial h(\boldsymbol{X})}{\partial \boldsymbol{X}})\right|_{\boldsymbol{X}=\hat{\boldsymbol{X}}} \tag{4.21}$$

$$\boldsymbol{H} = \left.\frac{\partial^2 h(\boldsymbol{X})}{\partial \boldsymbol{X} \partial \boldsymbol{X}^{\mathrm{T}}}\right|_{\boldsymbol{X}=\hat{\boldsymbol{X}}} \tag{4.22}$$

At the mode, $\boldsymbol{X} = \hat{\boldsymbol{X}}$, the first order derivative, $\boldsymbol{g}$, must be $\mathbf{0}$ yielding

$$h(\boldsymbol{X}) \approx h(\hat{\boldsymbol{X}}) + \frac{1}{2}(\boldsymbol{X} - \hat{\boldsymbol{X}})\boldsymbol{H}(\boldsymbol{X} - \hat{\boldsymbol{X}})^{\mathrm{T}} \tag{4.23}$$

If we insert (4.19) into (4.23) and raise to the power of e, we get

$$f(\boldsymbol{X}) \approx f(\hat{\boldsymbol{X}}) \exp\left(\frac{1}{2}(\boldsymbol{X} - \hat{\boldsymbol{X}})\boldsymbol{H}(\boldsymbol{X} - \hat{\boldsymbol{X}})^{\mathrm{T}}\right) \tag{4.24}$$

yielding the integral

$$I = \int_A f(\boldsymbol{X}) d\boldsymbol{X} \approx f(\hat{\boldsymbol{X}})(2\pi)^{d/2} |-\boldsymbol{H}|^{-1/2} \tag{4.25}$$

where d is the dimension (number of parameters) of $f(\boldsymbol{X})$. This technique is also known as **Laplace's method of approximation**, see for example (de Bruijn, 1981) and (Kass and Raftery, 1994). As an example, let us derive the approximation for the marginal likelihood, i.e. we set

$$h(\boldsymbol{\theta}) \equiv \log\left[p(\boldsymbol{\theta}|M)p(\boldsymbol{D}|\boldsymbol{\theta}, M)\right] \tag{4.26}$$

$$= \sum_{i=1}^{N} \log\left[p(\boldsymbol{\theta}|M)\right] + \log\left[p(\boldsymbol{d}_i|\boldsymbol{\theta}, M)\right] \tag{4.27}$$

as our data set contains N examples assumed to be i.i.d. Furthermore, let $\hat{\boldsymbol{\theta}}$ be the value of $\boldsymbol{\theta}$ that maximizes $h(\boldsymbol{\theta})$. This value also maximizes the posterior distribution, $p(\boldsymbol{\theta}|\boldsymbol{D})$, according to (4.10), and is the MAP estimate. Inserting (4.26) in (4.24) yields

$$p(\boldsymbol{\theta}|M)p(\boldsymbol{D}|\boldsymbol{\theta}, M) \approx p(\hat{\boldsymbol{\theta}}|M)p(\boldsymbol{D}|\hat{\boldsymbol{\theta}}, M) \exp\left(\frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\boldsymbol{H}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^{\mathrm{T}}\right) \tag{4.28}$$

- a multi-variate Gaussian with mean $\hat{\boldsymbol{\theta}}$ and covariance matrix $\hat{\Sigma} = (-\boldsymbol{H})^{-1}$. This gives us the approximate marginal likelihood according to (4.25)

$$p(\boldsymbol{D}|M)_{Laplace} = \int p(\boldsymbol{\theta}|M)p(\boldsymbol{D}|\boldsymbol{\theta}, M) d\boldsymbol{\theta} \tag{4.29}$$

$$= p(\hat{\boldsymbol{\theta}}|M)p(\boldsymbol{D}|\hat{\boldsymbol{\theta}}, M)(2\pi)^{d/2}|-\boldsymbol{H}|^{-1/2} \tag{4.30}$$

where d is the dimension of $\boldsymbol{\theta}$ or the number of rows/columns in $\boldsymbol{H}$. The dimensionality of the parameter space is the number of free parameters. For complete data this equals the number of parameters, but with hidden variables the dimensionality might be less, (Geiger et al., 1996).

The Laplace approximation consists of a likelihood term at the MAP setting, a penalty term from the prior, and a volume term calculated from the local curvature. The approximation is based on the assumption of a highly peaked integrand near its maximum $\hat{\boldsymbol{\theta}}$. This is usually true when the likelihood is highly peaked around $\hat{\boldsymbol{\theta}}$, and will often be the case for large sample sizes. In (Kass et al., 1990) the authors show that relative errors of this method are $\mathcal{O}(N^{-1})$ (under certain conditions), where N is the number of cases in the data set.

Using this method leaves us with two challenges: calculating the MAP estimate $\hat{\boldsymbol{\theta}}$ and the Hessian matrix $\boldsymbol{H}$. If we have a large data set, the effect of the prior

$p(\boldsymbol{\theta}|M)$ decreases as the sample size increases. In that case we can approximate the MAP estimate of $\hat{\boldsymbol{\theta}}$ by the MLE

$$\hat{\boldsymbol{\theta}}_{ML} = \arg\max_{\boldsymbol{\theta}} \left\{ p(\boldsymbol{D}|\boldsymbol{\theta}, M) \right\} \tag{4.31}$$

We can use gradient based methods to find local maxima of the posterior or the likelihood function. For example, (Buntine, 1996) discuss the case where the likelihood function belongs to the exponential family. Another solution would be to use the EM algorithm of Dempster et al. (1977). The volume term requires the calculation of $|-\boldsymbol{H}|$. It takes $\mathcal{O}(Nd^2)$ operations to compute the derivatives in the Hessian, see e.g. (Buntine, 1994) and Thiesson (1997), and then a further $\mathcal{O}(d^3)$ operations to calculate the determinant. This can be very demanding for high-dimensional models. An easy way of avoiding this is to approximate the true Hessian matrix with its diagonal elements, or assume a block-diagonal structure. However, this also implies independencies among the parameters, leading to an even worse approximation. Finally, the second derivatives themselves may be intractable to compute.

To overcome these computational difficulties, we can take the logarithm of (4.30)

$$\log\left[p(\boldsymbol{D}|M))\right] \approx \underbrace{\log\left[p(\hat{\boldsymbol{\theta}}|M)\right]}_{\mathcal{O}(1)} + \underbrace{\log\left[p(\boldsymbol{D}|\hat{\boldsymbol{\theta}}, M)\right]}_{\mathcal{O}(N)} + \underbrace{\frac{d}{2}\log\left(2\pi\right)}_{\mathcal{O}(1)} - \underbrace{\frac{1}{2}\log|-\boldsymbol{H}|}_{\mathcal{O}(d\log(N))} \tag{4.32}$$

and use only those terms that increase with N (the number of examples in $\boldsymbol{D}$) as indicated in (4.32)

$$\log\left[p(\boldsymbol{D}|M))\right] \approx \log\left[p(\boldsymbol{D}|\hat{\boldsymbol{\theta}}, M)\right] - \frac{1}{2}\log|-\boldsymbol{H}| \tag{4.33}$$

In (4.19) we note that the entries in $\boldsymbol{H}$, defined by (4.22), scales linearly with N, and so we can re-write the last term to get

$$\lim_{N\to\infty}\frac{1}{2}\log|-\boldsymbol{H}| = \frac{1}{2}\log|-N\boldsymbol{H}_0| = \frac{d}{2}\log(N) + \underbrace{\frac{1}{2}|-\boldsymbol{H}_0|}_{\mathcal{O}(1)} \tag{4.34}$$

For large N we can also use the MLE of $\hat{\boldsymbol{\theta}}$ to get

$$\log\left[p(\boldsymbol{D}|M)\right] \approx \log\left[p(\boldsymbol{D}|\hat{\boldsymbol{\theta}}_{ML}, M)\right] - \frac{d}{2}\log(N) \tag{4.35}$$

The approximation in (4.35) is known as the **Bayes Information Criterion approximation**, (Schwarz, 1978). The approximation makes intuitively sense. The first terms gives information on how well the model fits the data, while the second term increased and thus punishes the model complexity when d increases.

There exist numerous approximate inference techniques, such as Variational Bayes, (Jordan et al., 1998), (Wainwright and Jordan, 2005), (Geiger and Meek, 2005), (Attias, 2000), (Beal, 2003), (Jaakkola, 1997), Expectation Propagation, (Minka, 2001a), (Minka, 2001b), (Murphy, 2001b), Expectation Consistent approximate inference, (Opper and Winther, 2005), (Csato et al., 2003), (Opper and Winther, 2004), the Cluster Variation Method, (Kappen, 2002), and Variational Message Parsing, (Winn and Bishop, 2004), (Winn, 2003), and (Bishop et al., 2002), just to mention a few.

4.3.2 Bayes Net Toolbox

Kevin Murphy[3] has developed a very nice open source *Bayes Net Toolbox (BNT)* for Matlab, (Murphy, 2001a). It is very comprehensive and allows us to learn the structure and the parameters as well as do inference in a BN. We will not describe how to use the toolbox in details, a documentation is found at http://bnt.sourceforge.net/usage.html. In this work we will use and compare several different of the methods to learn the structure and the parameters of the BN. For more on parameter and structure learning in BNs, please refer to (Larsen, 2006), (, editor), (Heckerman, 1995), (Frey and Jojic, 2003), (MacKay, 2003), (Heckerman et al., 1995), (Chickering, 1996), (Cooper, 1995) and (Buntine, 1996).

4.3.2.1 Parameter Learning

We can divide the parameter estimation routines into 4 types, depending on whether we need a full (Bayesian) posterior over the parameters or a point estimate like ML or MAP, and whether or not we have missing data (partial observability). BNT supports point as well as Bayesian estimates for full observability and point estimates for partial observability using an EM algorithm.

4.3.2.2 Structure Learning

We can divide structure learning into constraint-based and search-and-score methods. In the constraint-based approach we have a fully connected graph and remove an edge between nodes if the data shows a conditional independency. However, repeated independence tests lose statistical power. Instead, we focus on the search-and-score methods, where we search the space of possible DAGs for the best model (a point estimate), or a sample of models (an approximation to the Bayesian posterior). Unfortunately, as shown in (Cooper, 1998), the number of DAGs as a function of the number of nodes, $G(N)$, is super-exponential in N, so we cannot exhaustively search the space. Instead, we use a local search algorithm like greedy hill climbing, (Bishop, 2006), or a global search algorithm like MCMC.

To compare the models we need a scoring function, either the BIC as described in Section 4.3.1.3, or the Bayesian score that integrates over the parameters, i.e. it is the marginal likelihood of the model. The BIC has the advantage of not

[3] http://bnt.sourceforge.net/

requiring a prior.

As with parameter learning, handling missing data is much harder than the CC. The structure learning routines in BNT can be classified into 4 types:

	Full observability	Partial observability
Point	K2	Structural EM
Bayes	MCMC	not supported

The brute-force approach to structure learning is to enumerate all possible DAGs and score each one, but in practice this is not feasible for more than 5 nodes. Instead, we focus on approximate algorithms.

If we know a total ordering of the nodes, we can find the best structure by searching for the best set of parents for each node independently. This is what the **greedy K2 search algorithm** by (Cooper and Herskovits, 1992) does. We initialize the algorithm with all nodes having no parents. In each step we add the parent that increases the score (of the resulting graph) most. When the addition of a single parent cannot increase the score, it stops adding parents to the node. Since we have a fixed ordering, we can choose the parents for each node independently and do not need to check for cycles.

In the **MCMC search algorithm** we use the MH algorithm described in Section 4.2.1 to search the space of all DAGs. For this purpose we need a proposal distribution. Normally, we would consider moving to all nearest neighbors. A neighbor is a graph that can be generated from the current graph by adding, deleting or reversing a single arc (subject to the acyclicity constraint). When there is partial observability, it is very difficult to compute the Bayesian score, because the parameter posterior is multi-modal (mixture distribution). Instead, we use the BIC approximation. However, to compute the score of each model we need the MLE. This implies running EM at each step of the algorithm, a computationally very expensive procedure. Alternatively, we can perform the local search steps inside the M-step, which is much cheaper as we have now filled in the missing values. This algorithm is the **structural EM algorithm** by (Friedman, 1998), (Friedman, 1997), (Cooper, 1998), and converges to a local maximum of the BIC score.

4.3.2.3 Inference

As explained in Section 4.3.1-4.3.1.3, there exist both exact and approximate inference techniques. BNT supports many inference engines of both types. We use the inference algorithms to estimate the missing values in our data set. We can choose between the joint distribution over the missing nodes and the *Most Probable Explanation (MPE).*

MPE is the most probable assignment of values to the hidden nodes, or the mode of the joint distribution. This gives us a single, complete “pseudo” data set that we can use as input to our survival algorithms.

If we compute the joint distribution, we can use it to sample values of the hidden nodes. This gives us the opportunity to do multiple imputation, i.e. that we can generate several complete “pseudo” data sets, where each set is a realization of the joint distribution for the missing values. We can also compute all realizations of the missing values and create an artificial data point for each missing data pattern. Then we assign a weight to each data point defined as the joint probability of this pattern. A fully observed subject has weight 1. We implement the latter solution in this work.

4.4 A Semi-Parametric Approach

Our second approach to the missing data problem is to use an EM algorithm, where we estimate the missing values in the E-step and update our (ML) parameter estimates in the M-step. To understand the algorithm we need to understand that data are not *just missing*. There is (perhaps) a reason why they are missing.

4.4.1 Missing Data Mechanisms

Missing data are divided into three categories depending on "why" they are missing, random processes, processes which are measured, and processes which are not measured.

4.4.1.1 Missing Completely at Random

Data are *Missing Completely At Random (MCAR)*, (Gelman et al., 2004), (Ibrahim et al., 2005), (Herring et al., 2004), if the probability that the data are missing is independent of any data, observed as well as missing. In survival analysis, let t_i be the observed failure/censoring time for the i'th subject, and $\boldsymbol{z}_i = \{\boldsymbol{z}_i^{\text{obs}}, \boldsymbol{z}_i^{\text{miss}}\}$ be the variable vector for the i'th subject where $\boldsymbol{z}_i^{\text{obs}}$ are the observed variables and $\boldsymbol{z}_i^{\text{miss}}$ the variables with missing values. The values. The missing variables, $\boldsymbol{z}_i^{\text{miss}}$, are MCAR if the probability of observing $\boldsymbol{z}_i$ does not depend on t_i, $\boldsymbol{z}_i^{\text{obs}}$, or the value of $\boldsymbol{z}_i^{\text{miss}}$. If data are MCAR, the observed data $\boldsymbol{z}_i^{\text{obs}}$ is a random sample of all the data, and missing cases are no different than non-missing cases in terms of the performed analysis. Hence, a CC analysis will not introduce bias, but is still inefficient if the proportion of missing data is significant, (Gelman et al., 2004), (Ibrahim et al., 2005).

4.4.1.2 Missing at Random

Data are *Missing At Random (MAR)*, (Gelman et al., 2004), (Ibrahim et al., 2005), (Herring et al., 2004), if the missingness is independent of the values of the missing data conditioned on the observed data. Hence, the conditional probability of missingness may depend on the observed data, and the un-conditional probability of missingness data may also depend on the unobserved data. The missing variables, $\boldsymbol{z}_i^{\text{miss}}$, are MAR if, conditional on $\boldsymbol{z}_i^{\text{obs}}$, the probability of observing $\boldsymbol{z}_i$ is independent of the values of $\boldsymbol{z}_i^{\text{miss}}$. However, this probability does not have to be independent of $\boldsymbol{z}_i^{\text{obs}}$ and t_i.

If data are MAR, and the missingness does not depend on t_i, the missing data are fully described by variables observed in the data set, and a CC analysis will give unbiased, yet inefficient, results, (Little and Rubin, 1987). However, if the missingness depends on t_i, the results will be biased. Although we observe t_i, we cannot account for its relation to $\boldsymbol{z}_i^{\text{miss}}$ when we set up a model to predict t_i given $\boldsymbol{z}_i$.

If we assume that the parameters of the missingness distribution are distinct from the parameters of the joint distribution of $(\boldsymbol{z}_i, t_i)$, and data are either MCAR or MAR, we have ignorable missing data. In this case we can ignore the missing data mechanism when we estimate the parameters of the joint distribution using (partial) likelihood inference.

4.4.1.3 Non-Ignorable Missing Data

If the missingness depends on the missing values that *would have been observed*, we have *Non-Ignorable (NI)* missing data, (Gelman et al., 2004), (Ibrahim et al., 2005), (Herring et al., 2004), e.g. if the probability of observing $\boldsymbol{z}_i$, conditional on $\boldsymbol{z}_i^{\text{obs}}$, depends on the values of $\boldsymbol{z}_i^{\text{miss}}$. When the missing data depends on events or items that are not measured, we have a problem. Again, a CC analysis will be inefficient, but lead to unbiased results if the missingness depends only on $\boldsymbol{z}_i$ and not t_i, (Gelman et al., 2004), (Ibrahim et al., 2005), (Herring et al., 2004). NI missing data is the most common situation, and a valid inference technique requires the specification of the correct model for the missing data mechanism and distributional assumptions for the variables with missing values. Martinussen (1999) used an EM algorithm to estimate missing values in Cox Regression, but assumed that the values were missing at random. In this work we use a semi-parametric approach to specify the joint distribution of the missing data mechanism R, the failure time T, and the variable vector $\boldsymbol{Z}$. A general approach would be to specify conditional distributions for $[R|T, \boldsymbol{Z}]$ and $[T|\boldsymbol{Z}]$, and a marginal distribution for $\boldsymbol{Z}$. In this approach, we place fully parametric distributions on $[R|T, \boldsymbol{Z}]$ and $\boldsymbol{Z}$, but use the CPH model on $[T|\boldsymbol{Z}]$. The method is completely general in terms of the type of missing data and the type and number of variables subject to missingness. The algorithm was first presented in (Herring and Ibrahim, 2001) omitting the missingness distribution, but was later incorporated by Herring et al. (2004) in a case-study of an international breast cancer study.

4.4.2 EM Estimation

Inserting (2.49) and (2.53) in (2.15), we get the failure time density, the probability of failure at time t, using the CPH model

$$\begin{aligned} f(t|\boldsymbol{z}_i, \boldsymbol{\beta}) &= h(t|\boldsymbol{z}_i, \boldsymbol{\beta}) S(t|\boldsymbol{z}_i, \boldsymbol{\beta}) \\ &= h_0(t) \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_i) \exp\big[-\exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_i) H_0(t)\big] \qquad (4.36) \end{aligned}$$

where we keep the baseline hazard function unspecified. We include the risk factor parameter vector, $\boldsymbol{\beta}$, in the conditions to highlight that the distribution changes with the parameter vector.

In this work we use right-censored data, where the observation of T is censored by a variable U so that the observable responses are $X = \min(T, U)$, and the failure indicator $\delta = I_{(T \leq X)}$ is equal to 1 if the observed event is a failure and 0 otherwise.

In the case of non-informative censoring, where the censoring distribution does not depend on the unknown parameters in the model, the probability distribution for the i'th data point conditional on the variables, $(x_i, \delta_i|\boldsymbol{z}_i)$, consists of a failure part and a censoring part. One part has weight 1 and the other 0 depending upon the indicator variable

$$\begin{aligned} p(x_i, \delta_i|\boldsymbol{z}_i, \boldsymbol{\beta}, H_0(x_i)) &\propto f(x_i|\boldsymbol{z}_i, \boldsymbol{\beta}, H_0(x_i))^{\delta_i} S(x_i|\boldsymbol{z}_i, \boldsymbol{\beta}, H_0(x_i))^{1-\delta_i} \qquad (4.37) \\ &= h(x_i|\boldsymbol{z}_i, \boldsymbol{\beta})^{\delta_i} S(x_i|\boldsymbol{z}_i, \boldsymbol{\beta}, H_0(x_i)) \qquad (4.38) \\ &= \Big[h_0(x_i) \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_i)\Big]^{\delta_i} \exp\big[-\exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_i) H_0(x_i)\big] \\ &\qquad (4.39) \end{aligned}$$

according to (4.36). The LL for the i'th individual is

$$l(\boldsymbol{\beta}) = \delta_i \Big[\log(h_0(x_i)) + \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_i\Big] - \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_i) H_0(x_i) \qquad (4.40)$$

As we would like to keep the (cumulative) baseline hazard function unspecified, we use the LPL in (2.56) to give

$$l_p(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left\{ \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_i - \log\left[\sum_{l=1}^{n} I_{(x_l \geq x_i)} \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_l)\right]\right\} \qquad (4.41)$$

using the score in (2.57), and the Newton-Raphson solution outlined in (2.58)-2.60. The only difference between (2.56) and (4.41) is notational: in (2.56) we sum over the k failure times whereas in (4.41) we sum over all n individuals, but each contribution is then multiplied by the indicator variable. Furthermore, the sum over the risk set corresponds to the sum over all individuals where $I_{(x_l \geq x_i)}$.

To further simplify the notation, we write the score equation, $\frac{\partial l(\boldsymbol{\beta})}{\boldsymbol{\beta}}$, as a stochastic integral using process notation, (Herring et al., 2004), (Andersen et al., 1993)

$$\boldsymbol{u}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_{0}^{\infty} \left[\boldsymbol{z}_i - \bar{\boldsymbol{Z}}(\boldsymbol{\beta}, u)\right] dN_i(u) \tag{4.42}$$

where

$$\bar{\boldsymbol{Z}}(\boldsymbol{\beta}, u) = \frac{\sum_{i=1}^{n} \boldsymbol{z}_i Y_i(u) \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_i)}{\sum_{i=1}^{n} Y_i(u) \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_i)} \tag{4.43}$$

and $N_i(t) = I_{(x_i \leq t, \delta_i = 1)}$ and $Y_i(t) = I_{(x_i \geq t)}$. The process $N_i(t)$ takes the value 1 if the i'th subject fails at or before time t, and 0 otherwise. The process $Y_i(t)$ takes the value 1 if the i'th individual is still at risk at time t, and 0 otherwise.

With missing risk factors (we assume fully observed responses), we need to specify parametric distributions for the risk factors with missing values. We refer to these distributions using the parameter vector $\boldsymbol{\alpha}$ for the values of the missing risk factors, and $\boldsymbol{\phi}$ for the parameters of the missing data mechanism.

Unfortunately, we also have to estimate the cumulative base-line hazard function, $\boldsymbol{H}_0(t)$. This gives us the following complete score equations

$$\boldsymbol{u}(\hat{\boldsymbol{\theta}}) = \begin{pmatrix} \boldsymbol{u}_{\beta}(\hat{\boldsymbol{\beta}}) \\ \boldsymbol{u}_{H_0}(\hat{\boldsymbol{H}}_0(t)) \\ \boldsymbol{u}_{\alpha}(\hat{\boldsymbol{\alpha}}) \\ \boldsymbol{u}_{\phi}(\hat{\boldsymbol{\phi}}) \end{pmatrix} = 0 \tag{4.44}$$

If we take expectations with respect to the conditional distribution of the missing variables given the observed data, we can estimate the parameters by solving the resulting estimation equations

$$\tilde{\boldsymbol{u}}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)}) = \begin{pmatrix} \frac{\partial E\left[l_{\beta}(\boldsymbol{\beta}|\boldsymbol{\theta}^{(m)})|\text{observed data}\right]}{\partial \boldsymbol{\beta}} \\ \frac{\partial E\left[l_{H_0}(\boldsymbol{H}_0|\boldsymbol{\theta}^{(m)})|\text{observed data}\right]}{\partial \boldsymbol{H}_0} \\ \frac{\partial E\left[l_{\alpha}(\boldsymbol{\alpha}|\boldsymbol{\theta}^{(m)})|\text{observed data}\right]}{\partial \boldsymbol{\alpha}} \\ \frac{\partial E\left[l_{\phi}(\boldsymbol{\phi}|\boldsymbol{\theta}^{(m)})|\text{observed data}\right]}{\partial \boldsymbol{\phi}} \end{pmatrix} = 0 \tag{4.45}$$

where the observed data are the observed risk factors, $\boldsymbol{z}^{\text{obs}}$, the event times, $\boldsymbol{X}$, the failure indicators, $\boldsymbol{\delta}$, and the current estimates of the missing variables, $\boldsymbol{z}^{\text{miss}}$. From now on we will omit the EM iteration index, (m), for clarity. The missing values can be categorical or continuous, and the missing data mechanism can be either MCAR, MAR, or NI; if we know they are MCAR, however, we do not need to model the missing data mechanism. Assume the missing variables are categorical. As in (Herring and Ibrahim, 2001) we approximate

$$\frac{\partial E\left[l_\beta(\boldsymbol{\beta}|\boldsymbol{\theta})|\text{observed data}\right]}{\partial \boldsymbol{\beta}} \tag{4.46}$$

to the first order using

$$\tilde{\boldsymbol{u}}_\beta(\boldsymbol{\beta}|\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{k=1}^{n_i}\int_0^\infty \left\{p_{ik}\left[\boldsymbol{z}_{ik} - \tilde{\boldsymbol{Z}}(\boldsymbol{\beta}, u)\right]\right\} dN_i(u) \tag{4.47}$$

where

$$\tilde{\boldsymbol{Z}}(\boldsymbol{\beta}, u) = \frac{\sum_{i=1}^{n}\sum_{k=1}^{n_i} p_{ik}\boldsymbol{z}_{ik}Y_i(u)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}_i)}{\sum_{i=1}^{n}\sum_{k=1}^{n_i} p_{ik}Y_i(u)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}_i)} \tag{4.48}$$

using $\boldsymbol{z}_{ik} = \{\boldsymbol{z}_i^{\text{obs}}, \boldsymbol{z}_{ik}^{\text{miss}}\}$ where $\boldsymbol{z}_{ik}^{\text{miss}}$ is the k'th of n_i possible missing value patterns for the i'th subject. The weights, p_{ik}, are the conditional (posterior) probabilities that the missing data for individual i takes on the pattern k where

$$\begin{aligned}
p_{ik} &= p_{ik}(\boldsymbol{\theta}) \\
&= p\left(\boldsymbol{z}_{ik}^{\text{miss}}|\boldsymbol{z}_i^{\text{obs}}, x_i, \delta_i, r_i, \boldsymbol{\theta}\right) \\
&= \frac{p(\boldsymbol{r}_i|x_i, \delta_i, \boldsymbol{\phi})p(x_i, \delta_i|\boldsymbol{z}_{ik}, \boldsymbol{\beta}, H_0(x_i))p(\boldsymbol{z}_{ik}|\boldsymbol{\alpha})}{\sum_{k=1}^{n_i} p(\boldsymbol{r}_i|x_i, \delta_i, \boldsymbol{\phi})p(x_i, \delta_i|\boldsymbol{z}_{ik}, \boldsymbol{\beta}, H_0(x_i))p(\boldsymbol{z}_{ik}|\boldsymbol{\alpha})} \\
&= \frac{p(\boldsymbol{r}_i|x_i, \delta_i, \boldsymbol{\phi})p(x_i, \delta_i|\boldsymbol{z}_{ik}^{\text{miss}}, \boldsymbol{z}_i^{\text{obs}}, \boldsymbol{\beta}, H_0(x_i))p(\boldsymbol{z}_{ik}^{\text{miss}}, \boldsymbol{z}_i^{\text{obs}}|\boldsymbol{\alpha})}{\sum_{k=1}^{n_i} p(\boldsymbol{r}_i|x_i, \delta_i, \boldsymbol{\phi})p(x_i, \delta_i|\boldsymbol{z}_{ik}, \boldsymbol{\beta}, H_0(x_i))p(\boldsymbol{z}_{ik}|\boldsymbol{\alpha})}
\end{aligned} \tag{4.49}$$

where $\boldsymbol{r}_i$ is the missingness vector for the i'th subject with $r_{ji} = 1$, if variable j is missing for the i'th subject, and $\sum_{k=1}^{n_i} p_{ik} = 1$.

For each case with missing variables we replace the individual with all possible combinations (patterns) of the missing values. Say, for example, we have a data set with 10 individuals and one of these individuals have one missing binary variable. The original data set is then replaced by a data set with 11 cases, where the missing value case is replaced by two complete cases: one where the missing variable has the value 0, and one where it has the value 1. Each case is then weighted with the probability that the missing values took on pattern j conditional on the observed variables. Cases with no missing values all have weight 1. Although the baseline hazard, $h_0(x_i)$, is used in

$$p(x_i, \delta_i|\boldsymbol{z}_{ik}, \boldsymbol{\beta}, H_0(x_i)) = \left[h_0(x_i)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}_{ik})\right]^{\delta_i}\exp\left[-\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}_{ik})H_0(x_i)\right] \tag{4.50}$$

we do not need to estimate it, since it only depends on the event time x_i and cancels out when we compute p_{ik}.

The algorithm proceeds as follows

1. Estimate $\boldsymbol{\theta}^{(0)} = \{\boldsymbol{\beta}, H_0(t), \boldsymbol{\alpha}, \boldsymbol{\phi}\}$ using the complete cases.

2. At the $(m+1)$'th EM iteration, compute p_{ik}, solve $\tilde{u}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)})$ for $\boldsymbol{\theta}^{(m+1)}$, and update the estimates of $\boldsymbol{\beta}$ and the nuisance parameters $H_0(t)$, $\boldsymbol{\alpha}$ and $\boldsymbol{\phi}$.

3. Iterate until convergence.

At each iteration of the EM algorithm we estimate the weights, p_{ik}, depending on the nuisance parameters $H_0(t)$, $\boldsymbol{\alpha}$ and $\boldsymbol{\phi}$, and the parameters of the weighted CPH model, $\boldsymbol{\beta}$. The estimates of $\boldsymbol{\beta}$ at each iteration of the EM algorithm are obtained by treating the weights as fixed and solving for $\tilde{\boldsymbol{\beta}}$.

4.4.3 Estimation of Nuisance Parameters

As we have chosen a semi-parametric approach to the missing values problem, we need to specify distributional assumptions for the risk factor distributions and the missingness distributions.

4.4.3.1 Variable Distributions

When a subject has missing values, we specify a distribution for the missing variables conditioned on the observed variables and estimate its parameters from the data. The parameters are *nuisance parameters*, i.e. they are not of inferential interest. Optimally, we would like to specify the joint distribution for all missing risk factors, but since we allow both categorical and continuous missing risk factors, there may not be a natural joint distribution. Instead, we allow the joint distribution to factorize, i.e. we assume we can write the joint distribution as a series of one-dimensional conditional distributions.

Let data have i.i.d. uncertain risk factor vectors $\boldsymbol{z}_i, i = 1, \ldots, n$ with density $p(\boldsymbol{z}_i|\boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ is distinct from $\boldsymbol{\beta}$, $H_0(t)$ and $\boldsymbol{\phi}$. Furthermore, write the risk factor vector $\boldsymbol{z}_i = \{\boldsymbol{z}_i^{\text{miss}}, \boldsymbol{z}_i^{\text{obs}}\} = \{(z_{i,1}^{\text{miss}}, \ldots, z_{i,p}^{\text{miss}}), \boldsymbol{z}_i^{\text{obs}}\}$ where $(z_{i,1}^{\text{miss}}, \ldots, z_{i,p}^{\text{miss}})$ are missing for at least one i, and $\boldsymbol{z}_i^{\text{obs}}$ are observed for all i. Then we write the joint distribution of the missing variables as

$$\begin{aligned}
p(z_{i,1}^{\text{miss}}, \ldots, z_{i,p}^{\text{miss}}|\boldsymbol{\alpha}) = p(z_{i,p}^{\text{miss}}&|z_{i,1}^{\text{miss}} \ldots, z_{i,p-1}^{\text{miss}}, \boldsymbol{z}_i^{\text{obs}}, \boldsymbol{\alpha}_p) \\
&\times p(z_{i,p-1}^{\text{miss}}|z_{i,1}^{\text{miss}} \ldots, z_{i,p-2}^{\text{miss}}, \boldsymbol{z}_i^{\text{obs}}, \boldsymbol{\alpha}_{p-1}) \times \ldots \\
&\times p(z_{i,1}^{\text{miss}}|\boldsymbol{z}_i^{\text{obs}}, \boldsymbol{\alpha}_1) \qquad (4.51)
\end{aligned}$$

where $\boldsymbol{\alpha}_j$ is the parameter vector for the j'th conditional distribution, the $\boldsymbol{\alpha}_j$'s are distinct and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p)$. Hence, we need only specify conditional distributions for risk factors not completely observed. The estimation equation for $\boldsymbol{\alpha}$ is $\boldsymbol{u}_\alpha(\hat{\boldsymbol{\alpha}}) = 0$ where

$$\boldsymbol{u}_\alpha(\boldsymbol{\alpha}|\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\partial\,[\log(p(\boldsymbol{z}_i|\boldsymbol{\alpha}))]}{\partial\boldsymbol{\alpha}} \tag{4.52}$$

With missing data, we use the expectation with respect to the conditional distribution of the missing data given the observed data

$$\tilde{\boldsymbol{u}}_\alpha(\boldsymbol{\alpha}|\boldsymbol{\theta}) = \frac{\partial}{\partial\boldsymbol{\alpha}} E\big[l_\alpha(\boldsymbol{\alpha}|\boldsymbol{\theta})|\text{observed data}\big] = \sum_{i=1}^{n}\sum_{k=1}^{n_i} p_{ik} \frac{\partial\big[\log(p(\boldsymbol{z}_{ik}|\boldsymbol{\alpha}))\big]}{\partial\boldsymbol{\alpha}} \tag{4.53}$$

where k is the index of the missing value pattern for the p missing variables, and not the variable index j. This technique has the obvious drawback that we need to specify a conditioning order, and compare various main effects and interaction models for the risk factor.

4.4.3.2 Missingness Distributions

Again, we use a sequence of one-dimensional conditional distributions to model the missing data mechanism

$$\begin{aligned} p(r_{i,1}, \ldots, r_{i,p}|x_i, \boldsymbol{z}_i, \boldsymbol{\phi}) &= p(r_{i,p}|r_{i,1} \ldots, r_{i,p-1}, x_i, \boldsymbol{z}_i, \boldsymbol{\phi}_p) \\ &\quad \times p(r_{i,p-1}|r_{,i,1} \ldots, r_{i,p-2}, x_i, \boldsymbol{z}_i, \boldsymbol{\phi}_{p-1}) \times \ldots \\ &\quad \times p(r_{i,1}|x_i, \boldsymbol{z}_i, \boldsymbol{\phi}_1) \end{aligned} \tag{4.54}$$

where $\boldsymbol{\phi}_j$ is the parameter vector for the j'th conditional distribution and $\boldsymbol{\phi} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_p)$.

