Machine Learning in Survival Analysis

Raphael Sonabend, Andreas Bender

Table of contents

List of Figures

List of Tables

This is the electronic version of the book… published by …

This version will be updated to correct mistakes (big and small) which will be incorporated into future editions of the published book.

The book is written by Raphael Sonabend and Andreas Bender…

Raphael Sonabend is…

Andreas Bender is…

Preface

Symbols and Notation

The most common symbols and notation used throughout this book are presented below; in rare cases where different meanings are intended within the book, this will be made clear.

Cases, Fonts, and Symbols

Lower-case letters, x, refer to fixed ('realised', 'observed') values and upper-case letters, X, refer to random variables. For example X is a random variable (r.v.) taking values in (t.v.i.) the set $\mathcal{X}$ if, $X : \Omega \to \mathcal{X}$ where Ω is the sample space of all possible outcomes; then $x \in \mathcal{X}$ is a possible realised value from X. A lower-case (Greek or Latin) letter, x, refers to either a single element or a vector, which will be clear from context. Calligraphic letters, $\mathcal{X}$, are used to refer to sets. A lower-case bold-face letter, $\mathbf{x}$, refers to a matrix. If x is a vector then x_i refers to the ith element in this vector. If $\mathbf{x}$ is a matrix then x_i refers to the ith row of the matrix, $x_{;j}$ refers to the jth column of the matrix, and x_{ij} refers to the ith row of the jth column of matrix $\mathbf{x}$. Unless otherwise stated, a 'vector' is used to refer to a column vector. An element with a 'hat', $\widehat{x}$, refers to the prediction or estimation of the variable without the hat, x. Inline code and datasets will use `this font` and package names will look like this `survival`. Finally, any dates will be presented in the ISO format: YYYY-MM-DD.

Italicised text emphasises a word or phrase that is the focus of the sentence or definition. 'Single quotation marks' are most often utilised to signify that the word or phase will either be defined later in the thesis, or to identify when a word should be taken in an English and not mathematical sense, for example 'a good model' would signify that the phrase does not refer to a particular mathematical definition of a model being good. ''Double quotation marks" are reserved for direct quotes and are always followed by the associated citation.

Distributions and Random Variables

Two separate notations are used to represent probability distributions and random variables. The first is the 'standard' notation: let X be a random variable following some distribution ζ, then f_X is the probability density function of X.

The second notation instead associates distribution functions directly with the distribution and not the variable. So if ζ is a distribution then $\zeta.f$ is the probability density function of ζ; analogously for other distribution defining functions. This notation is described in full detail when first introduced in the thesis.

Variables

The majority of variables will be defined when required however below are some that are commonly used throughout this thesis.

Variable	Definition
$\mathbb{R}$	Set of Reals.
$\mathbb{R}_{>0}$	Set of Positive Reals (excluding zero).
$\mathbb{R}_{\geq 0}$	Set of Non-Negative Reals (including zero).
$\bar{\mathbb{R}}$	Set of Extended Reals, equal to $\mathbb{R} \cup \{-\infty, +\infty\}$.
$\mathbb{N}_0$	Set of Naturals (including zero).
$\mathbb{N}_{>0}$	Set of Positive Naturals (excluding zero).
$\mathcal{N}$	Normal distribution.
$\mathcal{U}$	Uniform distribution.
$x, \mathbf{x}, X, \mathcal{X}$	Vector, matrix, random variable, and set of features.
$y, \mathbf{y}, Y, \mathcal{Y}$	Vector, matrix, random variable, and set of true outcomes.
$t, \mathbf{t}, T, \mathcal{T}$	Vector, matrix, random variable, and set of observed time outcomes.
δ, Δ	Vector and random variable of survival/censoring indicators.
β	Vector of model coefficients, or weights.
η	Linear predictor, $X\beta$.
$\zeta.f$	Probability density function of distribution ζ.
$\zeta.F$	Cumulative distribution function of distribution ζ.
$\zeta.h$	Hazard function of distribution ζ.
$\zeta.H$	Cumulative hazard function of distribution ζ.
$\zeta.S$	Survival function of distribution ζ.
$\mathcal{L}$	Likelihood function.

The indicator function, $\mathbb{I}(\cdot)$, expects a well-defined logical statement $(\cdot)$ and equals 1 when this statement is true, and 0 otherwise. Any distribution function with a '0' in the subscript refers to the 'baseline' function, e.g. h_0, S_0 are the baseline hazard and baseline survival functions respectively.

Function	Definition
$\text{Distr}(\mathcal{D})$	Space of distributions over the set $\mathcal{D}$.
$\lvert x \rvert$	Absolute value of x.
$\lVert x \rVert$	Euclidean norm of vector x, $\sqrt{\lvert x_1 \rvert^2 + ... + \lvert x_n \rvert^2}$.
$\bar{x}$	Sample mean of vector x, $\frac{1}{n}\sum_{i=1}^{n} x_i$.
$\mathbb{E}(X)$	Expectation of random variable X.
$\text{Var}(X)$	Variance of random variable X.

Let $f : \mathcal{X} \to \mathcal{Y}$ be any function with domain $\mathcal{X}$ and codomain $\mathcal{Y}$. Then the *function signature* of f is $\mathcal{X} \to \mathcal{Y}$. Arguments and parameters are separated in function signatures by a pipe, '|', where

variables to the left are parameters (free variables) and those to the right are arguments (fixed). For example let f be an indicator function that 'checks' if the parameter, ϕ, is below the fixed argument, θ, then f is fully defined by

$$f : \mathbb{R} \times \mathbb{R} \to \{0, 1\}; \quad (\phi|\theta) \mapsto \mathbb{I}(\phi < \theta)$$

Traditionally arguments are not included in the formal signature and the above could be expressed as: Let $\theta \in \mathbb{R}$ then $f : \mathbb{R} \to \{0, 1\}; \ (\phi) \mapsto \mathbb{I}(\phi < \theta)$. The first notation is preferred as it clearly specifies all variables included in the function with their domains, whether they are free or fixed, and cleanly extends to multiple parameters and arguments.

Acronyms

Below is a table of acronyms used throughout this thesis (styled as they appear in the text), these are all fully defined the first time they are used.

Acronym	Definition
AFT	Accelerated Failure Time
APT	Accessible, Performant, Transparent
ANN	Artificial Neural Network
AUC	Area Under the Curve
cdf	Cumulative Distribution Function
chf	Cumulative Hazard Function
CPH	Cox Proportional Hazards
GBM	Gradient Boosting Machine
GLM	Generalised Linear Model
IGS	Integrated Graf Score
IPC(W)	Inverse Probability of Censoring (Weighted)
I(S)LL	Integrated (Survival) Log Loss
KM	Kaplan-Meier
LHS	Left Hand Side
MAE	Mean Absolute Error
ML	Machine Learning
pdf	Probability Density Function
PH	Proportional Hazards
PO	Proportional Odds
RHS	Right Hand Side
(R)MSE	(Root) Mean Squared Error
ROC	Receiver Operating Characteristic
R(S)F	Random (Survival) Forest
r.v.	Random Variable
(S)SVM	(Survival) Support Vector Machine
s.t.	Such That
TNR	True Negative Rate
TPR	True Positive Rate
t.v.i.	Taking Values In
w.r.t.	With Respect To

Acronym	Definition
(W)(S)DLL	(Weighted) (Survival) Density Log Loss

1. Introduction

Writing in the middle of a global pandemic, applications of survival analysis are more relevant than ever. Predicting the time from onset of COVID-19 symptoms to hospitalisation, or the time from hospitalisation to intubation, or intubation to death, are all time-to-event predictions that are at the centre of survival analysis. As well as morbid applications, survival analysis predictions may be concerned with predicting the time until a customer cancels their gym membership, or the lifetime of a lightbulb; any event that is guaranteed (or at least very likely) to occur can be modelled by a survival analysis prediction. As these predictions can be so sensitive, for example a model predicting when a child should be taken off breathing support (Data Study Group Team 2020), the best possible predictions, evaluated to the highest standard, are a necessity. In other fields of predictive modelling, machine learning has made incredible breakthroughs (such as AlphaFold[1]), therefore applying machine learning to survival analysis is a natural step in the evolution of an important field.

Survival analysis is the field of Statistics focusing on modelling the distribution of an event, which may mean the time until the event takes place, the risk of the event happening, the probability of the event occurring at a single time, or the event's underlying probability distribution. Survival analysis ('survival') is a unique field of study in Statistics as it includes the added difficulty of 'censoring'. Censoring is best described through example: a study is conducted to determine the mortality rate of a group of patients after diagnoses with a particular disease. If a patient dies during this study then their outcome is 'death' and their time of death can be recorded. However if a patient drops-out of the study before they die, then their time of death (though guaranteed to occur) is unknown and the only available information is the time at which they left the study. This patient is now said to be *censored* at the time they drop out. The censoring mechanism allows as much outcome information (time and event) to be captured as possible for all patients (observations).

Machine learning (ML) is the field of Statistics primarily concerned with building models to either predict outputs from inputs or to learn relationships from data (Hastie, Tibshirani, and Friedman 2001; James et al. 2013). This thesis is limited to the former case, or more specifically supervised learning, as this is the field in which the vast majority of survival problems live. Relative to other areas of supervised learning, development in survival analysis has been slow – the majority of developments in machine learning for survival analysis have only been in the past decade (see chapters (**?@sec-review**)-(Chapter 11)). This appears to have resulted in less interest in the development of machine learning survival models (**?@sec-review**), less rigour in the evaluation of such models (Chapter 11), and fewer off-shelf/open-source implementations (R. Sonabend et al. 2021). This thesis seeks to set the foundations for clear workflows, good practice, and precise results for 'machine learning survival analysis'.

Section 1.1 will elaborate further on the motivation and objectives behind this PhD; research objectives and contributions are then presented in Section 1.2.

[1] https://deepmind.com/research/case-studies/alphafold

1.1. Motivations and Objectives

Experiments throughout the literature demonstrate that machine learning survival models often perform worse (or at least no better) than classical statistical models (Goli, Mahjub, Faradmal, and Soltanian 2016; KATTAN 2003; Ohno-Machado 1997; Puddu and Menotti 2012) (also see **?@sec-bench**).[2] This thesis sets out to explore why this is the case and how this has potential to be improved. The following questions, based on observations of the field, motivated this thesis:

1.1.1. Why are regression and classification more popular than survival analysis in machine learning?

There is no doubt that this is the case, for example the 'bibles of machine learning' (Bishop 2006; Hastie, Tibshirani, and Friedman 2001; James et al. 2013) discuss classification and regression in detail but survival analysis is never discussed. Survival analysis has important applications in healthcare, finance, engineering and more, all fields that directly impact upon individual lives on a day-to-day basis, and should perhaps be considered as important as classification and regression. The result of this gap in interest, is the erroneous assumption that one field can be directly applied to another. For example there is evidence of researchers treating censoring as a nuisance to be ignored and using regression models instead (Schwarzer, Vach, and Schumacher 2000). Censoring is indeed a challenge and may contribute to making survival analysis less accessible than other fields, but this need not be the case; a clear unification of terminology and presentation of methods may help make 'machine learning survival analysis' more accessible. Added accessibility could lead to more academics (and non-academics) engaging with the field and promoting good standards of practice, as well as developing more novel models and measures.

Why are probabilistic survival predictions important?

Development of survival models appears to be skewed towards 'ranking models', which predict the relative risk of an event occurring (Section 3.3). In many applications these predictions are sufficient, for example in randomised control trials if assessing the increased/decreased risk of an event after treatment. However, there are many use-cases where predicting an individual's survival probability distribution is required. Take, for example, an engineer calculating the lifetime of a plane's engine.[3] There are three important reasons to replace a jet engine at the optimal time:

- financial: jet engines are very expensive and replacing one sooner than required is a waste of money;
- environmental: an engine being replaced too early is a waste of potential usage;
- safety: if the engine is replaced too late then there is a risk to passengers.

Now consider examples for the three possible 'prediction types' the engineer can make:

i. A 'relative risk prediction': This engine is twice as likely to fail as another.
ii. A 'survival time prediction': The engine is expected to fail in 30 days.
iii. A 'survival distribution prediction': The lifetime of the engine is distributed according to the probability distribution ζ.

[2]The distinction between a 'classical' and 'machine learning' model used in this thesis is provided in **?@sec-sec-review**.

[3]In this engineering context, survival analysis is usually referred to as reliability analysis.

The first prediction type is not useful as the underlying relative risk may be unknown and the engineer is concerned with the individual lifetime. The second prediction type provides a useful quantity for the engineer to work with however there is no uncertainty captured in this prediction. The third prediction type can capture the uncertainty of failure over the entirety of the positive Reals (though usually only a small subset is possible and useful). With this final prediction type, the engineer can create safe decisions: 'replace the engine at time τ, where τ is the time when the predicted probability of survival drops below 60%, $S(\tau) = 0.6$'. There are ethical, economic, and environmental reasons for a good survival distribution prediction and this thesis considers a distribution prediction to be the most important prediction type.

How are survival models evaluated?

Evaluating predictions from survival models is of the utmost importance. This is especially important as survival models are often deployed in the public domain, particularly in healthcare. Physical products in healthcare, such as new vaccines, undergo rigorous testing and research in randomised control trials before being publically deployed; the same level of rigour should be expected for the evaluation of survival models that are used in life-and-death situations. Evaluation measures for regression and classification are well-understood with important properties, however survival measures have not undergone the same treatment. For example many survival models are still being evaluated solely with concordance indices that have been repeatedly crticised (**GonenHeller2005?**; Rahman et al. 2017; Schmid and Potapov 2012). This paper argues for the use of scoring rules (Section 11.7), which simultaneously assess predictions of distribution and relative risk.

Motivated by these questions, this thesis attempts to unify the two fields of machine learning and survival analysis to make the intersection of the two ('machine learning survival analysis') more concise and accessible. This aim is guided by three key themes: Accessibility, Transparency, and Performance. These are now briefly described to explain why they have been identified as key principles for this thesis.

1.1.2. Accessibility, Transparency, and Predictive Performance

In all critical analyses there must be a metric with which to judge the surveyed objects. For example, machine learning models may be judged by predictive performance, i.e. does one model outperform another? Or estimators may be judged according to bias and consistency properties. As this thesis compares multiple different types of objects, a more universal criteria is applied for the reviews, surveys, and comparisons. These are: Accessibility, Transparency, and (predictive) Performance. A model that satisfies all three criteria may be considered APT (accessible, transparent, performant). These key themes are now briefly described and then further discussion is given to why all must be satisfied for this thesis to consider a model or measure to be 'good'. These are primarily explained in terms of a 'model', though all extend naturally to other objects.

A model is termed *accessible* if there either exists an open-source implementation of the model, or sufficient infrastructure and published mathematics for the model to be implementable.[4] For example, a novel neural network without an open-source implementation can still be accessible if the model's architecture is clearly described and can therefore be implemented with neural network packages such as TensorFlow (Abadi et al. 2015).

[4]The term 'accessible' is slightly more general than terms such as 'off-shelf' as accessibility is defined to include objects that are not off-shelf but that can be implemented given information provided in the literature.

A model is called *transparent* if its properties are well-understood, its use and manipulation of data is clear, and its predictions have a precise interpretation. The word 'transparent' does not refer to the inner workings of the model and therefore a transparent model could still be a 'black-box'.[5] For example, random forests (Section 7.1) are built of hundreds or thousands of individual predictive models, thus making it impossible to fully identify how the final prediction is created. However the model is considered transparent as it is mathematically clear and intuitive how it utilises the individual components to produce its prediction.

A model has good predictive *performance* if its predictions are notably improved over some baseline model (Gressmann et al. 2018). Unlike transparency and accessibility, it is possible to quantify performance and compare this between models (Chapter 11). Whilst there is often a trade-off between predictive performance and model interpretability (e.g. compare neural networks and linear regression), this is not the case for predictive performance and transparency. When considering non-predictive objects, such as measures, then performance instead refers to verifying other established performance properties, for example consistency, unbiasedness, and robustness. An object with good performance is called 'performant'.

Performance is traditionally the primary metric by which models (and measures) are judged, but this thesis only considers a model to be 'good' (or APT) if all three of these themes are satisfied. In fact, it can be demonstrated that if even one of these conditions is not satisfied a model can be dishonest or inefficient.

By example, take the model that always predicts the height of a person as 42cm. This model is very accessible and transparent but has terrible predictive performance, the model is therefore useless. Now consider a patented model without open-source implementation that not only makes perfect predictions but is also clearly described. In this case as no accessible implementation exists, the model cannot be used and tested by the community and more importantly cannot be externally validated, leading to ethical questions about commercial implementation and even whether the results can be trusted. Finally, in the case of an accessible model with strong predictive performance but without clear description in a paper or reader-friendly code/documentation, there can only be limited trust in the model's performance, especially with respect to future performance.

1.2. Book Structure

- Chapter 3 introduces the survival and machine learning settings separately. First a mathematical overview to survival analysis is provided (Section 3.1) and then the scope of this thesis is identified and justified (Section 3.2). Survival prediction types are then mathematically defined for use throughout the thesis (Section 3.3). This is then mirrored for machine learning by first introducing supervised learning and important machine learning methods (Section 2.1) and then defining survival analysis as a machine learning task (Section 2.2).
- **?@sec-review** reviews classical survival models (Section 5.1) and surveys machine learning survival models (Section 6.1 - Section 10.1). Only a short review is provided for the classical setting as this has been covered extensively in the literature over the past few decades. This thesis takes a novel approach by focusing the classical review on model prediction types, in order to gain clarity in understanding how the models can and cannot be utilised. The rest of

[5] Therefore the term 'transparent' here does not refer to the concept of a 'glass-box' model, which is the opposite of a black-box model.

the chapter critically surveys the use of machine learning in survival analysis. The survey is first split by machine learning classes, and then further categorised again by model prediction types.

- Chapter 11 discusses how to evaluate the models introduced in the previous chapter. This starts (Section 11.1) with a general discussion about the importance of evaluation and how survival measures must be selected to relate to the correct survival task. The chapter continues with a full review of the different types of survival measures, how these relate, and what properties exist for the most common of these. Extensive discussion is given to survival scoring rules (Section 11.7) including introducing and completing novel definitions (Section 11.7.2) and proofs (Section 11.7.4).

1.2.1. Code and Reproducibility

Finally, some brief words on the programming present in this thesis.

Programming Languages This thesis includes simulations and figures generated in R and the benchmark experiments in **?@sec-bench** are also conducted in R. Some Python implementations are considered in **?@sec-review**. Only R and Python are considered as they are the two most popular open-source programming languages that intersect classical statistics and machine learning.

Reproducibility The R code for any figures or experiments in this thesis are freely available at https://github.com/RaphaelS1/MLSA under an MIT licence, all content on this website is available under CC BY 4.0. For any code that requires specific software packages, these are listed when required alongside version numbers. All R scripts have set seeds for reproducibility. The code used in this thesis was run using various R versions from 3.6 to 4.0.2 and whilst this should not affect reproducibility, this cannot be guaranteed.

2. Statistical Learning

TODO (150-200 WORDS)

2.1. Machine Learning

This section begins with a very brief introduction to machine learning and a focus on regression and classification; the survival machine learning task is then introduced (Section 2.2). Of the many fields within machine learning (ML), the scope of this thesis is narrowed to supervised learning. Supervised learning is the sub-field of ML in which predictions are made for outcomes based on data with observed dependent and independent variables. For example predicting someone's height is a supervised learning problem as data can be collected for features (independent variables) such as age and sex, and outcome (dependent variable), which is height. Predictive survival analysis problems fall naturally in the supervised learning framework as there are identifiable features and (multiple types of) outcomes.

2.1.1. Terminology and Methods

Common supervised learning methods are discussed in a simplified setting with features X *t.v.i.* $\mathcal{X}$ and outcomes Y *t.v.i.* $\mathcal{Y}$; usually outcomes are referred to as 'targets' (a 'target for prediction'). Let $\mathcal{D}_0 = \{(X_1, Y_1), ..., (X_n, Y_n)\}$ be a (training) dataset where $(X_i, Y_i) \overset{i.i.d.}{\sim} (X, Y)$. The methods below extend naturally to the survival setting.

Strategies and Models

In order to clearly separate between similar objects, several terms for machine learning are now introduced and clearly distinguished.

Let $g : \mathcal{X} \to \mathcal{Y}$ be the true (but unknown) mapping from the features to outcomes, referred to as the *true prediction functional*. Let $\mathcal{G}$ be the set of *prediction functionals* such that $\forall \Upsilon \in \mathcal{G}, \Upsilon : \mathcal{X} \to \mathcal{Y}$. A *learning* or *fitting algorithm* is defined to be any function of the form $\mathcal{A} : \mathcal{X}^n \times \mathcal{Y}^n \to \mathcal{G}$. The goal of supervised learning is to *learn* g with a learning algorithm *fit* on (i.e. the input to the algorithm is) training data, $\hat{g} := \mathcal{A}(\mathcal{D}_0) \in \mathcal{G}$. Note that $\hat{g}$ may take hyper-parameters that can be set or tuned (see below). The learning algorithm is 'good' if $\hat{g}(X) \approx g(X)$ (see 'Evaluation' below).

The learning algorithm is determined by the chosen *learning strategy* and *model*, where a model is a complete specification of a learning strategy including hyper-parameters. These terms are more clearly illustrated by example:

- Learning strategy – simple linear regression
- Model – $y = \beta_0 + \beta_1 x$ where $x \in \mathbb{R}$ is a single covariate, $y \in \mathbb{R}$ is the target, and $\beta_0, \beta_1 \in \mathbb{R}$ are model coefficients.
- Learning algorithm (model fitting) – Minimise the residual sum of squares: $(\hat{\beta}_0, \hat{\beta}_1) := \text{argmin}_{\beta_0,\beta_1}\{\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\}$ for $(x_i, y_i) \in \mathcal{D}_0, i = 1, ..., n$.
- Prediction functional – $\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$

To further illustrate the difference between learning strategy and model, note that the same learning strategy 'simple linear regression' could either utilise the model above or instead a model without intercept, $y = \beta x$, in which case the learning algorithm and prediction functional would also be modified.

The model in (ii) is called *unfitted* as the model coefficients are unknown and the model cannot be used for prediction. After step (iii) the model is said to be fit to the training data and therefore the model is *fitted*.[1] It is common to refer to the learning algorithm (and associated hyper-parameters) as the unfitted model and to refer to the prediction functional (and associated hyper-parameters) as the fitted model.

Evaluation

Models are *evaluated* by evaluation measures called *losses* or *scores*,[2] $L : \mathcal{Y} \times \mathcal{Y} \to \bar{\mathbb{R}}$. Let $(X^*, Y^*) \sim (X, Y)$ be test data (i.e. independent of $\mathcal{D}_0$) and let $\hat{g} : \mathcal{X} \to \mathcal{Y}$ be a prediction functional fit on $\mathcal{D}_0$, then these evaluation measures determine how closely predictions, $\hat{g}(X^*)$, relate to the truth, Y^*, thereby providing a method for determining if a model is 'good'.[3]

Task

A machine learning *task* is a simple mechanism to outline the problem of interest by providing: i) the data specification; ii) the definition of learning; iii) the definition of success (when is a prediction 'good'?) (Franz J. Király et al. 2021). All tasks in this paper have the same definitions of learning and success. For (ii), the aim is to learn the true prediction functional, g, by fitting the learning algorithm on training data, $\hat{g} := \mathcal{A}(\mathcal{D}_0)$. For (iii), a predicted functional is considered 'good' if the *expected generalization error*, $\mathbb{E}[L(Y^*, \hat{g}(X^*))]$, is low, where $(X^*, Y^*) \sim (X, Y)$ is independent of the training data $\mathcal{D}_0$, and L is some loss that is chosen according to the domain of interest (regression, classification, survival).

Resampling

Models are *tested* on their ability to make predictions. In order to avoid 'optimism of training error' (James et al. 2013) – overconfidence caused by testing the model on training data – models are tested on previously unseen or 'held-out' data. *Resampling* is the procedure of splitting one dataset

[1] The terms 'fitted' and 'unfitted' are used instead of 'fit' and 'unfit' to prevent confusion with words such as 'suitable' and 'unsuitable'.

[2] The term 'loss' is usually utilised to refer to evaluation measures to be minimised, whereas 'scores' should be maximised, this is returned to in (@sec-eval).