Because each r_{ji} is dichotomous, a sequence of logistic regressions may be used for (4.54). This greatly reduces the number of nuisance parameters and closely approximates a joint log-linear model for the missing data indicators. The estimation equation for $\boldsymbol{\phi}$ is $\boldsymbol{u}_\phi(\hat{\boldsymbol{\phi}}) = 0$ where

$$\boldsymbol{u}_\phi(\boldsymbol{\phi}|\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\partial\,[\log(p(\boldsymbol{r}_i|x_i, \boldsymbol{z}_i, \boldsymbol{\phi}))]}{\partial\boldsymbol{\phi}} \tag{4.55}$$

With missing data, we use the expectation with respect to the conditional distribution of the missingness given the observed data

$$\tilde{\boldsymbol{u}}_\phi(\boldsymbol{\phi}|\boldsymbol{\theta}) = \frac{\partial}{\partial\boldsymbol{\phi}} E\big[l_\phi(\boldsymbol{\phi}|\boldsymbol{X}, \boldsymbol{Z})\big] = \sum_{i=1}^{n}\sum_{k=1}^{n_i} p_{ik} \frac{\partial\big[\log(p(\boldsymbol{r}_i|x_i, \boldsymbol{z}_i, \boldsymbol{\phi}))\big]}{\partial\boldsymbol{\phi}} \tag{4.56}$$

4.4.3.3 Estimation of the Cumulative Baseline Hazard

We also need to estimate the cumulative hazard rate, $H_0(t)$, using the estimate proposed by (Breslow, 1974)

$$\hat{H}_0(t) = \sum_{i=1}^{n} \sum_{k=1}^{n_i} \int_0^t p_{ik} \frac{\sum_{i=1}^{n} dN_i(u)}{\sum_{i=1}^{n} Y_i(u) \exp(\tilde{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{z}_{i,k})} \tag{4.57}$$

to get an estimate of $p(x_i, \delta_i | \boldsymbol{z}_i, \boldsymbol{\beta}, H_0(x_i))$ used in the computation of the weight, p_{ik}.

4.4.3.4 Estimation of Weights

We get estimates of p_{ik} at the m'th iteration of the EM algorithm by inserting the m'th estimate of the parameters $\boldsymbol{\beta}$, $H_0(t)$, $\boldsymbol{\alpha}$ and $\boldsymbol{\phi}$, in (4.49).

4.4.4 Monte Carlo EM method for Continuous or Mixed Variables

In most practical applications we have categorical as well as continuous missing variables which complicates the situation even further. With continuous variables we integrate instead of summing over missing values

$$\tilde{\boldsymbol{u}}(\boldsymbol{\beta}|\boldsymbol{\theta}) = \sum_{i=1}^{n} \int_{\boldsymbol{z}_i^{\mathrm{miss}}} \left[\int_0^{\infty} \left\{ p\left(\boldsymbol{z}_i^{\mathrm{miss}} | \boldsymbol{z}_i^{\mathrm{obs}}, x_i, \delta_i, r_i, \boldsymbol{\theta}\right) \left[\boldsymbol{z}_i - \tilde{\boldsymbol{Z}}(\boldsymbol{\beta}, u)\right] \right\} dN_i(u) \right] d\boldsymbol{z}_i^{\mathrm{miss}} \tag{4.58}$$

In general, we cannot evaluate this integral analytically. However, as (4.58) is an expectation with respect to $p\left(\boldsymbol{z}_i^{\mathrm{miss}} | \boldsymbol{z}_i^{\mathrm{obs}}, x_i, \delta_i, r_i, \boldsymbol{\theta}\right)$ we may evaluate the integral by drawing samples from this distribution using an MCMC algorithm, see Section 4.2.

Furthermore, we can write $p\left(\boldsymbol{z}_i^{\mathrm{miss}} | \boldsymbol{z}_i^{\mathrm{obs}}, x_i, \delta_i, r_i, \boldsymbol{\theta}\right)$ in terms of the missingness distribution, the variable distribution, and the event distribution

$$p\left(\boldsymbol{z}_i^{\text{miss}}|\boldsymbol{z}_i^{\text{obs}}, x_i, \delta_i, \boldsymbol{\theta}\right)$$
$$= \frac{p\left(\boldsymbol{r}_i|x_i, \boldsymbol{z}_i^{\text{miss}}, \boldsymbol{z}_i^{\text{obs}}, \boldsymbol{\phi}\right) p\left(x_i, \delta_i|\boldsymbol{z}_i^{\text{miss}}, \boldsymbol{z}_i^{\text{obs}}, \boldsymbol{\beta}, H_0(t)\right) p\left(\boldsymbol{z}_i^{\text{miss}}, \boldsymbol{z}_i^{\text{obs}}|\boldsymbol{\alpha}\right)}{\int_{\boldsymbol{z}_i^{\text{miss}}} p\left(\boldsymbol{r}_i|x_i, \boldsymbol{z}_i^{\text{miss}}, \boldsymbol{z}_i^{\text{obs}}, \boldsymbol{\phi}\right) p\left(x_i, \delta_i|\boldsymbol{z}_i^{\text{miss}}, \boldsymbol{z}_i^{\text{obs}}, \boldsymbol{\beta}, H_0(t)\right) p\left(\boldsymbol{z}_i^{\text{miss}}, \boldsymbol{z}_i^{\text{obs}}|\boldsymbol{\alpha}\right) d\boldsymbol{z}_i^{\text{miss}}} \tag{4.59}$$
$$\propto p\left(\boldsymbol{r}_i|x_i, \boldsymbol{z}_i^{\text{miss}}, \boldsymbol{z}_i^{\text{obs}}, \boldsymbol{\phi}\right) p\left(x_i, \delta_i|\boldsymbol{z}_i^{\text{miss}}, \boldsymbol{z}_i^{\text{obs}}, \boldsymbol{\beta}, H_0(t)\right) \times p\left(\boldsymbol{z}_i^{\text{miss}}, \boldsymbol{z}_i^{\text{obs}}|\boldsymbol{\alpha}\right)$$

All these distributions are known (have been specified). In (Neal, 1993) it is shown that $p\left(x_i, \delta_i|\boldsymbol{z}_i^{\text{miss}}, \boldsymbol{z}_i^{\text{obs}}, \boldsymbol{\beta}, H_0(t)\right)$ is log-concave in the components of $\boldsymbol{z}_i$. If we select each one-dimensional conditional distribution in (4.51) and (4.54) from the exponential family, then $p\left(\boldsymbol{z}_i^{\text{miss}}, \boldsymbol{z}_i^{\text{obs}}|\boldsymbol{\alpha}\right)$ and $p\left(\boldsymbol{r}_i|x_i, \boldsymbol{z}_i^{\text{miss}}, \boldsymbol{z}_i^{\text{obs}}, \boldsymbol{\phi}\right)$ are also log-concave in the components of $\boldsymbol{z}_i$. A sum of log-concave densities is also concave, and we can use a MCMC method along with an adaptive rejection algorithm to sample the missing variables as explained in Section 4.2.

We can evaluate (4.58) as follows. For each subject i we take a sample $\boldsymbol{s}_{i,1}, \ldots, \boldsymbol{s}_{i,n_i'}$ of size n_i' from $p\left(\boldsymbol{z}_i^{\text{miss}}, \boldsymbol{z}_i^{\text{obs}}|\boldsymbol{\alpha}\right)$ and $p\left(\boldsymbol{r}_i|x_i, \boldsymbol{z}_i^{\text{miss}}, \boldsymbol{z}_i^{\text{obs}}, \boldsymbol{\phi}\right)$ where each $\boldsymbol{s}_{ik'}$, $k' = 1, \ldots, n_i'$ is a vector of length m_i, the length of $\boldsymbol{z}_i^{\text{miss}}$.

Obviously, each $\boldsymbol{s}_{ik'}$ also depends on the EM iteration number. Assigning equal weights $\frac{1}{n_i'}$ to each sample, we obtain the following E-step for $\boldsymbol{\beta}$

$$\tilde{\boldsymbol{u}}(\boldsymbol{\beta}|\boldsymbol{\theta}) = \sum_{i=1}^{n} \left[\frac{1}{n_i'} \sum_{k'=1}^{n_i'} \int_0^{\infty} \left\{ \boldsymbol{z}_{ik'} - \tilde{\boldsymbol{Z}}(\boldsymbol{\beta}, u) \right\} dN_i(u) \right] \tag{4.60}$$

where $\boldsymbol{z}_{ik'} = \{\boldsymbol{z}_i^{\text{obs}}, \boldsymbol{s}_{ik'}\}$, $k' = 1, \ldots, n_i$ is the joint vector of the observed risk factors and the sampled values of the missing variables for the i'th individual and

$$\tilde{\boldsymbol{Z}}(\boldsymbol{\beta}, u) = \frac{\sum_{i=1}^{n} \left[\frac{1}{n_i'} \sum_{k'=1}^{n_i'} \boldsymbol{z}_{ik'} Y_i(u) \exp(\boldsymbol{\beta}^{\text{T}} \boldsymbol{z}_{ik'}) \right]}{\sum_{i=1}^{n} \left[\frac{1}{n_i'} \sum_{k'=1}^{n_i'} Y_i(u) \exp(\boldsymbol{\beta}^{\text{T}} \boldsymbol{z}_{ik'}) \right]} \tag{4.61}$$

Having drawn the samples, we proceed to the maximization step. If we have both categorical as well as continuous missing variables, we still write the joint distribution of the missing variables as a product of one-dimensional conditional distributions. For example, if Z_2 is binary and Z_1 follows a normal distribution, we use logistic regression models for $p(Z_2|Z_1, \boldsymbol{\alpha}_2)$, $p(R_2|X, Z_1, Z_2, R_1, \boldsymbol{\phi}_2)$, and $p(R_1|X, Z_1, \boldsymbol{\phi}_1)$ and a normal linear regression model for $p(Z_1|\boldsymbol{\alpha}_1)$.

To simplify the sampling process, we separate the missing continuous and categorical variables and write $\boldsymbol{z}_i^{\text{miss}} = \{\boldsymbol{z}_i^{\text{miss},d}, \boldsymbol{z}_i^{\text{miss},c}\}$ using d for discrete (categorical) and c for continuous variables. For each missing categorical pattern we sample the continuous variables. By summing over the categorical patterns, we can avoid sampling values from their distribution. If we take a sample $\{\boldsymbol{s}_{i,1}^{c}, \ldots, \boldsymbol{s}_{i,n'_i}^{c}\}$ of size n'_i from $p\left(\boldsymbol{z}_i^{\mathrm{miss},c} | \boldsymbol{z}_i^{\mathrm{obs}}, x_i, \delta_i, \boldsymbol{\theta}\right)$ we get the Monte Carlo expectation of the score distribution

$$\tilde{\boldsymbol{u}}(\boldsymbol{\beta}|\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{n_i} \left[\frac{1}{n'_i} \sum_{k'=1}^{n'_i} \int_0^{\infty} p_{ikk'} \left(\boldsymbol{z}_{ikk'} - \tilde{\boldsymbol{Z}}(\boldsymbol{\beta}, u) \right) dN_i(u) \right] \tag{4.62}$$

where

$$\tilde{\boldsymbol{Z}}(\boldsymbol{\beta}, u) = \frac{\sum_{i=1}^{n} \sum_{k=1}^{n_i} \left[\frac{1}{n'_i} \sum_{k'=1}^{n'_i} \boldsymbol{z}_{ikk'} Y_i(u) \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_{ikk'}) \right]}{\sum_{i=1}^{n} \sum_{k=1}^{n_i} \left[\frac{1}{n'_i} \sum_{k'=1}^{n'_i} Y_i(u) \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_{ikk'}) \right]} \tag{4.63}$$

and $\boldsymbol{z}_{ikk'} = \{\boldsymbol{z}_i^{\mathrm{obs}}, \boldsymbol{z}_{ik}^{\mathrm{miss},d}, \boldsymbol{s}_{ik'}\}$, $k = 1, \ldots, n_i$, $k' = 1, \ldots, n'_i$ is the joint vector of the observed variables, the discrete missing value patterns, and the sampled values of the missing variables for the i'th individual. The weights are given by

$$p_{ikk'} = p\left(\boldsymbol{z}_{ik}^{\mathrm{miss},d} | \boldsymbol{s}_{ik'}^{c}, \boldsymbol{z}_i^{\mathrm{obs}}, x_i, \delta_i, \boldsymbol{\theta}\right) \tag{4.64}$$

Chapter 5

Databases

In this chapter we present two real-life data sets, the multiple myeloma patients data set, and the Copenhagen Stroke Study database used in the experimental sections.

5.1 Survival of Multiple Myeloma Patients

The *Multiple Myeloma Patients (MMP)* data set is presented in (Collet, 2003), and relates to the survival of 48 patients with multiple myeloma.

Multiple myeloma, (Lonial, 2005), also known as MM, myeloma, plasma cell myeloma, or as Kahler's disease after Otto Kahler, is a malignant disease, where abnormal plasma cells accumulate rapidly in the bone marrow causing pain, destruction of bone tissue, anaemia, haemorrhages, infections, and weakness. The affected cells are plasma cells (a type of white blood cell), which are our antibody- (immunoglobulin-) producing cells. Myeloma is called “multiple”, since there are frequently multiple patches or areas in the bones, where tumors or lesions have developed. Its prognosis, despite therapy, is generally poor, and treatment may involve chemotherapy and stem cell transplant. With no treatment, the disease is fatal. The study was performed out at the Medical Center of University of West Virginia, USA, and the aim was to investigate the

patient	time	status	age	sex	bun	ca	hb	cells	protein
1	13	1	66	1	25	10	14.6	18	1
2	52	0	66	1	13	11	12.0	100	0
3	6	1	53	2	15	13	11.4	33	1
4	40	1	69	1	10	10	10.2	30	1
5	10	1	65	1	20	10	13.2	66	0
6	7	0	57	2	12	8	9.9	45	0
7	66	1	52	1	21	10	12.8	11	1
8	10	0	60	1	41	9	14.0	70	1
9	10	1	70	1	37	12	7.5	47	0
10	14	1	70	1	40	11	10.6	27	0
11	16	1	68	1	39	10	11.2	41	0
12	4	1	50	2	172	9	10.1	46	1
13	65	1	59	1	28	9	6.6	66	0
14	5	1	60	1	13	10	9.7	25	0
15	11	0	66	2	25	9	8.8	23	0
16	10	1	51	2	12	9	9.6	80	0
17	15	0	55	1	14	9	13.0	8	0
18	5	1	67	2	26	8	10.4	49	0
19	76	0	60	1	12	12	14.0	9	0
20	56	0	66	1	18	11	12.5	90	0
21	88	1	63	1	21	9	14.0	42	1
22	24	1	67	1	10	10	12.4	44	0
23	51	1	60	2	10	10	10.1	45	1
24	4	1	74	1	48	9	6.5	54	0

Table 5.1: Survival of multiple myeloma patients data set - first half.

role of a number of risk factors on the survival time in months from diagnosis to death from multiple myeloma. The data are presented in Table 5.1 and 5.2, and are available for download, see (Collet, 2003).

When the study ended, some patients were still alive and thus (right)-censored, indicated by $status = 0$ in Table 5.1 and 5.2. When the patient was diagnosed, the values of a number of possible risk factors were noted (name in parenthesis is the abbreviation used in the experimental work): Age of patient in years (*age*), sex of patient (*sex*), level of blood urea nitrogen (*bun*), serum calcium (*ca*), haemoglobin (*hb*), percentage of plasma cells in the bone marrow (*pcells*), and an indicator of whether or not Bence-Jones protein was present in the urine (*protein*). Bence-Jones protein is a protein often found in the blood and urine of patients with multiple myeloma. The proteins are produced by defective plasma cell function. Table 5.3 shows the range and mean of the possible risk factors,

patient	time	status	age	sex	bun	ca	hb	cells	protein
25	40	0	72	1	57	9	12.8	28	1
26	8	1	55	1	53	12	8.2	55	0
27	18	1	51	1	12	15	14.4	100	0
28	5	1	70	2	130	8	10.2	23	0
29	16	1	53	1	17	9	10.0	28	0
30	50	1	74	1	37	13	7.7	11	1
31	40	1	70	2	14	9	5.0	22	0
32	1	1	67	1	165	10	9.4	90	0
33	36	1	63	1	40	9	11.0	16	1
34	5	1	77	1	23	8	9.0	29	0
35	10	1	61	1	13	10	14.0	19	0
36	91	1	58	2	27	11	11.0	26	1
37	18	0	69	2	21	10	10.8	33	0
38	1	1	57	1	20	9	5.1	100	1
39	18	0	59	2	21	10	13.0	100	0
40	6	1	61	2	11	10	5.1	100	0
41	1	1	75	1	56	12	11.3	18	0
42	23	1	56	2	20	9	14.6	3	0
43	15	1	62	2	21	10	8.8	5	0
44	18	1	60	2	18	9	7.5	85	1
45	12	0	71	2	46	9	4.9	62	0
46	12	1	60	2	6	10	5.5	25	0
47	17	1	65	2	28	8	7.5	8	0
48	3	0	59	1	90	10	10.2	6	1

Table 5.2: Survival of multiple myeloma patients data set - second half.

Variable	Range	Mean
age	50 - 77	62.9
sex	0 = male, 1 = female	1.4
bun	6 - 172	33.9
ca	8 - 15	9.9
hb	4.9 - 14.6	10.3
cells	3 - 100	42.9
protein	0 = absent, 1 = present	0.3

Table 5.3: Risk factors in the multiple myeloma patients data set.

Min	Mean	Max	25%	50% (median)	75%
1	23.4	91	6.5	14.5	38

Table 5.4: Distribution of survival times (in months) in the multiple myeloma patients data set.

Table 5.4 the distribution of survival times, and Figure 5.1 the histogram of survival times.

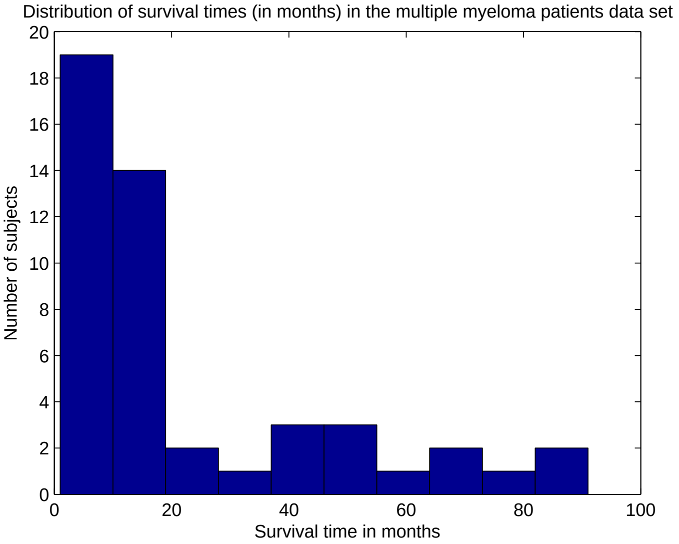

Figure 5.1: Histogram of survival times (in months) in the multiple myeloma patients data set.

5.2 Copenhagen Stroke Study Database

The *COpenhagen Stroke Study (COST)*, see e.g. (Andersen et al., 2005b), (Andersen et al., 2005a), (Nakayama et al., 1994), (Jørgensen et al., 1995a), (Jørgensen et al., 1995b), (Jørgensen et al., 1994b), (Jørgensen et al., 1994a), (Jørgensen et al., 1996b), (Jørgensen et al., 1997), (Jørgensen et al., 1996a), (Jørgensen et al., 1999b), (Jørgensen et al., 1999c), (Reith et al., 1997), (Jørgensen et al., 1999a), (Kammersgaard et al., 2004), (Kammersgaard et al., 2002), (Kammersgaard and Olsen, 2006), (Nakayama et al., 1996), is a prospective, community-based study of consecutive acute stroke patients treated on a single, 63-bed stroke unit within the neurological ward of Bispebjerg Hospital from the time (in days) of acute admission to the end of rehabilitation. This stroke unit receives all stroke patients admitted from a well-defined catchment area (population, 238,886) of Copenhagen City, Denmark. The stroke unit handles all stages of acute care, workup, and all stages of rehabilitation in all patients, regardless of the age of the patient, the severity of the stroke, and the condition of the patient prior to the stroke. The stroke admission rate in the area is high, 88%. All persons from the community who have an acute cerebrovascular disease that requires admission are referred to the neurological department of Bispebjerg Hospital. Those not admitted are patients who die before they reach the hospital and some patients with mild strokes. Inclusion started September 1, 1991, and ended September 30, 1993. Data were kindly provided by Dr. Tom Skyhøj Olsen from the stroke unit at Hvidovre Hospital.

On admission, all patients underwent a standardized examination program including CT-scan, electrocardiography, and a thorough cardiovascular risk factor evaluation using a standardized questionnaire. If patients were unable to communicate sufficiently, information was obtained from relatives or care givers. Stroke was defined according to the World Health Organization criteria, (WHO, 1993). Subarachnoid hemorrhage (bleeding into the subarachnoid space surrounding the brain) was not included.

In total, we have 999 patients with more than 1000 attributes. Some of these attributes are categorized versions of other attributes and many of the attributes are only sparsely recorded. The following prognostic factors were suggested as relevant to our work by Dr. Olsen (name in parenthesis is the abbreviation used in the experimental work): age (age), gender (sex), hypertension (hyp), ischemic heart disease (ihd), previous stroke (apo), other disabling disability (odd), daily alcohol consumption ($alco$), diabetes mellitus (dm), smoking ($smoke$), atrial fibrillation (af), type of stroke ($hemo$), initial stroke severity (sss), intermittent claudication (cla), and body temperature on admission ($temp$).

hyp Hypertension was present if a patient received antihypertensive treatment before admission, or if hypertension was diagnosed during hospital stay by repeated detection of blood pressure = 160/95 mmHg.

ihd Ischemic heart disease was present if a patient had a history of IHD, or had IHD diagnosed during the hospital stay.

apo Previous stroke was recorded if the patient had previously experienced a stroke.

odd Information concerning other disabling disease was obtained on admission and included disabling diseases other than previous stroke (e.g., amputation, multiple sclerosis, severe dementia, heart failure, latent or persistent respiratory insufficiency). Thus, various diseases were not registered separately, and the influence of specific diseases was not evaluated.

alco Alcohol was coded if a patient was drinking on a daily basis. Ex-alcohol consumers were coded as non-alcohol consumers.

dm Patients with known diabetes before stroke and patients with diabetes diagnosed after stroke onset either during the hospital stay or because admission plasma glucose was ¿0.11 mmol/L, in accordance with the World Health Organization diagnostic criteria for diabetes.

smoke Smoking was coded if a patient smoked any kind of tobacco on a daily basis. Ex-smokers were coded as non-smokers.

af Atrial fibrillation was diagnosed if the Electrocardiogram obtained on admission revealed AF.

hemo A CT scan determined stroke type as hemorrhage (bleeding) or infarct (an artery is blocked by some obstruction, e.g. a blood clot or cholesterol deposit).

sss The initial neurological stroke severity was assessed with the *Scandinavian Stroke Scale (SSS)* at the time of the acute admission. The SSS evaluates level of consciousness; eye movement; power in arm, hand, and leg; orientation; aphasia; facial paresis; and gait on a total score from 0 (worst) to 58 (best) [1].

cla Intermittent claudication was present if a patient had a history of intermittent claudication, or had intermittent claudication diagnosed during the hospital stay. Intermittent claudication is a cramping sensation in the

[1] The scale is available at `http://www.strokecenter.org/Trials/scales/scandinavian.html`.

legs that is present during exercise or walking and occurs as a result of decreased oxygen supply [2].

temp Body temperature on acute admission was recorded with a Diatek model 9000 infrared aural thermometer (Diatek); this device registers tympanic membrane temperature, which correlates well with core body temperature. Body temperature was coded as $temperature < 37.0°$ C or $\geq 37.0°$ C.

The long-term follow-up data on mortality and date of death were obtained from the Danish Central Registry of Persons, where date of death for all residents in Denmark is recorded through a unique 10-digit identification code containing information on birth date. Another experienced neurologist (L.P. Kammersgaard) who was blinded to data obtained on admission prospectively recorded the follow-up data. Follow-up was performed during the year 1999 with December 29, 1999, as censoring date. Furthermore, the database was updated with survival information during the year 2004 with November 3, 2003, as censoring date. Six patients had immigrated to another country and were lost on follow-up (reducing the sample size to 993).

The distribution of the categorial risk factors, and the mean values (standard deviations) of the continuous factors (of the 993 patients included) are shown in Table 5.5 and Table 5.6 respectively. All risk factors, except for *age* and *sex*, have missing values. 552 patients had no missing values in any of the selected variables.

Table 5.7 and Figure 5.2 show the distribution of survival times. In Table 5.8 we compute the percentage of subjects failing (dead) within a given time from admission. We note that the survival times are distributed over a time interval spanning more than 11 years, and just 19 subjects, or less than 2%, were dead on arrival. However, more than 10% died within the first week, more than 22% within 3 months, and almost 1/3 of all patients were dead within a year, indicating very high short-term mortality rate. However, as indicated in Table 5.7, the 25% quartile is 153 days or 5 months, while the median survival time is 1259 days or 4 years! As the histogram also illustrates, the mortality is very high in the days and weeks after stroke onset, but once a patient has survived this critical stage, the chances of surviving on longer terms increase rapidly.

The strength of this study is that it is prospective and community-based includ-

[2] This cramping usually occurs in the calf, but may also occur in the feet. When intermittent claudication is discussed it is measured by the number of "blocks" (e.g. 1 or 2 blocks) one can walk comfortably. It often indicates severe atherosclerosis. One of the hallmarks of this clinical entity is that it occurs intermittently. It disappears after a brief rest and the patient can start walking again until the pain recurs. Intermittent claudication is often a symptom of severe atherosclerotic disease of the peripheral vascular system.

Variable	Yes	No
sex (men/women)	438 (44.1%)	555
hyp	306 (32.9%)	625
ihd	189 (20.6%)	728
apo	195 (20.7%)	748
odd	205 (21.5%)	747
alco	261 (31.5%)	569
dm	148 (15.6%)	795
smoke	364 (44.4%)	455
af	162 (16.5%)	820
cla	107 (12.2%)	773
	infarct	hemorrhage
hemo	61 (7.6%)	744
	$< 37.0°$ C	$\geq 37.0°$ C
temp	502 (58.3%)	358

Table 5.5: Distribution of categorical COST variables.

Variable	Mean	Std.
age	74.3	11.0
sss	38.0	17.5

Table 5.6: Distribution of continuous COST variables.

$\sharp$ subjets with $t = 0$	Min	Mean	Max	25%	50% (median)	75%
19	0	1587	4262	153	1259	2828

Table 5.7: Distribution of survival times (in days) in the COST data set.

1 week	2 weeks	3 weeks	1 month	3 months	6 months	1 year
10.5	13.3	15.2	16.7	22.1	26.7	32.0

Table 5.8: Percentage of subjects failing (dead) within a given time from admission in the COST data set.

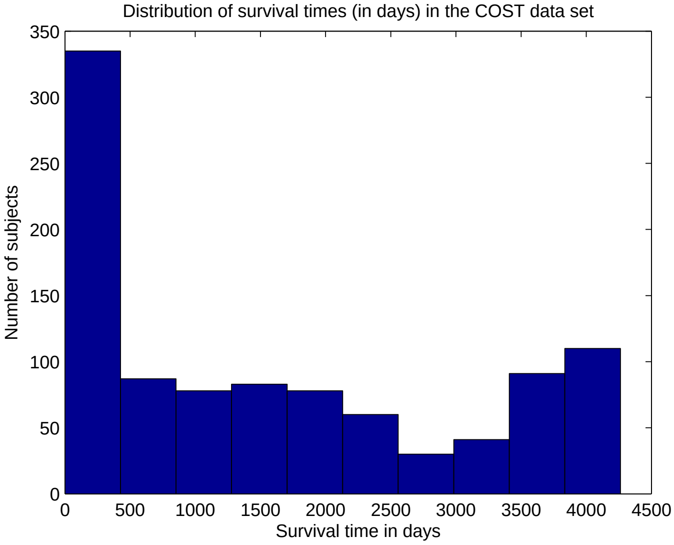

Figure 5.2: Histogram of survival times (in days) in the COST data set.

ing all patients in a well-defined community hospitalized with stroke regardless age, stroke severity, or other complicating diseases. Moreover, the stroke admittance rate in the area is high and close to the incidence reported in population-based studies. A limitation is that patients who die at home are not included and this may underestimate mortality. However, the small number of patients with minor strokes not being admitted to hospital may counterbalance it. Finally, because we have a sizeable study population and a lengthy follow-up, we consider bias to be of no major importance for the main conclusions of this study.

Chapter 6

Comparison of Stepwise Selection and Bayesian Model Averaging Applied to Real Life Data

In our first analysis we compare stepwise selection to BMA using the real life data sets presented in the previous chapter.

6.1 Survival of Multiple Myeloma Patients

The multiple myeloma patients data set is the simpler of the two data sets: It has much fewer subjects, a small number of risk factors and no missing values. To validate the proportional hazards assumption before any models are fitted, we use the log-cumulative hazard plots from Section 2.4.4.1. We compute histograms for each of the continuous variables, and decide to categorize *age*, *bun*, *hb*, and *pcells* using quartiles, [.25 .50 .75 1.0], while the distribution of *ca* shows that two categories (below and above the median value) are sufficient. In Figure 6.1 - 6.7 we show the log-cumulative hazard plots for all variables indicating that the proportional hazard assumption is not violated.

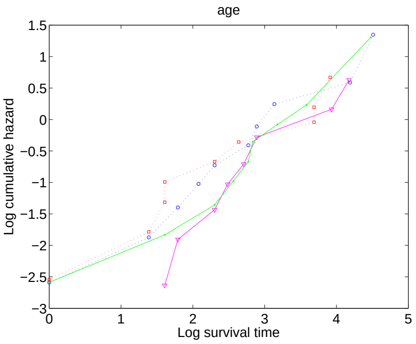

Figure 6.1: Log-cumulative hazard plot for *age*. Quartiles: 58.5 (blue, dotted), 62.5 (magenta, solid), 68.5 (green, solid), 77.0 (red, dotted).

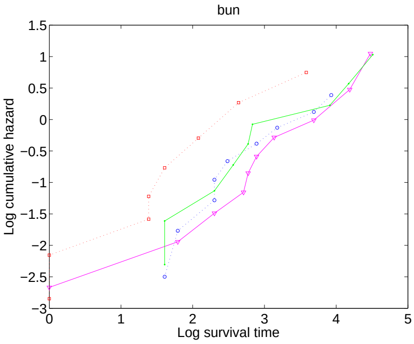

Figure 6.2: Log-cumulative hazard plot for *bun*. Quartiles: 13.5 (blue, dotted), 21.0 (magenta, solid), 39.5 (green, solid), 172.0 (red, dotted).

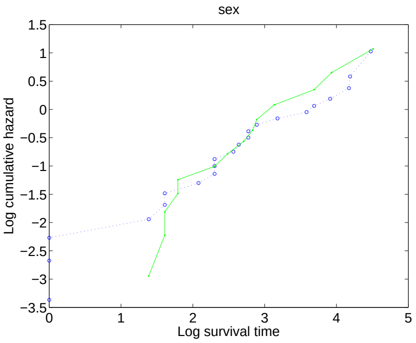

Figure 6.3: Log-cumulative hazard plot for *sex*. Male (green, solid), women (blue, dotted).

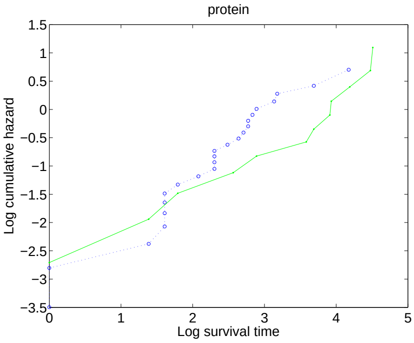

Figure 6.4: Log-cumulative hazard plot for *protein*. Yes (green, solid), no (blue, dotted).

Having validated the proportional hazards assumption, we proceed to the model fitting stage and use the stepwise selection algorithm implemented in a commercial statistical software package (SPSS, Statistical Package for the Social Sciences, SPSS Inc, Chicago, IL) to validate the results of our own implementation. Significance of predictors was based on the probability of the Wald test statistic and a significance level of $\alpha = 0.05$.

Results are presented in Table 6.1. For each variable, the second column shows the p-value at the time of removal, and in parenthesis the iteration in which

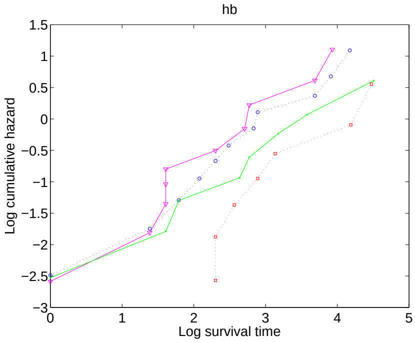

Figure 6.5: Log-cumulative hazard plot for *hb*. Quartiles: 8.5 (blue, dotted), 10.2 (magenta, solid), 12.7 (green, solid), 14.6 (red, dotted).

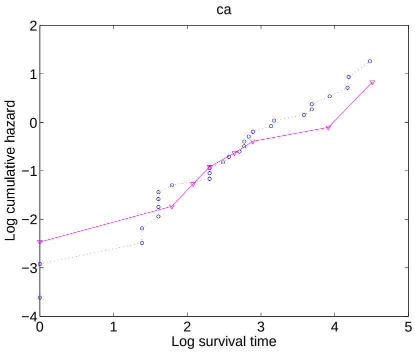

Figure 6.6: Log-cumulative hazard plot for *ca*. ≤ 10 (green, solid), > 10 (blue, dotted).

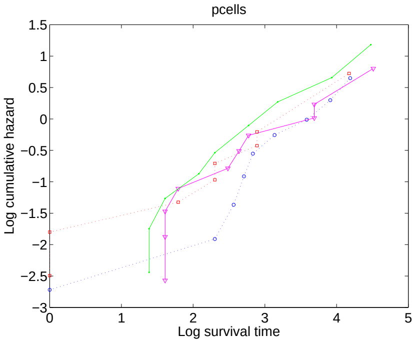

Figure 6.7: Log-cumulative hazard plot for *pcells*. Quartiles: 20.5 (blue, dotted), 33.0 (magenta, solid), 64.0 (green, solid), 100.0 (red, dotted).

it was removed. If the variable is included in the final model, the value in the third column is the HR, $\exp(\beta)$, defined as the change in hazard given a one unit increase from the mean value (continuous variables), or a shift in category (binary variables), e.g. from $sex = male$ to $sex = female$. Categorial variables are transformed into a set of binary variables, each including the reference category and one other category, e.g. $sex_1 = \{male, female\}$ and $sex_2 = \{male, hermaphrodite\}$. A HR > 1 corresponds to an increased hazard and vice versa.

The analysis shows that the level of blood urea nitrogen, *bun*, and haemoglobin, *hb*, are the only significant risk factors. A unit increase in *bun* increases the HR by a factor 1.02, while a unit increase in *hb* decreases the HR by a factor 0.87. Hence, a possible treatment should attempt to lower the level of blood urea nitrogen or increase the level of haemoglobin. A naive comparison of the hazard ratios shows that an increased level of haemoglobin is more profitable (pr. unit).

Variable	p-value	HR
age	0.56 (4)	
sex	0.53 (3)	
bun	< 0.01	1.02
ca	0.92 (1)	
hb	0.03	0.87
cells	0.81 (2)	
protein	0.13 (5)	

Table 6.1: p-values and HRs using stepwise selection in the MMP data set.

Age of patient in years, *age*, sex of patient, *sex*, serum calcium, *ca*, percentage of plasma cells in the bone marrow, *pcells*, and the Bence-Jones protein indicator, *protein*, are not significant predictors of the survival time using $\alpha = 0.05$.

After fitting the CPH model, we can use the Schoenfeld from Section 2.4.4.2 plots to validate the proportional hazards assumption. In Figure 6.8 - 6.14 we plot the scaled Schoenfeld residuals vs. survival times for each variable. For each variable we fit a linear model (shown on plot) and calculate 95% CIs for the slope presented in Table 6.2. Each CI includes the value zero, i.e. we cannot reject a linear model with slope zero using $\alpha = 0.05$, and we accept the proportional hazards assumption. One drawback of this method is that even though we cannot reject a linear model with slope zero, there could still be a higher order non-linear time dependency. Furthermore, as we have already stressed, we do not applaud the use of (artificial) significance levels. However, the method is a useful indicator of time dependence like the log-cumulative hazard plot.

Next, we apply our BMA method and include all possible models. With seven variables we have $2^7 = 128$ models to analyze and with no prior information available, we use a flat prior, i.e. we let all models be equally likely a priori. Results are presented in Table 6.3. The table shows the Top5 models in terms of PMP and for each model (row), included risk factors are indicated by a $\bullet$

The final model in stepwise selection is also the best model in terms of PMP, and we would use this model if we had performed *model selection*. However, the PMP is not more than 17%. So, given equal priors, we can explain less than 1/5

Variable	95% CI 1.0e-002 ×
age	-0.22 ; 0.29
sex	-4.48 ; 2.85
bun	-0.08 ; 0.03
ca	-1.81 ; 0.74
hb	-0.67 ; 0.50
cells	-0.07 ; 0.04
protein	-2.41 ; 5.78

Table 6.2: 95% CIs for the value of the slope fitting a linear model to Schoenfeld residuals for the variables in the MMP data set.

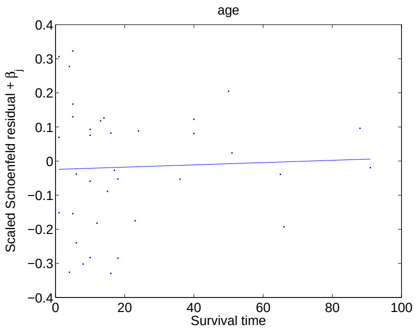

Figure 6.8: Schoenfeld plot for *age*.

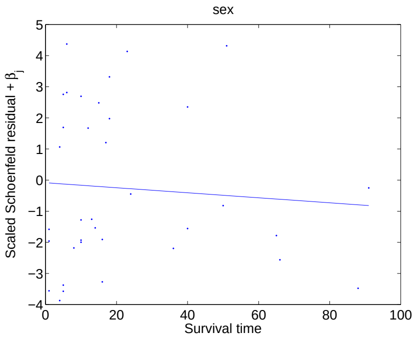

Figure 6.9: Schoenfeld plot for *sex*.

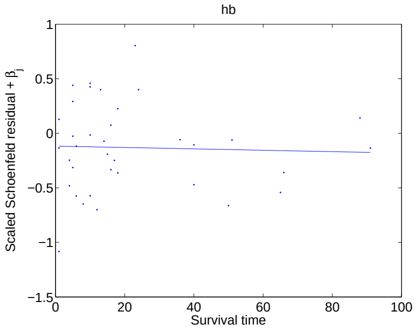

Figure 6.10: Schoenfeld plot for *hb*.

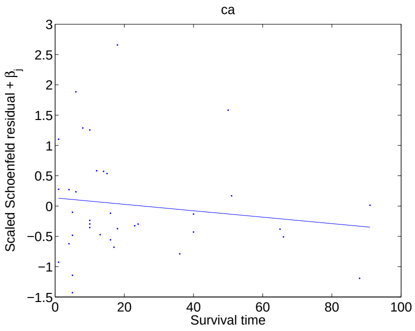

Figure 6.11: Schoenfeld plot for *ca*.

of the data with this model implying severe model uncertainty. In fact, the Top5 models were just assigned about half of the posterior probability mass. Thus, we expect an average model to be much better than the single "best" model at

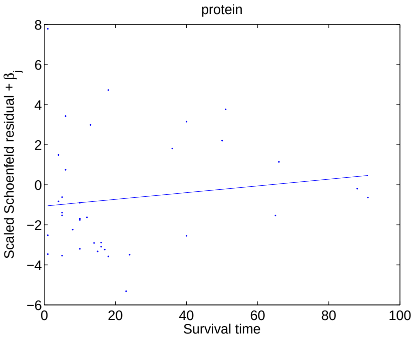

Figure 6.12: Schoenfeld plot for *protein*.

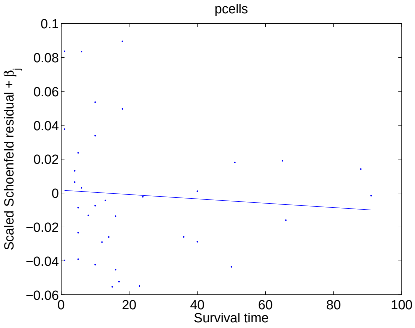

Figure 6.13: Schoenfeld plot for *pcells*.

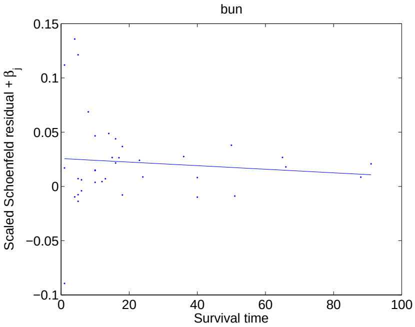

Figure 6.14: Schoenfeld plot for *bun*.

explaining the observed data, and hopefully also at predicting the survival time for new subjects.

We also note that the "most significant" variable in stepwise selection, *bun*, appears in all the Top5 models with 93.0 in PPP corresponding to positive evidence for an effect according to Table 3.1. In stepwise selection, *hb* is also significant, but appears only in three of the Top5 models with 55.7 in PPP implying just *weak* evidence for an effect.

The remaining variables all have PPPs corresponding to positive evidence *against* an effect and do not appear in the final stepwise selection model. Note that the variable with highest PPP, *protein*, was also the last variable to be excluded in the stepwise selection process.

The HR for *bun* corresponds to the HR in stepwise selection, while the HR for *hb* is 0.93 and thus closer to one (corresponding to no effect) compared to the HR of 0.87 in stepwise selection. We conclude that *hb* is indeed an explanatory variable, but the BMA estimate of the effect is more conservative due to the model uncertainty.

Model	age	sex	bun	ca	hb	cells	protein	PMP_A	PMP_O
1			●		●			.17	.20
2			●				●	.12	.14
3			●					.10	.12
4			●		●		●	.09	.11
5		●	●		●			.03	.04
PPP_O	10.3	10.5	96.9	9.6	55.1	10.0	42.1		
PPP_A	15.4	15.7	93.0	14.6	55.7	14.9	42.1		
HR_B			1.02		0.87				
HR_O	1.00	0.99	1.02	1.00	0.93	1.00	0.75		
HR_A	1.00	0.98	1.02	1.00	0.93	1.00	0.75		

Table 6.3: Top5 models using BMA on the MMP data set (**A**ll, **O**ccam, and **B**est). Total PMP for Top5: 0.52 (all) and 0.61 (Occam). 22 models included in Occam's window.

Table 6.3 also includes the results of a BMA analysis using Occam's window subset selection. The algorithm selects just 22 of 128 potential models, but the results are comparable with the results using all models. The Top5 models are identical, but the PMPs have increased slightly as a result of the reduced model domain. The HRs are also comparable. The PPPs, however, have decreased for *age*, *sex*, *ca*, and *pcells*, while the PPP for *bun* has increased from positive to strong evidence for an effect. Just *hb* and *protein* have preserved their posterior probabilities. The reduced model domain is now limited to models with PMP within range of the best model and will most likely strengthen the evidence for variables with high PPP and weaken those with low PPP. Variables in-between, *hb* and *protein*, will be more or less unaffected. If we use all available models, we are able to capture more of the model uncertainty and consequently more conservative parameter estimates. However, the main trends are easily captured using Occam's window subset selection. If we have many variables, subset selection is critical for obvious computational reasons.

Earlier, we used the final model from stepwise selection to calculate the scaled Schoenfeld residuals. We can also average over the set of BMA models to calculate the corresponding scaled Schoenfeld residuals of the average model, but we saw no indications of a violation.

In Table 6.4 we compare the p-values to the PPPs. Significant variables in stepwise selection are highlighted. We plot the PPPs (all models and Occam) against the (log) p-values in Figure 6.15 and Figure 6.16 respectively. Vertical lines indicate significance levels $\alpha = 0.05$, 0.01, and 0.001, while horizontal lines indicate the PPP levels in Table 3.1.

As the results from the stepwise selection and the BMA analysis are in agreement, significant variables ($\alpha = 0.05$) have PPPs above 50 and vice versa as indicated by the solid lines, but BMA does not label risk factors as in or out according to a more or less arbitrary significance criteria. As mentioned, the stepwise selection has a tendency to give overconfident estimates, and we cannot use the p-values to measure the evidence of an effect. Variables *bun* and *hb* are simple in; the rest are out. In BMA, however, we can comment on all risk factors and evaluate them relative to each other. At least the two methods agree on the best model, but BMA also revealed that several other models have significant posterior probability. In this experiment the evidence for an effect of *hb*, although above the 50% level, is not as strong as the evidence for *bun*. In fact, *hb* is close to the insignificant variable *protein* in terms of PPP. These aspects are not captured in stepwise selection.

Method	age	sex	bun	ca	hb	cells	protein
p-value	0.56	0.53	< **0.01**	0.92	**0.03**	0.81	0.13
PPP_O	10.3	10.5	96.9	9.6	55.1	10.0	42.1
PPP_A	15.4	15.7	93.0	14.6	55.7	14.9	42.1

Table 6.4: PPPs vs. p-values for variables in the MMP data set.

6.1.1 Predictive Performance

Stepwise selection can infer significant explanatory variables, but does not consider the predictive performance of the model. When a model has predictive power, it can be used to inform the patient of his expected lifetime and, if possible, how to increased the expected lifetime, e.g. with a change in lifestyle. The predictive power is also, as discussed earlier, a valuable tool for comparing competing models.

To evaluate the predictive power we randomly split the data set into a training set (70%) that we use to estimate the parameters, and a test set (remaining 30%) that we use for evaluation. Therefore the parameter estimates will be different from earlier, where we used the complete data set to estimate the parameters. We evaluate the models using the PPS and the predictive $\mathcal{Z}$-score from Section

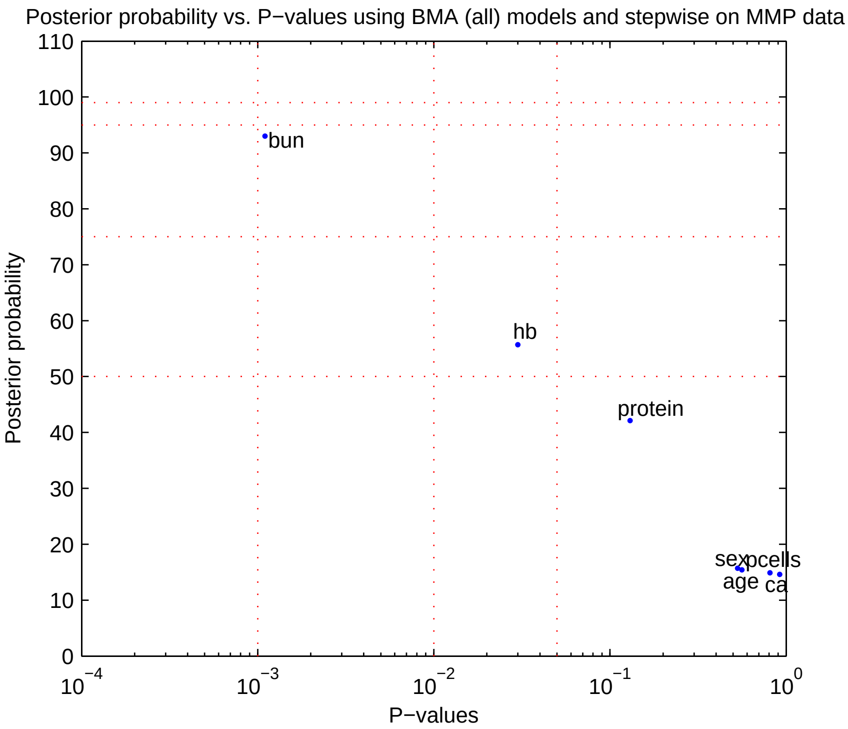

Figure 6.15: PPPs (BMA, all) vs. p-values (stepwise selection) for variables in the MMP data set.

3.3.2.1 and 3.3.2.2 respectively. To evaluate the latter score we pretend that the un-censored survival times in the test set are censored, and the actual size of the test set will therefore alternate depending un the number of subjects that are not censored. We average over 500 runs, and use the (log) median survival time in the evaluation of the predictive $\mathcal{Z}$-score (mean was comparable).

The PPP and HR for each variable are presented in Table 6.5. In stepwise selection, the PPP is the fraction of runs the variable appears in the final model, and the HR is $\exp(\bar{\hat{\boldsymbol{\beta}}})$, where $\bar{\hat{\boldsymbol{\beta}}}$ is the average (over runs) of the estimated coefficient vector. If risk factor j is not included in the final model in the i'th run, the estimated coefficient, $\hat{\beta}_{ji}$, will be zero. As expected, the "highly significant" variable *bun* is included in about 90% of the final models, while *hb* is in roughly 40%. As the BMA analysis showed data are much more uncertain about the effect of *bh*. With less data available for training, stepwise selection

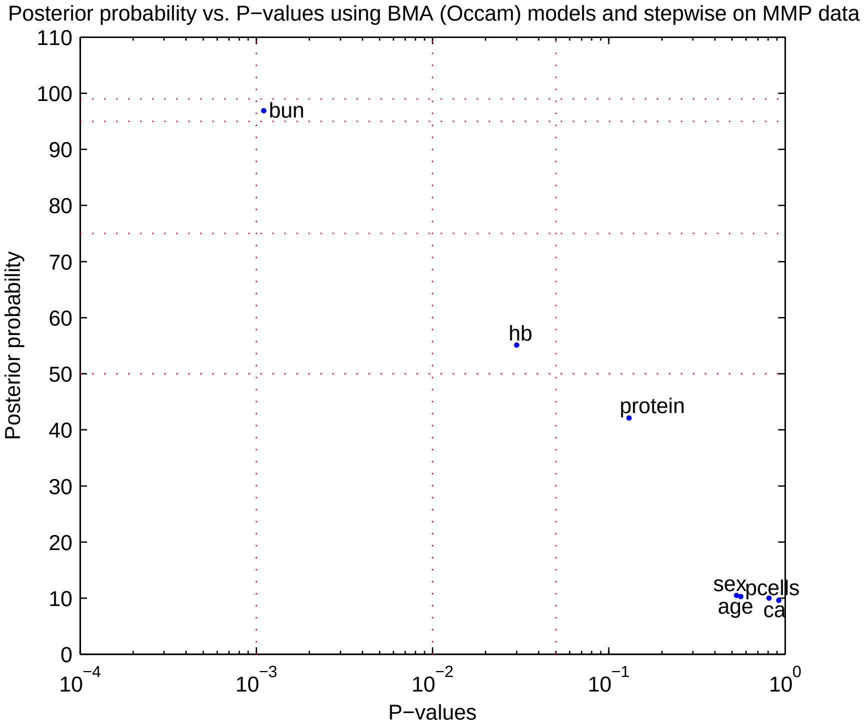

Figure 6.16: PPPs (BMA, Occam) vs. p-values (stepwise selection) for variables in the MMP data set.

does not include *hb* in more than 4 out of 10 runs although it was significant when we used all available data to estimate the model. Same argument applies to *protein* appearing in just 22% of the final models although it was "almost" significant. As expected, the remaining variables all have very low PPP. The HRs have changed accordingly and we note that the average stepwise selection differs from the other models with respect to the HRs for *sex*, *hb*, and especially *protein* reflecting the large differences in PPPs.

The BMA results on other hand are much more consistent, and the PPPs confirm our earlier findings. However, we note that the a large amount of the evidence for an effect, especially for *bun*, has been transferred to the remaining variables, because we have lost confidence in this "strong" variable given the reduced data set. This induces more radical HRs (moving away from the neutral value 1) for the "weaker" variables, *sex* and *protein*. We also see that the estimate of the HR for *hb* in the best model is much more conservative and in line with the other methods. This is because the reduced data set has introduced more model uncertainty implying smaller PMP for the best model and thus more conservative parameter estimates. Less data means less evidence to support the parameter estimates and the estimates of the posterior model probabilities. Remember, the assumption is that data are generated by a single model within our model domain. With unlimited data, this model will have $PMP = 1$. Increasing the size of the data set induces fewer models with high PMP, while less data induce more models with lower PMP. Using all available data, we included 28 models in Occam's window, now we include 32 on average. With less data, more models are able to explain the data "well enough" to be included in Occam's window.

	age	sex	bun	ca	hb	cells	protein
$\mathrm{PPP_S}$	2.4	2.8	89.2	1.6	39.8	0.8	22.0
$\mathrm{PPP_B}$	2.8	4.4	85.6	1.4	42.6	1.0	33.0
$\mathrm{PPP_O}$	17.2	18.8	78.2	15.4	49.7	15.3	38.5
$\mathrm{PPP_A}$	22.1	23.6	75.0	20.7	50.0	20.6	40.0
$\mathrm{HR_S}$	1.00	0.96	1.02	1.00	0.93	1.00	0.78
$\mathrm{HR_B}$	1.00	0.95	1.02	1.00	0.92	1.00	0.70
$\mathrm{HR_O}$	1.00	0.93	1.02	1.00	0.92	1.00	0.71
$\mathrm{HR_A}$	1.00	0.93	1.02	1.00	0.92	1.00	0.71

Table 6.5: PPPs and HRs using **S**tepwise selection and BMA (**A**ll, **O**ccam, and **B**est) on the MMP data set averaged over 500 runs. Mean number of models included in using Occam's window: 32.

To explore the predictive power we compute the mean of the PPS, the IC, and σ_{pred} in Table 6.6. In Table 6.7 we compare the methods with respect to the mean of the differences in PPS, IC, and σ_{pred}.

Method	PPS	IC	σ_{pred}
Stepwise	-19.9	0.87	1.7
$\mathrm{BMA_B}$	-20.0	0.86	1.7
$\mathrm{BMA_O}$	-19.4	0.93	1.6
$\mathrm{BMA_A}$	-19.3	0.93	1.6

Table 6.6: PPS, IC, and σ_{pred} using stepwise selection, BMA (**A**ll, **O**ccam, and **B**est) on the MMP data set averaged over 500 runs.

In the PPS column, the number in parenthesis is the increase in predictive

Method	PPS	IC	σ_{pred}
BMA_O – stepwise	0.47 (6.1%)	0.06	-0.12
BMA_A – stepwise	0.54 (6.9%)	0.06	-0.12
$\text{BMA}_\text{O} - \text{BMA}_\text{B}$	0.66 (8.2%)	0.08	-0.14
$\text{BMA}_\text{A} - \text{BMA}_\text{B}$	0.73 (9.1%)	0.08	-0.15
$\text{BMA}_\text{A} - \text{BMA}_\text{O}$	0.07 (0.7%)	0.00	0.00

Table 6.7: Difference in PPS, IC, and σ_{pred} using stepwise selection, BMA (**A**ll, **O**ccam, and **B**est) on the MMP data set averaged over 500 runs.

performance pr. event

$$\exp\left(\frac{\Delta PPS}{n_{cases}}\right) \tag{6.1}$$

Since PPS is a log score (transforming the product of predictive densities in the test set into a sum), we use *exp* to get a predictive performance score pr. event. As mentioned in Section 3.3.2.1, we only get non-zero contributions for failures (deaths) in the test set. In the i'th run, n^i_{cases} is the number of subjects failing in the test set. As we split the data randomly, this number may change in each run, so we use

$$\frac{1}{N}\sum_i \exp\left(\frac{\Delta PPS_i}{n^i_{cases}}\right) \tag{6.2}$$

where $N = \sum_i n^i_{cases}$ to compute the predictive performance pr. event. The results show that the BMA methods have more predictive power, indicated by a higher PPS, a higher IC, and a lower σ_{pred}. On average, BMA is 6-7% (vs. stepwise selection) and 8-9% (vs. best model) better pr. event, when we use the PPS as indicator of predictive power. We also see significant improvements in IC on the scale 6-8% which makes it obvious that model uncertainty is an important aspect of survival analysis. Note that although the improvements in PPS and IC seem to be of a similar order, the two scores are very different, and we cannot make a naive comparison.

A 95% CI on the predictive median survival time with $\sigma_{\text{pred}} = 1.6$ (BMA, all models/Occam) is $\bar{t} \pm 3.1$, i.e. we are able to predict the true survival time within a $\sim$ 6 month interval in more than 9 of 10 cases using BMA. With a predicted median survival time of, say 10 months, the predicted 95% CI is $[6.9; 13.1]$ months. The values of σ_{pred} (and the CI) should be viewed in light of the distribution of the true survival times presented in Table 5.4. As the table shows, half the subjects have survival times less than 14.5 (months), but with a minimum survival time of 1 month and a maximum survival time of 91 months (7.6 years), we find the predictive CIs acceptable. Using stepwise selection, the average CI is $\bar{t} \pm 3.4$, but although the average CI is wider, it does not include the true survival times in more than 87% of the time.

We also note that using all models is just slightly better than using Occam's window subset selection, even though we just include 32 of the 128 possible models (on average). The increased computational effort does not justify the 0.7% increase in predictive performance pr. event, and there is no measurable difference in terms of IC. Both methods clearly outperform stepwise selection showing that an average over (a subset) of models also improves the predictive power. We note that the best model has predictive scores close to the scores using stepwise selection, and at least in this case stepwise selection obtains results comparable with the results we would get using model selection, but we still see a significant gain in predictive power using an average over models.

In conclusion, all experiments indicate the importance of accounting for model uncertainty, even for a small data set. We improve the predictive power and the evaluation of the risk factors using BMA to compute a true probability to evaluate the evidence of an effect for each risk factor. We do not need an arbitrary significance level, factors are not "in'" or "out", and more data will only strengthen the PPP and PMP estimates rather than make all variables significant.

6.2 Copenhagen Stroke Study

In the next analysis we apply our methods to the COST data set. Of the 993 patients, only 552 have no missing values in any of the risk factors. Using stepwise selection we can choose between two strategies. The standard method (implement in statistical software packages) is to exclude all subjects with missing values in any of the considered variables to get a CC data set that we can use no matter which variables are in the model. However, this leaves us with just 55.6% of the original data.

Another approach is to begin with the CC data set. Then, in each step of the selection process, if the set of variables changes, we review the original data set and build a new CC data set, removing only subjects with missing values in one or more of the *remaining* variables. In this way we include the maximum number of subjects in each step of the algorithm. However, we could also argue that once we have removed a subject, it is not correct to let it re-enter the data set later in the selection process, as the variables that we decided to remove were evaluated using a data set where the subject was *not* included. We investigate how the data inclusion strategy affects the results of the analysis. In BMA we cannot alter the data set as we never exclude any variables. Otherwise, the analysis is comparable with the analysis on the multiple myeloma patients.

The results using stepwise selection and not allowing re-entry of data (column 2-3) respectively allowing re-entry (column 4-5) are shown in Table 6.8. Both methods select *age*, *sex*, *apo*, *odd*, *dm*, *cla*, and *sss* as significant variables using $\alpha = 0.05$. They also agree on the HRs for *age*, *sss*, and *cla*, while they are comparable for *sex* and *dm*. For *apo* and *odd*, the HRs decrease when we allow re-entry of data. If we look at the p-values we see that when we allow re-entry of data, almost all variables have the same or smaller p-value than when we do not allow re-entry of data. Variables af and *temp* are now significant, and the remaining variables have p-values much closer to the significance level. As shown, the results of stepwise selection depend heavily on the amount of available data. More data seem to imply more significant variables with smaller significance values, and also illustrates the importance of the significance level! In further analysis, and to compare with BMA, we use the CC results not allowing re-entry of data to avoid any discussions.

Next, we apply BMA using all possible models. With fourteen possible variables, we have $2^{14} = 16384$ models to analyze. Again, we do not use any prior information making all models equally likely a priori. The Top10 models are shown in Table 6.9.

As in the analysis of the multiple myeloma patients, the results show that we

	No re-entry of data		Re-entry of data	
Variable	p-value	HR	p-value	HR
age	<0.001	1.05	<0.001	1.05
sex	<0.01	1.37	<0.001	1.40
hyp	0.28 (4)		0.30 (3)	
ihd	0.64 (2)		0.51 (2)	
apo	<0.01	1.43	<0.01	1.30
odd	0.02	1.34	0.02	1.26
alco	0.60 (3)		0.17 (4)	
dm	0.04	1.30	<0.01	1.32
smoke	0.15 (6)		0.06 (5)	
af	0.14 (5)		0.02	1.30
hemo	0.77 (1)		0.67 (1)	
cla	0.02	0.72	<0.01	0.72
temp	0.10 (7)		<0.01	1.24
sss	<0.001	0.97	<0.001	0.97

Table 6.8: p-values and HRs using stepwise selection in the COST data set. Column 2-3 (no re-entry), column 4-6 (re-entry).

have a large amount of model uncertainty. The Top10 models account for just 37% of the posterior model probability, and the best model has a PMP of just 0.08! This time, the final model in stepwise selection is *not* the model with the highest PMP, but is ranked 10'th with a PMP as low as 0.02, and we expect that using an average model is much better than using stepwise selection or the best (in terms of PMP) model alone.

The variables *age*, *sex*, and *sss* appear in all Top10 models and have high PPPs as shown in Table 3.1. The data indicate positive evidence for an effect of *sex* and very strong evidence for an effect of *age* and *sss*. These three variables all have p-values < 0.01 in stepwise selection, and the estimated HRs shown in Table 6.11 are almost identical, except for *sex*, where the BMA estimate is a little more conservative, and we feel confident that these variables are important risk factors with reliable HR estimates.

Variables *apo*, *odd*, *dm*, and *cla* are also significant risk factors in stepwise selection, but in the BMA analysis we have just weak evidence for an effect of *apo*, *odd*, and *cla*, while there is positive evidence *against* an effect of *dm*! This also affects the estimate of the HRs which are much more conservative (closer to 1) when we use BMA.

Variables *af* and *temp*, who were significant in stepwise selection allowing re-entry of data, have just 11.5 and 18.2 in PPP. These values correspond to positive evidence against an effect and are far from the positive evidence for an effect level. Hence, we are confident that af and $temp$ are *not* important explanatory variables.

The remaining variables all have PPPs corresponding to positive evidence against an effect in agreement with the exclusion of these variables in stepwise selection, regardless of whether or not we allow re-entry of data.

The results using Occam's window to select a subset of just 47 of the possible 16384 models are also presented in Table 6.9-6.11. As in the MMP analysis, results using Occam's window subset selection are comparable with the results using all models. The Top10 models are identical, but the PMPs have increased, and the total PMP for Top10 is 0.58. The HRs are also comparable, with slightly increased/decreased HR for variables with high/low PPP. The PPP values have changed such that variables with low PPP (hyp, ihd, $alco$, dm, $smoke$, af, $hemo$, and $temp$) have decreased, and variables with high PPP (age, sex, apo, and sss) have increased as a result of the reduced model domain, and sex has even moved from positive to strong evidence for an effect. The PPPs for odd and cla were close to 0.5 and remain so.

Again, when we use all models in the BMA analysis, the evaluation of such a large number of models induce more conservative estimates of the PPPs and the HRs, but still Occam's window subset selection is able to identify the very few important models that account for the majority of the PMP mass and give reliable estimates of the PPPs and the HRs using much less computational resources. In fact, we have reduced the number of models with more than a factor 348!

In Table 6.10 we compare p-values and PPPs for each variable. Significant variables in stepwise selection are highlighted. PPPs from the BMA analysis (all models and Occam) are plotted against the (log) p-values in Figure 6.17 and Figure 6.18 respectively. Vertical lines indicate the common significance levels $\alpha = 0.05$, 0.01, and 0.001, while horizontal lines indicate the PPP levels in Table 3.1.

Significant risk factors ($\alpha = 0.05$) should give confidence values above 50 and vice versa as indicated by the solid lines. However, as we have a large amount of model uncertainty, this is not the case. As already mentioned, dm has a significant p-value, but does not have a PPP above 50. Otherwise, all variables are "classified" alike and close to the "optimal" straight line from PPP=100 to p-value = 1.

Model	age	sex	hyp	ihd	apo	odd	alco	PMP_A	PMP_O
1	●	●			●	●		.08	.13
2	●	●			●	●		.06	.10
3	●	●			●			.05	.08
4	●	●			●	●		.03	.05
5	●	●			●			.03	.04
6	●	●				●		.03	.04
7	●	●			●			.03	.04
8	●	●				●		.02	.03
9	●	●			●	●		.02	.03
10	●	●			●	●		.02	.03
step	●	●			●	●		.02	.03
	dm	smoke	af	hemo	cla	temp	sss		
1							●		
2					●		●		
3					●		●		
4	●						●		
5	●				●		●		
6							●		
7							●		
8					●		●		
9						●	●		
10	●				●		●		
step	●				●		●		

Table 6.9: Top10 models using BMA (**A**ll and **O**ccam) on the COST data set. Total PMP for Top10: 0.37 (all) and 0.58 (Occam). 47 models included in Occam's window.

Method	age	sex	hyp	ihd	apo	odd	alco
p-value (step)	<**0.001**	<**0.01**	0.28	0.64	<**0.01**	**0.02**	0.60
PPP_O	100	95.0	2.8	3.5	82.5	61.7	0
PPP_A	100	87.8	9.2	9.5	74.3	57.3	4.9
	dm	smoke	af	hemo	cla	temp	sss
p-value (step)	**0.04**	0.15	0.14	0.77	**0.02**	0.10	<**0.001**
PPP_O	25.0	7.1	4.7	0	52.6	12.9	100
PPP_A	30.4	14.3	11.5	4.7	53.0	18.2	100

Table 6.10: PPPs vs. *p*-values for variables in the COST data set.

	age	sex	hyp	ihd	apo	odd	alco
HR_S	1.05	1.37	1.00	1.00	1.43	1.34	1.00
HR_B	1.05	1.41	1.00	1.00	1.43	1.43	1.00
HR_O	1.05	1.36	1.00	1.01	1.35	1.23	1.00
HR_A	1.05	1.33	1.01	1.01	1.31	1.21	1.00
	dm	smoke	af	hemo	cla	temp	sss
HR_S	1.30	1.00	1.00	1.00	0.72	1.00	0.97
HR_B	1.00	1.00	1.00	1.00	1.00	1.00	0.98
HR_O	1.07	1.01	1.01	1.00	0.83	1.02	0.98
HR_A	1.09	1.03	1.02	1.00	0.82	1.03	0.98

Table 6.11: HRs using **S**tepwise selection and BMA (**A**ll, **O**ccam, and **B**est) on the COST data set.

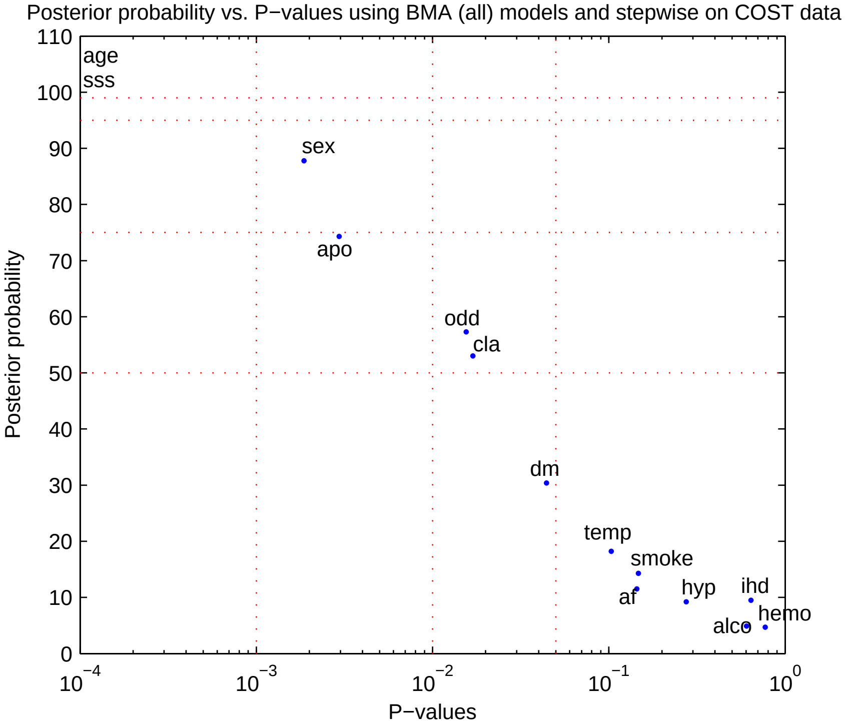

Figure 6.17: PPP (BMA, all models) vs. p-values (stepwise selection) for variables in the COST data set.

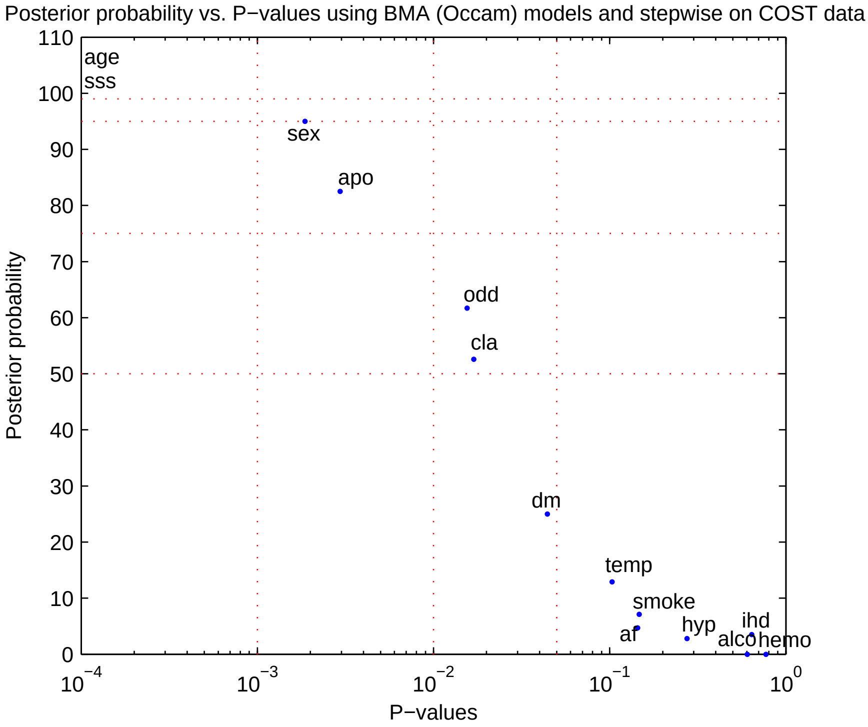

Figure 6.18: PPP (BMA, Occam) vs. p-values (stepwise selection) for variables in the COST data set.