[3] Here evaluation refers specifically to predictive ability; other forms of evaluation and further discussion of the area are provided in (@sec-eval).

into two or more for separated training and testing. In this paper only two resampling methods are utilised: *holdout* and *cross-validation*. Holdout is the process of splitting a primary dataset into training data for model fitting and testing data for model predicting. This is an efficient method but may not accurately estimate the expected generalisation error for future model performance, instead this is well-estimated by K-fold cross-validation (KCV) (Hastie, Tibshirani, and Friedman 2001). In KCV, data is split into $K \in \mathbb{N}_{>0}$ 'folds' such that $K-1$ of the folds are used for model training and the final Kth fold for testing. The testing fold is iterated over all K folds, so that each at some point is used for testing and then training (though never at the same time). In each iteration the model is fit on the training folds, and predictions are made and evaluated on the testing fold, giving a loss $L_k := L(\hat{g}(X^k), Y^k)$, where (X^k, Y^k) are data from the kth fold. A final loss is defined by, $L^* := \frac{1}{K}\sum_{k=1}^{K} L_k$. Commonly $K = 5$ or $K = 10$ (Breiman and Spector 1992; Kohavi 1995).

Model Performance Benchmarking

Whilst *benchmarking* often refers to speed tests, i.e. the time taken to complete an operation, it can also refer to any experiment in which objects (mathematical or computational) are compared. In this report, a benchmark experiment will either refer to the comparison of multiple models' predictive abilities, or comparison of computational speeds and object sizes for model fitting; which of these will be clear from context.

Model Comparison

Models can be analytically compared on how well they make predictions for new data. Model comparison is a complex topic with many open questions (Demšar 2006; Dietterich 1998; Nadeau and Bengio 2003) and as such discussion is limited here. When models are compared on multiple datasets, there is more of a consensus in how to evaluate models (Demšar 2006) and this is expanded on further in (R. E. B. Sonabend 2021). Throughout this thesis there are small simulation experiments for model comparison on single datasets however as these are primarily intended to aid exposition and not to generalise results, it suffices to compare models with the conservative method of constructing confidence intervals around the sample mean and standard error of the loss when available (Nadeau and Bengio 2003).

Hyper-Parameters and Tuning

A *hyper-parameter* is a model parameter that can be set by the user, as opposed to coefficients that are estimated as part of model fitting. A hyper-parameter can be set before training, or it can be tuned. *Tuning* is the process of choosing the optimal hyper-parameter value via automation. In the simplest setting, tuning is performed by selecting a range of values for the hyper-parameter(s) and treating each choice (combination) as a different model. For example if tuning the number of trees in a random forest (Section 7.1), m_r, then a range of values, say $100, 200, 500$ are chosen, and three models $m_{r100}, m_{r200}, m_{r500}$ are benchmarked. The optimal hyper-parameter is given by whichever model is the best performing. *Nested resampling* is a common method to prevent overfitting that could occur from using overlapping data for tuning, training, or testing. Nested resampling is the process of resampling the training set again for tuning.

2.1.2. Machine Learning in Classification and Regression

Before introducing machine learning for survival analysis, which is considered 'non-classical', the more standard classification and regression set-ups are provided; these are referenced throughout this thesis.

2.1.2.1. Classification

Classification problems make predictions about categorical (or discrete) events, these may be *deterministic* or *probabilistic*. Deterministic classification predicts which category an observation falls into, whereas probabilistic classification predicts the probability of an observation falling into each category. In this brief introduction only binary single-label classification is discussed, though the multi-label case is considered in **??**. In binary classification, there are two possible categories an observation can fall into, usually referred to as the 'positive' and 'negative' class. For example predicting the probability of death due to a virus is a probabilistic classification task where the 'positive' event is death.

A probabilistic prediction is more informative than a deterministic one as it encodes uncertainty about the prediction. For example it is clearly more informative to predict a 70 chance of rain tomorrow instead of simply 'rain'. Moreover the latter prediction implicitly contains an erroneous assumption of certainty, e.g. 'it will rain tomorrow'.

Classification Task

Box 2.1. *Let (X, Y) be random variables t.v.i. $\mathcal{X} \times \mathcal{Y}$ where $\mathcal{X} \subseteq \mathbb{R}^p$ and $\mathcal{Y} = \{0, 1\}$. Then,*

- *The* probabilistic classification task *is the problem of predicting the probability of a single event taking place and is specified by $g : \mathcal{X} \to [0, 1]$.*
- *The* deterministic classification task *is the problem of predicting if a single event takes place and is specified by $g : \mathcal{X} \to \mathcal{Y}$.*

The estimated prediction functional $\hat{g}$ is fit on training data $\backslash(X_1, Y_1), ..., (X_n, Y_n) \overset{i.i.d.}{\sim} (X, Y)$ and is considered 'good' if $\mathbb{E}[L(Y^*, \hat{g}(X^*))]$ is low, where $(X^*, Y^*) \sim (X, Y)$ is independent of $(X_1, Y_1), ..., (X_n, Y_n)$ and $\hat{g}$.
In the probabilistic case, the prediction $\hat{g}$ maps to the estimated probability mass function $\hat{p}_Y$ s.t. $\hat{p}_Y(1) = 1 - \hat{p}_Y(0)$.

2.1.2.2. Regression

A regression prediction is one in which the goal is to predict a continuous outcome from a set of features. For example predicting the time until an event (without censoring) occurs, is a regression problem.

Regression Task

Box 2.2. *Let* (X, Y) *be random variables t.v.i.* $\mathcal{X} \times \mathcal{Y}$ *where* $\mathcal{X} \subseteq \mathbb{R}^p$ *and* $\mathcal{Y} \subseteq \mathbb{R}$*. Let* $\mathcal{S} \subset \text{Distr}(\mathcal{Y})$ *be a convex set of distributions on* $\mathcal{Y}$*. Then,*

- *The* probabilistic regression task *is the problem of predicting a conditional distribution over the Reals and is specified by* $g : \mathcal{X} \to \mathcal{S}$*.*
- *The* deterministic regression task *is the problem of predicting a single continuous value in the Reals and is specified by* $g : \mathcal{X} \to \mathcal{Y}$*.*

The estimated prediction functional $\hat{g}$ *is fit on training data* $\backslash(X_1, Y_1), ..., (X_n, Y_n) \overset{i.i.d.}{\sim} (X, Y)$ *and is considered 'good' if* $\mathbb{E}[L(Y^*, \hat{g}(X^*))]$ *is low, where* $(X^*, Y^*) \sim (X, Y)$ *is independent of* $(X_1, Y_1), ..., (X_n, Y_n)$ *and* $\hat{g}$*.*

Whilst regression can be either probabilistic or deterministic, the latter is much more common and therefore in this thesis 'regression' refers to the deterministic case unless otherwise stated.

2.2. Survival Analysis Task

The survival prediction problems identified in (Section 3.3) are now formalised as machine learning tasks.

Survival Task

Box 2.3. *Let* (X, T, Δ) *be random variables t.v.i.* $\mathcal{X} \times \mathcal{T} \times \{0, 1\}$ *where* $\mathcal{X} \subseteq \mathbb{R}^p$ *and* $\mathcal{T} \subseteq \mathbb{R}_{\geq 0}$*. Let* $\mathcal{S} \subseteq \text{Distr}(\mathcal{T})$ *be a convex set of distributions on* $\mathcal{T}$ *and let* $\mathcal{R} \subseteq \mathbb{R}$*. Then,*

- *The* probabilistic survival task *is the problem of predicting a conditional distribution over the positive Reals and is specified by* $g : \mathcal{X} \to \mathcal{S}$*.*
- *The* deterministic survival task *is the problem of predicting a continuous value in the positive Reals and is specified by* $g : \mathcal{X} \to \mathcal{T}$*.*
- *The* survival ranking task *is specified by predicting a continuous ranking in the Reals and is specified by* $g : \mathcal{X} \to \mathcal{R}$*.*

The estimated prediction functional $\hat{g}$ *is fit on training data* $\backslash(X_1, T_1, \Delta_1), ..., (X_n, T_n, \Delta_n) \overset{i.i.d.}{\sim} (X, T, \Delta)$ *and is considered 'good' if* $\backslash\mathbb{E}[L(T^*, \Delta^*, \hat{g}(X^*))]$ *is low, where* $(X^*, T^*, \Delta^*) \sim (X, T, \Delta)$ *is independent of* $(X_1, T_1, \Delta_1), ..., (X_n, T_n, \Delta_n)$ *and* $\hat{g}$*.*

Any other survival prediction type falls within one of these tasks above, for example predicting log-survival time is the deterministic task and predicting prognostic index or linear predictor is the ranking task. Removing the separation between the prognostic index and ranking prediction types is due to them both making predictions over the Reals; their mathematical difference lies in interpretation only. In general, the survival task will assume that $\mathcal{T} \subseteq \mathbb{R}_{\geq 0}$, and the terms

'discrete' or 'reduced survival task' will refer to the case when $\mathcal{T} \subseteq \mathbb{N}_0$. Unless otherwise specified, the 'survival task', will be used to refer to the probabilistic survival task.[4]

Survival Analysis and Regression

Survival and regression tasks are closely related as can be observed from their respective definitions. Both are specified by $g : \mathcal{X} \to \mathcal{S}$ where for probabilistic regression $\mathcal{S} \subseteq \text{Distr}(\mathbb{R})$ and for survival $\mathcal{S} \subseteq \text{Distr}(\mathbb{R}_{\geq 0})$. Furthermore both settings can be viewed to use the same generative process. In the survival setting in which there is no censoring then data is drawn from (X, Y) *t.v.i.* $\mathcal{X} \times \mathcal{T}, \mathcal{T} \subseteq \mathbb{R}_{\geq 0}$ and in regression from (X, Y) *t.v.i.* $\mathcal{X} \times \mathcal{Y}, \mathcal{Y} \subseteq \mathbb{R}$, so that the only difference is whether the outcome data ranges over the Reals or positive Reals.

These closely related tasks are discussed in more detail in (Chapter 12), with a particular focus on how the more popular regression setting can be used to solve survival tasks. In (**?@sec-review**) the models are first introduced in a regression setting and then the adaptations to survival are discussed, which is natural when considering that historically machine learning survival models have been developed by adapting regression models.

[4]These definitions are given in the most general case where the time variable is over $\mathbb{R}_{\geq 0}$. In practice, all models instead assume time is over $\mathbb{R}_{>0}$ and any death at $T_i = 0$ is set to $T_i = \epsilon$ for some very small $\epsilon \in \mathbb{R}_{>0}$. Analogously for the discrete survival task. This assumption may not reflect reality as a patient could die at the study start however models cannot typically include this information in training.

3. Survival Analysis

TODO (150-200 WORDS)

In their broadest and most basic definitions, survival analysis is the study of temporal data from a given origin until the occurrence of one or more events or 'end-points' (Collett 2014), and machine learning is the study of models and algorithms that learn from data in order to make predictions or find patterns (Hastie, Tibshirani, and Friedman 2001). Reducing either field to these definitions is ill-advised.

This chapter collects terminology utilised in survival analysis (Section 3.1) and machine learning (Section 2.1) in order that this thesis can cleanly discuss 'machine learning survival analysis' (Section 2.2). Once the mathematical setting is set up, the thesis scope is fully presented in (Section 3.2). Whilst the content of this chapter is not novel with respect to either survival analysis or machine learning separately, this does appear to be the first formulation of the survival analysis machine learning 'task' (Franz J. Király et al. 2021).

3.1. Survival Analysis

Survival analysis is the field of Statistics concerned with the analysis of time-to-event data, which consists of covariates, a categorical (often binary) outcome, and the time until this outcome takes place (the 'survival time'). As a motivating example of time-to-event data, say 100 patients are admitted to a COVID-19 ward and for each patient the following covariate data are collected: age, weight and sex; additionally for each patient the time until death or discharge is recorded. In the time-to-event dataset, which takes a standard tabular form, each of the 100 patients is a row, with columns consisting of age, weight, and sex measurements, as well as the outcome (death or discharge) and the time to outcome.

Survival analysis is distinct from other areas of Statistics due to the incorporation of 'censoring', a mechanism for capturing uncertainty around when an event occurs in the real-world. Continuing the above example, if a patient dies of COVID-19 five dies after admittance, then their outcome is exactly known: they *died* after five days. Consider now a patient who is discharged after ten days. As death is a guaranteed event they have a true survival time but this may be decades later, therefore they are said to be *censored* at ten days. This is a convenient method to express that the patient survives up to ten days and their survival status at any time after this point is unknown. Censoring is a unique challenge to survival analysis that attempts to incorporate as much information as possible without knowing the true outcome. This is a 'challenge' as statistical models usually rely on learning from observed, i.e. known, outcome data; therefore censoring requires special treatment.

Whilst survival analysis occurs in many fields, for example as 'reliability analysis' in engineering and 'duration analysis' in economics, in this thesis the term 'survival' will always be used. Moreover the following terminology, analogous to a healthcare setting, are employed: survival analysis (or 'survival' for short) refers to the field of study; the event of interest is the 'event', or 'death'; an observation that has not experienced an event is 'censored' or 'alive'; and observations are referred to as 'observations', 'subjects', or 'patients'.

Some of the biggest challenges in survival analysis stem from an unclear definition of a 'survival analysis prediction' and different (sometimes conflicting) common notations. This thesis attempts to make discussions around survival analysis clearer and more precise by first describing the mathematical setting for survival analysis in (Section 3.1.1) and only then defining the prediction types to consider in (Section 3.3).

3.1.1. Survival Data and Definitions

Survival analysis has a more complicated data setting than other fields as the 'true' data generating process is not directly modelled but instead engineered variables are defined to capture observed information. Let,

- X *t.v.i.* $\mathcal{X} \subseteq \mathbb{R}^p, p \in \mathbb{N}_{>0}$ be the generative random variable representing the data *features/covariates/independent variables*.
- Y *t.v.i.* $\mathcal{T} \subseteq \mathbb{R}_{\geq 0}$ be the (unobservable) *true survival time*.
- C *t.v.i.* $\mathcal{T} \subseteq \mathbb{R}_{\geq 0}$ be the (unobservable) *true censoring time*.

It is impossible to fully observe both Y and C. This is clear by example: if an observation drops out of a study then their censoring time is observed but their event time is not, whereas if an observation dies then their true censoring time is unknown. Hence, two engineered variables are defined to represent observable outcomes. Let,

- $T := \min\{Y, C\}$ be the *observed outcome time*.
- $\Delta := \mathbb{I}(Y = T) = \mathbb{I}(Y \leq C)$ be the *survival indicator* (also known as the *censoring* or *event* indicator).[1]

Together (T, Δ) is referred to as the *survival outcome* or *survival tuple* and they form the dependent variables. The survival outcome provides a concise mechanism for representing the time of the *observed* outcome and indicating which outcome (death or censoring) took place.

Now the full generative template for survival analysis is given by $\setminus$ (X, Δ, C, Y, T) *t.v.i.* $\mathcal{X} \times \{0,1\} \times \mathcal{T} \times \mathcal{T} \times \mathcal{T}$ and with $(X_i, \Delta_i, C_i, Y_i, T_i)$ jointly i.i.d. A *survival dataset* is defined by $\mathcal{D} = \{(X_1, T_1, \Delta_1), ..., (X_n, T_n, \Delta_n)\}$ where $(X_i, T_i, \Delta_i) \overset{i.i.d.}{\sim} (X, T, \Delta)$ and X_i is a p-vector, $X_i = (X_{i;1}, ..., X_{i;p})$. Though unobservable, the true outcome times are defined by $(Y_1, C_1), ..., (Y_n, C_n)$ where $(Y_i, C_i) \overset{i.i.d.}{\sim} (Y, C)$.

(1) exemplifies a random survival dataset with n observations (rows) and p features.

[1] Indicators are usually named to reflect a positive condition in the function (in this case the event when $Y = T$), but counter to this convention the 'censoring indicator' is possibly the most common term.

X	X	X	T	Δ	Y	C
X_{11}	$\cdots$	X_{1p}	T_1	Δ_1	Y_1	C_1
$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_{n1}	$\cdots$	X_{np}	T_n	Δ_n	Y_n	C_n

Table 3.1.: Theoretical time-to-event dataset. (Y, C) are 'hypothetical' as they can never be directly observed. Rows are individual observations, X columns are features, T is observed time-to-event, Δ is the censoring indicator, and (Y, C) are hypothetical true survival and censoring times.

(2) exemplifies an observed survival dataset with a modified version of the `rats` dataset (Therneau 2015).

litter ($X_{,1}$) | **rx** ($X_{,2}$) | **sexF** ($X_{,3}$) | **time** | **status** | **survTime** | **censTime** |
X | X | X | T | Δ | Y | C |
– | – | — | – | –| – | – |
1 | 1 | 1 | 101 | 0 | 105 | 101 |
1 | 0 | 1 | 49 | 1 | 49 | 55|
1 | 0 | 1 | 104 | 0 | 200 | 104 |
2 | 1 | 0 | 91 | 0 | 92 | 91 |
2 | 0 | 0 | 104 | 1 | 104 | 104 |
2 | 0 | 0 | 102 | 1 | 102 | 120 |

: `rats` (Therneau 2015) time-to-event dataset with added hypothetical columns (Y, C). Rows are individual observations, X columns are features, T is observed time-to-event, Δ is the censoring indicator, and (Y, C) are hypothetical (here arbitrary values dependent on (T, Δ)) true survival and censoring times. {#tbl-surv-data-rats}

Both datasets includes two extra columns, on the right of the triple vertical line, which imagine hypothetical data for the unobserved true survival and censoring times. \\ Finally the following terms are used frequently throughout this report. Let $(T_i, \Delta_i) \overset{i.i.d.}{\sim} (T, \Delta), i = 1, ..., n$, be random survival outcomes. Then,

- The *set of unique* or *distinct time-points* refers to the set of time-points in which at least one observation dies or is censored, $\mathcal{U}_O := \{T_i\}_{i \in \{1,...,n\}}$.
- The *set of unique death times* refers to the set of unique time-points in which death (and not censoring) occurred, $\mathcal{U}_D := \{T_i : \Delta_i = 1\}_{i \in \{1,...,n\}}$.
- The *risk set* at a given time-point, τ, is the set of subjects who are known to be alive (not dead or censored) just before that time, $\mathcal{R}_\tau := \{i : T_i \geq \tau\}$ where i is a unique row/subject in the data.
- The *number of observations alive* at τ is the cardinality of the risk set, $|\mathcal{R}_\tau|$, and is denoted by $n_\tau := \sum_i \mathbb{I}(T_i \geq \tau)$.
- The *number of observations who die* at τ is denoted by $d_\tau := \sum_i \mathbb{I}(T_i = \tau, \Delta_i = 1)$.
- The Kaplan-Meier estimate of the average survival function of the training data *survival distribution* is the Kaplan-Meier estimator (Section 5.1.1) fit (Section 2.1.1) on training data (T_i, Δ_i) and is denoted by $\hat{S}_{KM}$.

- The Kaplan-Meier estimate of the average survival function of the training data *censoring distribution* is the Kaplan-Meier estimator fit on training data $(T_i, 1 - \Delta_i)$ and is denoted by $\hat{G}_{KM}$.

Notation and definitions will be recapped at the start of each chapter for convenience.

3.1.2. Censoring

Censoring is now discussed in more detail and important concepts introduced. Given the survival generating process (X, T, Δ) with unobservable (Y, C), the event is experienced if $Y \leq C$ and $\Delta = 1$ or censored if $\Delta = 0$.

Censoring 'Location' {.unnumbered .unlisted}

Right-censoring is the most common form of censoring in survival models and it occurs either when a patient drops out (but doesn't experience the event) of the study before the end and thus their outcome is unknown, or if they experience the event at some unknown point after the study end. Formally let $[\tau_l, \tau_u]$ be the study period for some, $\tau_l, \tau_u \in \mathbb{R}_{\geq 0}$. Then right-censoring occurs when either $Y > \tau_u$ or when $Y \in [\tau_l, \tau_u]$ and $C \leq Y$. In the first case $T = C = \tau_u$ and censoring is due to the true time of death being unknown as the observation period has finished. In the latter case, a separate censoring event, such as drop-out or another competing risk, is observed.

Left-censoring is a rarer form of censoring and occurs when the event happens at some unknown time before the study start, $Y < \tau_l$. Interval-censoring occurs when the event takes place in some interval within the study period, but the exact time of event is unknown. (Figure 3.1) shows a graphical representation of right-censoring.

Censoring 'Dependence'

Censoring is often defined as *uninformative* if $Y \perp\!\!\!\perp C$ and *informative* otherwise however these definitions can be misleading as the term 'uninformative' appears to be imply that censoring is independent of both X and Y, and not just Y. Instead the following more precise definitions are used in this report.

Definition 3.1 (Censoring)**.** Let (X, T, Δ, Y, C) be defined as above, then

- If $C \perp\!\!\!\perp X$, censoring is *feature-independent*, otherwise censoring is *feature-dependent*.
- If $C \perp\!\!\!\perp Y$, then censoring is *event-independent*, otherwise censoring is *event-dependent*.
- If $(C \perp\!\!\!\perp Y)|X$, censoring is conditionally independent of the event given covariates, or *conditionally event-independent*.
- If $C \perp\!\!\!\perp (X, Y)$ censoring is *uninformative*, otherwise censoring is *informative*.

Non-informative censoring can generally be well-handled by models as true underlying patterns can still be detected and the reason for censoring does not affect model inference or predictions. However in the real-world, censoring is rarely non-informative as reasons for drop-out or missingness in outcomes tend to be related to the study of interest. Event-dependent censoring is a tricky case that, if not handled appropriately (by a competing-risks framework), can easily lead to poor

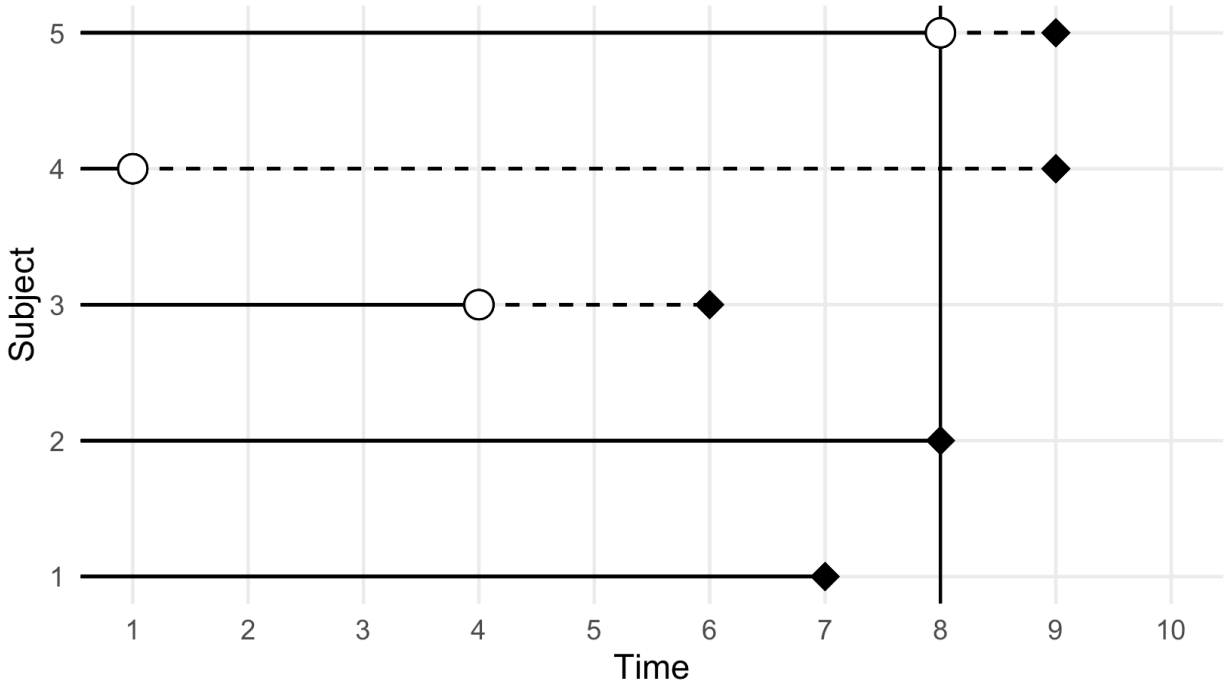


Figure 3.1.: Dead and censored subjects (y-axis) over time (x-axis). Black diamonds indicate true death times and white circles indicate censoring times. Vertical line is the study end time. Subjects 1 and 2 die in the study time. Subject 3 is censored in the study and (unknown) dies within the study time. Subject 4 is censored in the study and (unknown) dies after the study. Subject 5 dies after the end of the study.

model development; the reason for this can be made clear by example: Say a study is interested in predicting the time between relapses of stroke but a patient suffers a brain aneurysm due to some separate neurological condition, then there is a high possibility that a stroke may have occurred if the aneurysm had not. Therefore a survival model is unlikely to distinguish the censoring event (aneurysm) from the event of interest (stroke) and will confuse predictions. In practice, the majority of models and measures assume that censoring is conditionally event-independent and hence censoring can be predicted by the covariates whilst not directly depending on the event. For example if studying the survival time of ill pregnant patients in hospital, then dropping out of the study due to pregnancy is clearly dependent on how many weeks pregnant the patient is when the study starts (for the sake of argument assume no early/late pregnancy due to illness).