6.2.1 Time Dependent Variables

We are aware that we could validate the proportional hazards assumption before applying any models - and we did. However, to compare the results of the two methods, we show the log-cumulative hazard plots together with the plots of the Schoenfeld residuals that we calculate *after* a model has been fitted.

6.2.1.1 Log-Cumulative Hazard Plots

In Figure 6.19 - 6.32 we show the log-cumulative hazard plots for each variable. For the continuous variables we use duo-deciles, [.2 .4 .6 .8 1.0]. Most plots indicate that the proportional hazard assumption is not violated, but for *sss*, *alco*, *smoke*, and *hemo* there are indications of a time dependency as the log-cumulative hazard for subjects in the first duo-decile (data points marked 'o' and connected by a dotted, blue line) do not seem to be proportional to the log-cumulative hazards for the remaining subjects. However, for *hemo* we have very few data point for $hemo = yes$, so it is quite difficult to conclude whether or not we have proportional hazards.

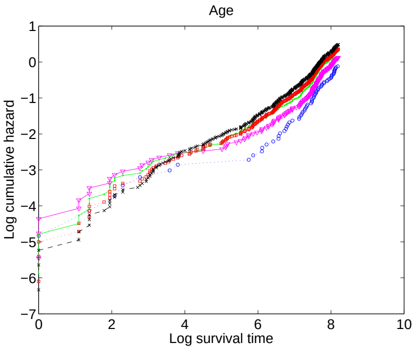

Figure 6.19: Log-cumulative hazard plot for *age*. Duo-deciles: 66 (blue, dotted), 73 (magenta, solid), 78 (green, solid), 83 (red, dotted), 98 (black, dashed).

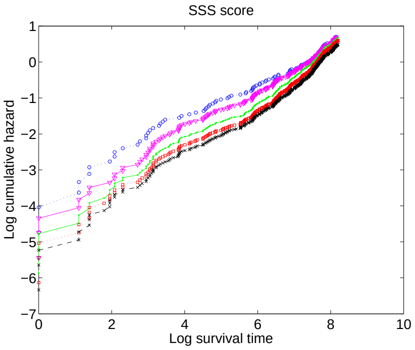

Figure 6.20: Log-cumulative hazard plot for *sss*. Duo-deciles: 26 (blue, dotted), 42 (magenta, solid), 49 (green, solid), 54 (red, dotted), 58 (black, dashed).

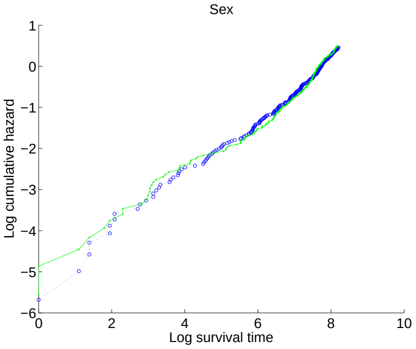

Figure 6.21: Log-cumulative hazard plot for *sex*. Male (green, solid), women (blue, dotted).

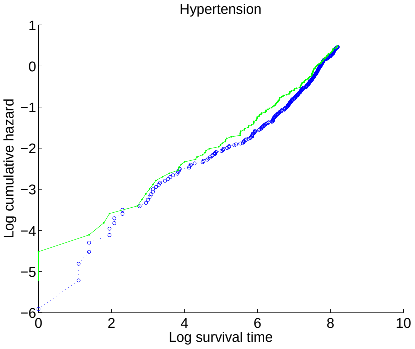

Figure 6.22: Log-cumulative hazard plot for *hyp*. Yes (green, solid), no (blue, dotted).

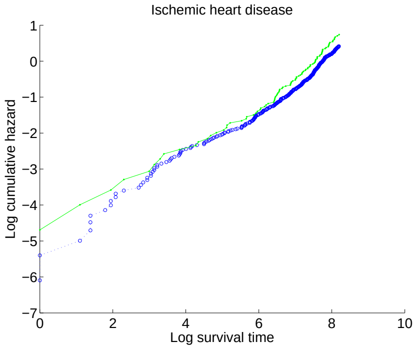

Figure 6.23: Log-cumulative hazard plot for *ihd*. Yes (green, solid), no (blue, dotted).

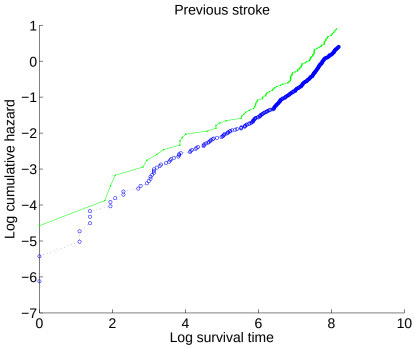

Figure 6.24: Log-cumulative hazard plot for *apo*. Yes (green, solid), no (blue, dotted).

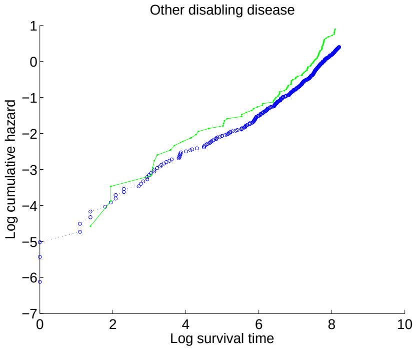

Figure 6.25: Log-cumulative hazard plot for *odd*. Yes (green, solid), no (blue, dotted).

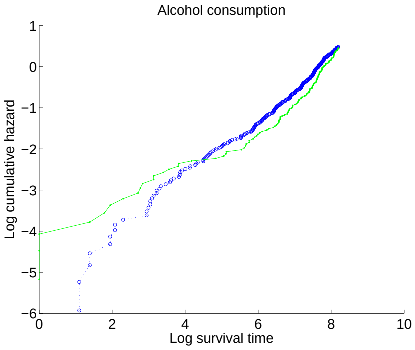

Figure 6.26: Log-cumulative hazard plot for *alco*. Yes (green, solid), no (blue, dotted).

6.2.1.2 Schoenfeld Plots

To get another opinion we plot the Schoenfeld residuals vs. the survival times for each variable after fitting a CPH model in Figure 6.33 - 6.46. For each variable we fit a linear model (shown on plot) and calculate 95% CIs for the slope presented in Table 6.12. For CIs including the value zero we cannot reject a linear model with slope zero using $\alpha = 0.05$, and we accept the proportional hazards assumption for those variables. As the log-cumulative hazard plots indicated, we believe the *sss* variable to be time dependent, since the CI does not include 0. Although the log-cumulative hazard plots did not indicate a time

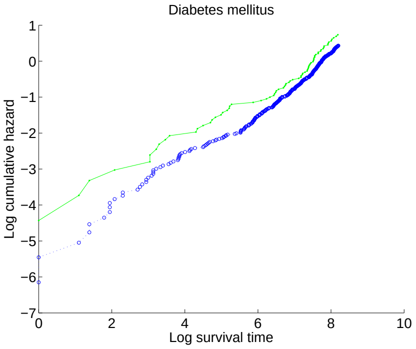

Figure 6.27: Log-cumulative hazard plot for *dm*. Yes (green, solid), no (blue, dotted).

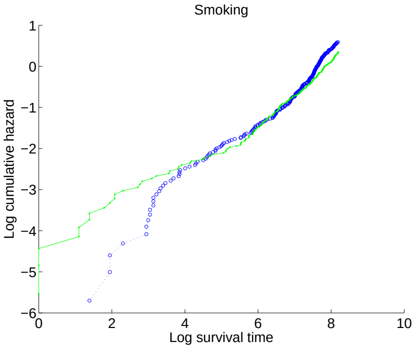

Figure 6.28: Log-cumulative hazard plot for *smoke*. Yes (green, solid), no (blue, dotted).

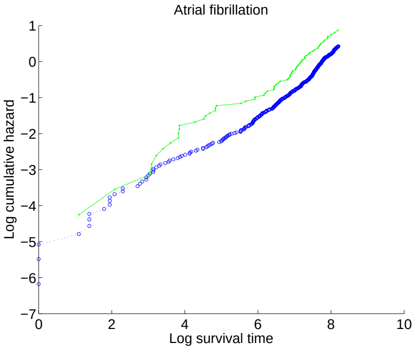

Figure 6.29: Log-cumulative hazard plot for *af*. Yes (green, solid), no (blue, dotted).

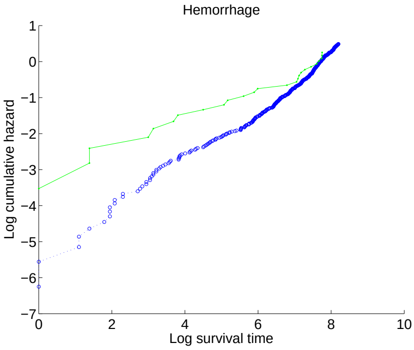

Figure 6.30: Log-cumulative hazard plot for *hemo*. Yes (green, solid), no (blue, dotted).

dependency for *alco* and *smoke*, the Schoenfeld residuals suggest otherwise, and we include a time dependent term for *alco* and *smoke*. As we expect subjects to quit drinking and smoking once they have experienced a stroke, it seems plausible that the effect of *alco* and *smoke* might change (decrease) over time. For *sss* we might expect that the stroke severeness has a great influence on the short-term survival, while the effect on the long-term survival is decreasing, as also reported in Andersen et al. (2006c) For *hemo*, the Schoenfeld residuals can not reject a linear model with slope, and we do not include a time dependent term for *hemo*.

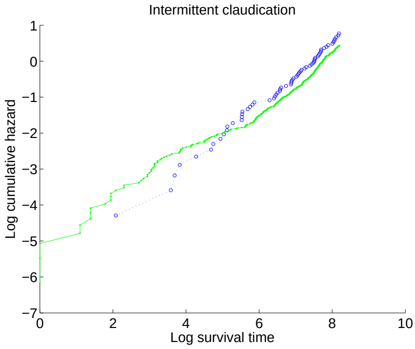

Figure 6.31: Log-cumulative hazard plot for *cla*. Yes (green, solid), no (blue, dotted).

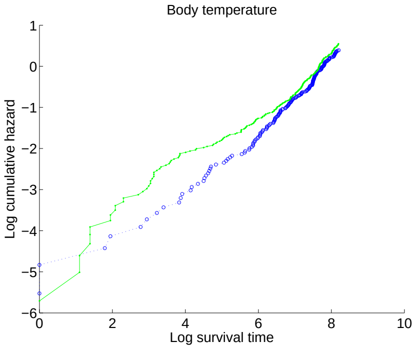

Figure 6.32: Log-cumulative hazard plot for *temp*. $< 37.0°$ C (green, solid), $\geq 37.0°$ C (blue, dotted).

Variable	95% CI, 1.0e-003 ×
age	-0.0179 ; 0.0041
sex	-0.2615 ; 0.1711
hyp	-0.3904 ; 0.0163
ihd	-0.3313 ; 0.1779
apo	-0.1402 ; 0.3450
odd	-0.0514 ; 0.4666
alco	**0.0410 ; 0.4890**
dm	-0.3858 ; 0.1522
smoke	**-0.5219 ; -0.0839**
af	-0.3892 ; 0.1613
hemo	-0.7959 ; 0.0303
sss	**0.0092 ; 0.0236**
cla	-0.2700 ; 0.3037
temp	-0.2591 ; 0.1161

Table 6.12: 95% CIs for the value of the slope fitting a linear model to Schoenfeld residuals for variables in the COST data set.

6.2.1.3 Including a Time Dependent Variable

To include time dependent variables we distinguish between *internal* and *external* variables. Internal variables relate to a particular subject, and are measured while the subject is alive, e.g. blood pressure. External variables can be measured without the subject being alive, and their values at any future time are sometimes known in advance, e.g. the subject's age or gender. It could also be

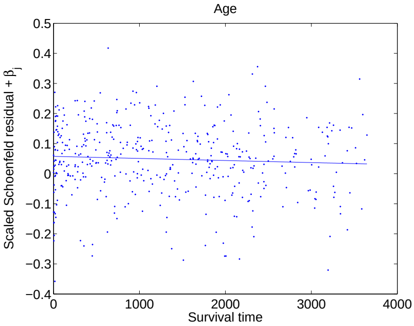

Figure 6.33: Schoenfeld plot for *age*.

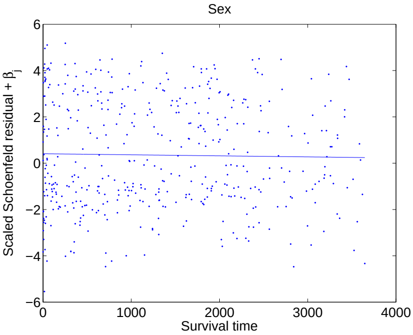

Figure 6.34: Schoenfeld plot for *sex*.

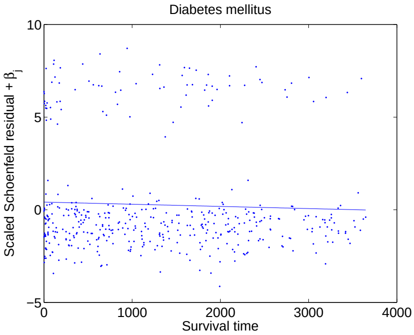

Figure 6.35: Schoenfeld plot for *dm*.

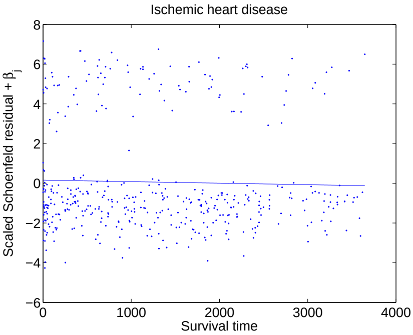

Figure 6.36: Schoenfeld plot for *ihd*.

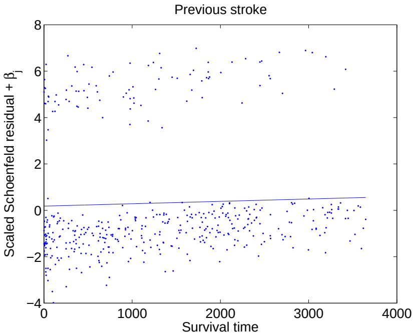

Figure 6.37: Schoenfeld plot for *apo*.

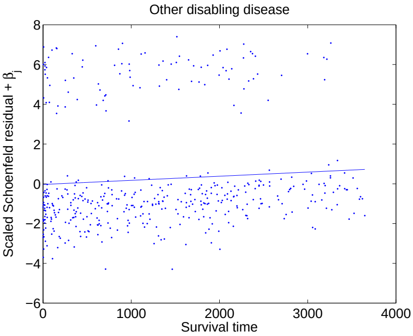

Figure 6.38: Schoenfeld plot for *odd*.

a variable that is not subject related, e.g. air temperature.

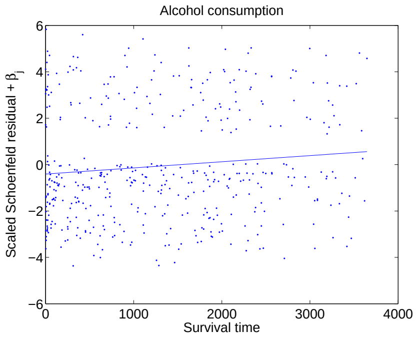

Figure 6.39: Schoenfeld plot for *alco*.

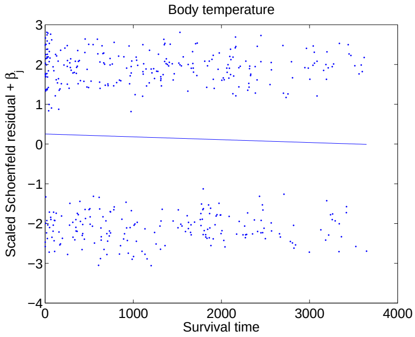

Figure 6.40: Schoenfeld plot for *temp*.

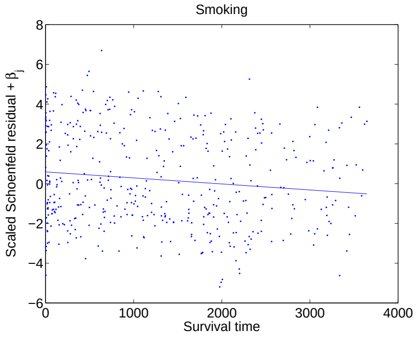

Figure 6.41: Schoenfeld plot for *smoke*.

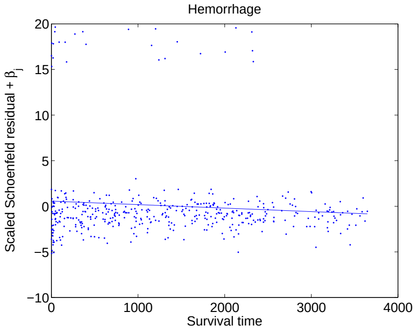

Figure 6.42: Schoenfeld plot for *hemo*.

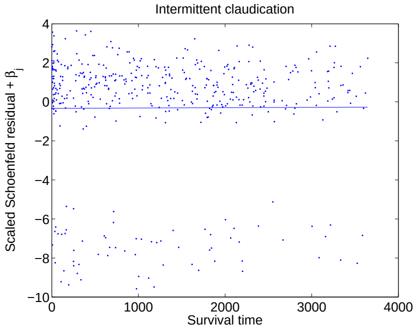

Figure 6.43: Schoenfeld plot for *cla*.

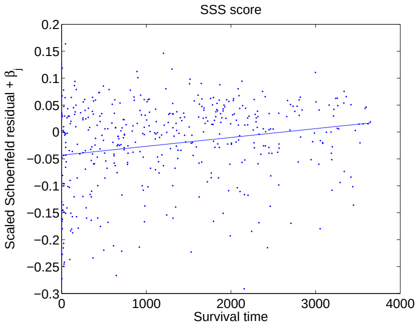

Figure 6.44: Schoenfeld plot for *sss*.

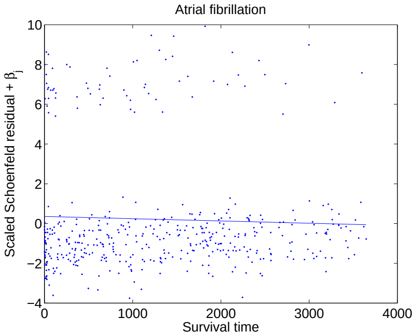

Figure 6.45: Schoenfeld plot for af.

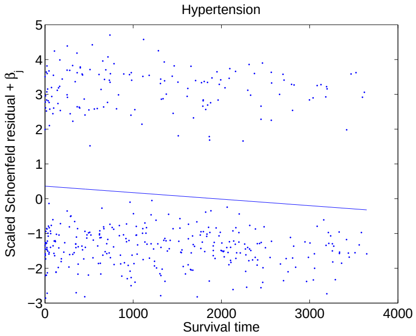

Figure 6.46: Schoenfeld plot for hyp.

We also have time dependent variables if the *coefficient* of a time constant variable varies with time, e.g. the sss variable is constant over time (measured on admittance), but is suspected to have a time dependent coefficient. As explained in Section 2.4, the coefficient of a variable in the CPH model is a log-hazard ratio, i.e. the hazard ratio is constant over time. If this ratio varies with time, the coefficient is a time varying coefficient, and we refer to such a model as a *Cox Regression (CR)* model.

Suppose a variable Z_j with coefficient β_j is a linear function of time, t. Then we can write this term as $\beta_j z_j t$ or $\beta_j Z_j(t)$, where $Z_j(t) = Z_j t$ is a time dependent variable. If the model has a variable Z_j with time varying coefficient $\beta_j(t)$, the model term is $\beta_j(t) Z_j$ or $\beta_j Z_j(t)$, i.e. a time varying coefficient can be expressed as a time dependent variable with a constant coefficient.

According to (2.69) the hazard of death at time t for the i'th subject is

$$h(t|\boldsymbol{z}_i) = h_0(t)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}_i) \tag{6.3}$$

$$= h_0(t)\exp\left(\sum_{j=1}^{p}\beta_j z_{ji}\right) \tag{6.4}$$

If some of the variables are time dependent we write $z_{ji}(t)$ for the j'th variable at time t for the i'th subject. The hazard of death at time t for the i'th subject in the CR model is then

$$h(t|\boldsymbol{z}_i) = h_0(t)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}_i) = h_0(t)\exp\left(\sum_{j=1}^{p}\beta_j z_{ji}(t)\right) \tag{6.5}$$

The baseline hazard, $h_0(t)$, is the hazard for an individual where all variables are zero (have baseline values) at the time origin and remain so throughout time. The ratio of hazards at time t for the r'th and the s'th individual is

$$\frac{h(t|\boldsymbol{z}_r)}{h(t|\boldsymbol{z}_s)} = \exp\left\{\beta_1\left[z_{1r}(t) - z_{1s}(t)\right] + \ldots + \beta_p\left[z_{pr}(t) - z_{ps}(t)\right]\right\} \tag{6.6}$$

and we can interpret the coefficient $\beta_j, j = 1, \ldots, p$ as the log-hazard ratio for two subjects whose values of the j'th variable at time t differ by one unit, while the other $p - 1$ variables have the same value for the two subjects.

The LPL function from (2.56) can be generalized to include time dependent variables

$$\sum_{i=1}^{n} \delta_i \left\{ \sum_{j=1}^{p} \beta_j z_{ji}(t_i) - \log \sum_{l \in R(t_i} \exp\left(\sum_{j=1}^{p} \beta_j z_{jl}(t_i) \right) \right\} \tag{6.7}$$

where $R(t_i)$ is the risk set at time t_i, the failure time of the i'th subject, and δ_i is the failure indicator for the i'th subject.

As in the CPH model, we would like to maximize this expression with respect to the β parameters using (2.57), but we need to know the values of each variable at each failure time for all subjects in the risk set at time t_i. This is no problem for external variables with preordained values, but for external variables with values that are independent of the subjects, and for internal variables, it is a problem. If, for example, we measure the blood pressure on admittance and at regular intervals hereafter, the value of this variable is time dependent. When the i'th subject dies at time t_i, we need the value of the blood pressure variable for the i'th subject and all other subjects in the risk set at time t_i. If these blood pressure values have not been measured, we need to estimate them. Several ways of doing this approximation is described in (Collet, 2003).

Luckily, we have no such variables in the COST data set. All our variables are measured on admittance, and their values do not change over time, except for *age* which is an explicit (linear) function of time, and we know its value at any point in time.

Having fitted a CR model, we can estimate the cumulative baseline hazard function, $H_0(t)$, and the corresponding baseline survival function, $S_0(t)$. We use the results for the CPH model in (2.65) and (2.66) respectively, and modify them to include time dependent variables.

The Breslow estimate of the cumulative baseline hazard function is

$$\hat{H}_0(t) = \sum_{t_i \le t} \frac{m_i}{\sum_{l \in R(t_i)} \exp(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{z}_l(t))} \tag{6.8}$$

and the baseline survival function is

$$\hat{S}_0(t) = \exp\left[-\hat{H}_0(t)\right] = \prod_{t_i \le t} \left\{ \exp\left[- \frac{m_i}{\sum_{l \in R(t_i)} \exp(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{z}_l(t))} \right] \right\} \tag{6.9}$$

where $\boldsymbol{z}_l(t)$ are the risk factors for the l'th subject at time t, and m_i is the number of failures at the i'th ordered failure time, $t_i, i = 1, \ldots, r$.

The estimate of the survival function for a particular subject is difficult to estimate, because $S_i(t)$ cannot be expressed as a power of $S_0(t)$ as in (2.67) for the CPH model. Instead, we need to integrate over time because the values of the variables change over time. The estimated survival function for the i'th subject is then

$$\hat{S}_i(t) = \exp\left\{ - \int_0^t \exp\left(\sum_{j=1}^{p} \beta_j z_{ji}(u) \right) h_0(u) du \right\} \tag{6.10}$$

The survival function depends on the time-dependent variables over the interval from 0 to t, which could be future, unknown values. To handle this, we can estimate the conditional probability that a subject survives in a certain time interval using a method described in (Collet, 2003). Fortunately, we do not have this problem in the COST data set. However, since the COST is a long-term study, and our survival times are measured in days, we have quite large values for t. To avoid numerical problems we use $\log t$ to model the time-dependent variables.

In Table 6.13 and 6.14 we show the stepwise selection results, and the BMA using Occam's window subset selection results. Based on the results from the previous experiments, we did not perform an analysis using all models.

First, we note that although we include more variables, we actually include fewer models in the Occam window (37 vs. 47), and we have also increased the maximum and Top10 PMP from 0.13/0.58 to 0.16/0.63, i.e. the inclusion of time dependent variables has decreased the amount of model uncertainty. The reason is that we now include the $sss{*}t$ term in all models, making it possible to explain the data much better than earlier. Actually, all time dependent variables are significant in stepwise selection, while BMA suggests positive evidence *against* an effect for $alco * t$ with PPP=2.5 and $smoke * t$ with PPP=0. On the other hand, the PPP for $sss * t$ is 100, showing strong evidence *for* an effect. The HR is 1.0019 for a 100 unit increase in $sss * t$, and compared to the HR of 0.95 for sss alone, it implies that a higher sss value (baseline value is 0) decreases the relative hazard, but that the effect decreases slightly with time.

For *apo*, *dm*, and especially *odd*, the PPPs and HRs have increased, while they have decreased for all other risk factors, especially *cla*. This is as a result of the entry of $sss{*}t$ that we now believe is able to explain phenomena in the data that we earlier explained otherwise. This has also lead to changes for other variables, most significantly are the increased PPPs for *odd* and *dm*. Consequently, *cla* has moved from weak evidence *for* an effect, to positive evidence *against* an effect, while *odd* has moved from weak to positive evidence for an effect.

On the other hand, in stepwise selection we do not see any effect for *odd*, while the p-value for *cla* has just increased from 0.02 to 0.04, and we are not able to capture the consequences that $sss{*}t$ introduced. However, as already mentioned, stepwise selection also finds $alco * t$ and $smoke * t$ significant, with the effect that *alco* and *smoke* themselves, along with af, also become significant! It is needless to say that stepwise selection and BMA do not agree on the results, and again we have much more confidence in the BMA results, where we include all variables throughout the analysis, and calculate real probabilities of an effect for each variable. This makes it possible to measure the effect of including new variables, both in terms of model uncertainty and parameter uncertainty.

Method	age	sex	hyp	ihd	apo	odd
p-value (step) PPP_O	<0.001 100	<0.01 96.9	0.33 (3) 2.0	0.52 (2) 1.4	<0.01 77.9	0.02 75.1
	alco	dm	smoke	af	hemo	cla
p-value (step) PPP_O	0.02 0	0.12 (5) 28.9	0.03 5.7	<0.01 3.6	0.76 (1) 0	0.04 39.3
	temp	sss	alco*t	smoke*t	sss*t	
p-value (step) PPP_O	0.13 (4) 9.3	<0.001 100	0.01 2.5	0.03 0	<0.001 100	

Table 6.13: p-values and PPPs using stepwise selection and BMA (**O**ccam) on the COST data set with time dependent variables. Max. PMP: 0.16. Total PMP for Top10: 0.63. 37 models included in Occam's window.

To explore the predictive power, we randomly split the data set into a training set (90%) that we use to estimate the parameters, and a test set (remaining 10%) that we use for evaluation. We evaluate the mean of the PPS, the IC, and σ_pred averaged over 200 runs in Table 6.15. In Table 6.16 we compare the methods with respect to the mean of the differences in PPS, IC, and σ_pred. In the PPS column, the number in parenthesis is the increase in predictive performance pr. event

Again, the results show that BMA (Occam) has more predictive power indicated by higher PPS, higher IC, and lower σ_pred, and is on average more than 3-4% (vs. stepwise selection) and ~5% (vs. best model) better pr. event. BMA (Occam)

Method	age	sex	hyp	ihd	apo	odd	alco
HR_S	1.05	1.36			1.45	1.34	0.65
HR_B	1.05	1.41			1.42	1.45	
HR_O	1.05	1.38	1.00	1.00	1.32	1.31	1.00
	dm	smoke	af	hemo	cla	temp	sss
HR_S		1.33	1.62		1.35		0.96
HR_B							0.96
HR_O	1.09	1.01	1.01	1.00	1.15	1.02	0.96
	alco*t	smoke*t	sss*t				
HR_S	1.03	0.98	1.0019				
HR_B			1.0018				
HR_O	1.00	1.00	1.0017				

Table 6.14: HRs using **S**tepwise selection and BMA (**O**ccam and **B**est) on the COST data set with time dependent variables. Max. PMP: 0.16. Total PMP for Top10: 0.63. 37 models included in Occam's window. Hazard ratio for $xxx * t$ is pr. 100 unit increment.

Method	PPS	IC	σ_{pred}
Stepwise	-252.0	0.74	35.9
BMA_B	-253.0	0.73	36.1
BMA_O	-249.5	0.78	30.8

Table 6.15: PPS, IC, and σ_{pred} using stepwise selection, BMA (**O**ccam and **B**est) on the COST data set averaged over 200 runs.

Method	PPS	IC	σ_{pred}
BMA_O − stepwise	2.45 (3.6%)	0.04	-5.1
$BMA_O - BMA_B$	3.48 (5.2%)	0.05	-5.3

Table 6.16: Difference in PPS, IC, and σ_{pred} using stepwise selection, BMA (**O**ccam and **B**est) on the COST data set averaged over 200 runs.

produces 95% CIs that contain the true survival times in 78% of the runs. Although it is a reasonable sized database for a medical study, we have only used 90% of 863 subjects ($\sim$ 777) to train the algorithms. CIs that contain the true survival time in almost 4/5 of the time is quite acceptable. On average, BMA produces CIs that include the true survival times 3.6% and 5.2% more often compared to stepwise selection and the best model respectively.

A 95% CI on the predictive median survival time with $\sigma_{\text{pred}} = 30.8$ (BMA, Occam) is $\bar{t} \pm 60.3$. With a predicted median survival time of, say 1095 days (3 years), the predicted 95% CI is (on average) $[1034.7; 1155.3]$ days or $[2.8; 3.2]$ years, i.e. with 78% in IC, we are able to predict the survival time within a ~ 4 month interval in more than 3 of 4 cases using BMA. As shown in Table 5.7, half the subjects (in the complete data set!) have survival times less than 1259 (days), but with a minimum survival time of 0 (19 patients were dead on arrival or died shortly after) and a maximum survival time of 4262 days (11.7 years), the survival times are scattered on a very large interval, and we find the predictive CIs acceptable. Using stepwise selection, the average CI is $\bar{t} \pm 70.3$ corresponding a 4.5 month interval. Although the CIs are wider, they do not include the true survival times in more than 74% of the time.

Chapter 7

Estimation of Missing Values in the COST Data Set

So far, we have seen the effect of including model uncertainty. It is obvious that the model as well as the parameter uncertainty should decrease when we see more data. To increase the amount of available data we use three different approaches. First, we remove variables that are clearly not important explanatory variables, and then we use BNs and a semi-parametric approach to estimate the missing values of the remaining variables.

7.1 Increasing the Amount of Available Data by Removing Variables

In the COST data set, 552 patients have no missing values in any of the 14 remaining variables, while 441 patients have one or more missing values distributed as shown in Table 7.1.

In the last chapter we concluded that there was positive evidence against an effect for several of the risk factors, and for some of the variables the PPPs were close to zero. Hence, we see no harm in removing these variables, and by doing so we can increase the amount of available data significantly. To remove variables

	age	sex	hyp	ihd	apo	odd	alco
Missing cases	0	0	62	76	50	41	163
	dm	smoke	af	hemo	cla	temp	sss
Missing cases	50	174	11	188	113	133	7

Table 7.1: Number of missing values for each variable in the COST data set.

one at a time, we adopt the principle of stepwise selection, and remove the variable with lowest PPP, if the PPP is very low, i.e. in the area of $0 - 5\%$, and preferable has a lot of missing values. Since we combine BMA with principles of stepwise selection, we refer to this technique as *stepwise BMA* ;-)

Both *alco* and *hemo* have PPP=0, but since *hemo* has a larger number of missing values and no time dependent term, we decide to remove this risk factor first. The results of removing *hemo*, and increasing the size of the data set to 641 subjects are presented in Table 7.2 and 7.3.

Method	age	sex	hyp	ihd	apo	odd
p-value (step) PPP_O	<0.001 100	<0.001 96.9	0.65 (1) 0.5	0.29 (2) 13.0	<0.01 64.5	0.02 39.8
	alco	dm	smoke	af	cla	temp
p-value (step) PPP_O	0.13 (5) 0	0.02 39.0	0.21 (4) 1.1	0.06 (8) 15.0	<0.01 70.0	0.13 (7) 6.8
	sss	alco*t	smoke*t	sss*t		
p-value (step) PPP_O	<0.001 100	0.21 (6) 2.4	0.18 (3) 0	<0.001 100		

Table 7.2: p-values and PPPs from stepwise BMA with *hemo* removed.

Method	age	sex	hyp	ihd	apo	odd	alco	dm
HR_S	1.05	1.37			1.35	1.29		1.32
HR_B	1.05	1.35			1.36			
HR_O	1.05	1.35	1.00	1.03	1.22	1.12	1.00	1.12
	smoke	af	cla	temp	sss	alco*t	smoke*t	sss*t
HR_S			1.39		0.95			1.0020
HR_B			0.68		0.95			1.0019
HR_O	1.00	1.04	1.30	1.01	0.95	1.0002	1.0000	1.0019

Table 7.3: HRs from stepwise BMA with *hemo* removed. Max. PMP: 0.09. Total PMP for Top10: 0.47. 62 models included in Occam's window. HR for $xxx * t$ is pr. 100 unit increment.

Now, since *alco* still has PPP=0, we decide to remove *alco* as the next variable. The results of removing *alco* (and $alco * t$), and increasing the size of the data set to 655 subjects are presented in Table 7.4 and 7.5.

Method	age	sex	hyp	ihd	apo
p-value (step)	<0.001	<0.001	0.62 (2)	0.46 (3)	<0.01
PPP_O	100	97.1	0.5	6.3	72.3
	odd	dm	smoke	af	cla
p-value (step)	0.02	<0.01	0.27 (4)	0.11 (5)	<0.01
PPP_O	43.3	59.1	0.8	7.9	69.3
	temp	sss	smoke*t	sss*t	
p-value (step)	0.10 (6)	<0.001	0.89 (1)	<0.001	
PPP_O	11.7	100	3.4	100	

Table 7.4: p-values and PPPs from stepwise BMA with *alco* removed.

Method	age	sex	hyp	ihd	apo	odd	dm
HR_S	1.05	1.36			1.36	1.29	1.35
HR_B	1.05	1.34			1.36		1.37
HR_O	1.05	1.35	1.00	1.01	1.25	1.13	1.21
	smoke	af	cla	temp	sss	smoke*t	sss*t
HR_S			1.39		0.95		1.0019
HR_B			0.69		0.95		1.0019
HR_O	1.00	1.02	1.30	1.02	0.95	1.00	1.0019

Table 7.5: HRs from stepwise BMA with *alco* removed. Max. PMP: 0.10. Total PMP for Top10: 0.55. 54 models included in Occam's window. HR for $xxx * t$ is pr. 100 unit increment.

Next, *hyp* is actually the risk factor with lowest PPP, but since the PPP for *smoke* is just marginally higher, and we have 174 missing values for *smoke* versus 62 for *hyp*, we decide to remove *smoke*. This will also have the positive side-effect of removing $smoke{*}t$. The results of removing *smoke* (and $smoke{*}t$), and increasing the size of the data set to 725 subjects are presented in Table 7.6 and 7.7.