Type I Censoring

Type I and Type II censoring are special-cases of right-censoring, only Type I is discussed in this thesis as it is more common in simulation experiments. Type I censoring occurs if a study has a set end-date, or maximum survival time, and a patient survives until the end of the study. If survival times are dependent on covariates (i.e. not random) and the study start date is known (or survival times are shifted to the same origin) then Type I censoring will usually be informative as censored patients will be those who survived the longest.

3.2. Thesis Scope

Now that the mathematical setting has been defined, the thesis scope is provided. For time and relevance the scope of this thesis is narrowed to the most parsimonious setting that is genuinely useful in modelling real-world scenarios. This is the setting that captures all assumptions made by the majority of proposed survival models and therefore is practical both theoretically and in application. This setting is defined by the following assumptions (with justifications):

- Let p be the proportion of censored observations in the data, then $p \in (0, 1)$. This open interval prevents the case when $p = 0$, which is simply a regression problem (Section 2.1.2.2), or the case when $p = 1$, in which no useful models exist (as the event never occurs).
- Only right-censoring is observed in the data, no left- or interval-censoring. This accurately reflects most real-world data in which observations that have experienced the event before the study start (left-censoring) are usually not of interest, and close monitoring of patients means that interval-censoring is unlikely in practice. It is acknowledged that left-truncation is a common problem in medical datasets though this is often handled not by models but by data pre-processing, which is not part of the workflow discussed in this thesis.
- There is only one event of interest, an observation that does not experience this event is censored. This eliminates the 'competing risk' setting in which multiple events of interest can be modelled.
- The event can happen at most once. For example the event could be death or initial diagnosis of a disease however cannot be recurrent such as seizure. In the case where the event could theoretically happen multiple times, only the time to one (usually the first) occurrence of the event is modelled.

- The event is guaranteed to happen at least once. This is an assumption implicitly made by all survival models as predictions are for the time until the true event, Y, and not the observed outcome, T.

For both the multi-event and recurrent-event cases, simple reductions exist such that these settings can be handled by the models discussed in this paper however this is not discussed further here.

No assumptions are made about whether censoring is dependent on the data but when models and measures make these assumptions, they will be explicitly discussed. \\ The purpose of any statistical analysis is dependent on the research question. For example techniques are available for data analysis, imputation, exploration, prediction, and more. This thesis focuses on the predictive setting; other objectives, such as model inspection and data exploration can be achieved post-hoc via interpretable machine learning techniques (Molnar 2019). \\ Finally, the methods in this thesis are restricted to frequentist statistics. Bayesian methods are not discussed as the frequentist setting is usually more parsimonious and additionally there are comparatively very few off-shelf implementations of Bayesian survival methods. Despite this, it is noted that Bayesian methods are particularly relevant to the research in this thesis, which is primarily concerned with uncertainty estimates and predictions of distributions. Therefore, a natural extension to the work in this thesis would be to fully explore the Bayesian setting.

3.3. Survival Prediction Problems

This section continues by defining the survival problem narrowed to the scope described in the previous section. Defining a single 'survival prediction problem' (or 'task') is important mathematically as conflating survival problems could lead to confused interpretation and evaluation of models. Let (X, T, Δ) and $\mathcal{D}$ be as defined above. A general survival prediction problem is one in which:

- a survival dataset, $\mathcal{D}$, is split (Section 2.1.1) for training, $\mathcal{D}_0$, and testing, $\mathcal{D}_1$;
- a survival model is fit on $\mathcal{D}_0$; and
- the model predicts some representation of the unknown true survival time, Y, given $\mathcal{D}_1$.

The process of 'fitting' is model-dependent, and can range from simple maximum likelihood estimation of model coefficients, to complex algorithms. The model fitting process is discussed in more abstract detail in (Section 2.1) and then concrete algorithms are discussed in (**?@sec-review**). The different survival problems are separated by 'prediction types' or 'prediction problems', these can also be thought of as predictions of different 'representations' of Y. Four prediction types are discussed in this paper, these may be the only possible survival prediction types and are certainly the most common as identified in chapters (**?@sec-review**) and (Chapter 11). They are predicting:

- The *relative risk* of an individual experiencing an event – A single continuous ranking.
- The *time until an event* occurs – A single continuous value.
- The *prognostic index* for a model – A single continuous value.
- An individual's *survival distribution* – A probability distribution.

The first three of these are referred to as *deterministic* problems as they predict a single value whereas the fourth is *probabilistic* and returns a full survival distribution. Definitions of these are expanded on below. \\ Survival predictions differ from other fields in two respects. Firstly, the

predicted outcome, Y, is a different object than the outcome used for model training, (T, Δ). This differs from, say, regression in which the same object (a single continuous variable) is used for fitting and predicting. Secondly, with the exception of the time-to-event prediction, all other prediction types do not predict Y but some other related quantity.

Survival prediction problems must be clearly separated as they are inherently incompatible. For example it is not meaningful to compare a relative risk prediction from one observation to a survival distribution of another. Whilst these prediction types are separated above, they can be viewed as special cases of each other. Both (1) and (2) may be viewed as variants of (3); and (1), (2), and (3) can all be derived from (4); this is elaborated on below.

Relative Risk/Ranking

This is perhaps the most common survival problem and is defined as predicting a continuous rank for an individual's 'relative risk of experiencing the event'. For example, given three patients, $\{i, j, k\}$, a relative risk prediction may predict the 'risk of event' as $\{0.1, 0.5, 10\}$ respectively. From these predictions, the following types of conclusions can be drawn:

- Conclusions comparing patients. e.g. i is at the least risk; the risk of j is only slightly higher than that of i but the risk of k is considerably higher than j; the corresponding ranks for i, j, k, are $1, 2, 3$.
- Conclusions comparing risk groups. e.g. thresholding risks at 1.0 places i and j in a 'low-risk' group and k in a 'high-risk' group

So whilst many important conclusions can be drawn from these predictions, the values themselves have no meaning when not compared to other individuals. Interpretation of these rankings has historically been conflicting in implementation, with some software having the interpretation 'higher ranking implies higher risk' whereas others may indicate 'higher ranking implies lower risk' **??**. In this thesis, a higher ranking will always imply a higher risk of event (as in the example above).

Time to Event

Predicting a time to event is the problem of predicting the deterministic survival time of a patient, i.e. the amount of time for which they are predicted to be alive after some given start time. Part of the reason this problem is less common in survival analysis is because it borders regression – a single continuous value is predicted – and survival – the handling of censoring is required – but neither is designed to solve this problem directly. Time-to-event predictions can be seen as a special-case of the ranking problem as an individual with a predicted longer survival time will have a lower overall risk, i.e. if t_i, t_j and r_i, r_j are survival time and ranking predictions for patients i and j respectively, then $t_i > t_j \rightarrow r_i < r_j$.

Prognostic Index

Given covariates, $x \in \mathbb{R}^{n \times p}$, and a vector of model coefficients, $\beta \in \mathbb{R}^p$, the linear predictor is defined by $\eta := x\beta \in \mathbb{R}^n$. The 'prognostic index' is a term that is often used in survival analysis papers that usually refers to some transformation (possibly identity), ϕ, on the linear predictor, $\phi(\eta)$. Assuming a predictive function (for survival time, risk, or distribution defining function (see

below)) of the form $g(\varphi)\phi(\eta)$, for some function g and variables φ where $g(\varphi)$ is constant for all observations (e.g. Cox PH (Section 5.1.2)), then predictions of η are a special case of predicting a relative risk, as are predictions of $\phi(\eta)$ if ϕ is rank preserving. A higher prognostic index may imply a higher or lower risk of event, dependent on the model structure.

Survival Distribution

Predicting a survival distribution refers specifically to predicting the distribution of an individual patient's survival time, i.e. modelling the distribution of the event occurring over $\mathbb{R}_{\geq 0}$. Therefore this is seen as the probabilistic analogue to the deterministic time-to-event prediction, these definitions are motivated by similar terminology in machine learning regression problems (Section 2.1). The above three prediction types can all be derived from a probabilistic survival distribution prediction (Chapter 12).

A survival distribution is a mathematical object that is estimated by predicting a *representation* of the distribution. Let W be a continuous random variable t.v.i. $\mathbb{R}_{\geq 0}$ with probability density function (pdf), $f_W : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, and cumulative distribution function (cdf), $F_W : \mathbb{R}_{\geq 0} \to [0,1]; (\tau) \mapsto P(W \leq \tau)$. The pdf, $f_W(\tau)$, is the likelihood of an observation dying in a small interval around time τ, and $F_W(\tau) = \int_0^\tau f_W(\tau)$ is the probability of an observation being dead at time τ (i.e. dying at or before τ). In survival analysis, it is generally more interesting to model the risk of the event taking place or the probability of the patient being alive, leading to other distribution representations of interest.

The survival function is defined as

$$S_W : \mathbb{R}_{\geq 0} \to [0,1]; \quad (\tau) \mapsto P(W \geq \tau) = \int_\tau^\infty f_W(u)\ du$$

and so $S_W(\tau) = 1 - F_W(\tau)$. This function is known as the survival function as it can be interpreted as the probability that a given individual survives until some point $\tau \geq 0$.

Another common representation is the hazard function,

$$h_W : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}; \quad (\tau) \mapsto \frac{f_W(\tau)}{S_W(\tau)}$$

The hazard function is interpreted as the instantaneous risk of death given that the observation has survived up until that point; note this is not a probability as h_W can be greater than one.

The cumulative hazard function (chf) can be derived from the hazard function by

$$H_W : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}; \quad (\tau) \mapsto \int_0^\tau h_W(u)\ du$$

The cumulative hazard function relates simply to the survival function by

$$H_W(\tau) = \int_0^\tau h_W(u)\ du = \int_0^\tau \frac{f_W(u)}{S_W(u)}\ du = \int_0^\tau -\frac{S'_W(u)}{S_W(u)}\ du = -\log(S_W(\tau))$$

Any of these representations may be predicted conditionally on covariates for an individual by a probabilistic survival distribution prediction. Once a function has been estimated, predictions can be made conditional on the given data. For example if n survival functions are predicted, $\hat{S}_1, ..., \hat{S}_n$, then $\hat{S}_i$ is interpreted as the predicted survival function given covariates of observation i, and analogously for the other representation functions.

4. Survival Models

TODO (150-200 WORDS)

This chapter provides a brief review of classical survival models and then a critical survey of machine learning survival models. The terms 'classical', 'machine learning', and even 'model' have hazy definitions that will be further specified to make clear how they apply in this paper.

Recall (Section 2.1.1) the separation between the following terms:

- Learning strategy – A method for estimating the true prediction functional, g
- Fitting algorithm, $\mathcal{A}$ – A function mapping the training data, $\mathcal{D}_0$, to an estimate of the true prediction functional, $\hat{g} := \mathcal{A}(\mathcal{D}_0)$. The choice of fitting algorithm is determined by the learning strategy.
- (Unfitted) Model – The complete specification of a learning strategy with hyper-parameters and any other components such as pre-processing
- Fitted Model/Prediction functional, $\hat{g} : \mathcal{X} \to \mathcal{Y}$ – Function, possibly with hyper-parameters, for making predictions on unseen data

'Classical' models are defined with a very narrow scope in this thesis: low-complexity models that are either non-parametric or have parameters that can be fit with maximum likelihood estimation (or an equivalent method). In contrast, 'machine learning' (ML) models require more intensive model fitting procedures such as recursion or iteration. The classical models in this paper are fast to fit and highly interpretable, though can be inflexible and may make unreasonable assumptions. Whereas the ML models are more flexible with hyper-parameters however are computationally more intensive (both in terms of speed and storage), require tuning to produce 'good' results, and are often a 'black-box' with difficult interpretation.

This chapter investigates models for predictive survival analysis with a focus on whether a model is APT (Section 1.1.2). As classical survival models have been studied extensively for decades, these are separated from the ML models in this chapter and reduced to a smaller literature review in (Section 5.1). The rest of this chapter then surveys each of the primary machine learning classes separately. The scope of the models discussed in this chapter is limited to the general thesis scope (Section 3.2), i.e. single event with right-censoring and no competing-risks, though in some cases these are discussed.

Novel adaptations for each of the ML models are suggested at the end of each section, these primarily serve as interesting avenues to explore for future research but none have been studied for theoretical properties or implemented in software packages, though most have been informally explored to demonstrate some 'proof-of-concept'.

Notation and Terminology

The notation introduced in (Chapter 3) is recapped for use in this chapter: the generative template for the survival setting is given by (X, T, Δ, Y, C) *t.v.i.* $\mathcal{X} \times \mathcal{T} \times \{0,1\} \times \mathcal{T} \times \mathcal{T}$ where $\mathcal{X} \subseteq \mathbb{R}^p$ and $\mathcal{T} \subseteq \mathbb{R}_{\geq 0}$, where C, Y are unobservable, $T := \min\{Y, C\}$, and $\Delta = \mathbb{I}(Y = T)$. Random survival data is given by $(X_i, T_i, \Delta_i, Y_i, C_i) \overset{i.i.d.}{\sim} (X, T, \Delta, Y, C)$. Usually data will instead be presented as a training dataset, $\mathcal{D}_0 = \{(X_1, T_1, \Delta_1), ..., (X_n, T_n, \Delta_n)\}$ where $(X_i, T_i, \Delta_i) \overset{i.i.d.}{\sim} (X, T, \Delta)$. For simplicity only a single testing observation needs to be defined to effectively write about the prediction functional, this test observation is given by $\mathcal{D}_1 = (X^*, T^*, \Delta^*) \sim (X, T, \Delta)$.

For regression models the generative template is given by (X, Y) t.v.i. $\mathcal{X} \subseteq \mathbb{R}^p$ and $Y \subseteq \mathbb{R}$. As with the survival setting, a regression training set is given by $\{(X_1, Y_1), ..., (X_n, Y_n)\}$ where $(X_i, Y_i) \overset{i.i.d.}{\sim} (X, Y)$ and a testing observation by $\mathcal{D}_1 = (X^*, Y^*) \sim (X, Y)$.

Finally recall: the set of unique time-points, $\mathcal{U}_O := \{T_i\}_{i \in \{1,...,n\}}$, the set of unique death times, $\mathcal{U}_D := \{T_i : \Delta_i = 1\}_{i \in \{1,...,n\}}$, the risk set at τ is $\mathcal{R}_\tau := \{i : T_i \geq \tau\}$, the number of observations alive or at risk at τ is $n_\tau := \sum_i \mathbb{I}(T_i \geq \tau)$, and the number of observations that die at τ is $d_\tau := \sum_i \mathbb{I}(T_i = \tau, \Delta_i = 1)$.

5. Classical Models

TODO (150-200 WORDS)

5.1. A Review of Classical Survival Models

This section provides a literature review of 'classical' models proposed for survival analysis. There are several possible taxonomies for categorising statistical models, these include:

- Parametrisation Type: One of non-, semi-, or fully-parametric. \ Non-parametric models assume that the data distribution cannot be specified with a finite set of parameters. In contrast, fully-parametric models assume the distribution can be specified with a finite set of parameters. Semi-parametric models are a hybrid of the two and are formed of a finite set of parameters *and* an infinite-dimensional 'nuisance' parameter.
- Conditionality Type: One of unconditional or conditional. A conditional prediction is one that makes use of covariates in order to condition the prediction on each observation. Unconditional predictors, which are referred to below as 'estimators', ignore covariate data and make the same prediction for all individuals.
- Prediction Type: One of ranking, survival time, or distribution (Section 3.3).

(3) summarises the models discussed below into the taxonomies above for reference. Note that the Cox model is listed as predicting a continuous ranking, and not a survival distribution, which may appear inconsistent with other definitions. The reason for this is elaborated upon in (Chapter 12). Though the predict-type taxonomy is favoured throughout this thesis, it is clearer to review classical models in increasing complexity, beginning with unconditional estimators before moving onto semi-parametric continuous ranking predictions, and finally conditional distribution predictors. The review is brief with mathematics limited to the model fundamentals but not including methods for parameter estimation. Also the review is limited to the 'basic' model specification and common extensions such as regularization are not discussed though they do exist for many of these models.

All classical models are highly transparent and accessible, with decades of research and many off-shelf implementations. Predictive performance of each model is briefly discussed as part of the review and then again in (R. E. B. Sonabend 2021).

Model[1]	Parametrisation[2]	Prediction[3]	Conditionality
Kaplan-Meier	Non	Distr.	Unconditional
Nelson-Aalen	Non	Distr.	Unconditional

Model[1]	Parametrisation[2]	Prediction[3]	Conditionality
Akritas	Non	Distr.	Conditional
Cox PH	Semi	Rank	Conditional
Parametric PH	Fully	Distr.	Conditional
Accelerated Failure Time	Fully	Distr.	Conditional
Proportional Odds	Fully	Distr.	Conditional
Flexible Spline	Fully	Distr.	Conditional

Table 5.1.: Table of models discussed in this literature review, classified by parametrisation, prediction type, and conditionality.

* 1. All models are implemented in the R package **survival** (Therneau 2015) with the exception of flexible splines, implemented in **flexsurv** (C. Jackson 2016), and the Akritas estimator in **survivalmodels** (R. Sonabend 2020). * 2. Non = non-parametric, Semi = semi-parametric, Fully = fully-parametric. * 3. Distr. = distribution, Rank = ranking.

5.1.1. {Non-Parametric Distribution Estimators

Unconditional Estimators

Unconditional non-parametric survival models assume no distribution for survival times and estimate the survival function using simple algorithms based on observed outcomes and no covariate data. The two most common methods are the Kaplan-Meier estimator (**KaplanMeier1958?**), which estimates the average survival function of a training dataset, and the Nelson-Aalen estimator (Aalen 1978; Nelson 1972), which estimates the average cumulative hazard function of a training dataset.

The Kaplan-Meier estimator of the survival function is given by

$$\hat{S}_{KM}(\tau|\mathcal{D}_0) = \prod_{t\in\mathcal{U}_O, t\leq\tau} \left(1 - \frac{d_t}{n_t}\right) \tag{5.1}$$

As this estimate is so important in survival models, this thesis will always use the symbol $\hat{S}_{KM}$ to refer to the Kaplan-Meier estimate of the average survival function fit on training data (T_i, Δ_i). Another valuable function is the Kaplan-Meier estimate of the average survival function of the *censoring* distribution, which is the same as above but estimated on $(T_i, 1-\Delta_i)$, this will be denoted by $\hat{G}_{KM}$.

The Nelson-Aalen estimator for the cumulative hazard function is given by

$$\hat{H}(\tau|\mathcal{D}_0) = \sum_{t\in\mathcal{U}_O, t\leq\tau} \frac{d_t}{n_t} \tag{5.2}$$

The primary advantage of these models is that they rely on heuristics from empirical outcomes only and don't require any assumptions about the form of the data. To train the models they only require (T_i, Δ_i) and both return a prediction of $\mathcal{S} \subseteq \text{Distr}(\mathcal{T})$ ((**box-task-surv?**)). In addition, both simply account for censoring and can be utilised in fitting other models or to estimate unknown censoring

distributions. The Kaplan-Meier and Nelson-Aalen estimators are both consistent estimators for the survival and cumulative hazard functions respectively.

Utilising the relationships provided in (Section 3.3), one could write the Nelson-Aalen estimator in terms of the survival function as $\hat{S}_{NA} = \exp(-\hat{H}(\tau|\mathcal{D}_0))$. It has been demonstrated that $\hat{S}_{NA}$ and $\hat{S}_{KM}$ are asymptotically equivalent, but that $\hat{S}_{NA}$ will provide larger estimates than $\hat{S}_{KM}$ in smaller samples (Colosimo et al. 2002). In practice, the Kaplan-Meier is the most widely utilised non-parametric estimator in survival analysis and is the simplest estimator that yields consistent estimation of a survival distribution; it is therefore a natural, and commonly utilised, 'baseline' model (Harald Binder and Schumacher 2008; Herrmann et al. 2021; Huang et al. 2020a; P. Wang, Li, and Reddy 2019): estimators that other models should be 'judged' against to ascertain their overall performance (Chapter 11).

Not only can these estimators be used for analytical comparison, but they also provide intuitive methods for graphical calibration of models (Section 11.6.2). These models are never stuidied for prognosis directly but as baselines, components of complex models (Chapter 12), or graphical tools (Habibi et al. 2018; Jager et al. 2008; Moghimi-dehkordi et al. 2008). The reason for this is due to them having poor predictive performance as a result of omitting explanatory variables in fitting. Moreover, if the data follows a particular distribution, parametric methods will be more efficient (P. Wang, Li, and Reddy 2019).

Conditional Estimators

The Kaplan-Meier and Nelson-Aalen estimators are simple to compute and provide good estimates for the survival time distribution but in many cases they may be overly-simplistic. Conditional non-parametric estimators include the advantages described above (no assumptions about underlying data distribution) but also allow for conditioning the estimation on the covariates. This is particularly useful when estimating a censoring distribution that may depend on the data (Chapter 11). However predictive performance of conditional non-parametric estimators decreases as the number of covariates increases, and these models are especially poor when censoring is feature-dependent (Gerds and Schumacher 2006).

The most widely used conditional non-parametric estimator for survival analysis is the Akritas estimator (Akritas 1994) defined by[1]

$$\hat{S}(\tau|X^*, \mathcal{D}_0, \lambda) = \prod_{j:T_j \leq \tau, \Delta_j = 1} \Big(1 - \frac{K(X^*, X_j|\lambda)}{\sum_{l=1}^{n} K(X^*, X_l|\lambda)\mathbb{I}(T_l \geq T_j)}\Big)$$

where K is a kernel function, usually $K(x, y|\lambda) = \mathbb{I}(|\hat{F}_X(x) - \hat{F}_X(y)| < \lambda), \lambda \in (0, 1]$, $\hat{F}_X$ is the empirical distribution function of the training data, $X_1, ..., X_n$, and λ is a hyper-parameter. The estimator can be interpreted as a conditional Kaplan-Meier estimator which is computed on a neighbourhood of subjects closest to X^* (Blanche, Dartigues, and Jacqmin-Gadda 2013). To account for tied survival times, the following adaptation of the estimator is utilised (Blanche, Dartigues, and Jacqmin-Gadda 2013)

[1] Arguments and parameters are separated in function signatures by a pipe, '|', where variables to the left are parameters (free variables) and those to the right are arguments (fixed). In this equation, τ is a parameter to be set by the user, and $X^*, \mathcal{D}_0, \lambda$ are fixed arguments. This could therefore be simplified to $\hat{S}(\tau)$ to only include free variables.

$$\hat{S}(\tau|X^*, \mathcal{D}_0, \lambda) = \prod_{t \in \mathcal{U}_O, t \leq \tau} \Big(1 - \frac{\sum_{j=1}^{n} K(X^*, X_j|\lambda)\mathbb{I}(T_j = t, \Delta_j = 1)}{\sum_{j=1}^{n} K(X^*, X_j|\lambda)\mathbb{I}(T_j \geq t)}\Big) \tag{5.3}$$

If $\lambda = 1$ then $K(\cdot|\lambda) = 1$ and the estimator is identical to the Kaplan-Meier estimator.

The non-parametric nature of the model is highlighted in (Equation 5.3), in which both the fitting and predicting stages are combined into a single equation. A new observation, X^*, is compared to its nearest neighbours from a training dataset, $\mathcal{D}_0$, without a separated fitting procedure. One could consider splitting fitting and predicting in order to clearly separate between training and testing data. In this case, the fitting procedure is the estimation of $\hat{F}_X$ on training data and the prediction is given by (Equation 5.3) with $\hat{F}_X$ as an argument. This separated fit/predict method is implemented in **survivalmodels** (R. Sonabend 2020). As with other non-parametric estimators, the Akritas estimator can still be considered transparent and accessible. With respect to predictive performance, the Akritas estimator has more explanatory power than non-parametric estimators due to conditioning on covariates, however this is limited to a very small number of variables and therefore this estimator is still best placed as a conditional baseline.

5.1.2. Continuous Ranking and Semi-Parametric Models: Cox PH

The Cox Proportional Hazards (CPH) (Cox 1972), or Cox model, is likely the most widely known semi-parametric model and the most studied survival model (Habibi et al. 2018; Moghimi-dehkordi et al. 2008; Reid 1994; P. Wang, Li, and Reddy 2019). The Cox model assumes that the hazard for a subject is proportionally related to their explanatory variables, $X_1, ..., X_n$, via some baseline hazard that all subjects in a given dataset share ('the PH assumption'). The hazard function in the Cox PH model is defined by

$$h(\tau|X_i) = h_0(\tau)\exp(X_i\beta)$$

where h_0 is the non-negative *baseline hazard function* and $\beta = \beta_1, ..., \beta_p$ where $\beta_i \in \mathbb{R}$ are coefficients to be fit. Note the proportional hazards (PH) assumption can be seen as the estimated hazard, $h(\tau|X_i)$, is directly proportional to the model covariates $\exp(X_i\beta)$. Whilst a form is assumed for the 'risk' component of the model, $\exp(X_i\beta)$, no assumptions are made about the distribution of h_0, hence the model is semi-parametric.