This time we remove *hyp*, since *hyp* and *ihd* have a comparable number of missing subjects. Furthermore, we expect *ihd* to be removed next, since BMA has shown quite consistent results so far. Removing a handful of subjects will probably not alter the results for *ihd* markedly. The results of removing *hyp*, and increasing the size of the data set to 730 subjects are presented in Table 7.8 and 7.9.

Method	age	sex	hyp	ihd	apo	odd
p-value (step)	<0.001	<0.001	0.46 (2)	0.69 (1)	<0.01	0.02
PPP$_{\mathrm{O}}$	100	100	0.6	0.7	78.8	49.0
	dm	af	cla	temp	sss	sss*t
p-value (step)	<0.01	0.05	<0.01	0.03	<0.001	<0.001
PPP$_{\mathrm{O}}$	72.2	16.7	64.7	31.7	100	100

Table 7.6: p-values and PPPs from stepwise BMA with *smoke* removed.

Method	age	sex	hyp	ihd	apo	odd
HR$_{\mathrm{S}}$	1.05	1.42			1.34	1.27
HR$_{\mathrm{B}}$	1.05	1.36			1.35	
HR$_{\mathrm{O}}$	1.05	1.39	1.00	1.00	1.27	1.14
	dm	af	cla	temp	sss	sss*t
HR$_{\mathrm{S}}$	1.36	1.25	1.37	1.21	0.95	1.0019
HR$_{\mathrm{B}}$	1.39		0.69		0.95	1.0019
HR$_{\mathrm{O}}$	1.26	1.04	1.25	1.06	0.95	1.0019

Table 7.7: HRs from stepwise BMA with *smoke* removed. Max. PMP: 0.11. Total PMP for Top10: 0.55. 46 models included in Occam's window. HR for $sss * t$ is pr. 100 unit increment.

Method	age	sex	ihd	apo	odd	dm
p-value (step)	<0.001	<0.001	0.77 (1)	<0.01	0.02	<0.01
PPP$_{\mathrm{O}}$	100	100	0.6	75.1	52.5	73.5
	dm	af	cla	temp	sss	sss*t
p-value (step)	0.05 (2)	0.01	0.03	<0.001	<0.001	
PPP$_{\mathrm{O}}$	15.7	59.9	36.4	100	100	

Table 7.8: p-values and PPPs from stepwise BMA with *hyp* removed.

Rightfully so, *ihd* still has very low PPP, and the results of removing *ihd*, and increasing the size of the data set to 742 subjects are presented in Table 7.10 and 7.11.

Now all PPPs are (significantly) different from zero, and we decide not to remove any more variables. We notice that throughout the selection process, the PPPs have changed significantly, while the p-values are more or less the same. In Table 7.12 and 7.13 we summarize the changes in p-values and PPPs for each of the remaining variables. In stepwise selection, only the final p-values for *dm* and *temp* (marked in bold) are noticeably different from their "starting" values,

Method	age	sex	ihd	apo	odd	dm
HR_{S}	1.05	1.40		1.36	1.27	1.36
HR_{B}	1.05	1.37		1.34		1.39
HR_{O}	1.05	1.40	1.00	1.25	1.15	1.27
	dm	af	cla	temp	sss	sss*t
HR_{S}		1.35	1.20	0.95	1.0019	
HR_{B}		0.70		0.95	1.0019	
HR_{O}	1.03	1.22	1.07	0.95	1.0019	

Table 7.9: HRs from stepwise BMA with *hyp* removed. Max. PMP: 0.09. Total PMP for Top10: 0.52. 49 models included in Occam's window. HR for $sss * t$ is pr. 100 unit increment.

Method	age	sex	apo	odd	dm
p-value (step)	<0.001	<0.001	<0.01	0.02	<0.01
$\mathrm{PPP}_{\mathrm{O}}$	100	100	73.5	56.9	78.5
	af	cla	temp	sss	sss*t
p-value (step)	0.03	0.02	0.03	<0.001	<0.001
$\mathrm{PPP}_{\mathrm{O}}$	26.5	51.9	33.4	100	100

Table 7.10: p-values and PPPs from stepwise BMA with *ihd* removed.

Method	age	sex	apo	odd	dm
HR_{S}	1.05	1.41	1.33	1.28	1.37
$\sigma_{\mathrm{HR,\ S}}$	~ 0	0.12	0.13	0.13	0.15
HR_{B}	1.05	1.40	1.35	1.34	1.38
HR_{O}	1.05	1.40	1.24	1.17	1.30
$\sigma_{\mathrm{HR,\ O}}$	~ 0	0.12	0.19	0.18	0.22
	af	cla	temp	sss	sss*t
HR_{S}	1.27	1.34	1.20	0.95	1.0019
$\sigma_{\mathrm{HR,\ S}}$	0.14	0.16	0.10	~ 0	~ 0
HR_{B}				0.95	1.0019
HR_{O}	1.07	1.18	1.06	0.95	1.0019
$\sigma_{\mathrm{HR,\ O}}$	0.13	0.22	0.11	~ 0	~ 0

Table 7.11: HRs from stepwise BMA with *ihd* removed. Max. PMP: 0.09. Total PMP for Top10: 0.49. 49 models included in Occam's window. HR for $sss * t$ is pr. 100 unit increment.

and all we have learned is that *dm* and *temp* are also significant explanatory risk factors.

age	sex	apo	odd	dm	af	cla	temp	sss	sss*t
0.001	0.01	0.01	0.02	**0.12**	0.01	0.04	**0.13**	0.001	0.001
0.001	0.001	0.01	0.02	0.02	0.06	0.01	0.13	0.001	0.001
0.001	0.001	0.01	0.02	0.01	*0.11*	0.01	0.10	0.001	0.001
0.001	0.001	0.01	0.02	0.01	0.05	0.01	0.03	0.001	0.001
0.001	0.001	0.01	0.02	0.01	0.05	0.01	0.03	0.001	0.001
0.001	0.001	0.01	0.02	**0.01**	0.03	0.02	**0.03**	0.001	0.001

Table 7.12: Change in p-values using stepwise BMA.

age	sex	apo	odd	dm	af	cla	temp	sss	sss*t
100	96.9	77.9	75.1	28.9	3.6	39.3	9.3	100	100
100	96.9	64.5	39.8	39.0	15.0	70.0	6.8	100	100
100	97.1	72.3	43.3	59.1	7.9	69.3	11.7	100	100
100	100	78.8	49.0	72.2	16.7	64.7	31.7	100	100
100	100	75.1	52.5	73.5	15.7	59.9	36.4	100	100
100	100	73.5	56.9	78.5	26.5	51.9	33.4	100	100

Table 7.13: Change in PPPs using stepwise BMA.

On the other hand, using BMA we constantly update the evidence of an effect for each variable, reflecting the changes in the data set as well as the variable set, and thus reflecting the parameter as well as the model uncertainty. Inspecting the values, we learn that the data show very strong evidence for an effect of *age*, *sex*, and *sss*. Although the PPPs for *sex* have increased a little throughout the selection process, we were, and remain, confident of an effect of these variables. Although the PPP for *apo* varied, the changes were within a 15% interval around the borderline between weak and positive confidence for an effect, and the extra data has not added significantly to our knowledge about an effect of *apo*.

On the other hand, we started out with positive evidence for an effect of *odd*, but the removal of *hemo* changed the "relative strength" of *odd* and *cla*. The extra data induced positive evidence against an effect of *odd*, while *cla* moved from positive evidence against an effect, to weak and almost positive evidence for an effect. This evened out in the end, however, and both *odd* and *cla* ended up with PPPs indicating (very) weak evidence for an effect. For *dm*, there was positive evidence *against* an effect when we used all variables, but the extra evidence and fewer variables has induced positive evidence *for* an effect. Especially the removal of *hemo*, *alco*, and *smoke* gave extra data that increased the evidence for an effect of *dm*.

The PPPs for af and *temp* were both very low when we included all variables and had little data available, but this changed throughout the selection process, and the extra data increased the evidence for an effect of both variables, but they ended up with PPPs around 30 still indicating positive evidence against an effect. Finally, the PPP for $sss * t$ started and remained at 100.

If we look at the model uncertainty in terms of the number of models included in Occam's window, the maximum PMP, and the Top10 PMP, the maximum PMP has been fairly stable around 10%, and the Top10 PMP increased from 0.47 to 0.55, when we removed *hemo*, but ended up at 0.49, i.e. about half of the posterior probability mass was assigned to 10 models at any stage. On the other hand, the number of models included in Occam's window decreased significantly from 62 to 49, i.e. that fewer models were within reasonable range of the best model in terms of PMP. Remembering that BMA assumes that data is generated by a single model within the model domain, it will assign full PMP to this model with unlimited data available and, eventually, fewer and fewer models will be included in Occam's window.

The important point is that all these aspects of the parameter and model uncertainty are not discovered in regular stepwise selection using p-values and significance levels. Finally, we calculate the standard deviation of the HRs

$$\sigma(\mathrm{HR}_j) = \sigma(\exp(\beta_j)) = \sqrt{\mathrm{V}(\exp(\beta_j))} \tag{7.1}$$

using the second-order Taylor expansion to approximate the variance of a function

$$\mathrm{V}[f(x)] \approx \left(\frac{\partial f(x)}{\partial x} \Big|_{x=\mathrm{E}(x)} \right)^2 \mathrm{V}(x) \tag{7.2}$$

where $f(x) = \exp(\beta_j)$, and we use (3.31) to calculate $\mathrm{V}(x) = \mathrm{V}(\beta_j)$ in BMA. The results are presented in Table 7.11, and for all variables except *age*, *sex*, *sss*, and *af*, the estimated variances using stepwise selection are smaller than using BMA. As explained in Section 3.3.1.4, the regression coefficient variance in BMA includes the model uncertainty. By ignoring the model uncertainty, stepwise selection underestimates the total uncertainty leading to overconfident parameter estimates.

7.2 Using Bayesian Networks to Estimate Missing Values

Although we have shown the advantages of (stepwise) BMA, we cannot use the information stored in subjects with missing values. With fewer variables we have limited the number of missing values, and with the removal of *hyp*, *ihd*, *alco*, *smoke* and *hemo*, we have 742 subjects with no missing values, and just 251 subjects with one or more missing values. Our next approach is to use BNs to estimate these values.

If we can make reliable estimates, we can increase the data set with another 17.5%. If we had included all 14 potential risk factors, only 552 subjects have no missing values. In that case we would have increased the size of the data set by 79.9% by estimating the missing values!

7.2.1 Estimation of Simulated Missing Values

Of these 251 subjects, no subjects have missing *age* values, and only 7 subjects have missing *sss* values (of which 5 were missing just the *sss* value and no other value). The *sss* values are probably missing because the patients were in a very bad condition and died shortly after admission (survival times were 0, 1, 1, 1, 4, 6, and 10 days after admission), making it impossible to record the complex SSS score. To avoid estimating a continuous variable, we simply replace the missing *sss* values with a mean value. However, we use the mean *sss* value of patients with survival times less than or equal to 10 (13.9), instead of the mean value for all patients (38.0), to reflect our assumption that missing *sss* values are related to short survival times.

However, as the number of DAGs is super-exponential in the number of variables, (Heckerman, 1995), 9 variables are still a lot. Furthermore, we would like to use information from the discarded variables to estimate the missing values of the remaining variables. To address this problem, we split it in two. First, we learn the structure and parameters of a network connecting the remaining variables with the restriction that the fully observed variables, *age*, *sex*, and now *sss*, do not have any incoming connections. In that way we limit the number of possible DAGs, and it also allows us to use a tabular distribution for *sex*, and continuous distributions other than the soft-max distribution, (Bishop, 2006), for *age* and *sss*. For the other variables we use the soft-max conditional probability distribution which allows both discrete and continuous parents. We refer to this network as the *blue network*, and the remaining variables are the *blue nodes*.

Next, we learn the structure and parameters of a network connecting the discarded variables with those remaining variables that have missing values. We apply the restriction that the discarded variables cannot have any incoming arcs to greatly limit the number of possible DAGs, but also to avoid using our limited amount of data to learn the parameters of the distributions connecting the discarded variables. Our main concern is the inference of the missing values of the remaining variables, and we simply estimate the tabular distributions of the discarded variables using a prior distribution over these variables. Since all variables in this network are discrete, all distributions are tabular distributions. We refer to this network as the *red network*, and the discarded variables are the *red nodes*.

Of the 251-5 = 246 subjects with missing values in one or more of the discrete, blue nodes, 136 (blue) subjects also have missing values in some of the red nodes, so we use the blue network to estimate the missing values of these subjects. This leaves 110 (red) subjects with no missing values in the red nodes, and we use a combination of the two networks to estimate the missing values.

First, we use the CC data set to see if we can estimate simulated missing values in the blue nodes. We split the CC data set into a training set (90%), and a test set (10%), where we pretend that some of the values are missing in the test set. We also use the test set for training, which is valid, as we do not use the (known) values of the missing values during training. We create separate test sets for the blue and the combined network, where we do not allow missing values for the red nodes in the test set for the combined blue network.

The missing values are selected at random, but we remove values for *apo*, *odd*, *dm*, af, *cla*, and *temp* only, as we are not interested in estimating variables that are always observed. We remove values such that the total fraction of missing values in the test set is 10%. This implies that some subjects can have more missing values than others.

7.2.1.1 Structure and Parameter Learning

We begin by learning the structure. Actually, we should learn the structure in each run using the new training (+ test) set, but as the estimated structure hardly ever changed, we decide to estimate the structure once and for all to save a lot of computation time. We can choose between the 8 different methods for learning the structure and the parameters (CC) outlined in Figure 7.1.

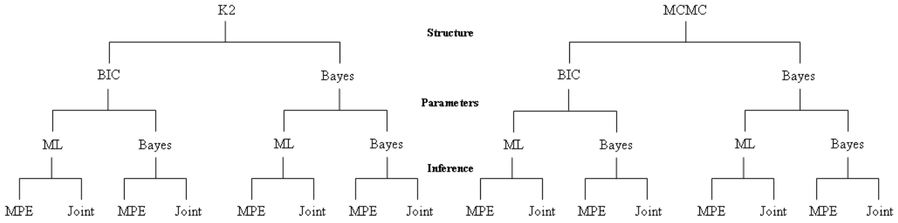

Figure 7.1: Methods available in BNT for structure/parameter learning (CC), and inference of missing values.

However, using the K2 algorithm, the final structure proved to depended heavily on the chosen ordering. Furthermore, the BNT version of the Bayesian scoring metric currently only works for tabular conditional probability distributions.

Using the MCMC algorithm with $N = 30000$ samples, and a burn-in of 500 samples to avoid that the results are influenced by the choice of initial structure, we get 30000 sampled DAGs distributed between 1000-1200 different DAGs for both the blue and the red network. Then we assign a weight, ω_k, to each of the K different sampled structures

$$\omega_k = \frac{freq(M_k)}{N}, \qquad k = 1, \ldots, K \tag{7.3}$$

defined as the frequency of the sampled structure divided by the total number of samples. The weights did not change significantly for $N > 25000$ samples. We saw no significant differences in parameter estimates, comparing point estimates (ML) to the full (Bayesian posterior) over parameters.

For each of the K structures we estimate the missing values, giving us K different estimates of the joint distribution of the missing values. The probability of subject i having missing value pattern j, given that we use structure k to estimate the missing values, is then

$$p(\boldsymbol{x}_{ijk}) = p(\boldsymbol{x}_{ijk}|M_k, \boldsymbol{\theta})p(\boldsymbol{\theta}|M_k), \qquad j = 1, \ldots, 2^J, \qquad k = 1, \ldots, K \tag{7.4}$$

where J is the number of missing binary variables for subject i.

Using MCMC we could take advantage of the entire sample of models (an approximation to the Bayesian posterior), using the approximated posterior to first sample a DAG, then learn the parameters, and finally estimate the missing values. This would give a new “sampled, CC data set”. Using a large number of samples, we could obtain a very large sampled data set with no missing values. However, this is beyond the scope of this thesis.

With missing values in our data set, we can also use the structural EM algorithm. The algorithm is able to learn the structure and parameters interchangeably using the in-complete data set with missing values (training + test set). However, as mentioned in Section 4.3.2, we cannot use the Bayesian scoring metric for this purpose. The structural EM algorithm also needs a starting point, i.e. an initial structure. Our approach is to use the MCMC sampled structures as initial structures. This gives us a new set of (EM) samples that we can use to compute augmented data sets, using the sampled structures to estimate the missing values. The pattern weights are identical to the weights in the MCMC samples.

This leaves us with 2 sets (2 blue and 2 red) of sampled structures and parameters, MCMC (BIC) and EM (BIC), using point estimates of the parameters. We combine these sets to give a combined MCMC (BIC), and a combined EM (BIC) network that we can use along with the blue networks to estimate the missing values. For the purpose of validating/comparing the structures and parameters, we begin with the simple MPE, allowing us to make a single estimate of each missing value that we can use to calculate the percentage of correctly estimated values in our simulation.

Using 500 runs we get the results in Table 7.14 - more or less independent of which method we use - when we use the MPE to estimate the missing values. However, when we inspected the estimated parameters (conditional probability distributions), we realized that in all structures, discrete nodes with missing values have a very strong preference for the value *no*, i.e. that a patient *does not* suffer from diabetes etc. This is reasonable, but also implies that the MPE will, in the vast majority of cases, be a set of *no*'s, which explains why there is no difference in estimation performance, and also explains why we get about 80% correctly estimated values, as this is roughly the percentage of missing values whose correct value is *no*! Hence, we cannot decide which method to use based on this experiment.

min	median	mean	max	std
0.79	0.84	0.84	0.88	0.02

Table 7.14: Simulation of missing values in the COST data set. Distribution of correctly estimated missing values using MPE.

Although we also expect the joint distribution to have a strong preference for the *no* pattern, all patterns are weighted and included in the augmented data set. Hence, using the joint distribution instead of the MPE to estimate the CPH model(s), we expect to get better estimates of the missing values. To compare the two sets/methods, we compare the estimated joint distributions. We use

500 runs, and in each run we compute the probability of the correct missing value pattern for each subject in the test set. Then we average over the test set giving us a "mean probability of the correct pattern score" for each method to average over the 500 runs. We rank the model that assigns higher probabilities to the correct pattern highest. The distribution of the scores for each model is listed in Table 7.15.

Structure	min	mean	max
$\mathrm{MCMC}_{\mathrm{BIC}}$	0.71	0.75	0.79
$\mathrm{EM}_{\mathrm{BIC}}$	0.77	0.81	0.85

Table 7.15: Simulation of missing values in the COST data set. Distribution of probabilities for the correct missing value patterns using the joint distribution.

The more complicated structural EM method performs better, taking advantage of the additional information stored in the subjects with missing values to obtain better structure and parameter estimates. Based on this experiment we keep the structural EM samples (BIC) to estimate the missing values in the COST data set. The MAP structures (most frequent samples) are shown in Figure 7.2-7.4.

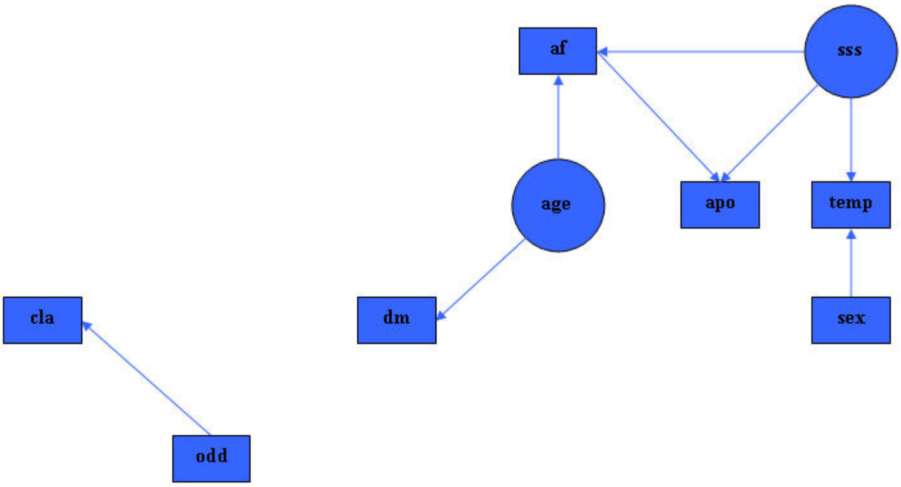

Figure 7.2: BN combining risk factors remaining after application of stepwise BMA (blue network).

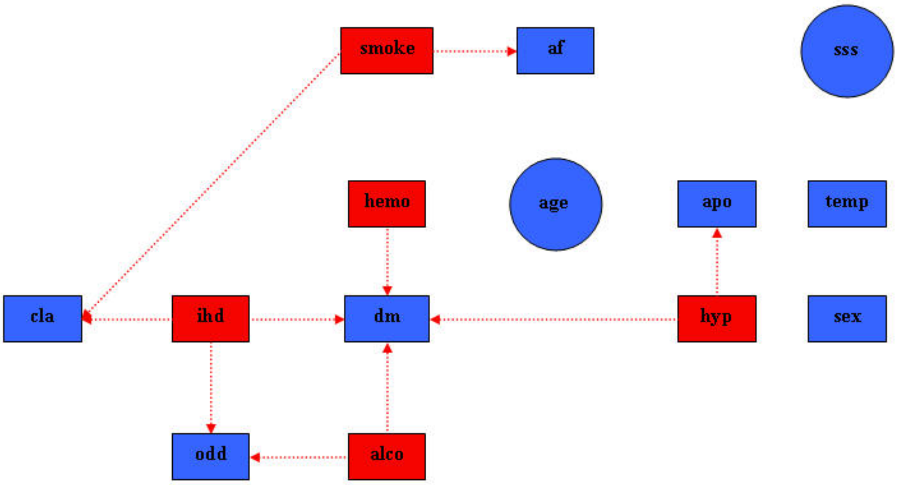

Figure 7.3: BN combining discarded risk factors with remaining risk factors after application of stepwise BMA (red network).

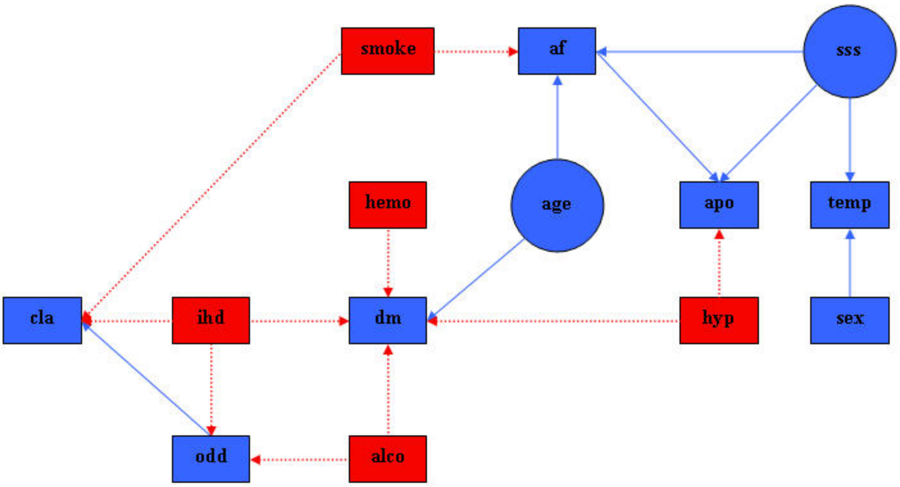

Figure 7.4: Combination of blue and red network.

7.2.2 Using an Augmented Data Set to Estimate CPH Models

To estimate the (true) missing values, we use the joint distribution of the missing values to get a data set where subject i is replaced with 2^{N_i} pseudo cases, each with a different combination of missing values, where N_i is the number of missing values for subject i. Furthermore, each subject has assigned a weight, 1 for a fully observed case and the joint posterior probability otherwise. In Table 7.16 and 7.17 we present the results using stepwise selection, and the results using BMA (Occam) on the augmented COST data set. To compare, we also include the CC results from the previous section using the same set of variables.

Method	age	sex	apo	odd	dm
p-value (step)	<0.001	<0.001	<0.01	0.02	<0.01
p-value (step, CC)	<0.001	<0.001	<0.01	0.02	<0.01
PPP_O	100	100	80.1	65.7	90.5
$\text{PPP}_\text{O,CC}$	100	100	73.5	56.9	78.5
	af	cla	temp	sss	sss*t
p-value (step)	0.03	0.02	0.03	<0.001	<0.001
p-value (step, CC)	0.03	0.02	0.03	<0.001	<0.001
PPP_O	31.6	57.3	40.0	100	100
$\text{PPP}_\text{O,CC}$	26.5	51.9	33.4	100	100

Table 7.16: p-values and PPPs using BNs to estimate missing values. Max. PMP: 0.15. Total PMP for Top10: 0.60. 39 models included in Occam's window. Hazard ratio for $sss * t$ is pr. 100 unit increment.

Although we have increased the size of the data set with more than 15%, we see no significant changes in the estimated p-values. Hence, according to stepwise selection, all variables are "as significant" as they were in the CC analysis. However, we do see slightly different HRs. For *apo*, *odd*, *dm*, af, *cla*, and *temp*, i.e. all variables that do not have $p < 0.001$, the HRs have increased. At the same time, the standard deviations of the HR estimates have not increased, in fact they have decreased for *dm* and af. All in all, the augmented data set has provided new information, leading to more accurate HR estimates that indicate a stronger influence on the survival time than we expected in the CC analysis.

In line with the HR rise in stepwise selection, the BMA results show increased PPPs for all variables that do not have full PPP. For *apo*, *odd*, and *dm*, the PPPs have increased about 10%, while the increase is about $5 - 7\%$ for af, *cla*, and *temp*. Hence, the estimated values provide information that confirms or increases the evidence for an effect of all the remaining variables. This illustrates

Method	age	sex	apo	odd	dm
HR_{S}	1.05	1.41	1.34	1.29	1.41
$\text{HR}_{\text{S,CC}}$	1.05	1.41	1.33	1.28	1.37
$\sigma_{\text{HR, S}}$	~ 0	0.12	0.13	0.13	0.14
$\sigma_{\text{HR, S, CC}}$	~ 0	0.12	0.13	0.13	0.15
HR_{O}	1.05	1.40	1.27	1.20	1.35
$\text{HR}_{\text{O,CC}}$	1.05	1.40	1.24	1.17	1.30
$\sigma_{\text{HR, O}}$	~ 0	0.11	0.18	0.16	0.19
$\sigma_{\text{HR, O, CC}}$	~ 0	0.12	0.19	0.18	0.22
	af	cla	temp	sss	sss*t
HR_{S}	1.30	1.36	1.21	0.95	1.0019
$\text{HR}_{\text{S,CC}}$	1.27	1.34	1.20	0.95	1.0019
$\sigma_{\text{HR, S}}$	0.13	0.16	0.10	~ 0	~ 0
$\sigma_{\text{HR, S, CC}}$	0.14	0.16	0.10	~ 0	~ 0
HR_{O}	1.08	1.21	1.08	0.95	1.0019
$\text{HR}_{\text{O,CC}}$	1.07	1.18	1.06	0.95	1.0019
$\sigma_{\text{HR, O}}$	0.11	0.20	0.09	~ 0	~ 0
$\sigma_{\text{HR, O, CC}}$	0.13	0.22	0.11	~ 0	~ 0

Table 7.17: HRs using BNs to estimate missing values. Max. PMP: 0.15. Total PMP for Top10: 0.60. 39 models included in Occam's window. Hazard ratio for $sss * t$ is pr. 100 unit increment.

one of the great advantages of BMA compared to stepwise selection: When we receive new evidence (more data), the information is reflected in updated PPPs. If the evidence is in favor of an effect, the PPP increases, and if the evidence speaks against an effect, the PPP decreases. In stepwise selection, new evidence may alter the p-values, but unless they cross the significance level, we will not note any difference. Furthermore, what if the p-value is 0.04999, and an extra data point makes it 0.05001? Then we have not seen much new evidence, but we have to change our classification from significant to not significant. Is that fair?

The changes in PPP are accompanied by increased HRs for most variables, some more than other, although there is no reason that the HRs should not remain unchanged, or even decrease. The PPP just reflects the *probability* of an effect - it does not say anything about the size or the sign of the effect. We also note that the HRs for age, sex, sss, and $sss * t$ have not changed, although the new evidence probably also provide evidence of an effect. However, since these variables already appear in all models included in Occam's window, the PPPs can not increase.

Finally, we note that the standard deviations of the HR estimates have all decreased or remain at ~ 0. Smaller standard deviations give smaller confidence intervals and imply more confident estimates - a positive consequence of the additional data. We also see an indication of reduced model uncertainty, as the maximum PMP has increased from 0.09 to 0.15, the total PMP for Top10 has increased from 0.49 to 0.60, and we just include 39 models compared to 49 in the CC analysis.

All in all, we conclude that BNs has proven a valuable tool for estimating the missing values in the COST data set, and that the models estimated using the augmented data set are different from the methods obtained using the CC data set. As we saw indications of more accurate parameter estimates, we also expect increased predictive performance if our missing value estimates are reliable. We compare the predictive performances in Section 7.4. However, the estimation of simulated missing values indicated that our estimates are fairly reliable.

BNs are also attractive in the sense that we can easily incorporate prior knowledge on the structure and/or the parameters, and we have access to a general applicable toolbox, BNT, that includes structure learning, parameter learning, and inference tools for complete as well as in-complete data sets. Hence, we do not need to spend valuable implementation time when we face a new problem. Instead, we can rely on the methods implemented in BNT to create multifarious models, and possibly combine them with other methods as we have done in this work using BMA and CPH models.

7.3 Using a Semi-Parametric Approach to Estimate Missing Values

Our final approach is to use a semi-parametric approach to estimate the missing values in the COST data set.

7.3.1 Estimation of Simulated Missing Values

First, however, we make a simulation with $n_{subjects} = 500$ to validate and investigate the modeling strategy in a simpler environment where we also know the ground truth. We generate survival times according to, (Bender et al., 2006)

$$t_i = \frac{-\log(u_i)}{\exp(h_0 + \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_i)} \tag{7.5}$$

where u_i is the i'th sample from a uniform distribution on the unit interval. We use a constant baseline hazard of $h_0 = -6$, and the samples are censored according to a Bernoulli distribution with $P_{succes} = 0.9$. We have three discrete variables, Z_1, Bernoulli distributed with $P_{succes} = 0.6$, and Z_2, Z_3 distributed according to logistic distributions

$$p(Z_2 = 1) = \frac{\exp(\alpha_{20} + \alpha_{21} Z_1)}{1 + \exp(\alpha_{20} + \alpha_{21} Z_1)} \tag{7.6}$$

and

$$p(Z_3 = 1) = \frac{\exp(\alpha_{30} + \alpha_{31} Z_1 + \alpha_{32} Z_2)}{1 + \exp(\alpha_{30} + \alpha_{31} Z_1 + \alpha_{32} Z_2)} \tag{7.7}$$

While Z_1 is always observed, the missingness of Z_2 and Z_3 denoted R_2 and R_3 is distributed according to logistic distributions

$$p(Z_2 = 1) = \frac{\exp(\phi_{20} + \phi_{21} Z_1)}{1 + \exp(\phi_{20} + \phi_{21} Z_1)} \tag{7.8}$$

and

$$p(Z_3 = 1) = \frac{\exp(\phi_{30} + \phi_{31} Z_1 + \phi_{32} Z_2)}{1 + \exp(\phi_{30} + \phi_{31} Z_1 + \phi_{32} Z_2)} \tag{7.9}$$

The true coefficients are shown in Table 7.18. The ϕ values imply a missing value percentage of around 20% for each variable with missing values. This gives a minimum of 20% and a maximum of 40% percent subjects with missing values depending on the overlap in each run.

We experiment using different α and ϕ distributions to see how they influence the estimates of the parameters and the missing values. As we do not know

α_{20}	α_{21}	α_{30}	α_{31}	α_{32}	ϕ_{20}	ϕ_{21}	ϕ_{30}	ϕ_{31}	ϕ_{32}	β_1	β_2	β_3
-1	1.2	-1	1	1.2	-2	1	-4	1	3	3	0.5	-1

Table 7.18: True model coefficients using a semi-parametric approach to estimate simulated missing values.

the true distributions in real life problems, it is interesting to see how a "false" choice of distribution affects the results. The simulations also serve as a mean of validating the implementations.

We set up 8 different experiments (referred to as "ex. 0" etc.).

Ex. 0) Using complete data set before deletion of values.

Ex. 1) Using CC data set after deletion of values.

Ex. 2) Using true α and ϕ distributions.

Hereafter we use the true distributions with the modifications listed below.

Ex. 3) Assume that Z_3 does not depend on Z_1.

Ex. 4) Assume that Z_3 does not depend on Z_2.

Ex. 5) Assume that Z_3 does not depend on either Z_1 nor Z_2.

Ex. 6) Assume missingness or Z_3 does not depend on Z_1.

Ex. 7) Assume missingness for Z_3 does not depend on Z_2.

Ex. 8) Assume missingness for Z_3 is MCAR.

In Table 7.19 - 7.21 we show the estimated values of the α, ϕ and β coefficients (with standard deviation in parenthesis) for each scenario, while Table 7.22 shows the percentage of correctly estimated missing values.

When we use the CC data set (ex. 1), our estimates are within the range of the true α values, but the results are not convincing. The estimates are greatly improved when we use the implemented method to estimate the missing values (ex. 2). We achieve estimates that are close to the true values and with standard deviations that are smaller than the standard deviations using the CC data set.

Experiment	$\alpha_{20} = -1$	$\alpha_{21} = 1.2$	$\alpha_{30} = -1$	$\alpha_{31} = 1$	$\alpha_{32} = 1.2$
0					
1	-1.06 (0.24)	1.64 (0.24)	-1.32 (0.20)	1.16 (0.25)	1.59 (0.24)
2	-0.97 (0.13)	1.05 (0.21)	-0.98 (0.13)	1.08 (0.19)	1.25 (0.12)
3	-1.11 (0.19)	1.09 (0.19)	-0.42 (0.14)		1.85 (0.23)
4	-1.17 (0.17)	1.12 (0.24)	-0.50 (0.11)	1.29 (0.14)	
5	-1.20 (0.11)	1.24 (0.20)	0.20 (0.04)		
6	-1.04 (0.12)	1.07 (0.19)	-0.95 (0.14)	1.18 (0.14)	1.59 (0.26)
7	-0.98 (0.12)	0.89 (0.19)	-0.98 (0.21)	1.25 (0.21)	1.50 (0.19)
8	-1.05 (0.10)	1.07 (0.15)	-1.19 (0.20)	1.49 (0.24)	1.64 (0.18)

Table 7.19: Estimated α coefficients using a semi-parametric approach to estimate simulated missing values.