The coefficients, β, are estimated by maximum likelihood estimation of the 'partial likelihood' (Cox 1975), which only makes use of ordered event times and does not utilise all data available (hence being 'partial'). The partial likelihood allows study of the informative β-parameters whilst ignoring the nuisance h_0. The predicted linear predictor, $\hat{\eta} := X^*\hat{\beta}$, can be computed from the estimated $\hat{\beta}$ to provide a ranking prediction.

Inspection of the model is also useful without specifying the full hazard by interpreting the coefficients as 'hazard ratios'. Let $p = 1$ and $\hat{\beta} \in \mathbb{R}$ and let $X_i, X_j \in \mathbb{R}$ be the covariates of two training observations, then the *hazard ratio* for these observations is the ratio of their hazard functions,

$$\frac{h(\tau|X_i)}{h(\tau|X_j)} = \frac{h_0(\tau)\exp(X_i\hat{\beta})}{h_0(\tau)\exp(X_j\hat{\beta})} = \exp(\hat{\beta})^{X_i - X_j}$$

If $\exp(\widehat{\beta}) = 1$ then $h(\tau|X_i) = h(\tau|X_j)$ and thus the covariate has no effect on the hazard. If $\exp(\widehat{\beta}) > 1$ then $X_i > X_j \rightarrow h(\tau|X_i) > h(\tau|X_i)$ and therefore the covariate is positively correlated with the hazard (increases risk of event). Finally if $\exp(\widehat{\beta}) < 1$ then $X_i > X_j \rightarrow h(\tau|X_i) < h(\tau|X_i)$ and the covariate is negatively correlated with the hazard (decreases risk of event). \\ Interpreting hazard ratios is known to be a challenge, especially by clinicians who require simple statistics to communicate to patients (Sashegyi and Ferry 2017; Spruance et al. 2004). For example the full interpretation of a hazard ratio of '2' for binary covariate X would be: 'assuming that the risk of death is constant at all time-points then the instantaneous risk of death is twice as high in a patient with X than without'. Simple conclusions are limited to stating if patients are at more or less risk than others in their cohort. Further disadvantages of the model also lie in its lack of real-world interpretabilitity, these include (Reid 1994):

- the PH assumption may not be realistic and the risk of event may not be constant over time;
- the estimated baseline hazard from a non-parametric estimator is a discrete step-function resulting in a discrete survival distribution prediction despite time being continuous; and
- the estimated baseline hazard will be constant after the last observed time-point in the training set (Gelfand et al. 2000).

Despite these disadvantages, the model has been demonstrated to have excellent predictive performance and routinely outperforms (or at least does not underperform) sophisticated ML models (Michael F. Gensheimer and Narasimhan 2018; Luxhoj and Shyur 1997; Vanya Van Belle, Pelckmans, Van Huffel, et al. 2011) (and (R. E. B. Sonabend 2021)). It's simple form and wide popularity mean that it is also highly transparent and accessible.

The next class of models address some of the Cox model disadvantages by making assumptions about the baseline hazard.

5.1.3. Conditional Distribution Predictions: Parametric Linear Models

Parametric Proportional Hazards

The CPH model can be extended to a fully parametric PH model by substituting the unknown baseline hazard, h_0, for a particular parameterisation. Common choices for distributions are Exponential, Weibull and Gompertz (Kalbfleisch and Prentice 2011; P. Wang, Li, and Reddy 2019); their hazard functions are summarised in ((**tab-survivaldists?**)) along with the respective parametric PH model. Whilst an Exponential assumption leads to the simplest hazard function, which is constant over time, this is often not realistic in real-world applications. As such the Weibull or Gompertz distributions are often preferred. Moreover, when the shape parameter, γ, is 1 in the Weibull distribution or 0 in the Gompertz distribution, their hazards reduce to a constant risk ((Figure 5.1)). As this model is fully parametric, the model parameters can be fit with maximum likelihood estimation, with the likelihood dependent on the chosen distribution.

Distribution[1]	$h_0(\tau)$[2]	$h(\tau\|X_i)$[3]
Exp(λ)	λ	$\lambda \exp(X_i\beta)$
Weibull(γ, λ)	$\lambda\gamma\tau^{\gamma-1}$	$\lambda\gamma\tau^{\gamma-1} \exp(X_i\beta)$
Gompertz(γ, λ)	$\lambda \exp(\gamma\tau)$	$\lambda \exp(\gamma\tau) \exp(X_i\beta)$

Distribution[1]	$h_0(\tau)$[2]	$h(\tau\|X_i)$[3]

Table 5.2.: Exponential, Weibull, and Gompertz hazard functions and PH specification.

* 1. Distribution choices for baseline hazard. γ, λ are shape and scale parameters respectively. * 2. Baseline hazard function, which is the (unconditional) hazard of the distribution. * 3. PH hazard function, $h(\tau|X_i) = h_0(\tau)\exp(X_i\beta)$.

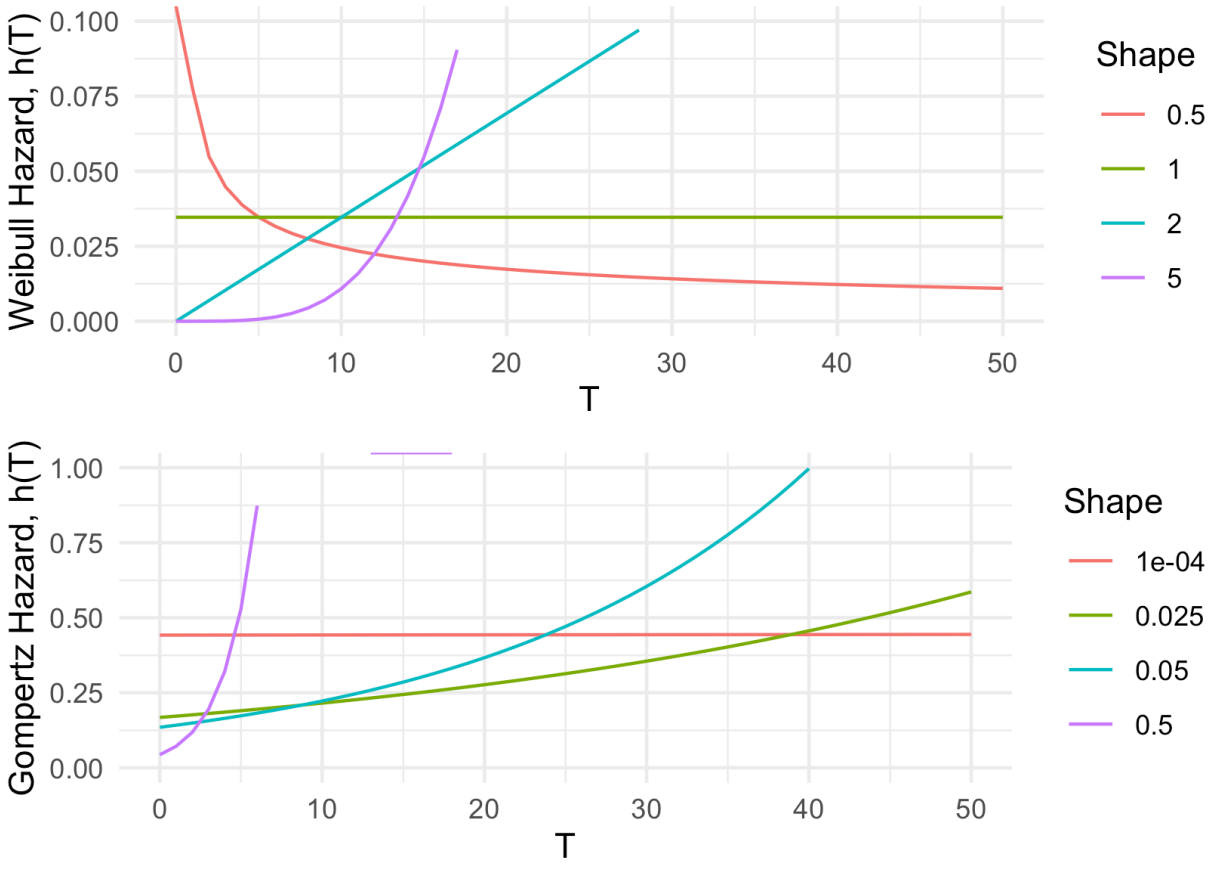


Figure 5.1.: Comparing the hazard curves under Weibull and Gompertz distributions for varying values of the shape parameter; scale parameters are set so that each parametrisation has a median of 20. x-axes are time and y-axes are Weibull (top) and Gompertz (bottom) hazards as a function of time.

In the literature, the Weibull distribution tends to be favoured as the initial assumption for the survival distribution (Michael F. Gensheimer and Narasimhan 2018; Habibi et al. 2018; Hielscher et al. 2010; R. and J. 1968; Rahman et al. 2017), though Gompertz is often tested in death-outcome models for its foundations in modelling human mortality (Gompertz 1825). There exist many tests for checking the goodness-of-model-fit (Section 11.3) and the distribution choice can even be treated as a model hyper-parameter. Moreover it transpires that model inference and predictions are largely insensitive to the choice of distribution (Collett 2014; Reid 1994). In contrast to the Cox model, fully parametric PH models can predict absolutely continuous survival distributions, they do not treat the baseline hazard as a nuisance, and in general will result in more precise and

interpretable predictions if the distribution is correctly specified (Reid 1994; Patrick Royston and Parmar 2002).

Whilst misspecification of the distribution tends not to affect predictions too greatly, PH models will generally perform worse when the PH assumption is not valid. PH models can be extended to include time-varying coefficients or model stratification (Cox 1972) but even with these adaptations the model may not reflect reality. For example, the predicted hazard in a PH model will be either monotonically increasing or decreasing but there are many scenarios where this is not realistic, such as when recovering from a major operation where risks tends to increase in the short-term before decreasing. Accelerated failure time models overcome this disadvantage and allow more flexible modelling, discussed next.

Accelerated Failure Time

In contrast to the PH assumption, where a unit increase in a covariate is a multiplicative increase in the hazard rate, the Accelerated Failure Time (AFT) assumption means that a unit increase in a covariate results in an acceleration or deceleration towards death (expanded on below). The hazard representation of an AFT model demonstrates how the interpretation of covariates differs from PH models,

$$h(\tau|X_i) = h_0(\exp(-X_i\beta)\tau)\exp(-X_i\beta)$$

where $\beta = (\beta_1, ..., \beta_p)$ are model coefficients. In contrast to PH models, the 'risk' component, $\exp(-X_i\beta)$, is the exponential of the *negative* linear predictor and therefore an increase in a covariate value results in a decrease of the predicted hazard. This representation also highlights how AFT models are more flexible than PH as the predicted hazard can be non-monotonic. For example the hazard of the Log-logistic distribution ((Figure 5.2)) is highly flexible depending on chosen parameters. Not only can the AFT model offer a wider range of shapes for the hazard function but it is more interpretable. Whereas covariates in a PH model act on the hazard, in an AFT they act on time, which is most clearly seen in the log-linear representation,

$$\log Y_i = \mu + \alpha_1 X_{i1} + \alpha_2 X_{i2} + ... + \alpha_p X_{ip} + \sigma\epsilon_i$$

where μ and σ are location and scale parameters respectively, $\alpha_1, ..., \alpha_p$ are model coefficients, and ϵ_i is a random error term. In this case a one unit increase in covariate X_{ij} means a α_j increase in the logarithmic survival time. For example if $\exp(X_i\alpha) = 0.5$ then i 'ages' at double the baseline 'speed'. Or less abstractly if studying the time until death from cancer then $\exp(X_i\alpha) = 0.5$ can be interpreted as 'the entire process from developing tumours to metastasis and eventual death in subject i is twice as fast than the normal', where 'normal' refers to the baseline when all covariates are 0.

Specifying a particular distribution for ϵ_i yields a fully-parametric AFT model. Common distribution choices include Weibull, Exponential, Log-logistic, and Log-Normal (Kalbfleisch and Prentice 2011; P. Wang, Li, and Reddy 2019). The Buckley-James estimator (Buckley and James 1979) is a semi-parametric AFT model that non-parametrically estimates the distribution of the errors however this model has no theoretical justification and is rarely fit in practice (Wei 1992). The fully-parametric model has theoretical justifications, natural interpretability, and can often provide a better fit than a PH model, especially when the PH assumption is violated (Patel, Kay, and Rowell 2006; Qi 2009; Zare et al. 2015).

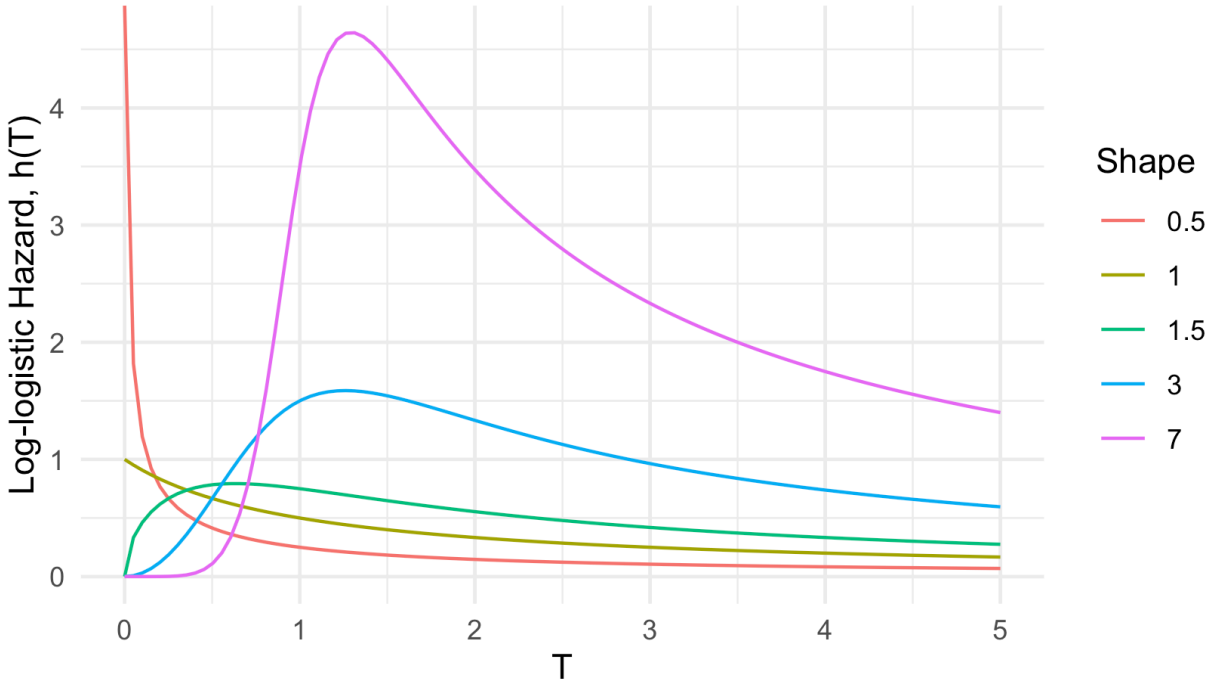


Figure 5.2.: Log-logistic hazard curves with a fixed scale parameter of 1 and a changing shape parameter. x-axis is time and y-axis is the log-logistic hazard as a function of time.

Proportional Odds

Proportional odds (PO) models (Bennett 1983) fit a proportional relationship between covariates and the odds of survival beyond a time τ,

$$O_i(\tau) = \frac{S_i(\tau)}{F_i(\tau)} = O_0(\tau)\exp(X_i\beta)$$

where O_0 is the baseline odds.

In this model, a unit increase in a covariate is a multiplicative increase in the odds of survival after a given time and the model can be interpreted as estimating the log-odds ratio. There is no simple closed form expression for the partial likelihood of the PO model and hence in practice a Log-logistic distribution is usually assumed for the baseline odds and the model is fit by maximum likelihood estimation on the full likelihood (Bennett 1983).

Perhaps the most useful feature of the model is convergence of hazard functions (Kirmani and Gupta 2001), which states $h_i(\tau)/h_0(\tau) \to 1$ as $\tau \to \infty$. This property accurately reflects real-world scenarios, for example if comparing chemotherapy treatment on advanced cancer survival rates, then it is expected that after a long period (say 10 years) the difference in risk between groups is likely to be negligible. This is in contrast to the PH model that assumes the hazard ratios are constant over time, which is rarely a reflection of reality.

In practice, the PO model is harder to fit and is less flexible than PH and AFT models, both of which can also produce odds ratios. This may be a reason for the lack of popularity of the PO model, in addition there is limited off-shelf implementations (Collett 2014). Despite PO models not

being commonly utilised, they have formed useful components of neural networks (Section 10.1) and flexible parametric models (below).

Flexible Parametric Models – Splines

Royston-Parmar flexible parametric models (Patrick Royston and Parmar 2002) extend PH and PO models by estimating the baseline hazard with natural cubic splines. The model was designed to keep the form of the PH or PO methods but without the semi-parametric problem of estimating a baseline hazard that does not reflect reality (see above), or the parametric problem of misspecifying the survival distribution.

To provide an interpretable, informative and smooth hazard, natural cubic splines are fit in place of the baseline hazard. The crux of the method is to use splines to model time on a log-scale and to either estimate the log cumulative Hazard for PH models, $\log H(\tau|X_i) = \log H_0(\tau) + X_i\beta$, or the log Odds for PO models, $\log O(\tau|X_i) = \log O_0(\tau) + X_i\beta$, where β are model coefficients to fit, H_0 is the baseline cumulative hazard function and O_0 is the baseline odds function. For the flexible PH model, a Weibull distribution is the basis for the baseline distribution and a Log-logistic distribution for the baseline odds in the flexible PO model. $\log H_0(\tau)$ and $\log O_0(\tau)$ are estimated by natural cubic splines with coefficients fit by maximum likelihood estimation. The standard full likelihood is optimised, full details are not provided here. Between one and three internal knots are recommended for the splines and the placement of knots does not greatly impact upon the fitted model (Patrick Royston and Parmar 2002).

Advantages of the model include being: interpretable, flexible, can be fit with time-dependent covariates, and it returns a continuous function. Moreover many of the parameters, including the number and position of knots, are tunable, although Royston and Parmar advised against tuning and suggest often only one internal knot is required (Patrick Royston and Parmar 2002). A recent simulation study demonstrated that even with an increased number of knots (up to seven degrees of freedom), there was little bias in estimation of the survival and hazard functions (Bower et al. 2019). Despite its advantages, a 2018 review (Ng et al. 2018) found only twelve instances of published flexible parametric models since Royston and Parmar's 2002 paper, perhaps because it is more complex to train, has a less intuitive fitting procedure than alternatives, and has limited off-shelf implementations; i.e. is less transparent and accessible than parametric alternatives. \\
The PH and AFT models are both very transparent and accessible, though require slightly more expert knowledge than the CPH in order to specify the 'correct' underlying probability distribution. Interestingly whilst there are many papers comparing PH and AFT models to one another using in-sample metrics (Section 11.3) such as AIC (Georgousopoulou et al. 2015; Habibi et al. 2018; Moghimi-dehkordi et al. 2008; Zare et al. 2015), no benchmark experiments could be found for out-of-sample performance. PO and spline models are less transparent than PH and AFT models and are even less accessible, with very few implementations of either. No conclusions can be drawn about the predictive performance of PO or spline models due to a lack of suitable benchmark experiments.

6. Machine Learning Survival Models

TODO (150-200 WORDS)

6.1. A Survey of Machine Learning Models for Survival Analysis

These next sections provide a technical, critical survey of machine learning models proposed for survival analysis with the focus on the 'simpler' setup of non-competing risks. Models are separated into their different 'classes' ((**tab-surv-ml-returns?**)), which exists as a natural taxonomy in machine learning. Each class review is then further separated by first discussing the simpler and more standard regression setting, before expanding into their survival framework. The focus is once again on the different predict types of the model, which enables clear exposition and discussion around how some areas have successfully dealt with the survival predictive problem, whereas others have fallen short.

This is not the first survey of machine learning models for survival analysis. A recent 2017 survey (P. Wang, Li, and Reddy 2019) focused on covering the breadth of machine learning models for survival analysis and this survey is recommended to the reader as a strong starting point to understand which ML models are available for survival analysis. However whilst this provides a comprehensive review and a 'big-picture' view, there is no discussion about how successful the discussed models are in solving the survival task.

A comprehensive survey of neural networks was presented by Schwarzer *et al.* (2000) (Schwarzer, Vach, and Schumacher 2000) in which the authors collected the many ways in which neural networks have been 'misused' in the context of survival analysis. This level of criticism is vital in the context of survival analysis and healthcare data as transparency and understanding are often prioritised over predictive performance. Whilst the survey in this thesis will try not to be as critical as the Schwarzer review, it will aim to discuss models and how well they actually solve the survival problem.

In line with the core topic of this thesis, this survey aims to demonstrate if each model is APT (Section 1.1.2). Historically, surveys have focused primarily on predictive performance, which is generally preferred for complex classification and regression tasks. However in the context of survival analysis, transparency is of the utmost importance and any model that does not solve the task it claims to, despite strong predictive performance, can be considered sub-optimal. The survey will also examine the accessibility of survival models. A model need not be open-source to be accessible, but it should be 'user-friendly' and not require expert cross-domain knowledge. For example, a neural network may require knowledge of complex model building, but if set-up correctly could be handled without medical or survival knowledge. Whereas a Gaussian Process

requires knowledge of the model class, simulation, (usually) Bayesian modelling, and also survival analysis. \\ ((**tab-surv-ml-returns?**)) provides information about the models reviewed in this survey, including a model reference for use in the (R. E. B. Sonabend 2021) benchmark experiment, the predict types of the model, and in which R package it is implemented.

Class[1]	Name[2]	Authors (Year)[3]	Task[4]	Implementation[5]
RF	RRT	LeBlanc and Crowley (1992) (LeBlanc and Crowley 1992)	Rank	**rpart** (Therneau and Atkinson 2019)
RF	RSDF-DEV	Hothorn *et al.* (2004) (Hothorn et al. 2004)	Prob.	**ipred** (Peters and Hothorn 2019)
RF	RRF	Ishwaran *et al.* (2004) (H. Ishwaran et al. 2004)	Rank	-
RF	RSCIFF	Hothorn *et al.* (2006) (Hothorn et al. 2005)	Det., Prob.	**party** (Hothorn, Hornik, and Zeileis 2006), **partykit** (Hothorn and Zeileis 2015)
RF	RSDF-STAT	Ishwaran *et al.* (2008) (B. H. Ishwaran et al. 2008)	Prob.	**randomForestSRC** (H. Ishwaran and Kogalur 2018), **ranger** (Wright and Ziegler 2017)
GBM	GBM-COX	Ridgeway (1999) (Ridgeway 1999) & Buhlmann (2007) (Buhlmann and Hothorn 2007)	Prob.	**mboost** (Hothorn et al. 2020), **xgboost** (T. Chen et al. 2020), **gbm** (Greenwell et al. 2019)
GBM	CoxBoost	Binder & Schumacher (2008) (Harald Binder and Schumacher 2008)	Prob.	**CoxBoost** (Harold Binder 2013)
GBM	GBM-AFT	Schmid & Hothorn (2008) (Schmid and Hothorn 2008b)	Det.	**mboost**, **xgboost**
GBM	GBM-BUJAR	Wang & Wang (2010) (Z. Wang and Wang 2010)	Det.	**bujar** (Z. Wang 2019)
GBM	GBM-GEH	Johnson & Long (2011) (B. A. Johnson and Long 2011)	Det.	**mboost**

Class[1]	Name[2]	Authors (Year)[3]	Task[4]	Implementation[5]
GBM	GBM-UNO	Mayr & Schmid (2014) (Mayr and Schmid 2014)	Rank	**mboost**
SVM	SVCR	Shivaswamy *et al.* (2007) (Shivaswamy, Chu, and Jansche 2007)	Det.	**survivalsvm** (Fouodo et al. 2018)
SVM	SSVM-Rank	Van Belle *et al.* (2007) (Vanya Van Belle et al. 2007)	Rank	**survivalsvm**
SVM	SVRc	Khan and Zubek (2008) (Khan and Bayer Zubek 2008)	Det.	-
SVM	SSVM-Hybrid	Van Belle (2011) (Vanya Van Belle, Pelckmans, Van Huffel, et al. 2011)	Det.	**survivalsvm**
SVM	SSVR-MRL	Goli *et al.* (2016) (Goli, Mahjub, Faradmal, and Soltanian 2016; Goli, Mahjub, Faradmal, Mashayekhi, et al. 2016)	Det.	-
ANN	ANN-CDP	Liestøl *et al.* (1994) (Liestol, Andersen, and Andersen 1994)	Prob.	-
ANN	ANN-COX	Faraggi and Simon (1995) (Faraggi and Simon 1995)	Rank	-
ANN	PLANN	Biganzoli *et al.* (1998) (E. Biganzoli et al. 1998)	Prob.	-
ANN	COX-NNET	Ching *et al.* (2018) (Ching, Zhu, and Garmire 2018)	Prob.	**cox-nnet*** (Ching 2015)
ANN	DeepSurv	Katzman *et al.* (2018) (J. L. Katzman et al. 2018)	Prob.	**survivalmodels** (R. Sonabend 2020)
ANN	DeepHit	Lee *et al.* (2018) (C. Lee et al. 2018)	Prob.	**survivalmodels**

Class[1]	Name[2]	Authors (Year)[3]	Task[4]	Implementation[5]
ANN	Nnet-survival	Gensheimer & Narasimhan (2019) (Michael F. Gensheimer and Narasimhan 2019)	Prob.	**survivalmodels**
ANN	Cox-Time	Kvamme *et al.* (2019) (Kvamme, Borgan, and Scheel 2019)	Prob.	**survivalmodels**
ANN	PC-Hazard	Kvamme & Borgan (2019) (**Kvamme2019?**)	Prob.	**survivalmodels**
ANN	RankDeepSurv	Jing *et al.* (2019) (Jing et al. 2019)	Det.	**RankDeepSurv***,† (Jing et al. 2018)
ANN	DNNSurv	Zhao & Fend (2020) (Zhao and Feng 2020)	Prob.	**survivalmodels**

Table 6.1.: Summarising the models discussed in (Section 6.1) by their model class and respective survival task.

* 1. Model Class. RSF – Random Survival Forest; GBM – Gradient Boosting Machine; SVM – Support Vector Machine; ANN – Artificial Neural Network. There is some abuse of notation here as some of the RSFs are actually decision trees and some GBMs do not use gradient boosting. * 2. Model identifier used in this section and (R. E. B. Sonabend 2021). * 3. Authors and year of publication, for RSFs this is the paper most attributed to the algorithm. * 4. Survival task type: Deterministic (Det.), Probabilistic (Prob.), Ranking (Rank). * 5. If available in R then the package in which the model is implemented, otherwise '$*$' signifies a model is only available in Python. With the exception of DNNSurv, all ANNs in **survivalmodels** are implemented from the Python package **pycox** (Kvamme 2018) with **reticulate** (Ushey, Allaire, and Tang 2020). * † – Code available to create model but not implemented 'off-shelf'.

7. Tree-Based Methods

TODO (150-200 WORDS)

7.1. Random Forests

7.1.1. Random Forests for Regression

Random forests are a composite algorithm built by fitting many simpler component models, decision trees, and then averaging the results of predictions from these trees. Decision trees are first briefly introduced before the key 'bagging' algorithm that composes these trees to a random forest. Woodland terminology is used throughout this subsection.

Decision Trees

Decision trees are a common model class in machine learning and have the advantage of being (relatively) simple to implement and highly interpretable. A decision tree takes a set of inputs and a given *splitting rule* in order to create a series of splits, or branches, in the tree that culminates in a final *leaf*, or *terminal node*. Each terminal node has a corresponding prediction, which for regression is usually the sample mean of the training outcome data. This is made clearer by example, (Figure 7.1) demonstrates a decision tree predicting the miles per gallon (`mpg`) of a car from the `mtcars` (Henderson and Velleman 1981) dataset. With this tree a new prediction is made by feeding the input variables from the top to the bottom, for example given new data, $x = \{\text{`}wt\text{`} = 3, \text{`}disp\text{`} = 250\}$, then in the first split the right branch is taken as $\texttt{wt} = 3 > 2.32$ and in the second split the left branch is taken as $\texttt{disp} = 250 \leq 258$, therefore the new data point 'lands' in the final leaf and is predicted to have an `mpg` of 20.8. This value of 20.8 arises as the sample mean of `mpg` for the 11 (which can be seen in the box) observations in the training data who were sorted into this terminal node. Algorithmically, as splits are always binary, predictions are simply a series of conditional logical statements.

Splitting Rules

Precisely how the splits are derived and which variables are utilised is determined by the splitting rule.[1] In regression, the most common splitting rule is to select the cut-off for a given variable

[1] Other methods for growing trees such as pruning are not discussed here as they are less relevant to random forests, which are primarily of interest. Instead see (e.g.) Breiman (1984) [@Breiman1984].

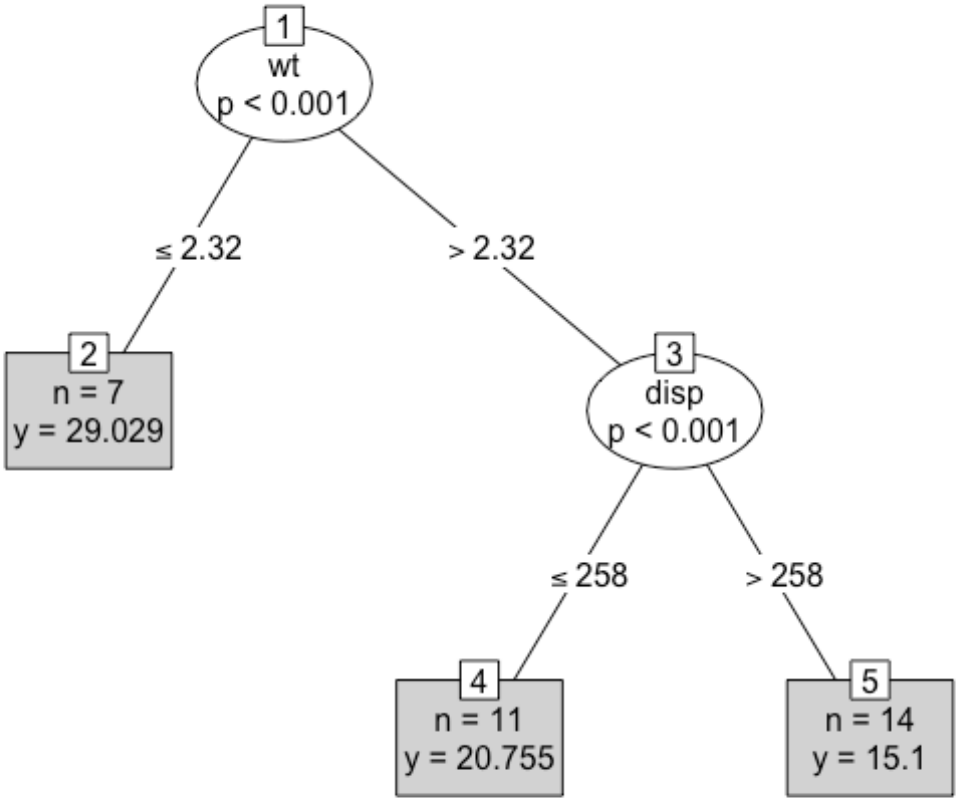


Figure 7.1.: Demonstrating classification trees using the `mtcars` (Henderson and Velleman 1981) dataset and the **party** (Hothorn, Hornik, and Zeileis 2006) package. Ovals are leaves, which indicate the variable that is being split. Edges are branches, which indicate the cut-off at which the variable is split. Rectangles are terminal nodes and include information about the number of training observations in the node and the terminal node prediction.

that minimises the mean squared error in each hypothetical resultant leaf. The goal is to find the variable and cutoff that leads to the greatest difference between the two resultant leaves and thus the maximal homogeneity within each leaf. For all decision tree and random forest algorithms going forward, let L denote some leaf, then let L_{xy}, L_x, L_y respectively be the set of observations, features, and outcomes in leaf L. Let $L_{y;i}$ be the ith outcome in L_y and finally let $L_{\bar{y}} = \frac{1}{n}\sum_{i=1}^{n} L_{y;i}$. To simplify notation, $i \in L$ is taken to be equivalent to $i \in \{i : X_i \in L_X\}$, i.e. the indices of the observations in leaf L.

Let $c \in \mathbb{R}$ be some cutoff parameter and let $L^a_{xy}(j,c) \coloneqq \{(X_i, Y_i) | X_{ij} < c, i = 1, ..., n\}, L^b_{xy}(j,c) = \{(X_i, Y_i) | X_{ij} \geq c, i = 1, ..., n\}$ be the two leaves containing the set of observations resulting from partitioning variable j at cutoff c. Then a split is determined by finding the arguments, (j^*, c^*) that minimise the sum of the mean squared errors (MSE) in both leaves (James et al. 2013),

$$(j^*, c^*) = \underset{j,c}{\operatorname{argmin}} \sum_{y \in L^a_y(j,c)} (y - L^a_{\bar{Y}}(j,c))^2 + \sum_{y \in L^b_y(j,c)} (y - L^b_{\bar{Y}}(j,c))^2 \tag{7.1}$$

This method is repeated from the first branch of the tree down to the very last such that observations are included in a given leaf L if they satisfy all conditions from all previous branches; features may be considered multiple times in the growing process. This is an intuitive method as minimising the above sum results in the set of observations within each individual leaf being as similar as possible, thus as an observation is passed down the tree, it becomes more similar to the subsequent leaves, eventually landing in a leaf containing homogeneous observations. Controlling how many variables to consider at each split and how many splits to make are determined by hyper-parameter tuning.

Decision trees are a powerful method for high-dimensional data as only a small sample of variables will be used for growing a tree, and therefore they are also useful for variable importance by identifying which variables were utilised in growth (other importance methods are also available). Decision trees are also highly interpretable, as demonstrated by (Figure 7.1). The recursive pseudo-algorithm in ((**alg-dt-fit?**)) demonstrates the simplicity in growing a decision tree (again methods such as pruning are omitted).

Algorithm 1 Fitting a decision tree.
Input Training data, $\mathcal{D}_0$. Splitting rule, SR.
Output Fitted decision tree, $\hat{g}$.

1: Compute (j^*, c^*) as the optimisers of SR (e.g. (@eq-dt-min)) to create the initial leaf and branches.
2: Repeat step 1 on all subsequent branches until a stopping rule is reached.
3: Return the fitted tree, $\hat{g}$, as the series of branches.

Stopping Rules

The 'stopping rule' in ((**alg-dt-fit?**)) is usually a condition on the number of observations in each leaf such that leaves will continue to be split until some minimum number of observations has been reached in a leaf. Other conditions may be on the 'depth' of the tree, which corresponds to the number of levels of splitting, for example the tree in (Figure 7.1) has a depth of 2 (the first level is not counted).

Random Forests

Despite being more interpretable than other machine learning methods, decision trees usually have poor predictive performance, high variance and are not robust to changes in the data. As such, *random forests* are preferred to improve prediction accuracy and decrease variance. Random forests utilise bootstrap aggregation, or *bagging* (Breiman 1996), to aggregate many decision trees. A pseudo fitting algorithm is given in ((**alg-rsf-fit?**)).

Algorithm 2 Fitting a random forest.
Input Training data, $\mathcal{D}_0$. Total number of trees, $B \in \mathbb{N}_{>0}$.
Output Fitted random forest, $\hat{g}$.

1: **for** $b = 1, ..., B$ **do**
2: Create a bootstrapped sample of the data, D_b.
3: Grow a decision tree, $\hat{g}_b$, on D_b with (@alg-dt-fit).
4: **end for**
5: $\hat{g} \leftarrow \{\hat{g}_b\}_{b=1}^{B}$ **return** $\hat{g}$

Prediction from a random forest follows by making predictions from the individual trees and aggregating the results by some function σ ((**alg-rsf-pred?**)); σ is usually the sample mean for regression,

$$\hat{g}(X^*) = \sigma(\hat{g}_1(X^*), ..., \hat{g}_B(X^*)) = \frac{1}{B}\sum_{b=1}^{B} \hat{g}_b(X^*)$$

where $\hat{g}_b(X^*)$ is the terminal node prediction from the bth tree and B are the total number of grown trees (`$B$' is commonly used instead of`N' to note the relation to bootstrapped data).

Algorithm 3 Predicting from a random forest.
Input Testing data $X^* \sim \mathcal{X}$, fitted forest $\hat{g}$ with $B \in \mathbb{N}_{>0}$ trees, aggregation method σ.
Output Prediction, $\hat{Y} \sim \mathcal{Y}$.

1: **for** $b = 1, ..., B$ **do**
2: 'Drop' X^* down the tree $\hat{g}_b$ individually to return a prediction $\hat{g}_b(X^*)$.
3: **end for**
4: $\hat{Y} \leftarrow \sigma(\hat{g}_1(X^*), ..., \hat{g}_B(X^*))$ **return** $\hat{Y}$

Usually many (hundreds or thousands) trees are grown, which makes random forests robust to changes in data and 'confident' about individual predictions. Other advantages include having several tunable hyper-parameters, including: the number of trees to grow, the number of variables to include in a single tree, the splitting rule, and the minimum terminal node size. Machine learning models with many hyper-parameters, tend to perform better than other models as they can be fine-tuned to the data, which is why complex deep learning models are often the best performing. Although as a caveat: too many parameters can lead to over-fitting and tuning many parameters can take a long time and be highly intensive. Random forests lose the interpretability of decision trees and are considered 'black-box' models as individual predictions cannot be easily scrutinised.

7.1.2. Random Forests for Survival Analysis

Given time constraints and the scope of this thesis, this survey of random forests for survival analysis will primarily focus on 'traditional' decision trees and random forests and will not look at other sub-fields such as causal forests. A comprehensive review of random survival forests (RSFs) is provided in Bou-Hamad (2011) (Bou-Hamad, Larocque, and Ben-Ameur 2011), which includes extensions to time-varying covariates and different censoring types. In order to prevent overlap, this survey will focus primarily on methods that have off-shelf implementations, their prediction types, and how successfully these methods handle the problem of censoring. Random forests and decision trees for survival are termed from here as Random Survival Forests (RSFs) and Survival Decision Trees (SDTs) respectively.

Unlike other machine learning methods that may require complex changes to underlying algorithms, individual components of a random forest can be adapted without altering the fundamental algorithm. The principle random forest algorithm is unchanged for RSFs, the difference is in the choice of splitting rule and terminal node prediction, which both must be able to handle censoring. Therefore instead of discussing individual algorithms, the different choices of splitting rules and terminal node predictions are discussed, then combinations of these are summarised into five distinct algorithms.

7.1.2.1. Splitting Rules

Survival trees and RSFs have been studied for the past four decades and whilst the amount of splitting rules to appear could be considered "numerous" (Bou-Hamad, Larocque, and Ben-Ameur 2011), only two broad classes are commonly utilised and implemented (H. Ishwaran and Kogalur 2018; Pölsterl 2020; Therneau and Atkinson 2019; Wright and Ziegler 2017). The first class rely on hypothesis tests, and primarily the log-rank test, to maximise dissimilarity between splits, the second class utilises likelihood-based measures. The first is discussed in more detail as this is common in practice and is relatively straightforward to implement and understand, moreover it has been demonstrated to outperform other splitting rules (Bou-Hamad, Larocque, and Ben-Ameur 2011). Likelihood rules are more complex and require assumptions that may not be realistic, these are discussed briefly.

Hypothesis Tests

The log-rank test statistic has been widely utilised as the 'natural' splitting-rule for survival analysis (Ciampi et al. 1986; B. H. Ishwaran et al. 2008; LeBlanc and Crowley 1993; Segal 1988). The log-rank test compares the survival distributions of two groups and has the null-hypothesis that both groups have the same underlying risk of (immediate) death, i.e. identical hazard functions.

Let L^A and L^B be two leaves then using the notation above let h^A, h^B be the (true) hazard functions derived from the observations in the two leaves respectively. The log-rank hypothesis test is given by $H_0 : h^A = h^B$ with test statistic (Segal 1988),

$$LR(L^A) = \frac{\sum_{\tau \in \mathcal{U}_D}(d_\tau^A - e_\tau^A)}{\sqrt{\sum_{\tau \in \mathcal{U}_D} v_\tau^A}}$$

where d_τ^A is the observed number of deaths in leaf A at τ,

$$d_\tau^A := \sum_{i \in L^A} \mathbb{I}(T_i = \tau, \Delta_i = 1)$$

e_τ^A is the expected number of deaths in leaf A at τ,

$$e_\tau^A := \frac{n_\tau^A d_\tau}{n_\tau}$$

and v_τ^A is the variance of the number of deaths in leaf A at τ,

$$v_\tau^A := e_\tau^A \Big(\frac{n_\tau - d_\tau}{n_\tau}\Big)\Big(\frac{n_\tau - n_\tau^A}{n_\tau - 1}\Big)$$

where $\mathcal{U}_D$ is the set of unique death times across the data (in both leaves), $\setminus\ n_\tau = \sum_i \mathbb{I}(T_i \geq \tau)$ is the number of observations at risk at τ in both leaves, $\setminus\ n_\tau^A = \sum_{i \in L^A} \mathbb{I}(T_i \geq \tau)$ is the number of observations at risk at τ in leaf A, and $\setminus\ d_\tau = \sum_i \mathbb{I}(T_i = \tau, \Delta_i = 1)$ is the number of deaths at τ in both leaves.

Intuitively these results follow as the number of deaths in a leaf is distributed according to Hyper$(n_\tau^A, n_\tau, d_\tau)$. The same statistic results if L^B is instead considered. ((**alg-dt-fit?**)) follows for fitting decision trees with the log-rank splitting rule, SR, to be maximised.

The higher the log-rank statistic, the greater the dissimilarity between the two groups, thereby making it a sensible splitting rule for survival, moreover it has been shown that it works well for splitting censored data (LeBlanc and Crowley 1993).[2] When censoring is highly dependent on the outcome, the log-rank statistic does not perform well and is biased (Bland and Altman 2004), which tends to be true of the majority of survival models. Additionally, the log-rank test requires no knowledge about the shape of the survival curves or distribution of the outcomes in either group (Bland and Altman 2004), making it ideal for an automated process that requires no user intervention.

The log-rank *score* rule (Hothorn and Lausen 2003) is a standardized version of the log-rank rule that could be considered as a splitting rule, though simulation studies have demonstrated non-significant predictive performance when comparing the two (B. H. Ishwaran et al. 2008). \\ Alternative dissimiliarity measures and tests have also been suggested as splitting rules, including modified Kolmogorov-Smirnov test and Gehan-Wilcoxon tests (Ciampi et al. 1988). Simulation studies have demonstrated that both of these may have higher power and produce 'better' results than the log-rank statistic (Fleming et al. 1980). Despite this, these do not appear to be in common usage and no implementation could be found that include these.

Likelihood Based Rules {.unnumbered .unlisted} Likelihood ratio statistics, or deviance based splitting rules, assume a certain model form and thereby an assumption about the data. This may be viewed as an advantageous strategy, as it could arguably increase interpretability, or a disadvantage as it places restrictions on the data. For survival models, a full-likelihood can be estimated with a Cox form by estimating the cumulative hazard function (LeBlanc and Crowley

[2]The results of this experiment are actually in LeBlanc's unpublished 1989 PhD thesis and therefore it has to be assumed that LeBlanc is accurately conveying its results in this 1993 paper.

1992). LeBlanc and Crowley (1992) (LeBlanc and Crowley 1992) advocate for selecting the optimal split by maximising the full PH likelihood, assuming the cumulative hazard function, H, is known,

$$\mathcal{L} := \prod_{m=1}^{M} \prod_{i \in L^m} h_m(T_i)^{\Delta_i} \exp(-H_m(T_i))$$

where M is the total number of terminal nodes, h_m and H_m are the (true) hazard and cumulative hazard functions in the mth node, and again L^m is the set of observations in terminal node m. Estimation of h_m and H_m are described with the associated terminal node prediction below.

The primary advantage of this method is that any off-shelf regression software with a likelihood splitting rule can be utilised without any further adaptation to model fitting by supplying this likelihood with required estimates. However the additional costs of computing these estimates may outweigh the benefits once the likelihood has been calculated, and this could be why only one implementation of this method has been found (Bou-Hamad, Larocque, and Ben-Ameur 2011; Therneau and Atkinson 2019).

Other Splitting Rules

As well as likelihood and log-rank spitting rules, other papers have studied comparison of residuals (Therneau, Grambsch, and Fleming 1990), scoring rules (H. Ishwaran and Kogalur 2018), and distance metrics (Gordon and Olshen 1985). These splitting rules work similarly to the mean squared error in the regression setting, in which the score should be minimised across both leaves. The choice of splitting rule is usually data-dependent and can be treated as a hyper-parameter for tuning. However if there is a clear goal in prediction, then the choice of splitting rule can be informed by the prediction type. For example, if the goal is to maximise separation, then a log-rank splitting rule to maximise homogeneity in terminal nodes is a natural starting point. Whereas if the goal is to estimate the linear predictor of a Cox PH model, then a likelihood splitting rule with a Cox form may be more sensible.

7.1.2.2. Terminal Node Prediction

Only two terminal node predictions appear in common usage.

Predict: Ranking

Terminal node ranking predictions for survival trees and forests have been limited to those that use a likelihood-based splitting rule and assume a PH model form (H. Ishwaran et al. 2004; LeBlanc and Crowley 1992). In model fitting the likelihood splitting rule model attempts to fit the (theoretical) PH model $h_m(\tau) = h_0(\tau)\theta_m$ for $m \in 1, ..., M$ where M is the total number of terminal nodes and θ_m is a parameter to estimate. The model returns predictions for $\exp(\hat{\theta}_m)$ where $\hat{\theta}_m$ is the estimate of θ_m. This is estimated via an iterative procedure in which in iteration $j+1$, $\hat{\theta}_m^{j+1}$ is estimated by

$$\hat{\theta}_m^{j+1} = \frac{\sum_{i \in L^m} \Delta_i}{\sum_{i \in L^m} \hat{H}_0^j(T_i)}$$

where as before L^m is the set of observations in leaf m and

$$\hat{H}_0^j(\tau) = \frac{\sum_{i:T_i \leq \tau} \Delta_i}{\sum_{m=1}^{M} \sum_{\{i:i \in \mathcal{R}_\tau \cap L^a\}} \hat{\theta}_m^j}$$

which is repeated until some stopping criterion is reached. The same cumulative hazard is estimated for all nodes however $\hat{\theta}_m$ varies across nodes. This method lends itself naturally to a composition to a full distribution (Chapter 12) as it assumes a PH form and separately estimates the cumulative hazard and relative risk (Section 7.1.3), though no implementation of this composition could be found.