Next, we investigate how a false choice of α distribution affects the results. In (ex. 3) we ignore the connection between Z_3 and the always observed variable Z_1. The consequence is that the estimates of Z_3's α coefficients are now worse than those obtained using the CC data set. The same effect is seen when we ignore the connection between Z_2 and Z_3 in (ex. 4). We also notice that the estimates of Z_2's α coefficients are slightly off in both (ex. 3) and (ex. 4). In (ex. 5) we assume that Z_3 is not connected to either Z_1 or Z_2, and the result is that all our estimates are inaccurate. Most of them worse than using the CC data set. We notice that since the true values of α_{31} and α_{32} are both positive, ignoring them causes the remaining α_{3x} values to increase to compensate for the missing link(s). For α_{30} this means moving towards zero, and even becoming positive in (ex. 5) where both α_{31} and α_{32} are missing.

In (ex. 6), (ex. 7) and (ex. 8) we see that the false ϕ distributions for Z_3 do no great harm to the α coefficient estimates for Z_2, but greatly affects the α coefficient estimates for Z_3.

The conclusion is that when we loose information on a given variable using incorrect distributions for either the value or the missingness of the variable, we obtain incorrect coefficient estimates. We also notice that incorrect α distributions, linking the values of the variables, seem to affect the estimates for both variables involved, whereas incorrect ϕ distributions, linking the missingness of variable i to the values of other variables, seem to do most harm to the α estimates for variable i. Furthermore, we see indications that the estimation error increases the more "incorrect" α and ϕ we use.

For the missingness distributions, ϕ, we cannot compare with the CC estimation, but we see that using the true distributions in (ex. 2) gives reliable coefficient

Scenario	$\phi_{20} = -2$	$\phi_{21} = 1$	$\phi_{30} = -4$	$\phi_{31} = 1$	$\phi_{32} = 3$
0					
1					
2	-2.05 (0.21)	1.01 (0.25)	-3.94 (0.31)	0.95 (0.28)	2.94 (0.25)
3	-2.00 (0.21)	0.98 (0.23)	-3.74 (0.29)	1.17 (0.30)	2.75 (0.27)
4	-2.06 (0.18)	1.09 (0.20)	-3.67 (0.34)	1.22 (0.32)	2.60 (0.29)
5	-2.07 (0.14)	1.13 (0.14)	-3.75 (0.51)	1.07 (0.21)	2.87 (0.46)
6	-2.27 (0.28)	1.30 (0.26)	-3.07 (0.25)		3.02 (0.28)
7	-2.06 (0.16)	1.15 (0.23)	-2.15 (0.18)	1.32 (0.24)	
8	-2.01 (0.13)	1.08 (0.16)			

Table 7.20: Estimated ϕ coefficients using a semi-parametric approach to estimate simulated missing values.

estimates. When we use incorrect α distributions in (ex. 3), (ex. 4) and (ex. 5), all ϕ estimates are affected, but most significantly for Z_3.

When we ignore the missingness link between Z_3 and Z_1 in (ex. 6), we still get a reliable estimate of ϕ_{32}, but the estimates of ϕ_{21} and especially ϕ_{31} have increased to compensate for the missing, positive link (increased probability of missingness). This has caused ϕ_{20} to decrease to compensate for the increased ϕ_{21}. When we ignore ϕ_{32} in (ex. 7), ϕ_{30} has moved even closer to zero, and ϕ_{31} has also increased to compensate for a missing link with a large, positive coefficient. Ignoring ϕ_{32} in (ex. 7) and (ex. 8) has removed the missingness link between Z_2 and Z_3, and as a result the ϕ_{2x} estimates are, surprisingly perhaps, quite reliable, presumably because ϕ_{2x} can no longer be used to compensate for the missing links for Z_3.

Scenario	$\beta_1 = 3$	$\beta_2 = 0.5$	$\beta_3 = -1$
0	3.01 (0.27)	0.53 (0.14)	-1.03 (0.19)
1	3.78 (0.43)	0.40 (0.12)	-1.16 (0.23)
2	3.01 (0.15)	0.55 (0.11)	-1.06 (0.06)
3	3.06 (0.16)	0.70 (0.10)	-1.09 (0.12)
4	2.96 (0.20)	0.61 (0.10)	-1.00 (0.12)
5	2.99 (0.12)	0.79 (0.10)	-1.02 (0.16)
6	3.02 (0.18)	0.65 (0.07)	-1.08 (0.08)
7	3.23 (0.19)	0.70 (0.12)	-1.12 (0.20)
8	3.14 (0.10)	0.81 (0.15)	-1.21 (0.14)

Table 7.21: Estimated β coefficients using a semi-parametric approach to estimate simulated missing values.

The estimates of β are by far the most important, as these are the coefficients we were originally looking for. As expected, using the true data set gives the best estimation of the β coefficients (ex. 0), while the CC data set gives estimates that are not acceptable (ex. 1) and have high standard deviations, i.e. unreliable estimates. Using the true α and ϕ distributions (ex. 2) gives estimates that are very close to the true values, and the estimates obtained using the true data set.

Using incorrect α distributions also affect the β estimates, but it seems that just the β_2 estimates are affected (ex. 3)-(ex. 5), even though we ignore just the link between Z_1 and Z_3 in (ex. 3). In any case, we need to compensate for the missing link, and it seems easier to estimate β_1 and β_3, perhaps because they are numerically larger (β_3 is also negative unlike the two other coefficients) and thus have greater influence on the survival times than β_2, and consequently the algorithm uses β_2 to compensate for the missing link. We also see that removing both α_{31} and α_{32} in (ex. 5) causes most damage. Still, the estimates of β_1 and β_3 are more accurate than the CC estimates.

Using incorrect missingness distributions (ex. 6)-(ex. 8) on the other hand seem to affect the estimates of all β coefficients. The estimate of β_2 is the most affected, and the estimate is already negatively affected when we remove ϕ_{31} in (ex. 6), but it is even worse when we remove ϕ_{32} in (ex. 8) and both in (ex. 9) where we also see loss of accuracy in the estimates of β_1 and β_3. However, we also see that the estimates of β_3 are still comparable with the CC estimate and much better when we compare the estimates of β_1. The estimates of β_2, though, are inaccurate. Again, this is probably because β_2 is the preferred coefficient to use as compensation coefficient. In this case, we achieve a more accurate estimate using the CC estimate. When we look at the percentage of

Scenario	Z_2	Z_3
0		
1	59.6 (8.4)	79.6 (7.4)
2	88.7 (5.1)	91.8 (4.2)
3	75.2 (4.8)	86.5 (4.2)
4	71.9 (5.2)	88.2 (4.4)
5	72.3 (5.0)	85.9 (4.1)
6	84.8 (4.6)	87.6 (4.2)
7	60.8 (5.1)	87.6 (4.7)
8	60.8 (5.3)	87.4 (4.5)

Table 7.22: Correctly estimated missing values using a semi-parametric approach to estimate simulated missing values.

correctly estimated values, we see that even though we do not specify the true distributions, we get estimates that are comparable with or better than the CC estimates. We also notice that especially the estimates of Z_2 are affected, when we do not use the true distributions. As mentioned earlier, Z_3 (and Z_1) is considered "more important" with respect to a greater influence on the survival times, making it easier to estimate its parameters and in turn its value. Furthermore, the incorrect α distributions affect the Z_2 estimates significantly, while the estimates of Z_3 have worsened, but not to the same extent. On the other hand, the missing ϕ_{31} distribution causes slightly decreased estimation performance for both Z_2 and Z_3, while missing ϕ_{32} (and ϕ_{31}) causes a dramatic decrease in performance for the estimation of Z_3, while the estimation of Z_2 is unaffected. However, the by far best performance is obtained when we use the true distributions.

We also experienced using different levels of missing values (using different missingness distributions). In conclusion, with higher levels of missingness, we see increased advantage of using our model to estimate the missing values. However, it also implies that the importance of using the true distributions increased.

All in all, we conclude that we can gain a lot by estimating the missing values using the implemented method. However, we also see that the advantage depends on how well we specify the α and ϕ distributions, especially for variables that do not have a large influence on the survival times (large β coefficients).

7.3.2 Using Augmented Data Set to Estimate CPH models

Next, we use the original COST data set and estimate the missing values using the joint distribution of the missing values. This gives us a data set where subject i is replaced with 2^{N_i} pseudo cases each with a different combination of missing values, where N_i is the number of missing values for subject i. Furthermore, each subject has assigned a weight; 1 for a fully observed case and the joint posterior probability otherwise.

7.3.2.1 Variable and Missingness Models

As mentioned in Section 4.4.3.1 and 4.4.3.2, we need to specify distributions for the variables, α-distributions of the form (4.51), and the missingness, ϕ-distributions of the form (4.54). We need to specify distributions for *apo*, *odd*, *dm*, *af*, *cla*, *temp*, and *sss* while *age* and *sex* are always observed and can be conditioned upon throughout the analysis. However, we also want to use the discarded variables to estimate the missing values of the remaining variables, but we do not want to use our limited amount of data to estimate conditional distributions for the values or the missingness of the discarded variables, and we do not include them in the CPH models. This would complicate the problem significantly, there would be a vast number of parameters to estimate, and it would also make it very difficult to propose an ordering of the variables. Hence, we model the values and the missingness for the discarded variables with simple logistic distributions that do not condition on any variables.

We still need to choose an ordering of the remaining variables though, allowing us to specify conditional distributions, where the i'th variable in the ordering may depend on the values of the variables $1, \ldots, i-1$. For this purpose we simply use the combined network in Figure 7.4 to give the ordering outlined in Figure 7.5. Since a BN is a DAG, it does not allow any cycles. This makes it a valid ordering. We estimate the values of sss using a simple, unconditional Gaussian. There are probably much better ways to model the distribution of the sss score, but with just 7 missing values, it will not influence the results significantly. We could also have used a simple mean value as we did in the BNs, but we model the sss to illustrate that the method can handle discrete as well as continuous variables.

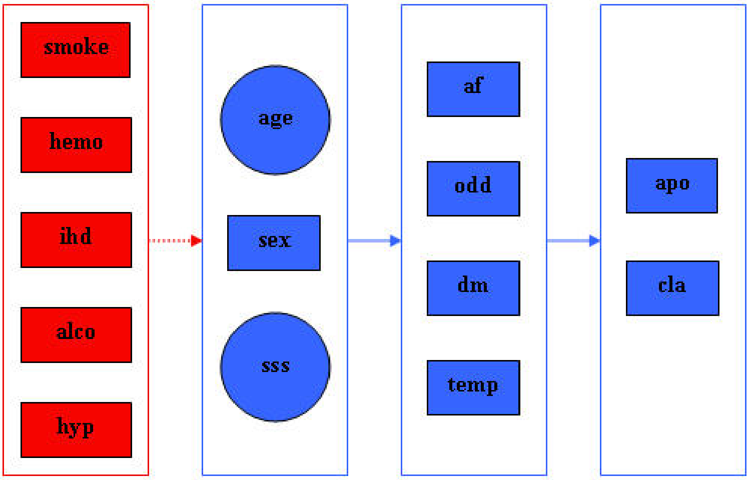

Figure 7.5: Illustration of $\boldsymbol{\alpha}$ structure using a semi-parametric approach to estimate missing values in the COST data set.

Next, we assume that the missingness of a variable does not depend on the values of other discrete variables with missing values. Instead, we believe that the missingness is a result of the short-time survival. We "model" the short-term survival by the severeness of the stroke, *sss*, and the patients *age*. If the stroke is severe and the patient is old, the patient is most likely in a very poor condition and often dies shortly after admission before the relevant patient information has been recorded. For the complicated SSS score, for example, missing values were observed for subjects with survival times 0, 1, 1, 1, 4, 6, and 10 days after admission, but for obvious reasons we cannot condition on the survival time.

The reason why we do not let the missingness of a variable depend on the variable itself, i.e. the missingness of the j'th discrete variable for the i'th individual, r_{ji}, depends on z_{ji}, is that the (by far) most *observed* value for all the remaining discrete variables with missing values is *no*. Hence, if we have a missing value, the most likely value will also be *no*, and we would then associate missingness with a *no*, and a *yes* with an observed variable. After all, there are about 15-20% observed *yes* values. In reality, however, we do not believe that there is correspondence between the missingness, and the value of a variable. The missingness must be a result of the short-time survival only.

The missingness model or missingness relations are illustrated in Figure 7.6. We have included the relation to the variable itself using a dotted line, as we will perform a separate experiment including these relations.

7.3.2.2 Estimation of Model Parameters

When the EM algorithm has converged, we have a new, augmented data set, and a set of parameter estimates, including p-values, $\boldsymbol{\alpha}$, $\boldsymbol{\phi}$, and $\boldsymbol{\beta}$ estimates that we can use to calculate HRs. As mentioned, the semi-parametric approach originally proposed by Herring et al. (2004) uses stepwise selection to estimate a single CPH model, and update the ML parameters in the M-step. Hence, an obvious improvement of this algorithm is to implement BMA as part of the M-step in the EM algorithm, and use an average model to estimate the survival times. This will include the model uncertainty, and give more accurate parameter estimates that in turn will improve the estimates of the missing values and vice versa. We refer to this implementation as the “extended” algorithm, and the original implementation as the “original” algorithm.

Estimation of α Parameters

In Table 7.23 we show the estimated α coefficients using the original algorithm, and in Table 7.24 the α estimates using the extended algorithm. For the discrete variables, all intercepts are negative and indicate a preference for *no*, or $< 37.0°$ C for *temp*. This is in line with our expectations, as there is an excess number of subjects with *no*'s respectively $< 37.0°$ C records in the database. The size of

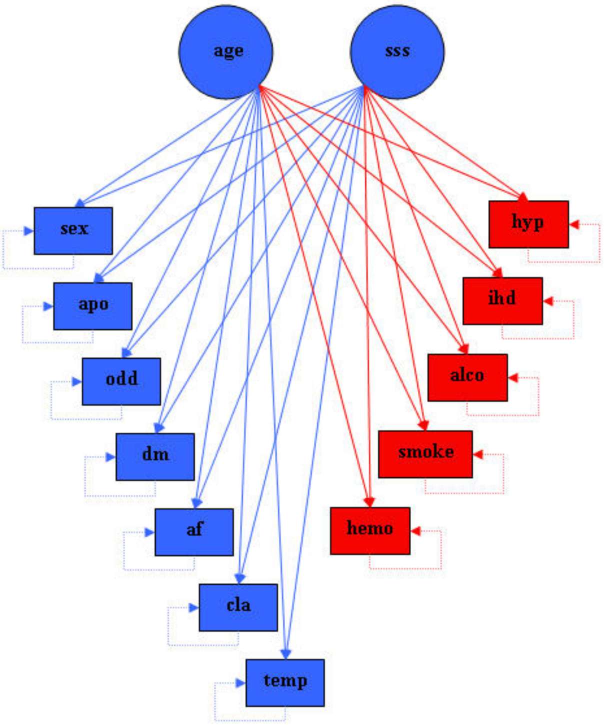

Figure 7.6: Illustration of ϕ structure using a semi-parametric approach to estimate missing values in the COST data set.

the intercepts is also in line with this distribution. There are no large differences between the estimates using the original and the extended algorithm, but the absolute values of the intercepts, expressing the a priori probabilities for the most likely values, have increased slightly.

If we look at the individual distributions, we note the following:

apo: High probability that the patient has not previously experienced a stroke. This probability increases if the stroke is mild (higher SSS score), but decreases if hypertension or atrial fibrillation is present.

odd: High probability that the patient does not suffer from another disabling disease. This probability increases if the patient consumes alcohol, but decreases if the patient has an ischemic heart disease.

Variable	apo	odd	dm	af	cla	temp	sss
K_0	-1.23	-1.47	-0.87	-4.29	-2.36	0.55	39.0
σ							4.1
age			-0.02	0.08			
sex						-0.30	
hyp	0.65		0.61				
ihd		0.59		1.02	0.88		
apo				0.46			
odd					0.89		
alco		-0.43	-0.67				
dm							
smoke				-1.03	0.75		
af	0.44						
hemo			-2.04				
cla							
temp							
sss	-0.02			-0.02		-0.02	

Table 7.23: Estimated α parameters using original algorithm in a semi-parametric approach to estimate missing values in the COST data set.

dm: High a priori probability that the patient does not have diabetes mellitus. This probability increases with the patients age, if the patient consumes alcohol, or the stroke is a hemorrhage, but decreases if hypertension is present.

af: High a priori probability that atrial fibrillation is not present. This probability increases if the stroke is mild (higher SSS score), or the patient is smoking, but it decrease with the patients age, if the patient has an ischemic heart disease, or has previously experienced a stroke.

cla: High a priori probability that the patient does not have intermittent claudication. This probability decreases if the patient is smoking, has an ischemic heart disease, or has previously experienced a stroke.

temp: Moderate a priori probability that the patients body temperature was $< 37.0°$ C. This probability decreases if the patient is male, or the stroke is mild (higher SSS score).

sss: The mean value of the Gaussian distribution is 39.0, and the standard deviation is 17.1. These values are close to the mean and standard deviation of *sss* for all subjects in the database.

Variable	apo	odd	dm	af	cla	temp	sss
K_0	-1.36	-1.44	-1.13	-4.30	-2.42	0.54	39.0
σ							17.1
age			-0.02	0.08			
sex						-0.24	
hyp	0.65		0.61				
ihd		0.59		1.02	0.88		
apo				0.47			
odd					0.79		
alco		-0.43	-0.67				
dm							
smoke				-1.03	0.75		
af	0.42						
hemo			-2.04				
cla							
temp							
sss	-0.01			-0.02		-0.02	

Table 7.24: Estimated α parameters using extended algorithm in a semi-parametric approach to estimate missing values in the COST data set.

Most of these relations seem plausible and make intuitively sense, e.g. that the probability of an earlier stroke increases, if hypertension is present, as hypertension is known to be increase the risk of a stroke [1]. Other relations might be a little surprising, such as ageing *decreasing* the probability that diabetes is present. This may seem odd at first, since we would expect diabetes to occur in older rather than younger people. However, if you have diabetes, you are at least twice as likely to have a heart disease or a stroke as someone who does not have diabetes. People with diabetes also tend to develop a heart disease or have strokes at an earlier age than other people. If you are middle-aged and have type 2 diabetes, some studies suggest that your chance of having a heart attack is as high as someone without diabetes who has already had a heart attack, (NDCI, 2005), (Andersen et al., 2006d), (Jørgensen et al., 1994b), (Tuomilehto et al., 1996). Hence, it makes sense that *dm* is an indicator of younger patients.

Estimation of ϕ Parameters

In Table 7.25 and 7.26 the estimated ϕ coefficients, using the original and the extended algorithm respectively, are presented. If we look at the individual estimates, we see that the missingness of all variables have, of course, a high a

[1] See e.g. the National Stroke Associations *stroke risk scorecard* at http://www.stroke.org/site/DocServer/scorecardQ.pdf?docID=601.

priori probability for *no*. Ageing increases the missingness probability, while it decreases with the SSS score, which makes sense, as older patients with more severe strokes are expected to be in a weaker condition making it less possible to obtain the relevant patient information. However, for *temp*, the missingness probability *increases* with the SSS score, i.e. that body temperature is more likely not recorded, if the patient experiences a mild stroke. The reason is that the body temperature needs to be recorded early after stroke onset. Otherwise, the body temperature in acute stroke can change very rapidly, even within 6 to 8 hours after onset as documented by Boysen and Christensen (2001). When a patient experiences a severe stroke, the patient is immediately admitted to hospital, while patients with mild strokes are often admitted much later. Perhaps because they were not even aware that they experienced a stroke at the time of onset. Hence, for these patients, body temperature is not recorded. Finally, as expected, we see a very high a priori probability of *no* for the missingness of *sss*.

	apo	odd	dm	af	cla	temp	sss
intercept	-5.72	-6.36	-3.47	-4.30	-3.69	-3.51	-102.57
age	0.06	0.07	0.03	0.02	0.05	0.01	
sss	-0.07	-0.08	-0.07	-0.05	-0.06	*0.04*	

Table 7.25: Estimated ϕ coefficients using original algorithm in a semi-parametric approach to estimate missing values in the COST data set.

	apo	odd	dm	af	cla	temp	sss
intercept	-6.17	-6.21	-4.23	-4.24	-3.89	-4.47	-102.57
age	0.07	0.06	0.04	0.01	0.05	0.01	
sss	-0.07	-0.07	-0.07	-0.05	-0.06	*0.06*	

Table 7.26: Estimated ϕ coefficients using extended algorithm in a semi-parametric approach to estimate missing values in the COST data set.

Just for the sake of it, we also implemented a ϕ structure where the missingness of a discrete variable was also conditional on the value of the variable itself. The corresponding estimates are shown in Table 7.27. As expected, the parameters are all large and negative because most of the observed values for z_ji are zero (*no*).

Estimation of β Parameters

Finally, we compare the estimates of the β coefficients in Table 7.28-7.29, and we include the CC results for comparison.

Algorithm	apo	odd	dm	af	cla	temp
Original	-7.36	-10.92	-8.42	-8.13	-4.32	-7.37
Extended	-6.08	-11.77	-7.30	-6.28	-3.95	-8.71

Table 7.27: Estimated ϕ_{jj} coefficients letting $p(r_{ji})$ be conditioned upon *age*, *sex*, and z_{ji}.

Method	age	sex	apo	odd	dm
p-value (org)	<0.001	<0.001	<0.01	0.02	<0.01
p-value (step, CC)	<0.001	<0.001	<0.01	0.02	<0.01
PPP_{ext}	100	100	81.3	64.9	88.7
$\text{PPP}_{\text{O,CC}}$	100	100	73.5	56.9	78.5
	af	cla	temp	sss	sss*t
p-value (org)	0.03	0.02	0.03	<0.001	<0.001
p-value (step, CC)	0.03	0.02	0.03	<0.001	<0.001
PPP_{ext}	29.6	57.1	38.2	100	100
$\text{PPP}_{\text{O,CC}}$	26.5	51.9	33.4	100	100

Table 7.28: p-values and PPPs using a semi-parametric method to estimate missing values. Max. PMP: 0.16. Total PMP for Top10: 0.64. 35 models included in Occam's window. Hazard ratio for $sss * t$ is pr. 100 unit increment.

Overall, the results are comparable with the results from the BN approach. Again, the changes in p-values using the original algorithm are so small that we cannot see them using two decimals, and all variables are "as significant" as they were in the CC analysis. For all variables that do not have $p < 0.001$, except for *dm*, the HRs have increased slightly, while the standard deviations of the HR estimates have not increased, neither have they decreased as we saw a few examples of in the BN solution. Hence, we conclude that the original semi-parametric approach - using these $\boldsymbol{\alpha}$ and $\boldsymbol{\phi}$ distributions - has not lead to more accurate HR estimates. All in all, the augmented data set has provided new information, leading to slightly altered HR estimates indicating stronger influence on the survival time compared to the CC results. The changes, however, are not extreme.

The results using the extended algorithm, with BMA incorporated, show increased PPPs for all variables, except for *age*, *sex*, *sss*, and $sss * t$ whose PPPs are already 100. The increase has been most significant for *apo*, *odd*, and *dm* with about $8 - 10\%$, while the increase i about $3 - 5\%$ for af, *cla*, and *temp*. Hence, we observe trends comparable with the BN solution, although the changes in PPP are smaller. The conclusion is the same though, namely

Method	age	sex	apo	odd	dm
HR_{org}	1.05	1.43	1.35	1.29	1.37
$\text{HR}_{\text{S,CC}}$	1.05	1.41	1.33	1.28	1.37
$\sigma_{\text{HR, org}}$	~ 0	0.12	0.13	0.13	0.15
$\sigma_{\text{HR, S, CC}}$	~ 0	0.12	0.13	0.13	0.15
HR_{ext}	1.05	1.40	1.27	1.20	1.34
$\text{HR}_{\text{O,CC}}$	1.05	1.40	1.24	1.17	1.30
$\sigma_{\text{HR, ext}}$	~ 0	0.11	0.18	0.16	0.20
$\sigma_{\text{HR, O, CC}}$	~ 0	0.12	0.19	0.18	0.22
	af	cla	temp	sss	sss*t
HR_{org}	1.29	1.35	1.21	0.95	1.0019
$\text{HR}_{\text{S,CC}}$	1.27	1.34	1.20	0.95	1.0019
$\sigma_{\text{HR, org}}$	0.14	0.16	0.10	~ 0	~ 0
$\sigma_{\text{HR, S, CC}}$	0.14	0.16	0.10	~ 0	~ 0
HR_{ext}	1.07	1.20	1.07	0.95	1.0019
$\text{HR}_{\text{O,CC}}$	1.07	1.18	1.06	0.95	1.0019
$\sigma_{\text{HR, ext}}$	0.12	0.20	0.09	~ 0	~ 0
$\sigma_{\text{HR, O, CC}}$	0.13	0.22	0.11	~ 0	~ 0

Table 7.29: HRs using a semi-parametric method to estimate missing values. Max. PMP: 0.16. Total PMP for Top10: 0.64. 35 models included in Occam's window. Hazard ratio for $sss * t$ is pr. 100 unit increment.

that the augmented data set provides information that confirms or increases the evidence for an effect of all variables. Again, the new evidence is reflected in the updated PPPs, and the results using a semi-parametric approach to estimate the missing values confirm our findings using BNs for the same purpose.

This time, the changes in PPPs are accompanied by increased HRs (compared to the CC estimates) for *apo*, *odd*, *dm*, *cla*, and *temp*, while the HRs for *age*, *sex*, *sss*, $sss * t$, and also af have not changed. These observations are also in line with the BN solution. Finally, the standard deviations of the HR estimates have all decreased or remain at ~ 0, indicating more confident estimates. We also see an indication of reduced model uncertainty, as the maximum PMP has increased from 0.09 to 0.16, the total PMP for Top10 has increased from 0.49 to 0.64, and we just include 35 models compared to 49 in the CC analysis.

All in all, the semi-parametric approach is also a valuable tool for estimating missing values. One of the advantages is that it combines three sources of information: How the value of a variable is related to the values of other variables, how the missingness of a variable is related to the values of other variables and/or the value of itself, and finally how the estimated values affect the estimated survival time. Hence, we do not base our estimates of the missing values on one source of information as we do in the BN approach, where we do not incorporate neither the missingness nor the survival time distributions. In the semi-parametric approach we also use an EM algorithm do update the parameter and missing value estimates in turn to improve the estimates iteratively. Using BNs, we simply estimate the values once and for all, and use the augmented data set to estimate the parameters once. However, we did use an EM algorithm to include subjects with missing values to estimate the network structure/parameters, and the missing values in turn. This turned out to be the best solution, and we have now presented two examples of the advantage of iteratively updating parameter and missing value estimates.

One of the drawbacks of the semi-parametric approach that we do not have when we use BNs, is that we need to specify how variables are connected a priori. Unless we have prior knowledge enabling us to do so, we need to look for other ways to order and connect the variables. Simulations showed that the results are clearly affected when we use different distributions, but the effects seem to vanish when we use the extended algorithm, at least we obtained results comparable with the BN solution, although we have to remember that we "borrowed" the α structure from the BN solution. However, we have probably also seen an effect of the improved CPH model estimates using BMA. Accurate β estimates are probably more important than the specification of true α and ϕ distributions, and makes the extended algorithm more robust to miss-specifications.

There are many ways to model variable distributions (especially continuous variables), inter-variable relations and missingness distributions, making the semi-parametric approach a comprehensive modeling area, and results should be thoroughly evaluated and compared, e.g. with respect to predictive power, and sensitivity analysis should always be part of the modeling. Using BNs, we have methods that enable us to learn the network structure using the available data - even including missing values. If we have prior knowledge that we would like to incorporate as required or perhaps illegal connections, we can easily do so. This makes BNs much more flexible.

Method	min	mean	max
Original	0.76	0.78	0.80
Extended	0.79	0.83	0.88

Table 7.30: Simulation of missing values in the COST data set. Distribution of probabilities for correct missing value patterns using the joint distribution.

We compare the semi-parametric approach to the CC and the BN solution in terms of predictive performance in Section 7.4, but we also experimented with the estimation of simulated missing values. In Table 7.30 we present the distribution of the probabilities for the correct missing value patterns using the joint $\boldsymbol{\alpha}$ distribution. Again, we randomly remove 10% of the observed discrete values, allowing us to compare the results with the corresponding simulation using BNs. The original algorithm obtain results that are worse than the results using BNs, while the extended algorithm shows improved estimates of the missing values - at least when we use this method to compare the algorithms.

7.4 Predictive Performance and Survival Plots

Having shown the standard CC approach using the full set of variables, how to increase the size of the database using stepwise BMA, and two ways to estimate missing values, we would like to evaluate the predictive performance of each method. This will allow us to compare the models in terms of generalization error by evaluating the models on data they have not seen before. We randomly split the data set into a training set (90%) that we use to estimate the parameters, and a test set (remaining 10%) that we use for evaluation. We evaluate the mean of the PPS, the IC, and σ_{pred} averaged over 200 runs in Table 7.31, and we compare each of the BMA approaches to the standard survival analysis method, using stepwise selection on the CC data set including all potential risk factors.

Since the data sets are different in each method, and the PPS is an evaluation of how well a method fits the data, the test set has to be part of the CC data set using the full set of variables. Otherwise, the PPS scores would not be comparable. The predictive $\mathcal{Z}$-score does not have this limitation, but it would not be fair to include subjects with estimated values in the test set, as the method responsible for the estimated values would have an obvious advantage. However, the training sets are different. We simple let the training set be the remaining part of the CC data set + the remaining complete cases for the limited set of variables + the subjects augmented with BN/semi-parametric estimates respectively. In other words, we use the largest possible training set for each method.

	BMA			Stepwise		
Method	PPS	IC	σ_{pred}	PPS	IC	σ_{pred}
Full	-249.5	0.78	30.8	-252.0	0.74	35.9
Drop	-246.2	0.82	26.4	-249.5	0.78	32.9
BN	-243.4	0.86	22.3	-247.5	0.81	28.2
Semi	-242.1	0.87	21.1	-247.2	0.81	28.4

Table 7.31: PPS, IC, and σ_{pred} using: CC data/all risk factors (*Full*), CC data set/risk factors remaining after application of stepwise BMA (*Drop*), remaining risk factors/BN estimates of missing values (*BN*), and remaining risk factors/semi-parametric estimates of missing values (*semi*) averaged over 200 runs.

Comparing the results in Table 7.31 (multi) column-wise, the results clearly show the superiority of the BMA methods compared to the corresponding stepwise selection implementations with regard to higher PPS, higher IC, and lower σ_{pred}. Comparing the results row by row, we see significant improvements in predictive

Method	PPS	IC	σ_{pred}
$\text{Full}_{\text{BMA}} - \text{Full}_{\text{step}}$	2.5 (3.6%)	0.04	-5.1
$\text{Drop}_{\text{BMA}} - \text{Full}_{\text{step}}$	5.8 (8.6%)	0.08	-9.5
$\text{BNT}_{\text{BMA}} - \text{Full}_{\text{step}}$	8.6 (12.8%)	0.12	-13.6
$\text{Semi}_{\text{BMA}} - \text{Full}_{\text{step}}$	9.9 (14.8%)	0.13	-14.8

Table 7.32: Difference in PPS, IC, and σ_{pred} compared to a CC data/all risk factors (*Full*) approach.

performance when we increase the amount of available information. As we cannot handle missing values when we fit our Cox models, we do it by removing variables, and by estimating the missing values.

The results in Table 7.32 are very interesting, and is the experimental "climax". The table shows the predictive improvements compared to the standard approach to the survival analysis problem: using all potential risk factors in a stepwise selection approach on the CC data set. As expected, or at least hoped, we see improved predictive performance compared to the standard solution regardless of whether we use the PPS or the predictive $\mathcal{Z}$-score. Using PPS, the predictions are on average 2.5% better pr. event and the IC 4% better when we use BMA on the CC data set using all potential risk factors. When we remove all risk factors with very low PPP, the average PPS improvement is 8.6% pr. event and 8% in IC. The results are even better when we augment the data using estimated missing values to include all potential subjects in the database. Using BN estimates, the improvement is 12.8% pr. event on average, and 12% in IC, while the corresponding results using the extended semi-parametric method are 14.8% pr. event and 13% in IC. We also note that at the same time the σ_{pred} decrease, i.e. that the estimated CIs are actually narrower, and still give more accurate estimates of the predicted survival times.

As the main objective in standard survival analysis is to evaluate how potential risk factors influence the survival time, and how significant this effect is, we summarize the estimated p-values, PPPs, and HRs for the implemented methods in Table 7.33-7.36 for the risk factors that were included after the application of stepwise BMA. The conclusion is that we have very strong evidence for an effect of age, sex, sss, and $sss * t$, and we are very confident that male patients, older patients, and patients experiencing severe strokes will have shorter survival times/increased hazard. Both the p-values, the PPPs, and the HRs for these variables have hardly changed at any time.

We also have positive evidence for an effect of apo and dm, and as expected, the hazard increases if the patient has already experienced a stroke, and/or if the patient has diabetes. We also find weak evidence of an effect of *odd* and *cla*, with increased hazard for patients suffering from other disabling diseases, and/or intermittent claudication. Finally, stepwise selection believes that af and *temp* are significant explanatory variables, while BMA permanently suggests positive evidence against an effect, although the PPPs have increased when we increased the size of the data set. Anyhow, the results indicate that a potential effect would be an increased hazard for patients with atrial fibrillation, and for patients with hyperthermia (body temperature $\geq 37.0°$ C).

All in all, results indicate that survival times are longer for younger, female patients who experience mild strokes, have not had an earlier stroke, and do not suffer from an other disabling disease, diabetes, or intermittent claudication, and perhaps atrial fibrillation and body temperature are also important explanatory variables, but the methods do not agree on whether data show evidence for an effect or not. We also note that hypertension, ischemic heart disease, alcohol consumption, smoking habits, and the type of stroke were *not* believed to have an effect on the survival time, or at least the implemented methods found no evidence suggesting otherwise.

Method	age	sex	apo	odd	dm
p-value (full)	<0.001	<0.01	<0.01	0.02	0.1180 (5)
p-value (drop)	<0.001	<0.001	<0.01	0.02	<0.01
p-value (BNT)	<0.001	<0.001	<0.01	0.02	<0.01
p-value (semi)	<0.001	<0.001	<0.01	0.02	<0.01
	af	cla	temp	sss	sss*t
p-value (full)	<0.01	0.04	0.13 (4)	<0.001	<0.001
p-value (drop)	0.03	0.02	0.03	<0.001	<0.001
p-value (BNT)	0.03	0.02	0.03	<0.001	<0.001
p-value (semi)	0.03	0.02	0.03	<0.001	<0.001

Table 7.33: Summary of p-value estimates using: CC data/all risk factors (*Full*), CC data set/risk factors remaining after application of stepwise BMA (*Drop*), remaining risk factors/BN estimates of missing values (*BN*), and remaining risk factors/semi-parametric estimates of missing values (*semi*).