Predict: Survival Distribution

The most common terminal node prediction appears to be predicting the survival distribution by estimating the survival function, using the Kaplan-Meier or Nelson-Aalen estimators, on the sample in the terminal node (Hothorn et al. 2004; B. H. Ishwaran et al. 2008; LeBlanc and Crowley 1993; Segal 1988). Estimating a survival function by a non-parametric estimator is a natural choice for terminal node prediction as these are natural 'baselines' in survival, similarly to taking the sample mean in regression. The prediction for SDTs is straightforward, the non-parametric estimator is fit on all observations in each of the terminal nodes. This is adapted to RSFs by bagging the estimator across all decision trees (Hothorn et al. 2004). Using the Nelson-Aalen estimator as an example, let m be a terminal node in an SDT, then the terminal node prediction is given by,

$$\hat{H}_m(\tau) = \sum_{\{i:i \in L^m \cap T_i \leq \tau\}} \frac{d_i}{n_i} \tag{7.2}$$

where d_i and n_i are the number of events and observations at risk at time T_i in terminal node m. Ishwaran (B. H. Ishwaran et al. 2008) defined the bootstrapped Nelson-Aalen estimator as

$$\hat{H}_{Boot}(\tau) = \frac{1}{B} \sum_{b=1}^{B} \hat{H}_{m,b}(\tau), \quad m \in 1, ..., M \tag{7.3}$$

where B is the total number of bootstrapped estimators, M is the number of terminal nodes, and $\hat{H}_{m,b}$ is the cumulative hazard for the mth terminal node in the bth tree. The bootstrapped Kaplan-Meier estimator is calculated analogously. More generally these can be considered as a uniform mixture of B distributions (Chapter 12). \\ All implemented RSFs can now be summarised into the following five algorithms: \\ **RRT** {#mod-rrt}\ LeBlanc and Crowley's (1992) (LeBlanc and Crowley 1992) survival decision tree uses a deviance splitting rule with a terminal node ranking prediction, which assumes a PH model form. These 'relative risk trees' (RRTs) are implemented in the package **rpart** (Therneau and Atkinson 2019). This model is considered the least accessible and transparent of all discussed in this section as: few implementations exist, it requires assumptions that may not be realistic, and predictions are harder to interpret than other models. Predictive performance of the model is expected to be worse than RSFs as this is a decision tree; this is confirmed in (R. E. B. Sonabend 2021). \\ **RRF** {#mod-rrf}\ Ishwaran *et al.* (2004) (H. Ishwaran et al. 2004) proposed a random forest framework for the relative risk trees, which makes a slight adaptation and applies the iteration of the terminal node prediction after the tree is grown as opposed to during the growing process. No implementation for these 'relative risk forests' (RRFs) could be found or any usage in the literature. Therefore RRFs are also considered not to be APT

for the same reasons given to the RRTs, except that in this case the predictive performance of RRFs is simply unknown (though can reasonably be expected to outperform an RRT). \\ **RSDF-DEV** {#mod-rsdfdev}\ Hothorn *et al.* (2004) (Hothorn et al. 2004) expanded upon the RRT by introducing a bagging composition thus creating a random forest with a deviance splitting rule, again assuming a PH form. However the ranking prediction is altered to be a bootstrapped Kaplan-Meier prediction in the terminal node. This is implemented in **ipred** (Peters and Hothorn 2019). This model improves upon the accessibility and transparency of the RRT by providing a more straightforward and interpretable terminal node prediction. However, as this is a decision tree, predictive performance is again expected to be worse than the RSFs. \\ **RSCIFF** {#mod-rsciff}\ Hothorn *et al.*[@Hothorn2005] studied a conditional inference framework in order to predict log-survival time. In this case the splitting rule is based on an IPC weighted loss function, which allows implementation by off-shelf classical random forests. The terminal node predictions are a weighted average of the log-survival times in the node where weighting is determined by the Kaplan-Meier estimate of the censoring distribution. This 'random survival conditional inference framework forest' (RSCIFF) is implemented in **party** (Hothorn, Hornik, and Zeileis 2006) and **partykit** (Hothorn and Zeileis 2015), which additionally includes a distribution terminal node prediction via the bootstrapped Kaplan-Meier estimator. The survival tree analogue (SDCIFT) is implemented in the same packages. Implementation of the RSCIFF is complex, which is likely why all implementations (in the above packages) are by the same authors. The complexity of conditional inference forests may also be the reason why several reviews, including this one, mention (or completely omit) RSCIFFs but do not include any comprehensive details that explain the fitting procedure (Bou-Hamad, Larocque, and Ben-Ameur 2011; H. Wang and Li 2017). In this regard, it is hard to claim that RSCIFFs are transparent or accessible. Moreover the authors of the model state that random conditional inference forests are for "expert user[s] only and [their] current state is rather experimental" (Hothorn and Zeileis 2015). Finally with respect to model performance, there is evidence that they can outperform RSDFs (below) dependent on the data type (Nasejje et al. 2017) however no benchmark experiment could be found that compared them to other models. \\ **RSDF-STAT** {#mod-rsdfstat}\ Finally Ishwaran *et al.* (2008) (B. H. Ishwaran et al. 2008) proposed the most general form of RSFs with a choice of hypothesis tests (log-rank and log-rank score) and survival measure (Brier, concordance) splitting rules, and a bootstrapped Nelson-Aalen terminal node prediction. These are implemented in **randomForestSRC** (H. Ishwaran and Kogalur 2018) and **ranger** (Wright and Ziegler 2017). This final class of RSFs are likely the only class that can be considered APT. There are several implementations of these models across programming languages, and extensive details for the fitting and predicting procedures, which makes them very accessible. The models utilise a standard random forest framework, which makes them transparent and familiar to those without expert Survival knowledge. Moreover they have been proven to perform well in benchmark experiments, especially on high-dimensional data (Herrmann et al. 2021; Spooner et al. 2020).

7.1.3. Novel Adaptations

Based on this survey of RSFs, a couple of novel adaptations may be considered as natural extensions.

Parametric Terminal Node Predictions

All probabilistic RSFs make use of a non-parametric estimator for the terminal node prediction. As an adaptation one could fit a semi- or fully-parametric model in the terminal nodes. However this could suffer from the problem of increased complexity/run-time, as well as overfitting, though is a sensible method worth considering. Alternatively a random forest for inference could be designed whereby a theoretical (say Weibull) survival distribution is assumed and the terminal node predictions are then MLE (or other inference method) estimates for the distribution parameters.

RRT and RRF Composition

As discussed above, Ishwaran's Relative Risk Forest makes a relative risk prediction in each terminal node (Section 7.1.2.2) by fitting

$$\hat{H}_{h;b}(\tau) = \hat{H}_{0;b}(\tau)\hat{\theta}_h$$

in which $\hat{H}_{0;b}(\tau)$ and $\hat{\theta}_h$ are iteratively updated and the final prediction is $\hat{\theta}_h$. A natural alternative would be to return the bootstrapped survival distribution prediction over $\hat{H}_{h;b}(\tau)$, instead of only returning $\hat{\theta}$. Ishwaran *et al.* allude to this prediction type in Section 3.2 of the 2004 paper (H. Ishwaran et al. 2004), however this is not formalised or implemented. It would be natural to first consider this for RRTs (before extension to RRFs) and implementation would likely be straightforward as any software must first estimate $\hat{H}_{0;b}(\tau)$ and $\hat{\theta}_h$.

7.1.4. Conclusions

Random forests are a highly flexible algorithm that allow the various components to be adapted and altered without major changes to the underlying algorithm. The result is that relatively few R implementations of RSFs cover almost half a century's worth of developments. The only algorithm that does not seem to be implemented is the relative risk forest.

Of the methods reviewed, only one can be considered APT for survival predictions. A lack of accessibility, transparency, or proven performance makes RRT and RSDF-DEV a poor choice for model fitting. RSCIFF is potentially a powerful method with promising results in benchmark experiments, but even the authors recognise its complexity prevents it from being accessible. Ishwaran's RSFs on the other hand are APT and suitable for model fitting and deployment. Simulation studies have demonstrated that RSFs can perform well even with high levels of censoring and there is evidence that on some datasets these can outperform a Cox PH (B. H. Ishwaran et al. 2008). Despite only one of the five models discussed here being APT, Ishwaran's model is highly flexible, and its implementation in software packages reflects this. Therefore one can still confidently conclude that random forests are a powerful algorithm in regression, classification, and survival analysis.

8. Support Vector Machines

TODO (150-200 WORDS)

8.1. Support Vector Machines

8.1.1. SVMs for Regression

In the simplest explanation, support vector machines (SVMs) (Cortes and Vapnik 1995) fit a hyperplane, g, on given training data and make predictions for new values as $\hat{g}(X^*)$ for some testing covariate X^*. One may expect the hyperplane to be fit so that all training covariates would map perfectly to the observed labels (a 'hard-boundary') however this would result in overfitting and instead an acceptable ('soft'-)boundary of error, the 'ϵ-tube', dictates how 'incorrect' predictions may be, i.e. how large an underestimate or overestimate. (Figure 8.1) visualises support vector machines for regression with a linear hyperplane g, and an acceptable boundary of error within the dashed lines (the ϵ-tube). SVMs are not limited to linear boundaries and *kernel* functions are utilised to specify more complex hyperplanes. Exact details of the optimization/separating procedure are not discussed here but many off-shelf 'solvers' exist in different programming languages for fitting SVMs. \\ In the regression setting, the goal of SVMs is to estimate the function

$$g : \mathbb{R}^p \to \mathbb{R}; \quad (x) \mapsto x\beta + \beta_0 \tag{8.1}$$

by estimation of the weights $\beta \in \mathbb{R}^p, \beta_0 \in \mathbb{R}$ via the optimisation problem

$$\begin{aligned} &\min_{\beta,\beta_0,\xi,\xi^*} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ &\text{s.t.} \begin{cases} Y_i - g(X_i) & \leq \epsilon + \xi_i \\ g(X_i) - Y_i & \leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* & \geq 0, \ i = 1, ..., n \end{cases} \end{aligned} \tag{8.2}$$

where $C \in \mathbb{R}$ is the regularization/cost parameter, ξ_i, ξ_i^* are slack parameters and ϵ is a margin of error for observations on the wrong side of the hyperplane, and g is defined in (Equation 8.1). The effect of the slack parameters is seen in (Figure 8.1) in which a maximal distance from the ϵ-tube is dictated by the slack variables.

In fitting, the dual of the optimisation is instead solved and substituting the optimised parameters into (Equation 8.1) gives the prediction function,

$$\hat{g}(X^*) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)K(X^*, X_i) + \beta_0$$

where α_i, α_i^* are Lagrangrian multipliers and K is some kernel function.[1] The Karush-Kuhn-Tucker conditions required to solve the optimisation for α result in the key property of SVMs, which is that values $\alpha_i = \alpha_i^* = 0$ indicate that observation i is 'inside' the ϵ-tube and if $\alpha_i \neq 0$ or $\alpha_i^* \neq 0$ then i is outside the tube and termed a *support vector*. It is these 'support vectors' that influence the shape of the separating boundary. \\ The choice of kernel and its parameters, the regularization parameter C, and the acceptable error ϵ, are all tunable hyper-parameters, which makes the support vector machine a highly adaptable and often well-performing machine learning method. However the parameters C and ϵ often have no clear apriori meaning (especially true when predicting abstract rankings) and thus require extensive tuning over a great range of values; no tuning will result in a very poor model fit.

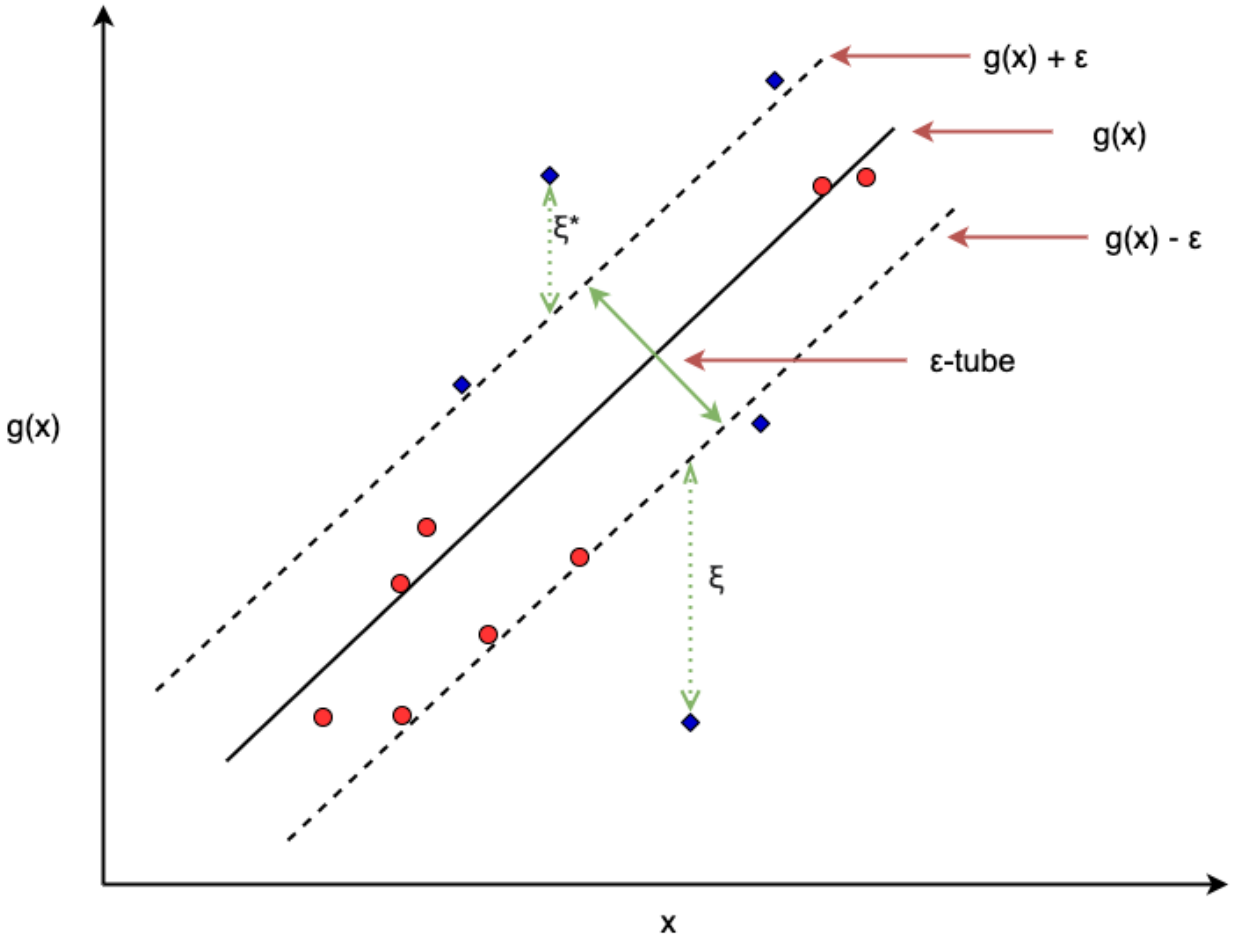


Figure 8.1.: Visualising a support vector machine with an ϵ-tube and slack parameters ξ and ξ^*. Red circles are values within the ϵ-tube and blue diamonds are values outside the tube; x-axis is single covariate, x, and y-axis is $g(x) = x\beta + \beta_0$.

[1] Discussion and derivation of these terms can be found in [@Hastie2013; @Vapnik1998]

8.1.2. SVMs for Survival Analysis

Similarly to random forests, all research for Survival Support Vector Machines (SSVMs) can be reduced to very few algorithms, in fact only one unique off-shelf algorithm is identified in this survey. No SSVM for distribution predictions exist, instead they either predict survival time, rankings, or a hybrid of the two.

Other reviews and surveys of SSVMs include a short review by Wang *et al.* (2017) (P. Wang, Li, and Reddy 2019) and some benchmark experiments and short surveys from Van Belle *et al.* (2011) (Vanya Van Belle, Pelckmans, Van Huffel, et al. 2011), Goli *et al.* (2016) (Goli, Mahjub, Faradmal, and Soltanian 2016) and Fouodo *et al.* (2018) (Fouodo et al. 2018). All the benchmark experiments in these papers indicate that the Cox PH performs as well as, if not better than, the SSVMs. \\ Initial attempts at developing SSVMs by Shivaswamy *et al.* (2007) (Shivaswamy, Chu, and Jansche 2007) took the most 'natural' course and attempt to treat the problem as a regression one with adjustments in the optimisation for censoring. These methods have a natural interpretation and are intuitive in their construction. Further development of these by Khan and Zubek (2008) (Khan and Bayer Zubek 2008) and Land *et al.* (2011) (Land et al. 2011) focused on different adjustments for censoring in order to best reflect a realistic survival data set-up. Simultaneously, ranking models were developed in order to directly optimise a model's discriminatory power. Developments started with the work of Evers and Messow (2008) (Evers and Messow 2008) but were primarily made by Van Belle *et al.* (2007)-(2011) (V. Van Belle et al. 2010; Vanya Van Belle et al. 2007, 2008; Vanya Van Belle, Pelckmans, Suykens, et al. 2011). These lack the survival time interpretation but are less restrictive in the optimisation constraints. Finally a hybrid of the two followed naturally from Van Belle *et al.* (2011) (Vanya Van Belle, Pelckmans, Van Huffel, et al. 2011) by combining the constraints from both the regression and ranking tasks. This hybrid method allows a survival time interpretation whilst still optimising discrimination. These hybrid models have become increasingly popular in not only SSVMs, but also neural networks (Section 10.1). Instead of presenting these models chronologically, the final hybrid model is defined and then other developments can be more simply presented as components of this hybrid. One model with an entirely different formulation is considered after the hybrid.

For all SSVMs defined in this section let: ξ_i, ξ_i^*, ξ_i' be slack variables; β, β_0 be model weights in $\mathbb{R}$; C, μ be regularisation hyper-parameters in $\mathbb{R}$; $(X_i, T_i, \Delta_i) \overset{i.i.d.}{\sim} (X, T, \Delta)$ be the usual training data; and $g(x) = x\beta + \beta_0$.

8.1.2.1. SSVM-Hybrid {.unnumbered .unlisted}

Van Belle *et al.* published several papers developing SSVMs, which culminate in the hybrid model here termed 'SSVM-Hybrid' (Vanya Van Belle, Pelckmans, Van Huffel, et al. 2011). The model is defined by the optimisation problem, \\ **SSVM-Hybrid**\

$$\min_{\beta,\beta_0,\xi,\xi',\xi^*} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n}\xi_i + \mu\sum_{i=1}^{n}(\xi_i' + \xi_i^*)$$

$$\text{s.t.} \begin{cases} g(X_i) - g(X_{j(i)}) \geq T_i - T_{j(i)} - \xi_i, \\ \Delta_i(g(X_i) - T_i) \leq \xi_i^* \\ T_i - g(X_i) \leq \xi_i' \\ \xi_i, \xi_i', \xi_i^* \geq 0, \quad \forall i = 1, ..., n \end{cases}$$

where $j(i) := \text{argmax}_{j \in 1,...n}\{T_j : T_j < T_i\}$ is an index discussed further below. A prediction for test data is given by,

$$\hat{g}(X^*) = \sum_{i=1}^{n} \alpha_i (K(X_i, X^*) - K(X_{j(i)}, X^*)) + \alpha_i^* K(X_i, X^*) - \Delta_i \alpha_i' K(X_i, X^*) + \beta_0$$

where $\alpha_i, \alpha_i^*, \alpha_i'$ are Lagrange multipliers and K is a chosen kernel function, which may have hyper-parameters to select or tune.

SVCR (Regression)

Examining the components of the SSVM-Hybrid model will help identify its relation to previously published SSVMs. First note the model's connection to the regression setting when on setting $C = 0$, removing the associated first constraint and ignoring Δ in the second constraint, the regression setting is exactly recovered:

$$\min_{\beta,\beta_0,\xi,\xi'} \frac{1}{2}\|\beta\|^2 + \mu \sum_{i=1}^{n} (\xi_i + \xi_i')$$
$$\text{s.t.} \begin{cases} g(X_i) - T_i \leq \xi_i \\ T_i - g(X_i) \leq \xi_i' \\ \xi_i, \xi_i' \geq 0, \quad \forall i = 1, ..., n \end{cases}$$

Note a slight difference in the formulation of this optimisation to the original regression problem, here no error component ϵ is directly included, instead this is part of the optimisation and considered as part of the slack parameters ξ_i, ξ_i'; effectively this is the same as setting $\epsilon = 0$. This formulation removes the ϵ-tube symmetry seen previously and therefore distinguishes more clearly between overestimates and underestimates, with each being penalised differently. Removing the ϵ parameter can lead to model overfitting as all points become support vectors, however careful tuning of other hyper-parameters can effectively control for this.

This formulation allows for clearer control over left-, right-, and un-censored observations. Clearly if an observation is uncensored then the true value is known and should be predicted exactly, hence under- and over-estimates are equally problematic and should be penalised the same. If an observation is right-censored then the true death time is greater than the observed time and therefore overestimates should not be heavily penalised but underestimates should be; conversely for left-censored observations.

This leads to the first SSVM for regression from Shivaswamy *et al.* (2007) (Shivaswamy, Chu, and Jansche 2007). \\ **SVCR**

$$\min_{\beta,\beta_0,\xi,\xi^*} \frac{1}{2}\|\beta\|^2 + \mu\Big(\sum_{i \in R} \xi_i + \sum_{i \in L} \xi_i^*\Big)$$
$$\text{s.t.} \begin{cases} g(X_i) - T_i \leq \xi_i^*, \quad \forall i \in R \\ T_i - g(X_i) \leq \xi_i, \quad \forall i \in L \\ \xi_i \geq 0, \forall i \in R; \xi_i^* \geq 0, \forall i \in L \end{cases}$$

where L is the set of observations who are either left- or un-censored, and R is the set of observations who are either right- or un-censored. Hence an uncensored observation is constrained on both sides as their true survival time is known, whereas a left-censored observation is constrained in the amount of 'over-prediction' and a right-censored observation is constrained by 'under-prediction'. This is intuitive as the only known for these censoring types are the lower and upper bounds of the actual survival time respectively.

Reducing this to the thesis scope of right-censoring only results in the optimisation:

$$\min_{\beta,\beta_0,\xi,\xi^*} \frac{1}{2}\|\beta\|^2 + \mu\Big(\sum_{i=1}^{n} \xi_i + \xi_i^*\Big)$$
$$\text{s.t.} \begin{cases} \Delta_i(g(X_i) - T_i) \leq \xi_i \\ T_i - g(X_i) \leq \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \\ \forall i \in 1, ..., n \end{cases}$$

which can be seen to be identical to SSVM-Hybrid when $C = 0$ and the first constraint is removed. Predictions are found by,

$$\hat{g}(X^*) = \sum_{i=1}^{n} \alpha_i^* K(X_i, X^*) - \Delta_i \alpha_i' K(X_i, X^*) + \beta_0$$

\\ The advantage of this algorithm is its simplicity. Clearly if no-one is censored then the optimisation is identical to the regression optimisation in (Equation 8.2). As there is no ϵ hyper-parameter, the run-time complexity is the same as, if not quicker than, a regression SVM. Both left- and right-censoring are handled and no assumptions are made about independent censoring. With respect to performance, benchmark experiments (Fouodo et al. 2018) indicate that the SVCR does not outperform a na"ive SVR (i.e. censoring ignored). The SVCR is implemented in the R package **survivalsvm** (Fouodo et al. 2018) and is referred to as 'regression'. \\ As discussed, the error margin for left- and right- censoring should not necessarily be equal and the penalty for each should not necessarily be equal either. Hence a natural extension to SVCR is to add further parameters to better separate the different censoring types, which gives rise to the SVRc (Khan and Bayer Zubek 2008). However this model is only briefly discussed as left-censoring is out of scope of this thesis and also the model is patented and therefore not easily accessible. The model is given by the optimisation, \\ **SVRc**

$$\min_{\beta,\beta_0,\xi,\xi^*} \frac{1}{2}\|\beta\|^2 + \sum_{i=1}^{n} C_i\xi_i + C_i^*\xi_i'$$
$$\text{s.t.} \begin{cases} g(X_i) - T_i \leq \epsilon_i' + \xi_i' \\ T_i - g(X_i) \leq \epsilon_i + \xi_i \\ \xi_i, \xi_i' \geq 0, \quad \forall i = 1, ..., n \end{cases}$$

Where $C_i = \Delta_i C_c + (1 - \Delta_i)C_n, \epsilon_i = \Delta_i\epsilon_c + (1 - \Delta_i)\epsilon_n$ and analogously for $C_i^*, C_C^*, \epsilon^*, ...$ The new hyper-parameters $C_c, C_n, \epsilon_c, \epsilon_n$ are the penalty for errors in censored predictions (c) and uncensored predictions (n) for left and right (*) censoring, and the acceptable margin of errors respectively. The rationale behind this algorithm is clear, by having asymmetric error margins the algorithm can penalise predictions that are clearly wrong whilst allowing predictions that may be correct

(but ultimately unknown due to censoring). Experiments indicate the model may have superior discrimination than the Cox PH (Khan and Bayer Zubek 2008) and SVCR (Du and Dua 2011). However these conclusions are weak as independent experiments do not have access to the patented model.

The largest drawback of the algorithm is a need to tune eight parameters. As the number of hyper-parameters to tune increases, so too does model fitting time as well as the risk of overfitting. The problem of extra hyper-parameters is the most common disadvantage of the model given in the literature (Fouodo et al. 2018; Land et al. 2011). Land *et al.* (2011) (Land et al. 2011) present an adaptation to the SVRc to improve model fitting time, termed the EP-SVRc, which uses Evolutionary Programming to determine the optimal values for the parameters. No specific model or algorithm is described, nor any quantitative results presented. No evidence can be found for this method being used since publication. The number of hyper-parameters in the SVRc, coupled with its lack of accessibility, outweigh the benefits of the claimed predictive performance and is therefore clearly not APT and will not be considered further.

8.1.2.2. SSVM-Rank {.unnumbered .unlisted}

The regression components of SSVM-Hybrid (8.1.2.1) have been fully examined, now turning to the ranking components and setting $\mu = 0$. In this case the model reduces to \\ **SSVM-Rank**

$$\min_{\beta,\beta_0,\xi} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n}\xi_i$$
$$\text{s.t.} \begin{cases} g(X_i) - g(X_{j(i)}) \geq T_i - T_{j(i)} - \xi_i, \\ \xi_i \geq 0, \quad \forall i = 1, ..., n \end{cases}$$

with predictions

$$\hat{g}(X^*) = \sum_{i=1}^{n} \alpha_i (K(X_i, X^*) - K(X_{j(i)}, X^*))$$

This formulation, termed here 'SSVM-Rank', has been considered by numerous authors in different forms, including Evers and Messow (Evers and Messow 2008) and Van Belle *et al.*[@VanBelle2007; @VanBelle2008; @VanBelle2011b]. The primary differences between the various models are in which observations are compared in order to optimise discrimination; to motivate why this matters, first observe the intuitive nature of the optimisation constraints. By example, define $k := T_i - T_{j(i)}$ and say $T_i > T_{j(i)}$. Then, in the first constraint, $g(X_i) - g(X_{j(i)}) \geq k - \xi_i$. As $k > 0$ and $\xi_i \geq 0$, it follows that $g(X_i) > g(X_{j(i)})$, hence creating a concordant ranking[2] which is the opposite to the between observations i (ranked higher) and $j(i)$; illustrating why this optimisation results in a ranking model.

This choice of comparing observations i and $j(i)$ (defined below) stems from a few years of research in an attempt to optimise the algorithm with respect to both speed and predictive performance. In the original formulation, RANKSVMC (Vanya Van Belle et al. 2007), the model ranks all possible pairs of observations. This is clearly infeasible as it increases the problem to a $\mathcal{O}(qn^2/2)$ runtime

[2]Note this ranking has the interpretation 'higher rank equals lower risk'.

where q is the proportion of non-censored observations out of a total sample size n (Vanya Van Belle et al. 2008). The problem was reduced by taking a nearest neighbours approach and only considering the kth closest observations (Vanya Van Belle et al. 2008). Simulation experiments determined that the single nearest neighbour was sufficient, thus arriving at $j(i)$, the observation with the largest observed survival time smaller than T_i,

$$j(i) := \underset{j \in 1,...n}{\operatorname{argmax}} \{T_j : T_j < T_i\}$$

This requires that the first observation is taken to be an event, even if it is actually censored. In practice, sorting observations by survival time then greatly speeds up the model run-time (Fouodo et al. 2018). The RANKSVMC and SSVM-RANK are implemented in **survivalsvm** (Fouodo et al. 2018) and referred to as 'vanbelle1' and 'vanbelle2' respectively.

The hybrid model is repeated below with the ranking components in blue, the regression components in red, and the common components in black, clearly highlighting the composite nature of the model.

$$\min_{\beta,\beta_0,\xi,\xi',\xi^*} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n}\xi_i + \mu\sum_{i=1}^{n}(\xi_i' + \xi_i^*)$$

$$\text{s.t.} \begin{cases} g(X_i) - g(X_{j(i)}) \geq T_i - T_{j(i)} - \xi_i \\ \Delta_i(g(X_i) - T_i) \leq \xi_i^* \\ T_i - g(X_i) \leq \xi_i' \\ \xi_i, \xi_i', \xi_i^* \geq 0, \quad \forall i = 1, ..., n \end{cases}$$

and predictions are made with,

$$\hat{g}(X^*) = \sum_{i=1}^{n} \alpha_i(K(X_i, X^*) - K(X_{j(i)}, X^*)) + \alpha_i^* K(X_i, X^*) - \Delta_i \alpha_i' K(X_i, X^*) + \beta_0$$

The regularizer hyper-parameters C and μ now have a clear interpretation. C is the penalty associated with the regression method and μ is the penalty associated with the ranking method. By always fitting the hybrid models and tuning these two parameters, there is never a requirement to separately fit the regression or ranking methods as these would be automatically identified as superior in the tuning procedure. Moreover, the hybrid model retains the interpretability of the regression method and predictions can be interpreted as survival times. The hybrid method is implemented in **survivalsvm** as 'hybrid'. By Van Belle's own simulation studies, these models do not outperform the Cox PH with respect to Harrell's C.

SSVR-MRL

Not all SSVMs can be considered a variant of the SSVM-Hybrid, though all prominent and commonly utilised suggestions do seem to have this formulation. One other algorithm of note is termed here the 'SSVM-MRL' (Goli, Mahjub, Faradmal, and Soltanian 2016; Goli, Mahjub, Faradmal,

Mashayekhi, et al. 2016), which is a regression SSVM. The algorithm is identical to SVCR with one additional constraint. \\ **SSVR-MRL**\

$$\min_{\beta,\beta_0,\xi,\xi^*,\xi'} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) + C^*\sum_{i=1}^{n}\xi_i'$$

$$\text{s.t.} \begin{cases} T_i - g(X_i) \leq \xi_i \\ \Delta_i(g(X_i) - T_i) \leq \xi_i^* \\ (1-\Delta_i)(g(X_i) - T_i - MRL(T_i|\hat{S})) \leq \xi_i' \\ \xi_i, \xi_i^*, \xi_i' \geq 0 \\ \forall i = 1, ..., n \end{cases}$$

where $MRL(T_i|\hat{S})$ is the 'mean residual lifetime' function (Klein and Moeschberger 2003)

$$MRL(\tau|\hat{S}) = \frac{\int_\tau^\infty \hat{S}(u)du}{\hat{S}(\tau)}$$

which is the area under the estimated survival curve (say by Kaplan Meier), $\hat{S}$, from point τ, weighted by the probability of being alive at point τ. This is interpreted as the expected remaining lifetime from point τ. On setting $C^* = 0$ and removing associated constraint three, this reduces exactly to the SVCR and similarly if there's no censoring then the standard regression setting is recovered. Unlike other strategies, no new hyper-parameters are introduced and Kaplan-Meier estimation should not noticeably impact run-time. There is no evidence of this model being used in practice, nor of any off-shelf implementation. Theoretically, the hybrid model could be expanded to include this extra penalty term and constraint (discussed below).

8.1.3. Novel Adaptations

Based on the above survey, one novel adaptation is proposed to merge the SSVM-Hybrid with SSVR-MRL. This is a simple addition in which one extra constraint (and associated penalty and slack parameter) is added in order to control for right-censored observations. The SSVM-Hybrid becomes, \\

$$\min_{\beta,\beta_0,\xi,\xi',\xi'',\xi^*} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n}\xi_i + \mu\sum_{i=1}^{n}(\xi_i' + \xi_i^*) + \gamma\sum_{i=1}^{n}\xi_i''$$

$$\text{s.t.} \begin{cases} g(X_i) - g(X_{j(i)}) \geq T_i - T_{j(i)} - \xi_i \\ (1-\Delta_i)(g(X_i) - T_i - MRL(T_i|\hat{S})) \leq \xi_i'' \\ \Delta_i(g(X_i) - T_i) \leq \xi_i^* \\ T_i - g(X_i) \leq \xi_i' \\ \xi_i, \xi_i', \xi_i^*, \xi_i'' \geq 0, \quad \forall i = 1, ..., n \end{cases}$$

Where the ranking (blue) and regression (red) components are unchanged but the additional MRL (magenta) constraint is added for censored observations. One additional parameter should not impact upon fitting time or overfitting too greatly, though this should be tested on large datasets. As with the combination of hybrid and ranking models, the additional constraint can be automatically 'tuned out' of the model, or just manually removed, by setting $\gamma = 0$.

8.1.4. Conclusions

Several SSVMs have been proposed for survival analysis. These can generally be categorised into 'regression' models that adapt SVMs to account for censoring and predict a survival time, 'ranking' models that predict a relative ranking in order to optimise measures of discrimination, and 'hybrid' models that optimise measures of discrimination but make survival time predictions. Other SSVMs that lie outside of these groupings are not able to solve the survival task (e.g. (Shiao and Cherkassky 2013)). Other SVM-type approaches could be considered, including relevance vector machines and import vector machines, however less work has been developed in these areas and further consideration is beyond the scope of this thesis.

The models that have received the most attention are SVCR, SSVM-Rank, and SSVM-Hybrid; the first two are special cases of SSVM-Hybrid. Judging if SSVM-Hybrid (and by extension SVCR and SSVM-Rank) is APT is not straightforward. On the one hand it could be considered transparent as SVMs have been studied for decades and the literature for SSVMs, especially from Van Belle, is extensive. On the other hand, the predictions from SSVM-Hybrid should be interpretable as survival times but first hand experience indicates that this is not the case (though this may be due to implementation), which calls into question whether the interpretation they claim to have is actually correct. For accessibility, there appears to be only one implementation of SSVMs in R[@pkgsurvivalsvm], and also only one in Python (Pölsterl 2020), which may be due to SSVMs being difficult to implement, even when several optimisation solvers exist off-shelf. Finally, there is no evidence that SSVMs outperform the Cox PH or baseline models and moreover they often perform worse (Fouodo et al. 2018; Vanya Van Belle, Pelckmans, Van Huffel, et al. 2011), which is also seen in (R. E. B. Sonabend 2021). Yet one cannot dismiss SSVMs outright as they often require extensive tuning to perform well, even in classification settings, and no benchmark experiment has yet to emerge for testing SSVMs with the required set-up.[3] Therefore SSVMs may not be APT for now but future developments will be worth paying attention to.

[3]Though one is in progress as a result of the work in [@Sonabend2021b].

9. Boosting Methods

TODO (150-200 WORDS)

9.1. Gradient Boosting Machines

9.1.1. Gradient Boosting Machines for Regression

Boosting is a machine learning strategy that can be applied to any model class. Similarly to random forests, boosting is an 'ensemble' method that creates a model from a 'committee' of learners. The committee is formed of 'weak' learners that make poor predictions individually, which creates a 'slow learning' approach (as opposed to 'greedy') that requires many iterations for a model to be a good fit to the data. Boosting models are similar to random forests in that both make predictions from a large committee of learners. However the two differ in how this committee is combined to a prediction. In random forest algorithms, each decision tree is grown independently and their predictions are combined by a simple mean calculation. In contrast, weak learners in a boosting model are fit sequentially and predictions are made by a linear combination of predictions from each learner. With respect to transparency, it is simpler to inspect 100 trees in a random forest, than it is to inspect 100 weak learners in a boosted model, though both are considered black-box models.

The best known boosting algorithm is likely AdaBoost (Freund and Schapire 1996), which is more generally a Forward Stagewise Additive Model (FSAM) with an exponential loss (Hastie, Tibshirani, and Friedman 2001). Today, the most widely used boosting model is the Gradient Boosting Machine (GBM) (J. H. Friedman 2001).

Training a GBM

Pseudo-code for training a componentwise GBM is presented in (7). The term 'componentwise' is explained fully below, only this variation of GBM is presented as it is the most common in implementation (Greenwell et al. 2019; Hothorn et al. 2020). Line 1: the initial function is initialized as $g_0 = 0$;[1] Line 2: iterate over boosting steps $m = 1, ..., M$ and; Line 3: randomly sample the training data, $\mathcal{D}_0$, to a smaller sample, $\mathcal{D}_0^*$, this may be ignored if $\phi = 1$; Line 4: for all training observations in the reduced dataset, $i \in \{i : X_i \in \mathcal{D}_0^*\}$, compute the negative gradient, r_{im}, of the differentiable loss function, L, with respect to predictions from the previous iteration,

[1] Some algorithms may instead initialize g_0 by finding the value that minimises the given loss function, however setting $g_0 = 0$ appears to be the most common practice for componentwise GBMs.

$g_{m-1}(X_i)$; Line 5: fit one weak learner for each feature, $j = 1, ..., p$, in the training data, where the feature, $X_{;j}$, is the single covariate and r_{im} are the labels; Line 6: select the optimal weak learner as the one that minimises the squared error between the prediction and the true gradient; Line 7: update the fitted model by adding the optimal weak learner with a shrinkage penalty, ν; Line 9: return the model updated in the final iteration as the fitted GBM.

Algorithm 4 Training a componentwise Gradient Boosting Machine.
Input Training data, $\mathcal{D}_0 = \{(X_1, Y_1), ..., (X_n, Y_n)\}$, where $(X_i, Y_i) \overset{i.i.d.}{\sim} (X, Y)$. Differentiable loss, L. Hyper-parameters: sampling fraction, $\phi \in (0, 1]$; step-size, $\nu \in (0, 1]$; number of iterations, $M \in \mathbb{R}_{>0}$.
Output Boosted model, $\hat{g}$.

1: Initialize $g_0 \leftarrow 0$
2: **for** $m = 1, ..., M$ **do**
3: $\quad \mathcal{D}_0^* \leftarrow$ Randomly sample $\mathcal{D}_0$ w.p. ϕ
4: $\quad r_{im} \leftarrow -[\frac{\partial L(y_i, g_{m-1}(X_i))}{\partial g_{m-1}(X_i)}], i \in \{i : X_i \in \mathcal{D}_0^*\}$
5: $\quad$ Fit p weak learners, w_j to $(X_i, r_{im}), j = 1, .., p$
6: $\quad j^* \leftarrow \text{argmin}_{j=1,..,p} \sum_{i \in \{i: X_i \in \mathcal{D}_0^*\}} (r_{im} - w_j(X_i))^2$
7: $\quad g_m \leftarrow g_{m-1} + \nu w_{j^*}$
8: **end for**
9: $\hat{g} \leftarrow g_M$ **return** $\hat{g}$

Predicting with a GBM

In general, predictions from a trained GBM are simple to compute as the fitted model (and all individual weak learners) take the same inputs, which are passed sequentially to each of the weak learners. In (7), the fitted GBM is a single model, which is a linear combination of weak learners. Instead one could think of the returned model as a collection of the optimal weak learners, i.e. let $w_{m;j^*}$ be the optimal weak learner from iteration m and let the fitted GBM (Line 9 (7)) be $\hat{g} := \{w_{m;j^*}\}_{m=1}^M$.[2] With this formulation, making predictions from the GBM can be demonstrated simply in ((**alg-surv-gbm-pred?**)).

Algorithm 5 Predicting from a Gradient Boosting Machine.
Input Fitted GBM, $\hat{g} := \{w_{m;j^*}\}_{m=1}^M$, trained with step-size ν. Testing data $X^* \sim \mathcal{X}$.
Output Prediction, $\hat{Y} \sim \mathcal{Y}$.

1: Initialize $\hat{Y} = 0$
2: **for** $m = 1, ..., M$ **do**
3: $\quad \hat{Y} \leftarrow \hat{Y} + \nu w_{m;j^*}(X^*)$
4: **end forreturn** $\hat{Y}$

The biggest advantages of boosting are firstly relatively few hyper-parameters, which all have a meaningful and intuitive interpretation, and secondly its modular nature means that, like random

[2] This formulation is computationally and mathematically identical to the formulation in (@alg-surv-gbm) and is practically more convenient for implementation, indeed this is the implementation in **mboost** [@pkgmboost]. Despite this, the formulation in (@alg-surv-gbm) is common in the literature, which often conflates model training and predicting.

forests, relatively few parts need to be updated to derive a novel model. First the model components will be discussed and then the hyper-parameters. Once this has been established, deriving survival variants can be simply presented.

9.1.1.1. Losses and Learners

Losses

Building a GBM requires selection of the loss to minimise, L, selection of weak learners, w_j, and a method to compare the weak learners to the loss gradient. The only constraint in selecting a loss, L, is that it must be differentiable w.r.t. $g(X)$ (Hastie, Tibshirani, and Friedman 2001). Of course a sensible loss should be chosen (a classification loss should not be used for regression) and different choices of losses will optimise different tasks. L_2-losses have been demonstrated to be effective for regression boosting, especially with high-dimensional data (Bühlmann and Yu 2003); this is referred to as L_2-boosting.

Weak Learners

(4) is specifically a *componentwise* GBM (Bühlmann and Yu 2003), which means that each of the p weak learners is fit on a single covariate from the data. This method simplifies selecting the possible choices for the weak learners to selecting the class of weak learner (below). Additionally, componentwise GBMs provide a natural and interpretable feature selection method as selecting the optimal learner ((7), line 6) corresponds to selecting the feature that minimises the chosen loss in iteration m.

Only three weak, or 'base', learner classes are commonly used in componentwise GBMs (Hothorn et al. 2020; Z. Wang and Wang 2010). These are linear least squares (J. H. Friedman 2001), smoothing splines (Bühlmann and Yu 2003), and decision stumps (Bühlmann and Yu 2003; J. H. Friedman 2001). Let L be a loss with negative gradient for observation i in the mth iteration, r_{im}, and let $\mathcal{D}_0$ be the usual training data. For linear least squares, an individual weak learner is fit by (J. H. Friedman 2001; Z. Wang and Wang 2010),

$$w_j(\mathcal{D}_0) = X_{;j} \frac{\sum_{i=1}^{n} X_{ij} r_{im}}{\sum_{i=1}^{n} (X_{ij})^2}$$

For smoothing splines, usually cubic splines are implemented, these fit weak learners as the minimisers of the equation (Bühlmann and Yu 2003),

$$w_j := \underset{g \in \mathcal{G}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (r_{im} - g(X_{ij}))^2 + \lambda \int (g''(u))^2 du$$

where g'' is the second derivative of g, $\mathcal{G}$ is the set of functions, $\setminus\, \mathcal{G} := \{g : g \text{ is twice continuously differentiable and } \infty\}$, and λ is a hyper-parameter usually chosen so that the number of degrees of freedom, df, is small, with df ≈ 4 suggested (Bühlmann and Yu 2003; Schmid and Hothorn 2008a; Z. Wang and Wang 2010).

Finally for decision stumps ((**?@fig-surv-stump**)), a decision tree, w_j, is grown ((**alg-dt-fit?**)) on $(X_{;j}, r_m)$ to depth one (equivalently to two terminal nodes) for each of the $j = 1, ..., p$ covariates (J. H. Friedman 2001).

9.1.1.2. Hyper-Parameters

The hyper-parameters in (7) are the 'step-size', ν, the sampling fraction, ϕ, and the number of iterations, M.

Number of iterations, M

The number of iterations is often claimed to be the most important hyper-parameter in GBMs and it has been demonstrated that as the number of iterations increases, so too does the model performance (with respect to a given loss on test data) up to a certain point of overfitting (Buhlmann 2006; Hastie, Tibshirani, and Friedman 2001; Schmid and Hothorn 2008a). This is an intuitive result as the foundation of boosting rests on the idea that weak learners can slowly be combined to form a single powerful model. This is especially true in componentwise GBMs as time is required to learn which features are important. Finding the optimal value of M is critical as a value too small will result in poor predictions, whilst a value too large will result in model overfitting. Two primary methods have been suggested for finding the optimal value of M. The first is to find the $M \in \mathbb{N}_{>0}$ that minimises a given measure based on the AIC (Akaike 1974), the second is the 'usual' empirical selection by nested cross-validation. In practice the latter method is usually employed.

Step-size, ν

The step-size parameter ((7), line 7), ν, is a shrinkage parameter that controls the contribution of each weak learner at each iteration. Several studies have demonstrated that GBMs perform better when shrinkage is applied and a value of $\nu = 0.1$ is often suggested (Buhlmann and Hothorn 2007; Hastie, Tibshirani, and Friedman 2001; J. H. Friedman 2001; D. K. K. Lee, Chen, and Ishwaran 2019; Schmid and Hothorn 2008a). The optimal values of ν and M depend on each other, such that smaller values of ν require larger values of M, and vice versa. This is intuitive as smaller ν results in a slower learning algorithm and therefore more iterations are required to fit the model. Accurately selecting the M parameter is generally considered to be of more importance, and therefore a value of ν is often chosen heuristically (e.g. the common value of 0.1) and then M is tuned by cross-validation and/or early-stopping.

Sampling Fraction, ϕ

Motivated by the success of bagging in random forests, stochastic gradient boosting (J. Friedman 1999) randomly samples the data in each iteration. It appears that subsampling performs best when also combined with shrinkage (Hastie, Tibshirani, and Friedman 2001) and as with the other hyper-parameters, selection of ϕ is usually performed by nested cross-validation.

9.1.2. Gradient Boosting Machines for Survival Analysis

In a componentwise GBM framework, adapting boosting to survival analysis requires only selecting a sensible choice of loss function L. Therefore fitting and predicting algorithms for componentwise survival GBMs are not discussed as these are fully described in algorithms (7) and ((**alg-surv-gbm-pred?**)) respectively. However, some GBMs in this section are not componentwise and therefore

require some more detailed consideration. Interestingly, unlike other machine learning algorithms that historically ignored survival analysis, early GBM papers considered boosting in a survival context (Ridgeway 1999); though there appears to be a decade gap before further considerations were made in the survival setting. After that period, several developments by Binder, Schmid, and Hothorn, adapted componentwise GBMs to a framework suitable for survival analysis. Their developments are covered exhaustively in the R packages **gbm** (Greenwell et al. 2019) and **mboost** (Hothorn et al. 2020). This survey continues with the predict type taxonomy.

9.1.2.1. Cox Survival Models

All survival GBMs make ranking predictions and none are able to directly predict survival distributions. However, the GBMs discussed in this section all have natural compositions to distributions as they are modelled in the semi-parametric proportional hazards framework (Chapter 12). The models discussed in the next section can also be composed to distributions though the choice of composition is less clear and therefore they are listed as pure 'ranking' models. \\ **GBM-COX** {#mod-gdcox} {#mod-gbmcox}\ The 'GBM-COX' aims to predict the distribution of data following the PH assumption by estimating the coefficients of a Cox model in a boosting framework (Ridgeway 1999). The model attempts to predict $\hat{g}(X^*) = \hat{\eta} := X^*\hat{\beta}$, by minimising a suitable loss function. As the model assumes a PH specification, the natural loss to optimise is the Cox partial likelihood (Cox 1972, 1975), more specifically to minimise the negative partial log-likelihood, $-l$, where

$$l(\beta) = \sum_{i=1}^{n} \Delta_i \Big[\eta_i \; - \; \log \Big(\sum_{j \in \mathcal{R}_{t_i}}^{n} \exp(\eta_i) \Big) \Big] \tag{9.1}$$

where $\mathcal{R}_{t_i}$ is the set of patients at risk at time t_i and $\eta_i = X_i\beta$. The gradient of $-l(\beta)$ at iteration m is

$$r_{im} := \Delta_i - \sum_{j=1}^{n} \Delta_j \frac{\mathbb{I}(T_i \geq T_j) \exp(g_{m-1}(X_i))}{\sum_{k \in \mathcal{R}_{t_j}} \exp(g_{m-1}(X_k))} \tag{9.2}$$

where $g_{m-1}(X_i) = X_i\beta_{m-1}$.

(5) now follows with the loss $L := -l(\beta)$.[3]

The GBM-COX is implemented in **mboost** (Hothorn et al. 2020) and has been demonstrated to perform well even when the data violates the PH assumption (B. A. Johnson and Long 2011). Despite being a black-box, GBMs are well-understood and individual weak learners are highly interpretable, thus making GBMs highly transparent. Several well-established software packages implement GBM-COX and those that do not tend to be very flexible with respect to custom implementations. GBM-COX is therefore considered an APT survival model. \\ **CoxBoost** {#mod-coxboost}\ The CoxBoost algorithm boosts the Cox PH by optimising the penalized partial-log likelihood; additionally the algorithm allows for mandatory (or 'forced') covariates (Harald Binder and Schumacher 2008). In medical domains the inclusion of mandatory covariates may be essential, either for model interpretability, or due to prior expert knowledge. This is not a feature usually

[3]Early implementations and publications of the GBM algorithm [@Friedman1999; @Friedman2001] included an additional step to the algorithm in which a step size is estimated by line search. More recent research has determined that this additional step is unneccesary [@Buhlmann2007] and the line search method does not appear to be used in practice.

supported by boosting. CoxBoost deviates from (7) by instead using an offset-based approach for generalized linear models (Tutz and Binder 2007). This model has a non-componentwise and componentwise framework but only the latter is implemented by the authors (Harold Binder 2013) and discussed here. Let $\mathcal{J}_{mand}$ be the indices of the mandatory covariates to be included in all iterations, $m = 1, ..., M$, then for an iteration m the indices to consider for fitting are the set

$$I_m = \{\{1\} \cup \mathcal{J}_{mand}, ..., \{p\} \cup \mathcal{J}_{mand}\}/\{\{j\} \cup \mathcal{J}_{mand} : j \in \mathcal{J}_{mand}\}$$

i.e. in each iteration the algorithm fits a weak learner on the mandatory covariates and one additional (non-mandatory) covariate (hence still being componentwise).

In addition, a penalty matrix $\mathbf{P} \in \mathbb{R}^{p \times p}$ is considered such that $P_{ii} > 0$ implies that covariate i is penalized and $P_{ii} = 0$ means no penalization. In practice this is usually a diagonal matrix (Harald Binder and Schumacher 2008) and by setting $P_{ii} = 0, i \in I_{mand}$ and $P_{ii} > 0, i \notin I_{mand}$, only optional (non-mandatory) covariates are penalized. The penalty matrix can be allowed to vary with each iteration, which allows for a highly flexible approach, however in implementation a simpler approach is to either select a single penalty to be applied in each iteration step or to have a single penalty matrix (Harold Binder 2013).