If we use an average of the final (average) models using BNs and the extended semi-parametric method to estimate missing values, the CR model for predicting the survival time of stroke patients is (including af and *temp*)

$$\begin{aligned} h(t) = h_0(t)\exp(&1.05age + 1.40sex + 1.27apo + 1.20odd + 1.35dm \\ &+ 1.08af + 1.21cla + 1.08temp + 0.95sss + 1.0019\text{e-}002sss * t) \end{aligned} \quad (7.10)$$

If we also use an average of the estimated cumulative baseline hazards, we can plot survival curves for hypothetical patients. These plots visualize the effect of

Method	age	sex	apo	odd	dm
PPP_{full}	100	96.9	77.9	75.1	28.9
PPP_{drop}	100	100	73.5	56.9	78.5
PPP_{BNT}	100	100	80.1	65.7	90.5
PPP_{semi}	100	100	81.3	64.9	88.7
	af	cla	temp	sss	sss*t
PPP_{full}	3.6	39.3	9.3	100	100
PPP_{drop}	26.5	51.9	33.4	100	100
PPP_{BNT}	31.6	57.3	40.0	100	100
PPP_{semi}	29.6	57.1	38.2	100	100

Table 7.34: Summary of PPP estimates using: CC data/all risk factors (*Full*), CC data set/risk factors remaining after application of stepwise BMA (*Drop*), remaining risk factors/BN estimates of missing values (*BN*), and remaining risk factors/semi-parametric estimates of missing values (*semi*).

Method	age	sex	apo	odd	dm
$\text{HR}_{\text{S(full)}}$	1.05	1.36	1.45	1.34	
$\text{HR}_{\text{S(drop)}}$	1.05	1.41	1.33	1.28	1.37
$\text{HR}_{\text{S(BNT)}}$	1.05	1.41	1.34	1.29	1.41
HR_{semi}	1.05	1.43	1.35	1.29	1.37
	af	cla	temp	sss	sss*t
$\text{HR}_{\text{S(full)}}$	1.62	1.35		0.96	1.0019
$\text{HR}_{\text{S(drop)}}$	1.27	1.34	1.20	0.95	1.0019
$\text{HR}_{\text{S(BNT)}}$	1.30	1.36	1.21	0.95	1.0019
$\text{HR}_{S(\text{semi})}$	1.29	1.35	1.21	0.95	1.0019

Table 7.35: Summary of stepwise selection HR estimates using: CC data/all risk factors (*Full*), CC data set/risk factors remaining after application of stepwise BMA (*Drop*), remaining risk factors/BN estimates of missing values (*BN*), and remaining risk factors/semi-parametric estimates of missing values (*semi*).

the risk factors, and give a better understanding than written numbers in a table. Let the "reference" subject have the following risk factor profile: $age = 74$, $sex = female$, $apo, odd, dm, af, cla = no$, $temp < 37°$ C, and $sss = 38$, i.e. a subject with "mean" values in all categories.

In Figure 7.7-7.15 the survival curves for the reference subject is plotted along with the survival curves for a subject, where we have changed the value of each of the risk factors. For all discrete risk factors, the new subject has *apo*, *odd*, *dm*, *af*, and *cla* = *yes* respectively. For *age*, we plot the reference subject along

Method	age	sex	apo	odd	dm
$HR_{O(full)}$	1.05	1.38	1.32	1.31	1.09
$HR_{O(drop)}$	1.05	1.40	1.24	1.17	1.30
$HR_{O(BNT)}$	1.05	1.40	1.27	1.20	1.35
HR_{semi}	1.05	1.40	1.27	1.20	1.34
	af	cla	temp	sss	sss*t
$HR_{O(full)}$	1.01	1.15	1.02	0.96	1.0017
$HR_{O(drop)}$	1.07	1.18	1.06	0.95	1.0019
$HR_{O(BNT)}$	1.08	1.21	1.08	0.95	1.0019
$HR_{S(semi)}$	1.07	1.20	1.07	0.95	1.0019

Table 7.36: Summary of BMA HR estimates using: CC data/all risk factors (*Full*), CC data set/risk factors remaining after application of stepwise BMA (*Drop*), remaining risk factors/BN estimates of missing values (*BN*), and remaining risk factors/semi-parametric estimates of missing values (*semi*).

with a subject of age 50, and a subject of age 90. For sss, we plot the reference subject along with a subject with $sss = 10$ (severe stroke), and a subject with $sss = 58$ (mild stroke). Finally, we show the survival curve of the reference subject along with the survival curve for the same subject, where all discrete variables are set to one (*yes*) and $temp \geq 37°$ C in Table 7.16, i.e. we compare the survival curves of a “healthy” and a seriously “unhealthy” subject with same age and SSS score.

The visualized survival plots all aid in the understanding of the effect each risk factor has on the survival time, indicated by the distance between the survival curves. It suddenly becomes crystal clear how important the patients age and the stroke severity are. Very old patients and patients with very severe strokes have very, very poor prospects and should not expect to live more than a few weeks. Most of these patients are also dead when they arrive at the hospital or shortly after. Of the 19 patients (1.9%) that are dead on arrival, the mean age is 77.7 years compared to the mean age 74.3 years of all patients, and the mean SSS score is 4.8 compared to the general mean of 38.0. Of the 104 patients (10.4%) that die within the first week of admittance, the mean age is 76.1 years and the mean SSS score is 13.4.

If the patient is younger or has a mild stroke, the survival chances are much better. For example, the median survival time for the reference subject with age changed to 50, *or* the SSS score changed to 58, is about 1 year. We also note that (in terms of survival time) it is better to be 50 years than having a very mild stroke, as the effect of the stroke severity decreases over time due to the $sss * t$ variable. This implies that the estimated probability of surviving 5 years is about 0.2 for the 50 year old reference subject, while it is about 0.13 for the reference subject with 58 points in SSS score.

Finally, the survival prospects are very poor for the very unhealthy reference subject answering *yes* to all discrete risk factors. In fact, the survival curve resembles that of the reference subject with $sss = 10$.

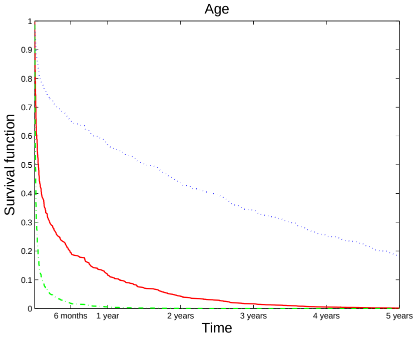

Figure 7.7: Survival plot for *age*. 74 years (red, solid), 50 years (blue, dotted), 90 years (green, dashdot).

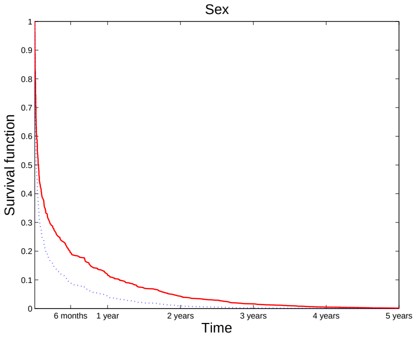

Figure 7.8: Survival plot for *sex*. Female (red, solid), male (blue, dotted).

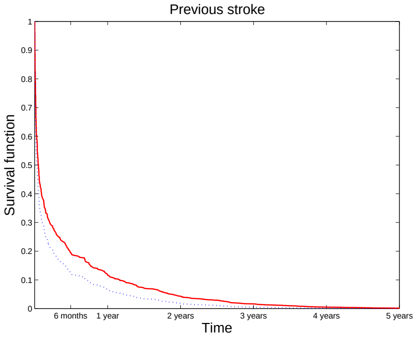

Figure 7.9: Survival plot for *apo*. No (red, solid), yes (blue, dotted).

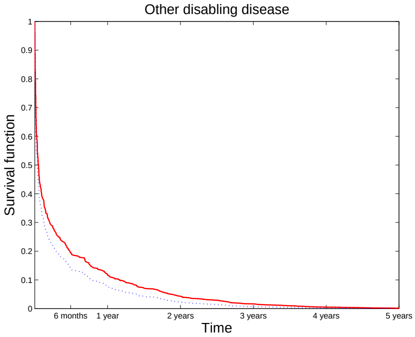

Figure 7.10: Survival plot for *odd*. No (red, solid), yes (blue, dotted).

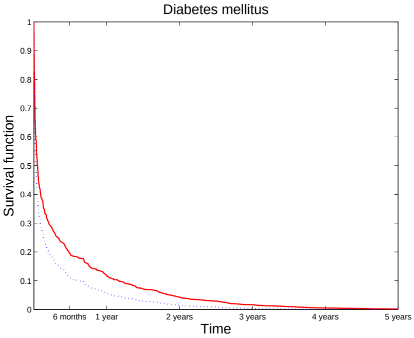

Figure 7.11: Survival plot for *dm*. No (red, solid), yes (blue, dotted).

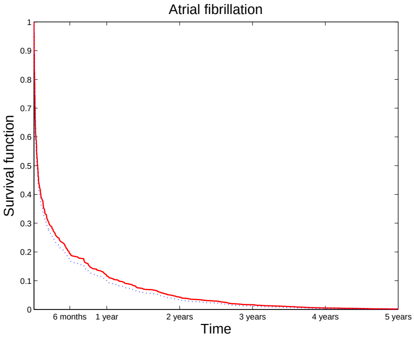

Figure 7.12: Survival plot for *af*. No (red, solid), yes (blue, dotted).

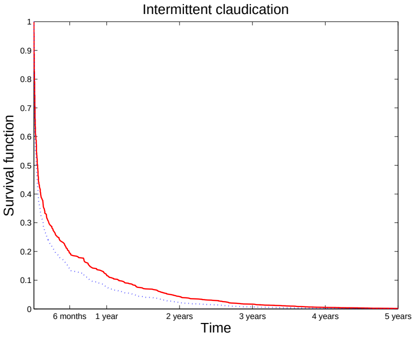

Figure 7.13: Survival plot for *cla*. No (red, solid), yes (blue, dotted).

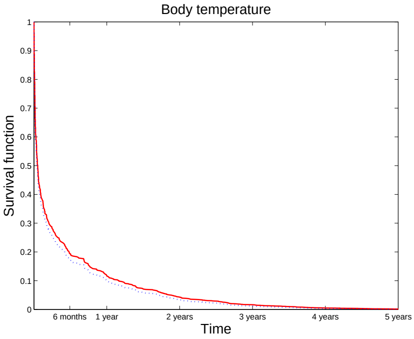

Figure 7.14: Survival plot for *temp*. $< 37°$ C (red, solid), $\geq 37°$ C (blue, dotted).

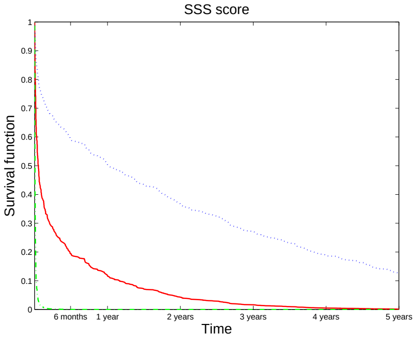

Figure 7.15: Survival plot for *sss*. 39 points (red, solid), 58 points (blue, dotted), 10 points (green, dashdot).

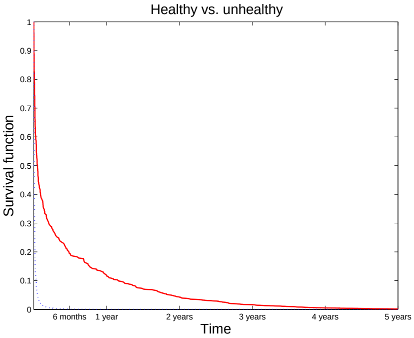

Figure 7.16: Survival plot for *all*. “Healthy” (red, solid), “unhealthy” (blue, dotted).

Chapter 8

Conclusion and Discussion

In this thesis we have thoroughly compared BMA with the most common approach to survival analysis, stepwise selection. BMA is rarely used, especially in medical studies, because the method is based on mathematical concepts such as Bayes' theorem, posterior probability, subset selection, and model averaging that physicians are not familiar with. In contrast, stepwise selection is a very simple method that uses simple statistical terms such as p-values and significance levels, and stepwise selection is an integrated part of commercial statistical packages. The result is that any physician can perform survival analysis. However, the concept of p-values is often misunderstood and believed to be the probability that the null hypothesis is true. Hence, physicians claim that a p-value below an artificial significance level provides evidence for an effect of a given risk factor.

We have identified several flaws of the stepwise selection method, among other factors the use of a significance level that is used to include/exclude variables. We have shown that BMA does not exclude any variables and provides real probabilities of an effect for each variable. Furthermore, we used BMA to estimate the model uncertainty and provide more reliable parameter estimates. The results showed that the final model in stepwise selection was not necessarily the model with highest PMP. Furthermore, there were several other models with significant PMP, and often the Top10 models in terms of PMP were not assigned more than half of the posterior probability mass. Stepwise selection does not take model uncertainty into account, and we have shown how this leads to overconfident estimates of the model parameter estimates and more significant variables.

The methods were compared on two real-life data sets, a small multiple myeloma patients data set, and the large Copenhagen Stroke study database, using CPH models to model the distribution of the survival times. The results showed that the conclusions of stepwise selection were not always in accordance with the conclusions of BMA. Often, stepwise selection found a risk factor significant, while BMA suggested that data showed evidence *against* an effect. These disagreements are the result of the model uncertainty. Furthermore, we showed how stepwise selection does not distinguish between significant variables, while BMA uses the posterior parameter probability to assess the probability of an effect. The experiments also indicated that we can obtain results that are comparable with an average over all potential models using a much smaller subset of models. The subsets were identified using Occam's window subset selection, and we managed to reduce the number of models to average over significantly, in some cases more than a factor 300.

The improved evaluation of the risk factors was not the only advantage of BMA. We also showed how the predictive performance increased. We used the existing PPS score to evaluate the predictive performance, and also suggested a novel evaluation score, the *predictive $\mathcal{Z}$-score*. The proposed score computes predictive intervals for the expected survival times. We used these intervals to compute the interval coverage, with the advantage that the score is interpretable, and we can easily relate to differences in interval coverage. Furthermore, we can use the size of the predictive intervals to estimate the accuracy of the predictions.

We also showed how to evaluate the assumption of proportional hazards using log-cumulative hazard plots and weighted Schoenfeld residuals. These methods allow us to evaluate the assumption of proportional hazards before and after model fitting. In the COST experiment, this assumption was violated, and we showed how to include time-dependent variables. Data showed very strong evidence of an effect of a new, time-dependent stroke severity variable. The results indicated that the effect of the stroke severity decreased with time. This aspect would not have been captured in most survival analysis studies.

Next, we proposed a *stepwise BMA* algorithm to remove risk factors whose posterior parameter probabilities were close to zero. Hence, we removed variables that we were confident did not have an effect on the survival times. The algorithm reduced the number of potential risk factors, and in turn significantly reduced the number of subjects with missing values. The result was more reliable and accurate parameter estimates and reduced model uncertainty.

We also showed how to estimate missing values in the data set using BNs to connect the risk factors. We showed that the best results were obtained using a structural EM algorithm to estimate the structure/parameters and the missing values in turn. The algorithm is able to use subjects with missing values to learn the structure and parameters of the network. Hence, it is able to use all available data compared to normal learning algorithms that are only able to use completely observed cases.

Finally, we used the BN structure to propose an ordering of the risk factors such that we could place parametric distributions on the risk factors as well as the missingness mechanisms. By modeling the missingness mechanisms, we do not have to assume that values are missing completely at random. Instead, we used the missingness as extra information to improve the estimates of the parameters and the missing values. The original algorithm use the CPH model to model the survival times, and we showed how improve the algorithm by replacing stepwise selection with BMA in the M-step of the EM algorithm. The modification lead to improved parameter and missing value estimates, and also increased the predictive power.

The results show that we are confident that the expected survival time of stroke patients is lower for male patients, or if the patient has previously experienced a stroke. The expected survival time also decreases with ageing, severity of stroke, presence of another disabling disease, diabetes mellitus, or intermittent claudication.

These observations are in line with other studies in the literature, e.g. the suggestions in (Andersen et al., 2005b) and (Andersen et al., 2006f): Short-term stroke survival is the same for men and women, whereas long-term stroke survival is markedly better for women, (Kammersgaard et al., 2004): Age per se is a strong predictor of outcome and mortality after stroke, (Andersen et al., 2006c): SSS is the single most important predictor of short-term mortality (1 year), (Andersen et al., 2006d), (Tuomilehto et al., 1996), (Jørgensen et al., 1994b), and (Knuiman et al., 1992): Diabetes mellitus is a strong, independent predictor of premature death following stroke, (Knuiman et al., 1992): For stroke mortality, intermittent claudication is a strong predictor of death and many others.

However, our results also show that the effect of stroke severity decreases with time. Andersen et al. (2006c) also suggest that the stroke severity score measured by means of the Scandinavian Stroke Scale is the single most important predictor of short-term mortality. For long-term mortality, cardiovascular risk factors were predominant. However, we have not found other studies that include a risk factor expressing the stroke severity as a function of time, and we also believe to be the first study to estimate *how* the effect of stroke severity on the survival time changes with time.

We also suspect the expected survival time to decrease with the presence of atrial fibrillation, or if the body temperature is $\geq 37.0°$ C on admission. In stepwise selection, these risk factors were significant, but using BMA, data did not show positive evidence of an effect. However, these effects would be in accordance with observations documented by Jørgensen et al. (1996a), suggesting that patients with AF have a higher mortality rate, and Kammersgaard et al. (2002) suggesting that hypothermia (body temperature $< 37.0°$ C) decreases the short- and long-term mortality risk.

Finally, data showed evidence of an effect for the presence of hypertension, ischemic heart disease, alcohol consumption, smoking habits, and the type of stroke when we adjusted for the other risk factors. This conclusion is also suggested by Andersen et al. (2005b). However, Andersen et al. (2005b) also show how the set of significant risk factors depend on the time of censoring. Using stepwise selection to fit a CPH model on the COST database, significant risk factors using $\alpha = 0.05$ were *age*, *sex*, *dm*, *af*, *hemo*, and *sss* using a 1 year censoring, *age*, *sex*, *apo*, *odd*, *dm*, *af*, and *sss* using 5 and 10 year censoring. Hence, *hemo* was only found to be a significant risk factor for the 1 year survival analysis, which is also suggested by Jørgensen et al. (1995a) and Anderson et al. (1994). In this thesis we use 10 year censoring. In (Andersen et al., 2005b), intermittent claudication and body temperature on admission are not included as potential risk factors.

Suggestions for future work We have many suggestions for improvements and interesting experiments. We include some of these in the list below to propose directions for future work.

Sensitivity analysis Using a semi-parametric approach to estimate missing values and model parameters, a thorough sensitivity analysis should be conducted, e.g. by changing the order of conditioning in the risk factor/missingness distributions, comparing various main effects and interaction models for the risk factor/missingness, and explore how the algorithms robustness is affected by the implementation of BMA. Furthermore, the effect of using different distributions to model continuous variables should be explored.

Explore new survival distribution We should explore alternatives to CPH and CR to model the distribution of survival times, e.g. using Weibull/ Lognormal distributions or Poisson regression models.

MI using BNs In this work we used the BN samples to infer a single, augmented data set with a set of estimated missing values weighted by posterior probabilities. Another approach would be to draw a sample (BN) using the posterior probability distribution, and then use this network to produce an augmented data set. Repeating this, we could obtain N augmented data sets that we could explore using BMA.

Neural Networks As shown in (Bakker et al., 2000), (Bakker et al., 2004), and Section 2.4.3, we can use neural networks to express the CPH model. With more hidden units, we can model non-proportional hazards, and use the network to estimate missing values in the data set.

MCMC solution We could use MCMC to sample everything! - or at least the baseline hazard, the risk factor coefficients, and the missing values. With enough samples, we could obtain approximations as accurate as we desire.

BN to model missingness In the same way we used BNs to model the risk factor distributions, we could use BNs to model the missingness.

Short- and long-term survival As shown by Andersen et al. (2005b), different risk factors predict the short- and long-term survival. Hence, we should explore our data set using different censoring dates, e.g. to identify risk factors responsible for instant survival (first week), short-term survival (1 year), medio-term survival (5 years), and long-term survival (10 years) to obtain a more varied solution.

Other risk factors Diet, exercise, stroke in relatives, social relations, marital status, work, cholesterol, Body Mass Index (BMI) are examples of potential risk factors that should definitively be explored as predictors of the survival time. Petersen et al. (2006) suggest that increasing BMI is associated with decreasing risk of post-stroke death; risk of death decreased 5% pr. unit increase in BMI up to a certain level.

Reference group The study should include a reference group of people from the same community, who had *not* experienced a stroke. Then we would compare the survival times of the stroke patients with the survival times of “normal” people to adjust for differences in survival times for males vs. females etc.

Appendix A

Occams' Razor Applied to Socrates

Taking up the philosophical dialogue between Doofus, Socrates, and Ignoramus[1], Doofus claims to have seen ducks down by the lake, but Ignoramus starts questioning what they really were. In response, Doofus says: “If it walks like a duck and talks like a duck, it is a duck! Surely this is evident”. However, Socrates suggests that it could have been men that have learned to walk and talk like ducks. Hence, we have two models that describe the observations equally well, but surely most people would agree that Doofus' duck model seems more plausible.

Ignoring all other factors such as height and weight, we cannot ignore Socrates' proposal, but let us see if the observations made by Doofus is in favor of the ducks or the men. We make the following assumptions:

- Men also knows how to walk and talk like humans (and not any other creature).
- Men are equally likely to walk and talk like men and ducks (remember that Socrates lived in 469-399 BC).
- Ducks only know how to walk and talk like ducks.

[1] see http://www.psych.upenn.edu/ fjgil/doofus.htm

- We do not want to favor any of the two models over one another.

Given this, we can setup the following Bayesian analysis:

$$\frac{p(M_{ducks}|\boldsymbol{D})}{p(M_{men}|\boldsymbol{D})} = \frac{p(M_{ducks})p(\boldsymbol{D}|M_{ducks})}{p(M_{men})p(\boldsymbol{D}|M_{men})} = \frac{0.5 \times 1}{0.5 \times 0.5} = 2$$

Hence, the observation is indeed in the favor of ducks, with a ratio of 2:1, and we can rephrase Doofus' statement: "If it walks like a duck and talks like a duck, it probably is a duck!". Case closed.

Appendix B

Sex Differences in Stroke Survival: 10-Year Follow-up of the Copenhagen Stroke Study Cohort

Sex Differences in Stroke Survival: 10-Year Follow-up of the Copenhagen Stroke Study Cohort

Morten Nonboe Andersen, MS,* Klaus Kaae Andersen, MS, PhD,†
Lars Peter Kammersgaard, MD,‡ and Tom Skyhøj Olsen, MD, PhD‡

Background: Although diverging, most studies show that sex has no significant influence on stroke survival. *Methods:* In a Copenhagen, Denmark, community all patients with stroke during March 1992 to November 1993 were registered on hospital admission. Stroke severity was measured using the Scandinavian Stroke Scale (0-58); computed tomography determined stroke type. A risk factor profile was obtained for all including ischemic heart disease, hypertension, diabetes mellitus, atrial fibrillation, previous stroke, smoking, and alcohol consumption. Date of death was obtained within a 10-year follow-up period. Predictors of death were identified using a Cox proportional hazards model. *Results:* Of 999 patients, 559 (56%) were women and 440 (44%) were men. Women were older (77.0 *v* 70.9 years; $P < .001$) and had more severe strokes (Scandinavian Stroke Scale: 36.1 *v* 40.5; $P < .001$). Age-adjusted risk factors showed no difference between sexes for ischemic heart disease, hypertension, atrial fibrillation, diabetes mellitus, and previous stroke. Men more often were smokers and alcohol consumers. Unadjusted survival in men and women did not differ: 70.3% versus 66.7% (1-year), 40.0% versus 38.9% (5-year), and 17.4% versus 18.7% (10-year), respectively. Adjusting for age, stroke severity, stroke type, and risk factors, women had a higher probability of survival at 1 year (hazard ratio 1.47, 95% confidence interval 1.10-2.00); 5 years (hazard ratio 1.47, 95% confidence interval 1.23-1.76); and 10 years (hazard ratio 1.49, 95% confidence interval 1.28-1.76). Before 9 months poststroke, no difference in survival was seen. Severity of stroke had the same effect on sex. *Conclusion:* Stroke is equally severe in men and women. Short-term survival is the same. Having survived stroke, women, however, live longer. **Key Words:** Stroke—sex—mortality—prognosis.

Sex has no significant influence on survival after stroke in most studies.[1-21] In a minority of studies, survival is significantly better for men than women[22-25] and vice versa.[26-29] This finding is surprising because women, because of their markedly longer life expectancy,[30] would be anticipated to encompass at least a better long-term survival. Controversies about the influence of sex on stroke outcome may reflect diversity among studies in respect to design, sample size, and follow-up.

Current recommendations usually support equal treatments for men and women, but recent research increasingly points to the need of individualization.[31,32] A clarification of possible differences in outcome between sexes is, therefore, still needed.

We hypothesized that a better survival of women would emerge if initial stroke severity measured by a validated stroke scale and a cardiovascular risk factor evaluation were encountered in a sizable study with a lengthy follow-up.

From *Informatics and Mathematical Modeling, Section for Intelligent Signal Processing, and †Statistics, Technical University of Denmark, Lyngby; and ‡The Stroke Unit, Hvidovre University Hospital, Denmark.

Received May 31, 2005; accepted June 10, 2005.

Address correspondence to Tom Skyhøj Olsen, MD, PhD, The Stroke Unit, Department of Neurorehabilitation, Hvidovre University Hospital, Kettegaard Allé 30, DK-2650 Hvidovre, Denmark. E-mail: Tom.Skyhoej.Olsen@hh.hosp.dk.

1052-3057/$—see front matter

doi:10.1016/j.jstrokecerebrovasdis.2005.06.002

In a community-based cohort of 999 patients hospitalized with acute stroke, we recorded prospectively data on 10-year survival from the acute admission on March 1992 until November 1993. Based on initial stroke severity measured by a validated stroke scale and a thorough cardiovascular risk factor profile, we studied the influence of sex on short- and long-term stroke survival.

Methods

The study was community-based and prospective. In a well-defined area of Copenhagen, Denmark, with 240,000 inhabitants, all having a stroke were admitted to a 62-bed stroke department at the same hospital. The inclusion period was March 1992 to November 1993. No preselection of patients was performed, as all who had a stroke in the area were brought to the stroke department of our hospital, regardless of age, stroke severity, or comorbid diseases. In our community, all who experience symptoms of a stroke or transient ischemic attack (including nursing home residents) are urged to go to the hospital immediately. General practitioners are instructed to hospitalize all patients with stroke or transient ischemic attack. Hospital care is free, and a very high proportion (88%)[33] of the patients with stroke in the area were admitted to this hospital during the time of inclusion. On admission, all underwent a standardized program including computed tomography scan, electrocardiography, and a cardiovascular risk factor evaluation using a standardized questionnaire. Information was obtained from relatives or caregivers if needed.

Stroke was defined according to the World Health Organization (WHO) criteria.[34] Transient ischemic attack or subarachnoid hemorrhage was not included. On admission, the Scandinavian Stroke Scale (SSS) was used to assess stroke severity. SSS evaluates level of consciousness; eye movement; power in arm, hand, and leg; orientation; aphasia; facial paresis; and gait on a total score from 0 (worst) to 58 (best).[35] Computed tomography determined stroke type (hemorrhage/infarct).

The following prognostic factors were investigated in the statistical analyses: age, sex, initial stroke severity (SSS), diabetes mellitus (DM), atrial fibrillation (AF), ischemic heart disease (IHD), hypertension, previous stroke, pre-existing disability, alcohol consumption, and smoking.

DM was considered present if a patient had known DM on admission or if plasma glucose level was greater than 11 mmol/L on admission or during the hospital stay. AF was diagnosed if present on admission electrocardiogram. Information concerning other disabling disease was obtained on admission and included disabling diseases other than previous stroke (e.g., amputation, multiple sclerosis, severe dementia, heart failure, latent or persistent respiratory insufficiency). IHD was present if a patient had a history of IHD, or had IHD diagnosed during the hospital stay. Hypertension was present if a patient received antihypertensive treatment before admission, or if hypertension was diagnosed during hospital stay by repeated detection of blood pressure 160/95 mm Hg or higher. Smoking was coded if a patient smoked any kind of tobacco on a daily basis. Ex-smokers were coded as nonsmokers. Intake of alcohol intake was coded if consumed daily.

Follow-up

For patients who had died, information on date of death within 10 years after the stroke onset was obtained from the Danish Central Registry of Persons. The follow-up was performed during the year 2003 ending November 3 (censoring date). Six patients had immigrated to another country and were lost on follow-up.

Statistical Analyses

Statistical analyses were performed with the a statistical software package (SPSS, Statistical Package for the Social Sciences, SPSS Inc, Chicago, IL). Difference in age and SSS score for sex was analysed using a standard *t* test. Logistic regression models were applied to calculate an age-adjusted estimate of the odds ratio between sex and all possible risk factors, each coded as binary variables. Independent predictors of death were identified using the Cox proportional hazards (CPH) model. Significance of predictors was based on the probability of the Wald statistic and a significance level of 5%. To assess whether the baseline hazard functions were proportional log-minus-log plots were performed for each variable. Log-linearity of age and SSS score was tested by elaborating these variables and performing a likelihood ratio test. The study was approved by the ethics committee.

Results

Of the 999 patients included, 559 (56%) were women and 440 (44%) were men. Mean age was higher in women (77.0 *v* 70.9 years; $P < .001$) and stroke severity expressed by the mean SSS score was more severe in women (36.1 *v* 40.5; $P < .001$). Table 1 shows the age-adjusted odds ratio values of potential cardiovascular risk factors for men relative to women. AF did not differ for sex (19.9 *v* 12.1; $P = .205$), but hemorrhage (9.4% *v* 5.2%; $P = .019$) was more often in women. Men were more often smokers (53.8% *v* 36.5%; $P = .014$) and had daily alcohol consumption (49% *v* 16.7%; $P < .001$). No significant difference between sexes was found for hypertension, IHD, previous stroke, DM, or AF.

Survival

Three subanalyses were done for end points 1, 5, and 10 years poststroke. Unadjusted survival in men and

Table 1. *Age-adjusted odds ratio of risk factors for men compared with women*

Variable	Women		Men		OR	95% CI
	Yes/no	Percentage	Yes/no	Percentage		
Hypertension	172/341	33.5	134/289	31.7	0.838	0.630-1.116
Known ischemic heart disease	102/404	20.2	87/329	20.9	1.165	0.836-1.623
Previous stroke	101/421	19.3	94/332	22.1	1.272	0.917-1.764
Other disabling disease	130/398	26.6	75/354	17.5	0.720	0.518-1.000
Alcohol consumption	75/375	16.7	188/196	49.0	0.962	0.949-0.976
DM	72/447	13.9	76/353	17.7	1.244	0.866-1.787
Smoking	162/282	36.5	204/175	53.8	1.463	1.081-1.982
Atrial fibrillation	109/440	19.9	53/385	12.1	0.786	0.542-1.141
Hemorrhage	42/404	9.4	19/345	5.2	0.497	0.278-0.890

CI, Confidence interval; *DM,* diabetes mellitus; *OR,* odds ratio.

women did not differ significantly: men 70.3%, women 66.7% (1-year); men 40.0%, women 38.9% (5-year); and men 17.4%, women 18.7% (10-year).

One-Year Survival

The variables in Table 2 were found significant in the CPH model for 1-year survival (*P* value, hazard ratio [HR], 95% confidence interval). Women had a significantly higher probability of survival (HR 1.465). A 10-year increase in age decreased the probability of survival (HR 1.460) whereas a 10-point increase in the SSS score increased the probability of survival (HR 0.621). DM (HR 2.085), AF (HR 1.438), and hemorrhage (HR 1.980) decreased the probability of 1-year survival.

Five-Year Survival

The variables in Table 3 were found significant in the CPH model for 5-year survival (*P* value, HR, 95% confidence interval). Women had a significantly higher probability of survival (HR 1.471). A 10-year increase in age decreased the survival probability (HR 1.649 per 10 years) whereas a 10-point increase SSS score increased the survival probability (HR 0.696 per 10 points). DM (HR 1.440), AF (HR 1.339), previous stroke (HR 1.334), and other disabling disease (HR 1.306) decreased the probability of 5-year survival.

Ten-Year Survival

The variables in Table 4 were found significant in the CPH model for 10-year survival. The results of this analysis are almost identical to the 5-year survival analysis and still display a significantly higher probability of survival for women (HR 1.490).

Fig 1 illustrates the sex-specific CPH survival plot for 10-year survival.

To identify the cut-off point in the Cox regression analysis (i.e., the survival censoring point where sex becomes a significant explanatory variable) we increase the censoring by 1 month starting with a 1-month censoring. The analysis showed that sex became significant using 9-month censoring.

Separate models have been applied to analyze whether there is an interaction between sex and SSS score. The results of this analysis are shown in Table 5. It is seen that there is no significant interaction between sex and SSS score (i.e., severity of the stroke has the same effect on each sex).

Table 2. *Significant variables in 1-year survival Cox proportional hazards regression model*

	Parameter estimate, B	*P* value	Hazard ratio, Exp(B)	95% CI for Exp(B)	
				Lower	Upper
Age (units of 10 years)	0.378	<.001	1.460	1.234	1.728
SSS (units of 10 points)	−0.476	<.001	0.621	0.567	0.680
Diabetes	0.735	<.001	2.085	1.458	2.981
Atrial fibrillation	0.363	.053	1.438	0.996	2.077
Hemorrhage	0.683	.005	1.980	1.224	3.201
Sex	0.382	.016	1.465	1.075	1.999

CI, Confidence interval; *SSS,* Scandinavian Stroke Scale.

Table 3. *Significant variables in 5-year survival Cox proportional hazards regression model*

				95% CI for Exp(B)	
	Parameter estimate, B	*P* value	Hazard ratio, Exp(B)	Lower	Upper
Age (units of 10 years)	0.500	<.001	1.649	1.492	1.822
SSS (units of 10 points)	−0.362	<.001	0.696	0.661	0.733
Previous stroke	0.288	<.005	1.334	1.090	1.631
Other disabling disease	0.267	.009	1.306	1.069	1.594
Diabetes	0.365	.001	1.440	1.151	1.802
Atrial fibrillation	0.292	.008	1.339	1.080	1.661
Sex	0.386	<.001	1.471	1.228	1.762

CI, Confidence interval; *SSS,* Scandinavian Stroke Scale.

Discussion

Two main findings emerged from this study. Men and women are at the same risk of dying from a stroke. Having survived the stroke, however, women live longer than men. In other words, short-term stroke survival is the same for men and women whereas long-term stroke survival is markedly better for women.

Women and men differed in respect to important confounders. Women were older and had more severe strokes. This explains that short-term survival at a first glance appears to be significantly better for men and that long-term survival looks equal for both sexes. Moreover, women more often had other disabling diseases and hemorrhagic strokes whereas men more often were consumers of alcohol and tobacco. The lack of adjustment for one or more of these variables explains much of the diverging conclusions among studies. However, as we found no interaction between stroke severities and sex, our study shows that severity of stroke is not influenced by sex per se and differences in survival between men and women are determined by other factors.

The strength of this study is that it is prospective and community-based including all patients in a well-defined community hospitalized with stroke regardless age, stroke severity, or other complicating diseases. Moreover, the stroke admittance rate in the area is high and close to the incidence reported in population-based studies. A limitation is that patients who die at home are not included and this may underestimate mortality. However, the small number of patients with minor strokes not being admitted to hospital may counterbalance it. Finally, as a multivariate analysis was applied and because we had a sizeable study population and a lengthy follow-up, we consider bias to be of no major importance for the main conclusion of this study.

The difference in survival between sexes became significant at 9 months poststroke. However, it appears from the sex-specific survival plots that a difference in survival between sexes takes effect much earlier. On the other hand, it is also evident from these plots that there is no difference between sexes in the very acute state. Thus, our study does not point to the presence of a sex-specific ability to survive stroke per se. Findings from other large-scale studies on short-term survival are diverging: in the WHO MONICA populations the age-adjusted 28-day case fatality is higher among women.[24] Stroke severity and other prognostic confounders are, however, not recorded in these studies. In several studies the level of consciousness or degree of paresis were used as markers of stroke severity and no difference in 1- to 3-month

Table 4. *Significant variables in 10-year survival Cox proportional hazards regression model*

				95% CI for Exp(B)	
	Parameter estimate, B	*P* value	Hazard ratio, Exp(B)	Lower	Upper
Age (units of 10 years)	0.481	<.001	1.618	1.490	1.757
SSS (units of 10 points)	−0.299	<.001	0.742	0.709	0.776
Previous stroke	0.248	.006	1.281	1.072	1.531
Other disabling disease	0.283	.002	1.328	1.114	1.583
Diabetes	0.357	<.001	1.429	1.178	1.734
Atrial fibrillation	0.290	.003	1.336	1.100	1.622
Sex	0.398	<.001	1.490	1.278	1.736

CI, Confidence interval; *SSS,* Scandinavian Stroke Scale.

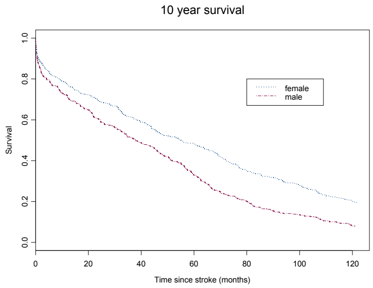

Figure 1. *Sex-specific survival curve for 10-year mortality.*

survival between sexes was observed when adjusting for these and other confounding variables.[1-7] In a study using the National Institutes of Health Stroke Scale as marker of stroke severity there was no sex-specific difference in 3-month survival when adjusting for this and other relevant confounders as done also in our study. In the Rochester, Minn, population stroke severity was determined retrospectively from hospital records and no sex-specific difference in 3-month survival was found.[10] On the other hand, in a Dutch study using Glasgow Coma Scale as marker of stroke severity women had a better 6-month survival,[26] whereas in a Polish study using level of consciousness as marker of stroke severity 2-week survival was poorer in women.[23]

In our study survival is markedly better in women 9 months poststroke and onward. Women continuously have a 1.5 better chance of being alive up to 10 years after the stroke. Several large-scale studies did not find any sex-specific difference in stroke mortality in studies with 1,[9] 10,[10] and even 20[17] years follow-up, but these studies, except for age, did not adjust for stroke severity or other confounders of importance for stroke survival. Two Swedish studies[13,19] used validated stroke severity scores and did not find any sex-specific difference in 1-year[13] and 3-year[19] survival when adjusting for stroke severity and other relevant confounders. In other studies stroke severity was estimated on the basis of consciousness or various neurologic deficits; 1-year[11,12,14] and 3-year[20,21] survival did not differ between sexes. In the Rochester, Minn, population[10,18] sex did not influence 5-year stroke survival whereas in the Framingham population[28] 5-year survival was better among women.

Table 5. *Test for interaction between sex and Scandinavian Stroke Scale scores*

	SSS (female) parameter estimate, B_1	SSS (male) parameter estimate, B_2	Wald test statistics	*P* value
1 y	0.622	0.619	0.0326	.97
5 y	0.728	0.757	−0.4601	.65
10 y	0.770	0.785	−0.2749	.78

SSS, Scandinavian Stroke Scale.

Our study is the first large-scale study with a follow-up as long as 10 years where stroke severity at stroke onset and a thorough cardiovascular risk factor profile were determined prospectively using a validated stroke severity scale. Other studies with a long follow-up either did not measure stroke severity, stroke severity was determined retrospectively from hospital records, or stroke severity was estimated without using a stroke scale. There was no interaction between stroke severity and sex, but stroke severity was a strong predictor not only of short-term survival, but of long-term survival as well. Information of stroke severity is, thus, important for analyzing predictors of stroke survival.

In the industrialized world women live 5 to 7 years longer than men,[30] which is in agreement with the result of our study. Women experienced stroke on average 6 years later than men. This is undoubtedly the key to the understanding of the better long-term survival of women with stroke. Women also experience myocardial infarction several years later than men.[36] Women, therefore, experience fatal cardiovascular diseases later than men and, hence, live longer than men even if they have had a stroke. Higher consumption of tobacco and alcohol further contributes to earlier occurrence of cardiovascular disease in men. The great diversity among studies in respect to design, sample size, follow-up, and results calls, however, for further study.

In conclusion, stroke is equally severe in men and women and short-term survival is the same for men and women. Having survived stroke, women, however, live longer than men, most certainly because of their lower risk of a subsequent cardiovascular event.

References

1. Collins TC, Petersen NJ, Menke TJ, et al. Short-term, intermediate-term, and long-term mortality in patients hospitalized for stroke. J Clin Epidemiol 2003;56:81-87.
2. Truelsen T, Grønbæk M, Schnohr P, et al. Stroke case fatality in Denmark from 1977 to 1992: The Copenhagen city heart study. Neuroepidemiology 2002;21:22-27.
3. Immonen-Räihä P, Mähönen M, Tuomilehto J, et al. Trends in case-fatality of stroke in Finland during 1983 to 1992. Stroke 1997;28:2493-2499.
4. Roquer J, Campello AR, Gomis M. Sex differences in first-ever acute stroke. Stroke 2003;34:1581-1585.
5. Mayo NE, Nevill D, Kirkland S, et al. Hospitalization and case fatality rates for stroke in Canada from 1982 through 1991. Stroke 1996;27:1215-1220.
6. Carlo AD, Lamassa M, Baldereschi M, et al. Sex differences in the clinical presentation, resource use, and

3-month outcome of acute stroke in Europe: Data from a multicenter multinational hospital-based registry. Stroke 2003;34:1114-1119.

7. Glader E-L, Stegmayr B, Norrving B, et al. Sex differences in management and outcome after stroke: A Swedish national perspective. Stroke 2003;34:1970-1975.
8. Weimar C, Ziegler A, König IR, et al. Predicting functional outcome and survival after acute ischemic stroke. J Neurol 2002;249:888-895.
9. Hollander M, Koudstaal PJ, Bots ML, et al. Incidence, risk, and case fatality of first ever stroke in the elderly population: The Rotterdam study. J Neurol Neurosurg Psychiatry 2003;74:317-321.
10. Petty GW, Brown RD, Whisnant JP, et al. Survival and recurrence after first cerebral infarction: A population-based study in Rochester, Minnesota, 1975 through 1989. Neurology 1998;50:208-216.
11. Anderson CS, Jamrozik KD, Broadhurst RJ, et al. Predicting survival for 1 year among different subtypes of stroke: Results from the Perth community stroke study. Stroke 1994;25:1935-1944.
12. Vemmos KN, Bots ML, Tsibouris PK, et al. Prognosis of stroke in the south of Greece: 1 year mortality, functional outcome and its determinants; the Arcadia stroke registry. J Neurol Neurosurg Psychiatry 2000;69:595-600.
13. Appelros P, Nydevik I, Viitanen M. Poor outcome after first-ever stroke: Predictors for death, dependency, and recurrent stroke within the first year. Stroke 2003;34:122-126.
14. Devroey D, Casteren VV, Buntinx F. Registration of stroke through the Belgian sentinental network and factors influencing stroke mortality. Cerebrovasc Dis 2003; 16:272-279.
15. Térent A. Trends in stroke incidence and 10-year survival in Söderham, Sweden, 1975-2001. Stroke 2003;34: 1353-1356.
16. Kiyohara Y, Kubo M, Kato I, et al. Ten-year prognosis of stroke and risk factors for death in a Japanese community: The Hisayama study. Stroke 2003;34:2343-2348.
17. Hart C, Hole DJ, Smith GD. Risk factors and 20-year stroke mortality in men and women in the Renfrew/ Paisely study in Scotland. Stroke 1999;30:1999-2007.
18. Vernino S, Brown RD, Sejvar JJ, et al. Cause-specific mortality after first cerebral infarction: A population-based study. Stroke 2003;34:1828-1832.
19. Elneihoum AM, Göransson M, Falke P, et al. Three-year survival and recurrence after stroke in Malmö, Sweden: An analysis of stroke registry data. Stroke 1998;29:2114-2117.
20. Bonita R, Ford MA, Stewart AW. Predicting survival after stroke: A three-year follow-up. Stroke 1998;19:669-673.
21. Loor HI, Groenier KH, Limburg M, et al. Risk and causes of death in a community-based stroke population: 1 month and 3 years after stroke. Neuroepidemiology 1999;18:75-84.
22. Arboix A, Oliveres M, Garcia-Eroles L, et al. Acute cerebrovascular disease in women. Eur Neurol 2001;45: 199-205.
23. Czlonkowska A, Niewada M, El-Baroni IS, et al. High early case fatality after ischemic stroke in Poland: Exploration of possible explanations in the international stroke trial. J Neurol Sci 2002;202:53-57.
24. Thorvaldsen P, Asplund K, Kuulasmaa K, et al. Stroke incidence, case fatality, and mortality in the WHO MONICA project. Stroke 1995;26:361-367.
25. Brønnum-Hansen H, Davidsen M, Thorvaldsen P. Long-term survival and causes of death after stroke. Stroke 2001;32:2131-2136.
26. van Straten A, Reitsma JB, Limburg M, et al. Impact of stroke type on survival and functional health. Cerebrovasc Dis 2001;12:27-33.
27. Holroyd-Leduc JM, Kapral MK, Austin PC, et al. Sex differences and similarities in the management and outcome of stroke patients. Stroke 2000;31:1833-1837.
28. Sacco RL, Wolf PA, Kannel WB, et al. Survival and recurrence following stroke: The Framingham study. Stroke 1982;13:290-295.
29. Gresham GE, Kelley-Hayes, Wolf PA, et al. Survival and functional status 20 or more years after first stroke: The Framingham study. Stroke 1998;29:793-797.
30. United Nations: Demographic yearbook, 2000. New York, NY: United Nations Publications; 2000.
31. Sacco RL, Benjamin EJ, Broderick JP, et al. Risk factors. Stroke 1997;28:1507-1517.
32. Ayala C, Croft JB, Greenlund KJ, et al. Sex differences in US mortality rates for stroke and stroke subtypes by race/ethnicity and age, 1995-1998. Stroke 2002;33:1197-1201.
33. Jørgensen HS, Plesner A-M, Hübbe P, et al. Marked increase of stroke incidence in men between 1972 and 1990 in Frederiksberg, Denmark. Stroke 1992;23:1701-1704.
34. Stroke–1989. Recommendations on stroke prevention, diagnosis, and therapy: Report of the WHO task force on stroke and other cerebrovascular disorders. Stroke 1989; 20:1407-1431.
35. Scandinavian Stroke Study Group. Multicenter trial of hemodilution in ischemic stroke–background and study protocol. Stroke 1985;16(5):885-890.
36. Williams RI, Fraser AG, West RR. Gender differences in management after acute myocardial infarction: Not 'sexism' but a reflection of age at presentation. J Public Health 2004;26:259-263.

Bibliography

A. Agresti and D. Hitchcock. Bayesian inference for categorical data analysis: A survey. *Journal of the Italian Statistical Society, Statistical Methods and Applications*, 14:297–330, 2005.

S.M. Aji and R.J. McEliece. The generalized distributive law. *IEEE Trans. on Information Theory*, 46(2):325–343, 2000.

K.K. Andersen, L.P. Kammersgaard, H.G. Petersen ad M.N. Andersen, and T.S. Olsen. Intracerebral hematomas versus infarction: Stroke severity and risk factor profile. a danish nation-wide evaluation of 25 839 patients with acute stroke. Joint World Congress on Stroke. Cape Town, South Africa, October 2006a.

K.K. Andersen, L.P. Kammersgaard, and T.S. Olsen. Time related differences in mortality of patients with intracerebral hematomas and cerebral infarcts. a 4 year follow-up of 25 839 patients with acute stroke. Joint World Congress on Stroke. Cape Town, South Africa, October 2006b.

K.K. Andersen, T.S. Olsen, H.G. Petersen, and M.N. Andersen. On the importance of a stroke severity score in modelling mortality of stroke patients. Joint World Congress on Stroke. Cape Town, South Africa, October 2006c.

K.K. Andersen, H.G. Petersen, M.N. Andersen, L.P. Kammersgaard, and T.S. Olsen. Stroke in diabetics: Frequency, clinical characteristics and survival. a 4-year follow-up study of 24 121 patients with acute stroke. Joint World Congress on Stroke. Cape Town, South Africa, October 2006d.

M.N. Andersen, K.K. Andersen, L.P. Kammersgaard, and T.S. Olsen. Gender differences in stroke survival: 10-year follow-up of the copenhagen stroke study cohort. 14'th European Stroke Conference. Bologna, May 2005a.

M.N. Andersen, K.K. Andersen, L.P. Kammersgaard, and T.S. Olsen. Sex differences in stroke survival: 10-year follow-up of the copenhagen stroke study cohort. volume 14, pages 215–220, May 2005b.

M.N. Andersen, K.K. Andersen, H.G. Petersen, and T.S. Olsen. Using bayesian statistics to account for model uncertainty in survival analysis. a study of risk factors in 25 839 patients with acute stroke. Joint World Congress on Stroke. Cape Town, South Africa, October 2006e.

M.N. Andersen, K.K. Andersen, H.G. Petersen, and T.S. Olsen. Women survive stroke better than men. a study of gender-specific differences in survival of 25 839 patients with acute stroke. Joint World Congress on Stroke. Cape Town, South Africa, October 2006f.

P.K. Andersen, Ø. Borgan, R.D. Gill, and N. Keidin. *Statistical models based on counting processes.* New York: Springer-Verlag, 1993.

P.K. Andersen and R.D. Gill. Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, 10(4):1100–1120, 1982.

C.S. Anderson, K.D. Jamrozik, R.J. Broadhurst, and E.G. Stewart-Wynne. Predicting survival for 1 year among different subtypes of stroke. results from the perth community stroke study. *Stroke*, 25:1935–1944, 1994.

S. Arnborg, D. G. Corneil, and A. Proskurowski. Complexity of finding embeddings in a k-tree. *SIAM J. Alg. Disc. Meth.*, 8:277–284, 1987.

H. Attias. A variational bayesian framework for graphical models. In *Advances in Neural Information Processing Systems*, volume 12, pages 209–215, 2000.

B. Bakker, T. Heskes, J. Neijt, and B. Kappen. Improving cox survival analysis with a neural-bayesian approach. *Statistics in Medicine*, 23(19):2989–3012, 2004.

B. Bakker, B. Kappen, and T. Heskes. Survival analysis: A neural-bayesian approach. In *Proc. Artificial Neural Networks In Medicine And Biology*, pages 162–167. Springer, 2000.

M.J. Beal. *Variational algorithms for approximate Bayesian inference.* PhD thesis, University College London, 2003.

I. Beinlich, G. Suermondt, R. Chavez, and G. Cooper. The alarm monitoring system: A case study with two probabilistic inference techniques for belief networks. In *Proc. of the 2'nd European Conf. on Artificial Intelligence and Medicine*, volume 38, pages 247–256. Springer-Verlag, 1989.

R. Bender, T. Augustin, and M. Blettner. Generating survival times to simulate cox proportional hazards models. *Statistics in Medicine*, 25(11):1978–1979, 2006.

C.M. Bishop. *Pattern Recognition and Machine Learning.* Springer Verlag, 2006.

C.M. Bishop, D. Spiegelhalter, and J.M. Winn. Vibes: A variational inference engine for bayesian networks. In S. Thrun S. Becker and K. Obermeyer, editors, *Advances in Neural Information Processing Systems*, volume 12, pages 793–800, 2002.

G. Boysen and H. Christensen. Stroke severity determines body temperature in acute stroke. *Stroke*, 32(1):413–417, 2001.

N. Breslow. Covariance analysis of censored survival data. *Biometrics*, 30(1): 89–99, 1974.

J.P. Broderick, S.J. Phillips, W.M. O'Fallon, R.L. Frye, and J.P. Whisnant. Relationship of cardiac disease to stroke occurrence, recurrence, and mortality. *Stroke*, 23:1250–1256, 1992.

I.D.J. Bross. *Critical Levels, Statistical Language and Scientific Inference.* Holt, Rinehart & Winston of Canada, Ltd., 1971.

W. Buntine. A guide to the literature on learning probabilistic networks from data. *IEEE Trans. On Knowledge And Data Engineering*, 8:195–210, 1996.

W.L. Buntine. Operations for learning with graphical models. *Journal of Artificial Intelligence Research*, 2:159–225, 1994.

D.M. Chickering. Learning bayesian networks is np-complete. In D. Fisher and H.J. Lenz, editors, *Learning from Data: Artificial Intelligence and Statistics V*, pages 121–130. Springer-Verlag, 1996.

D. Collet. *Modelling Survival Data in Medical Research.* Chapman and Hall/CRC, 2. edition, 2003. http://www.crcpress.com/e_products/downloads/download.asp?cat_no=C3251.

G.F. Cooper. A bayesian method for learning belief networks that contain hidden variables. *Journal of Intelligent Information Systems*, 4(1):71–88, 1995.

G.F. Cooper. The bayesian structural em algorithm. In *Proc. of the 14'th Conf. on Uncertainty in Artificial Intelligence.* Morgan Kaufmann, 1998.

G.F. Cooper and E. Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309–347, 1992.

D.R. Cox. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society*, 34:187–220, 1972.

D.R. Cox and D. Oakes. *Analysis of Survival Data.* Chapman and Hall, New York, 1984.

L. Csato, M. Opper, and O. Winther. Tractable inference for probabilistic data models. *Complexity*, 8(4):64–68, 2003.

J. Dallal. Why p=0.05?, 2007. http://www.tufts.edu/∼gdallal/p05.htm.

O. Davidsen. Netdoktor.dk - apopleksi (slagtilfælde), 2007. http://www.netdoktor.dk/sygdomme/fakta/blodprophjerne.htm.

N.G. de Bruijn. *Asymptotic Methods in Analysis.* Dover, New York, NY, 1981.

A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, B(39): 1–38, 1977.

D.W. Hosmer DW and S. Lemeshow. *Applied Survival Analysis.* New York: Wiley, 1999.

M.I. Jordan (editor). *Learning in Graphical Models.* Kluwer Academic Publishers, 1998.

T. El-Hay. Efficient methods for exact and approximate inference in discrete graphical models. Master's thesis, The Hebrew University of Jerusalem, 2001.

B. Frey and N. Jojic. A comparison of algorithms for inference and learning in probabilistic graphical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2003.

N. Friedman. Learning belief networks in the presence of missing values and hidden variables. In *Proc. of the 14'th Int. Conf. on Machine Learning*, 1997.

N. Friedman. The bayesian structural em algorithm. In *Proc. of the 14'th Conf. on Uncertainty in Artificial Intelligence*, pages 129–138. Morgan Kaufmann, 1998.

G.M. Furnival and R.W. Wilson. Regression by leaps and bounds. *Technometrics*, 42(1):69–79, 2000.

D. Geiger, D. Heckerman, and C. Meek. Asymptotic model selection for directed networks with hidden variables. In *Proc. of 12'th Conf. on Uncertainty in Artificial Intelligence*, pages 283–290. Morgan Kaufmann, 1996. Also appears as Technical Report MSR-TR-96-07, Microsoft Research.

D. Geiger and C. Meek. Structured variational inference procedures and their realizations. In Z. Ghahramani and R. Cowell, editors, *Proc. of 10'th Int. Workshop on Artificial Intelligence and Statistics.* The Society for Artificial Intelligence and Statistics, 2005.

A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis.* Chapman and Hall/CRC, 2004.

W.R. Gilks, D.J. Spiegelhalter, and S. Richardson. *Markov Chain Monte Carlo in Practice.* CRC Press, 1996.

P.M. Grambsch and T.M. Therneau. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81:515–526, 1994.

A.L. Gulløv, B.G. Koefoed, P. Petersen, T.S. Pedersen, E.D. Andersen, J. Godtfredsen, and G. Boysen. Fixed minidose warfarin and aspirin alone and in combination vs adjusted-dose warfarin for stroke prevention in atrial fibrillation second copenhagen atrial fibrillation, aspirin, and anticoagulation study. *Archives of Internal Medicine*, 158:1513–1521, 1998.

H. Guo and W. Hsu. A survey on algorithms for real-time bayesian network inference. In *The joint AAAI-02/KDD-02/UAI-02 workshop on Real-Time Decision Support and Diagnosis Systems, Edmonton, Alberta, Canada*, 2002.

D. Heckerman. A tractable inference algorithm for diagnosing multiple diseases. In *Proc. of 5'th Conf. on Uncertainty in Artificial Intelligence*, pages 163–171. Elsevier, 1989. Also appears as Technical Report KSL-89-36, Knowledge Systems Laboratory.

D. Heckerman. A tutorial on learning with bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research, 1995.

D. Heckerman and D. Chickering. A comparison of scientific and engineering criteria for bayesian model selection. *Statistics and Computing*, 10:55–62, 2000. Also appears as Technical Report MSR-TR-96-12, Microsoft Research.

D. Heckerman, D. Geiger, and D. Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995. Also appears as Technical Report MSR-TR-94-09, Microsoft Research.

A.H. Herring and J.G. Ibrahim. Likelihood-based methods for missing covariates in the cox proportional hazards model. *Journal of the American Statistical Association*, 96:292–302, 2001.

A.H. Herring, J.G. Ibrahim, and S.R. Lipsitz. Non-ignorable missing covariate data in survival analysis a case-study of an int. breast cancer study group trial. *Journal of the Royal Statistical Society*, 53(2):293–310, 2004.

J.A. Hoeting, D. Madigan, A.E. Raftery, and C.T. Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–417, 1999.

C. Huang and A. Darwiche. Inference in belief networks: A procedural guide. *Int. Journal of Approximate Reasoning*, 15(3):225–263, 1994.

R. Hubbard and J.S. Armstrong. Why we don't really know what "statistical significance" means: A major educational failure. 2005.

J.G. Ibrahim, M.H. Chen, and D. Sinha. *Bayesian Survival Analysis.* Springer, 2005.

T.S. Jaakkola. *Variational Methods for Inference and Estimation in Graphical Models.* PhD thesis, Massachusetts Institute of Technology, 1997.

F.V. Jensen. *An Introduction to Bayesian Networks.* Springer Verlag, New York, 1996.

R.A. Johnson. *Probability and Statistics for Engineers.* Prentice Hall, 7 edition, 2005.

M.I. Jordan, Z. Ghahramani, T.S. Jaakkola, and L.K. Saul. *An introduction to variational methods for graphical models.* MIT Press, 1998.

H.S. Jørgensen, L.P. Kammersgaard, H. Nakayama, H.O. Raaschou, K. Larsen, P. Hubbe P, and T.S. Olsen. Treatment and rehabilitation on a stroke unit improves 5-year survival: A community-based study. *Stroke*, 30(5):930–933, 1999a.

H.S. Jørgensen, H. Nakayama, L.P. Kammersgaard, H.O. Raaschou, and T.S. Olsen. Predicted impact of intravenous thrombolysis on prognosis of general population of stroke patients: simulation model. *BMJ*, 319:288–289, 1999b.

H.S. Jørgensen, H. Nakayama, H.O. Raaschou, and T.S. Olsen. Effect of blood pressure and diabetes on stroke in progression. *The Lancet*, 344:156–159, 1994a.

H.S. Jørgensen, H. Nakayama, H.O. Raaschou, and T.S. Olsen. Stroke in patients with diabetes. the copenhagen stroke study. *Stroke*, 25(10):1977–1984., 1994b.

H.S. Jørgensen, H. Nakayama, H.O. Raaschou, and T.S. Olsen. Intracerebral hemorrhage versus infarction: stroke severity, risk factors, and prognosis. *Annals of Neurology*, 38(1):45–50, 1995a.

H.S. Jørgensen, H. Nakayama, H.O. Raaschou, and T.S. Olsen. Progressive apoplexy. incidence, risk factors and prognosis: The copenhagen stroke study. *Ugeskrift for Læger*, 157(25):3619–3622, 1995b.

H.S. Jørgensen, H. Nakayama, H.O. Raaschou, and T.S. Olsen. Acute stroke care and rehabilitation: An analysis of the direct cost and its clinical and social determinants. *Stroke*, 28:1138–1141, 1997.

H.S. Jørgensen, H. Nakayama, J. Reith, H.O. Raaschou, and T.S. Olsen. Acute stroke with atrial fibrillation. *Stroke*, 27:1765–1769, 1996a.

H.S. Jørgensen, J. Reith, H. Nakayama, L.P. Kammersgaard, H.O. Raaschou, and T.S. Olsen. What determines good recovery in patients with the most severe strokes? *Stroke*, 30:2008–2012, 1999c.

H.S. Jørgensen, J. Reith, P.M. Pedersen, H. Nakayama, and T.S. Olsen. Body temperature and outcome in stroke patients. *The Lancet*, 347:422–425, 1996b.

L.P. Kammersgaard, H.S. Jørgensen, J. Reith, H. Nakayama, P.M. Pedersen, and T.S. Olsen. Short- and long-term prognosis for very old stroke patients: The copenhagen stroke study. *Age and Ageing*, 33(2):149–154, 2004.

L.P. Kammersgaard, H.S. Jørgensen, J.A. Rungby, H. Nakayama J. Reith, U.J. Weber, J. Houth, and T.S. Olsen. Admission body temperature predicts long-term mortality after acute stroke. *Stroke*, 33:1759–1762, 2002.

L.P. Kammersgaard and T.S. Olsen. Cardiovascular risk factors and 5-year mortality in the copenhagen stroke study. *Cerebrovascular Diseases*, 21:187–193, 2006.

H.J. Kappen. The cluster variation method for approximate reasoning in medical diagnosis. *Modeling Bio-medical signals*, pages 3–16, 2002.

R.E. Kass and A.E. Raftery. Bayes factors. Technical Report 254, 571, University of Washington, Carnegie-Mellon University, 1994.

R.E. Kass, L. Tierney, and J.B. Kadane. *Bayesian and Likelihood Methods in Statistics and Econometrics*, chapter The Validity of Posterior Expansions Based on Laplace's Method, pages 473–488. Amsterdam: North-Holland, 1990.

U. Kjaerulff. Triangulation of graphs – algorithms giving small total state space. Technical Report TR R 90-09, University of Aalborg, 1990.

M.W. Knuiman, T.A. Welborn, and D.E. Whittall. An analysis of excess mortality rates for persons with non-insulin-dependent diabetes mellitus in western australia using the cox proportional hazards model. *American Journal of Epidemiologi*, 135:638–648, 1992.

L. Kuo and A.F.M. Smith. Bayesian computations in survival models via the gibbs sampler. In J.P. Klein and P.K. Goel, editors, *Survival Analysis: State of the Art*, pages 11–24, 1992. Proceedings of the NATO Advanced Research Workshop on Survival Analysis and Related Topics, Columbus, Ohio.

J. Larsen. Basics of bayesian learning - basically bayes, September 2006.

S.L. Lauritzen and D.J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *Journal of the Royal Statistical Society*, 50:157–224, 1988.

J.F. Lawless and K. Singhal. Efficient screening of nonnormal regression models. *Biometrics*, 43:318–327, 1978.

E.T. Lee and J.W. Wang. *Statistical Methods for Survival Data Analysis.* Wiley Series in Probability and Statistics. Wiley, 3 edition, 2003.

I-M Lee, C.H. Hennekens, K. Berger, J.E. Buring, and J.E. Manson. Exercise and risk of stroke in male physicians. *Stroke*, 30:1–6, 1999.

T. Leong, S.R. Lipsitz, and J.G. Ibrahim. Incomplete covariates in the cox model with applications to biological marker data. *Journal of the Royal Statistical Society*, 50 (4):467–484, 2001.

D.V. Lindley. A statistical paradox. *Biometrika*, 44:187–192, 1957.

R.J.A. Little and D.B. Rubin. *Statistical Analysis with Missing Data.* J. Wiley & Sons, New York, 1987.

J. S. Liu. *Monte Carlo Strategies in Scientific Computing.* Springer-Verlag, New York, 2001.

S. Lonial. Intro to myeloma, 2005. http://www.multiplemyeloma.org/about_myeloma/.

D.J.C. MacKay. *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press, 2003.

D. Madigan and A.E. Raftery. Model selection and accounting for model uncertainty in graphical models using occam's window. *Journal of the American Statistical Association*, 89:1335–1346, 1994a.

D. Madigan and A.E. Raftery. Model selection and accounting for model uncertainty in graphical models using occam's window. Technical report, University of Washington, 1994b.

D. Madigan, A.E. Raftery, J.C. York, J.M. Bradshaw, and R.G. Almond. Strategies for graphical model selection. In *Proc. of the 4'th Int. Workshop on Artificial Intelligence and Statistics*, pages 361–366, 1993.

D. Madigan and J. York. Bayesian graphical models for discrete data. *Int. Statistical Review*, 63(2):215–232, 1995.

T. Martinussen. Cox regression with incomplete covariate measurements using the em-algorithm. *The Scandinavian Journal of Statistics*, 26:479–491, 1999.

A.J. Miller. *Subset Selection in Regression.* Chapman & Hall, 1990.

T.P. Minka. Bayesian model averaging is not model combination. MIT Media Lab note, 2000.

T.P. Minka. *A family of algorithms for approximate Bayesian inference.* PhD thesis, Massachusetts Institute of Technology, 2001a.

T.P. Minka. A family of algorithms for approximate bayesian inference - thesis defense, 2001b.

Q.D.J. Morris. *Practical Probabilistic Inference.* PhD thesis, Massachusetts Institute of Technology, 2002.

K. Murphy. An introduction to graphical models. 2001a.

K.P. Murphy. From belief propagation to expectation propagation. 2001b.

M.N. Rosenbluth. A.H. Teller N. Metropolis, A.W. Rosenbluth and E. Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1092, 1953.

H. Nakayama, H.S. Jørgensen, H.O. Raaschou, and T.S. Olsen. The influence of age on stroke outcome: The copenhagen stroke study. *Stroke*, 25(4):808–813, 1994.

N. Nakayama, H.S. Jørgensen, P.M. Pedersen, H.O. Raaschou, and T.S. Olsen. Prevalence and risk factors of incontinence after stroke. *Stroke*, 28:58–62, 1996.

National Diabetes Information Clearinghouse NDCI. Diabetes, heart disease, and stroke. NIH Publication No. 06 5094, 2005. http://diabetes.niddk.nih.gov/dm/pubs/stroke/DM_Heart_Stroke.pdf.

R.M. Neal. Probabilistic inference using markov chain monte carlo methods. Technical Report CRG-TR-93-1, University of Toronto, 1993.

S.F. Nielsen. Survival analysis with coarsely observed covariates. *SORT*, 27(1): 41–64, 2003.

T.S. Olsen, K.K. Andersen, M.N. Andersen, and H.G. Petersen. Hemorrhagic strokes in patients with atrial fibrillation: Frequency, clinical characteristics and prognosis. Joint World Congress on Stroke. Cape Town, South Africa, October 2006.

M. Opper and O. Winther. Expectation consistent free energies for approximate inference. In *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2004.

M. Opper and O. Winther. Expectation consistent approximate inference. *Journal of Machine Learning Research*, 2005.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, 1988.

H.G. Petersen, K.K. Andersen, M.N. Andersen, L.P. Kammersgaard, and T.S. Olsen. Body mass index (bmi) and survival after stroke. Joint World Congress on Stroke. Cape Town, South Africa, October 2006.

L.R. Rabiner and B.H. Juang. An introduction to hidden markov models. *IEEE ASSP Magazine*, pages 4–16, 1986.

A.E. Raftery. Approximate bayes factors and accounting for model uncertainty in generalized linear models. Technical Report 255, University of Washington, 1994.

A.E. Raftery. Bayesian model selection in social research (with discussion). *Sociological Methodology*, pages 111–196, 1995.

A.E. Raftery. Approximate bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika*, 83:251–266, 1996.

A.E. Raftery, D. Madigan, and C.T. Volinsky. Accounting for model uncertainty in survival analysis improves predictive performance (with discussion). *Bayesian Statistics*, 5:323–349, 1995.

M. Ramoni and P. Sebastiani. Robust learning with missing data. *Machine Learning*, pages 147–170, 2001.

J. Reith, H.S. Jørgensen, H. Nakayama, H.O. Raaschou, and T.S. Olsen. Seizures in acute stroke: Predictors and prognostic significance. *Stroke*, 28:1585–1589, 1997.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6: 461–464, 1978.

C. Shan. Model selection for belief networks when learning with incomplete data. 2001.

D.J. Spiegelhalter and S.L. Lauritzen. Sequential updating of conditional probabilities on directed graphical structures. *Networks*, 20:579–605, 1990.

J.A.C. Sterne. Sifting the evidence - what's wrong with significance tests? *BMJ*, 322(7280):226–231, 2001.

B. Thiesson. Score and information for recursive exponential models with incomplete data. In D. Geiger and P.P. Shenoy, editors, *Proc. of the 13'th Conf. on Uncertainty in Artificial Intelligence*, pages 453–463. Morgan Kaufmann, 1997.

B. Thiesson, C. Meek, and D. Heckerman. Accelerating em for large databases. Technical Report MSR-TR-99-31, Microsoft Research, 1999.

J. Tuomilehto, D. Rastenyt, P. Jousilahti, C. Sarti, and E. Vartiainen. Diabetes mellitus as a risk factor for death from stroke prospective study of the middle-aged finnish population. *Stroke*, 27:210–215, 1996.

V. Viallefont, A.E. Raftery, and S. Richardson. Variable selection and bayesian model averaging. *Statistics in Medicine*, 20:3215–3230, 2001.

C.T. Volinsky. *Bayesian Model Averaging for Censored Survival Models*. PhD thesis, University of Washington, 1997.

C.T. Volinsky, D. Madigan, A.E. Raftery, and R.A. Kronmal. Bayesian model averaging in proportional hazard models: Assessing the risk of a stroke. *Journal of the Royal Statistical Society*, 46:433–448, 1997.

C.T. Volinsky and A.E. Raftery. Bayesian information criterion for censored survival models. *Biometrics*, 56:256–262, 2000.

M.J. Wainwright and M.I. Jordan. *A variational principle for graphical models*. MIT Press, 2005.

D. Wang, W. Zhang, and A. Bakhai. Comparison of bayesian model averaging and stepwise methods for model selection in logistic regression. *Statistics in Medicine*, 23:3451–3467, 2004.

D.L. Weakliem. A critique of the bayesian information criterion for model selection. *Sociological Methods and Research*, 27 (3):359–397, 1999.

WHO. The icd-10 classification of mental and behavioral disorders: diagnostic criteria for research, 1993.

J.M. Winn. *Variational Message Parsing and its Applications*. PhD thesis, University of Cambridge, 2003.

J.M. Winn and C.M. Bishop. Variational message passing. *Submitted to Journal of Machine Learning Research*, 5, 2004.

www.vfhj.dk. Videnscenter for hjerneskade - apopleksi, 2007. http://www.vfhj.dk/default.asp?PageID=833.

Index