At the mth iteration and the kth set of indices to consider ($k = 1, ..., p$), the loss to optimize is the penalized partial-log likelihood given by

$$l_{pen}(\gamma_{mk}) = \sum_{i=1}^{n} \Delta_i \Big[\eta_{i,m-1} + X_{i,\mathcal{J}_{mk}} \gamma_{mk}^T\Big] -$$
$$\Delta_i \log\Big(\sum_{j=1}^{n} \mathbb{I}(T_j \leq T_i) \exp(\eta_{i,m-1} + X_{i,\mathcal{J}_{mk}} \gamma_{mk}^T\Big) - \lambda \gamma_{mk} \mathbf{P}_{mk} \gamma_{mk}^T$$

where $\eta_{i,m} = X_i \beta_m$, γ_{mk} are the coefficients corresponding to the covariates in $\mathcal{J}_{mk}$ which is the possible set of candidates for a subset of total candidates $k = 1, ..., p$, $\mathbf{P}_{mk}$ is the penalty matrix, and λ is a penalty hyper-parameter to be tuned or selected.[4] \\ In each iteration, all potential candidate sets (the union of mandatory covariates and one other covariate) are updated by

$$\hat{\gamma}_{mk} = \mathbf{I}_{pen}^{-1}(0) U(0)$$

where $U(\gamma) = \partial l / \partial \gamma(\gamma)$ and $\mathbf{I}_{pen}^{-1} = \partial^2 l / \partial \gamma \partial \gamma^T (\gamma + \lambda \mathbf{P}_{mk})$ are the first and second derivatives of the unpenalized partial-log-likelihood. The optimal set is then found as

$$k^* := \underset{k}{\operatorname{argmax}}\, l_{pen}(\gamma_{mk})$$

and the estimated coefficients are updated with

$$\hat{\beta}_m = \hat{\beta}_{m-1} + \gamma_{mk^*}, \quad k^* \in \mathcal{J}_{mk}$$

The step size, ν, is then one, but this could potentially be altered. \\ The algorithm deviates from (7) as l_{pen} is directly optimised and not its gradient, additionally model coefficients are iteratively updated instead of a more general model form. The algorithm is implemented in **CoxBoost** (Harold Binder 2013). Experiments suggest that including the 'correct' mandatory covariates may

[4] On notation, note that $\mathbf{P}_{ij}$ refers to the penalty matrix in the ith iteration for the jth set of indices, whereas P_{ij} is the (i, j)th element in the matrix $\mathbf{P}$.

increase predictive performance (Harald Binder and Schumacher 2008). CoxBoost is less accessible than other boosting methods as it requires a unique boosting algorithm, as such only one off-shelf implementation appears to exist and even this implementation has been removed from CRAN as of 2020-11-11. CoxBoost is also less transparent as the underlying algorithm is more complex, though is well-explained by the authors (Harald Binder and Schumacher 2008). There is good indication that CoxBoost is performant (R. E. B. Sonabend 2021). In a non-medical domain, where performance may be the most important metric, then perhaps CoxBoost can be recommended as a powerful model. However, when sensitive predictions are required, CoxBoost is currently not APT. Further papers studying the model and more off-shelf implementations could change this in the future.

9.1.2.2. Ranking Survival Models

The ranking survival models in this section are all unified as they make predictions of the linear predictor, $\hat{g}(X^*) = X^*\hat{\beta}$.[5] \\ **GBM-AFT** {#mod-gbmaft}\ Schmid and Hothorn (2008) (Schmid and Hothorn 2008b) published a GBM for accelerated failure time models in response to PH-boosted models that may not be suitable for non-PH data. Their model fits into the GBM framework by assuming a fully-parametric AFT and simultaneously estimating the linear predictor, $\hat{g}(X_i) = \hat{\eta}$, and the scale parameter, $\hat{\sigma}$, controlling the amount of noise in the distribution. The (fully-parametric) AFT is defined by

$$\log Y = \eta + \sigma W$$

where W is a random variable independent of the covariates that follows a given distribution and controls the noise in the model. By assuming a distribution on W, a distribution is assumed for the full parametric model. The full likelihood, $\mathcal{L}$, is given by

$$\mathcal{L}(\mathcal{D}_0|\mu,\sigma,W) = \prod_{i=1}^{n}\Big[\frac{1}{\sigma}f_W\Big(\frac{\log(T_i)-\mu}{\sigma}\Big)\Big]^{\Delta_i}\Big[S_W\Big(\frac{\log(T_i)-\mu}{\sigma}\Big)\Big]^{(1-\Delta_i)} \tag{9.3}$$

where f_W, S_W is the pdf and survival function of W for a given distribution. By setting $\mu := g(X_i)$, σ is then rescaled according to known results depending on the distribution (Klein and Moeschberger 2003). The gradient of the negative log-likelihood, $-l$, is minimised in the mth iteration where

$$l(\mathcal{D}_0|\hat{g},\hat{\sigma},W) = \sum_{i=1}^{n}\Delta_i\Big[-\log\sigma + \log f_W\Big(\frac{\log(T_i)-\hat{g}_{m-1}(X_i)}{\hat{\sigma}_{m-1}}\Big)\Big] + (1-\Delta_i)\Big[\log S_W\Big(\frac{\log(T_i)-\hat{g}_{m-1}(X_i)}{\hat{\sigma}_{m-1}}\Big)\Big]$$

where $\hat{g}_{m-1}, \hat{\sigma}_{m-1}$ are the location-scale parameters estimated in the previous iteration. Note this key difference to other GBM methods in which two estimates are made in each iteration step. In order to allow for this, (7) is run as normal but in addition, after updating $\hat{g}_m$, one then updates $\hat{\sigma}_m$ as

$$\hat{\sigma}_m := \underset{\sigma}{\operatorname{argmin}} -l(\mathcal{D}_0|g_m,\sigma,W)$$

σ_0 is initialized at the start of the algorithm with $\sigma_0 = 1$ suggested (Schmid and Hothorn 2008b). \\ This algorithm provides a ranking prediction without enforcing an often-unrealistic PH assumption

[5] This is commonly referred to as a 'linear predictor' as it directly relates to the boosted linear model (e.g. Cox PH), however it is more accurately a 'prognostic index' as the final prediction is not the true linear predictor.

on the data. This model is implemented in **mboost** and **xgboost**. Experiments indicate that this may outperform the Cox PH (Schmid and Hothorn 2008b). Moreover the model has the same transparency and accessibility as the GBM-COX and is therefore also considered APT. \\ **GBM-GEH** {#mod-gbmgeh}\ The concordance index is likely the most popular measure of discrimination, this in part due to the fact that it makes little-to-no assumptions about the data (Section 11.5). A less common measure is the Gehan loss, motivated by the semi-parametric AFT. Johnson and Long proposed the GBM with Gehan loss, here termed GBM-GEH, to optimise separation within an AFT framework (B. A. Johnson and Long 2011).

The semi-parametric AFT is defined by the linear model,

$$\log Y = \eta + \epsilon$$

for some error term, ϵ.

The D-dimensional Gehan loss to minimise is given by,

$$G_D(\mathcal{D}_0, \hat{g}) = -\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Delta_i(\hat{e}_i - \hat{e}_j)\mathbb{I}(\hat{e}_i \leq \hat{e}_j)$$

where $\hat{e}_i = \log T_i - \hat{g}(X_i)$. The negative gradient of the loss is,

$$r_{im} := \frac{\sum_{j=1}^{n}\Delta_j\mathbb{I}(\hat{e}_{m-1,i} \geq \hat{e}_{m-1,j}) - \Delta_i\mathbb{I}(\hat{e}_{m-1,i} \leq \hat{e}_{m-1,j})}{n}$$

where $\hat{e}_{m-1,i} = \log T_i - \hat{g}_{m-1}(X_i)$. \\ (7) then follows naturally substituting the loss and gradient above. The algorithm is implemented in **mboost**. Simulation studies on the performance of the model are inconclusive (B. A. Johnson and Long 2011) however the results in (R. E. B. Sonabend 2021) indicate strong predictive performance. Therefore this can tentatively be considered APT but further benchmark experiments would be preferred. \\ **GBM-BUJAR** {#mod-gbmbujar}\ GBM-BUJAR is another boosted semi-parametric AFT. However the algorithm introduced by Wang and Wang (2010) (Z. Wang and Wang 2010) uses Buckley-James imputation and minimisation. This algorithm is almost identical to a regression GBM (i.e. using squared loss or similar for L), except with one additional step to iteratively impute censored survival times. Assuming a semi-parametric AFT model, the GBM-BUJAR algorithm iteratively updates imputed outcomes with the Buckley-James estimator (Buckley and James 1979),

$$T^*_{m,i} := \hat{g}_{m-1}(X_i) + e_{m-1,i}\Delta_i + (1-\Delta_i)\Big[\hat{S}_{KM}(e_{m-1,i})^{-1}\sum_{e_{m-1,j}>e_{m-1,i}} e_{m-1,j}\Delta_j\hat{p}_{KM}(e_{m-1,j})\Big]$$

where $\hat{g}_{m-1}(X_i) = \hat{\eta}_{m-1}$, and $\hat{S}_{KM}, \hat{p}_{KM}$ are Kaplan-Meier estimates of the survival and probability mass functions respectively fit on some training data, and $e_{m-1,i} := \log(T_i) - g_{m-1}(X_i)$. Once $T^*_{m,i}$ has been updated, (7) continues from with least squares as with any regression model. \\ GBM-BUJAR is implemented in **bujar** (Z. Wang 2019) though without a separated fit/predict interface, its accessibility is therefore limited. There is no evidence of wide usage of this algorithm nor simulation studies demonstrating its predictive ability. Finally, there are many known problems with semi-parametric AFT models and the Buckey-James procedure (Wei 1992), hence GBM-BUJAR is also not transparent. \\ **GBM-UNO** {#mod-gbmuno}\ Instead of optimising models based on a given model form, Chen *et al.*[@Chen2013] studied direct optimisation of discrimination by Harrell's C whereas Mayr and Schmid (Mayr and Schmid 2014) focused instead on Uno's C. Only

an implementation of the Uno's C method could be found, this is therefore discussed here and termed 'GBM-UNO'. \\ The GBM-UNO attempts to predict $\hat{g}(X^*) := \hat{\eta}$ by optimising Uno's C (Section 11.5.1),

$$C_U(\hat{g}, \mathcal{D}_0) = \frac{\sum_{i \neq j} \Delta_i \{\hat{G}_{KM}(T_i)\}^{-2} \mathbb{I}(T_i < T_j) \mathbb{I}(\hat{g}(X_i) > \hat{g}(X_j))}{\sum_{i \neq j} \Delta_i \{\hat{G}_{KM}(T_i)\}^{-2} \mathbb{I}(T_i < T_j)}$$

The GBM algorithm requires that the chosen loss, here C_U, be differentiable w.r.t. $\hat{g}(X)$, which is not the case here due to the indicator term, $\mathbb{I}(\hat{g}(X_i) > \hat{g}(X_j))$. Therefore a smoothed version is instead considered where the indicator is approximated by the sigmoid function (Ma and Huang 2006),

$$K(u|\sigma) = (1 + \exp(-u/\sigma))^{-1}$$

where σ is a hyper-parameter controlling the smoothness of the approximation. The measure to optimise is then,

$$C_{USmooth}(\mathcal{D}_0|\sigma) = \sum_{i \neq j} \frac{k_{ij}}{1 + \exp\left[(\hat{g}(X_j) - \hat{g}(X_i))/\sigma)\right]} \tag{9.4}$$

with

$$k_{ij} = \frac{\Delta_i (\hat{G}_{KM}(T_i))^{-2} \mathbb{I}(T_i < T_j)}{\sum_{i \neq j}^{n} \Delta_i (\hat{G}_{KM}(T_i))^{-2} \mathbb{I}(T_i < T_j)}$$

The negative gradient at iteration m for observation i can then be found,

$$r_{im} := -\sum_{j=1}^{n} k_{ij} \frac{-\exp(\frac{\hat{g}_{m-1}(X_j) - \hat{g}_{m-1}(X_i)}{\sigma})}{\sigma(1 + \exp(\frac{\hat{g}_{m-1}(X_j) - \hat{g}_{m-1}(X_i)}{\sigma}))} \tag{9.5}$$

(6) can then be followed exactly by substituting this loss and gradient; this is implemented in **mboost**. One disadvantage of GBM-UNO is that C-index boosting is more insensitive to overfitting than other methods (Mayr, Hofner, and Schmid 2016), therefore stability selection (Meinshausen and Bühlmann 2010) can be considered for variable selection; this is possible with **mboost**. Despite directly optimising discrimination, simulation studies do not indicate that this model has better separation than other boosted or lasso models (Mayr and Schmid 2014). GBM-UNO has the same accessibility, transparency, and performance (R. E. B. Sonabend 2021) as previous APT boosting models and is therefore also considered APT.

9.1.3. Novel Adaptations

A clear theme emerging throughout this survey is a historical focus on predicting survival time or ranking, with less interest in direct optimisation of distributional predictions, which may be due to less off-shelf software for the task. Optimisation of a distribution is possible by considering a scoring rule (Section 11.7) as the GBM loss. The integrated Graf score (IGS) is discussed below but others are also possible.

The Integrated Graf Score (IGS) is given by,

$$L_{IGS}(t, \delta, \hat{S}|\tau^*) = \int_0^{\tau^*} \frac{\hat{S}(\tau)^2 \mathbb{I}(t \leq \tau, \delta = 1)}{\hat{G}_{KM}(t)} + \frac{\hat{F}(\tau)^2 \mathbb{I}(t > \tau)}{\hat{G}_{KM}(\tau)} \, d\tau$$

where τ^* is a threshold cut-off but in this case it is assumed $\tau^* = \max\{T_i : i = 1, .., n\}$. Differentiating with respect to $\hat{S}(\tau)$, the negative gradient in the mth iteration is given by

$$r_{im} := \int_0^{\tau^*} 2\hat{f}(\tau) \Big[\frac{\hat{F}(\tau)\mathbb{I}(t_i > \tau)}{\hat{G}_{KM}(\tau)} - \frac{\hat{S}(\tau)\mathbb{I}(t_i \leq \tau, \delta_i = 1)}{\hat{G}_{KM}(t_i)} \Big] \, d\tau \tag{9.6}$$

where $\hat{f}$ is the estimated probability density function.

(7) follows with these equations. The package **mboost** can be utilised to test these equations as a 'custom family'.

9.1.4. Conclusions

Componentwise gradient boosting machines are a highly flexible and powerful machine learning tool. They have proven particularly useful in survival analysis as minimal adjustments are required to make use of off-shelf software. The flexibility of the algorithm allows all the models above to be implemented in very few R (and other programming languages) packages.

Boosting is a method that often relies on intensive computing power and therefore dedicated packages, such as **xgboost** (T. Chen et al. 2020), exist to push CPU/GPUs to their limits in order to optimise predictive performance. This can be viewed as a strong advantage though one should be careful not to focus too much on predictive performance to the detriment of accessibility and transparency.

Boosting, especially with tree learners, is viewed as a black-box model that is increasingly difficult to interpret as the number of iterations increase. However, there are several methods for increasing interpretability, such as variable importance and SHAPs (Lundberg and Lee 2017). There is also evidence that boosting models can outperform the Cox PH (Schmid and Hothorn 2008b) (not something all ML models can claim) and in general survival GBMs are considered APT.

10. Neural Networks

TODO (150-200 WORDS)

10.1. Neural Networks

Before starting the survey on neural networks, first a comment about their transparency and accessibility. Neural networks are infamously difficult to interpret and train, with some calling building and training neural networks an 'art' (Hastie, Tibshirani, and Friedman 2001). As discussed in the introduction of this thesis, whilst neural networks are not transparent with respect to their predictions, they are transparent with respect to implementation. In fact the simplest form of neural network, as seen below, is no more complex than a simple linear model. With regard to accessibility, whilst it is true that defining a custom neural network architecture is complex and highly subjective, established models are implemented with a default architecture and are therefore accessible 'off-shelf'.

10.1.1. Neural Networks for Regression

(Artificial) Neural networks (ANNs) are a class of model that fall within the greater paradigm of *deep learning*. The simplest form of ANN, a feed-forward single-hidden-layer network, is a relatively simple algorithm that relies on linear models, basic activation functions, and simple derivatives. A short introduction to feed-forward regression ANNs is provided to motivate the survival models. This focuses on single-hidden-layer models and increasing this to multiple hidden layers follows relatively simply. \\ The single hidden-layer network is defined through three equations

$$Z_m = \sigma(\alpha_{0m} + \alpha_m^T X_i), \quad m = 1, ..., M \tag{10.1}$$
$$T = \beta_{0k} + \beta_k^T Z, \quad k = 1, .., K \tag{10.2}$$
$$g_k(X_i) = \phi_k(T) \tag{10.3}$$

where $(X_1, ..., X_n) \overset{i.i.d.}{\sim} X$ are the usual training data, α_{0m}, β_0 are bias parameters, and $\theta = \{\alpha_m, \beta\}$ $(m = 1, .., , M)$ are model weights where M is the number of hidden units. K is the number of classes in the output, which for regression is usually $K = 1$. The function ϕ is a 'link' or 'activation function', which transforms the predictions in order to provide an outcome of the correct return type; usually in regression, $\phi(x) = x$. σ is the 'activation function', which transforms outputs from each layer. The α_m parameters are often referred to as 'activations'. Different activation functions

may be used in each layer or the same used throughout, the choice is down to expert knowledge. Common activation functions seen in this section include the sigmoid function,

$$\sigma(v) = (1 + \exp(-v))^{-1}$$

tanh function,

$$\sigma(v) = \frac{\exp(v) - \exp(-v)}{\exp(v) + \exp(-v)} \tag{10.4}$$

and ReLU (Nair and Hinton 2010)

$$\sigma(v) = \max(0, v) \tag{10.5}$$

A single-hidden-layer model can also be expressed in a single equation, which highlights the relative simplicity of what may appear a complex algorithm.

$$g_k(X_i) = \sigma_0(\beta_{k0} + \sum_{h=1}^{H}(\beta_{kh}\sigma_h(\beta_{h0} + \sum_{m=1}^{M}\beta_{hm}X_{i;m})) \tag{10.6}$$

where H are the number of hidden units, β are the model weights, σ_h is the activation function in unit h, also σ_0 is the output unit activation, and $X_{i;m}$ is the ith observation features in the mth hidden unit.

An example feed-forward single-hidden-layer regression ANN is displayed in (Figure 10.1). This model has 10 input units, 13 hidden units, and one output unit; two bias parameters are fit. The model is described as 'feed-forward' as there are no cycles in the node and information is passed forward from the input nodes (left) to the output node (right).

Back-Propagation

The model weights, θ, in this section are commonly fit by 'back-propagation' although this method is often considered inefficient compared to more recent advances. A brief pseudo-algorithm for the process is provided below.

Let L be a chosen loss function for model fitting, let $\theta = (\alpha, \beta)$ be model weights, and let $J \in \mathbb{N}_{>0}$ be the number of iterations to train the model over. Then the back-propagation method is given by,

- **For** $j = 1, ..., J$: *[]* Forward Pass [i.] Fix weights $\theta^{(j-1)}$. *[ii.] Compute predictions* $\hat{Y} := \hat{g}_k^{(j)}(X_i|\theta^{(j-1)})$ *with (Equation 10.6).* [] *Backward Pass [iii.] Calculate the gradients of the loss* $L(\hat{Y}|\mathcal{D}_0)$. [] *Update* *[iv.] Update $\alpha^{(r)}, \beta^{(r)}$ with gradient descent.
- **End For**

In regression, a common choice for L is the squared loss,

$$L(\hat{g}, \theta|\mathcal{D}_0) = \sum_{i=1}^{n}(Y_i - \hat{g}(X_i|\theta))^2$$

which may help illustrate how the training outcome, $(Y_1, ..., Y_n) \overset{i.i.d.}{\sim} Y$, is utilised for model fitting.

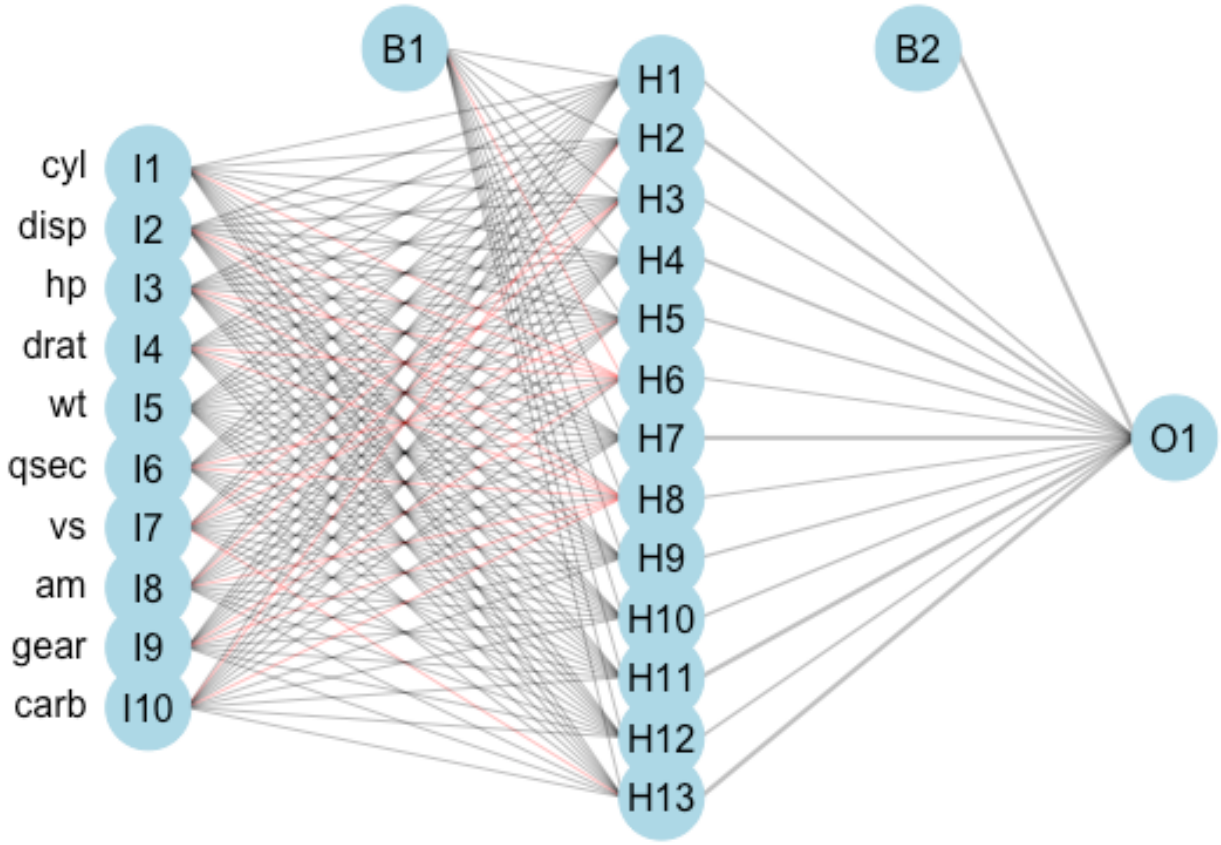


Figure 10.1.: Single-hidden-layer artificial neural network with 13 hidden units fit on the `mtcars` (Henderson and Velleman 1981) dataset using the **nnet** (N. Venables and D. Ripley 2002) package, and **gamlss.add** (Stasinopoulos et al. 2020) for plotting. Left column are input variables, I1-I10, second column are 13 hidden units, H1-H13, right column is single output variable, O1. B1 and B2 are bias parameters.

Making Predictions

Once the model is fitted, predictions for new data follow by passing the testing data as inputs to the model with fitted weights,

$$g_k(X^*) = \sigma_0(\hat{\beta}_{k0} + \sum_{h=1}^{H}(\hat{\beta}_{kh}\sigma_h(\hat{\beta}_{h0} + \sum_{m=1}^{M}\hat{\beta}_{hm}X_m^*))$$

Hyper-Parameters

In practice, a regularization parameter, λ, is usually added to the loss function in order to help avoid overfitting. This parameter has the effect of shrinking model weights towards zero and hence in the context of ANNs regularization is usually referred to as 'weight decay'. The value of λ is one of three important hyper-parameters in all ANNs, the other two are: the range of values to simulate initial weights from, and the number of hidden units, M.

The range of values for initial weights is usually not tuned but instead a consistent range is specified and the neural network is trained multiple times to account for randomness in initialization.

The regularization parameter and number of hidden units, M, depend on each other and have a similar relationship to the learning rate and number of iterations in the GBMs (Section 9.1). Like the GBMs, it is simplest to set a high number of hidden units and then tune the regularization parameter (Bishop 2006; Hastie, Tibshirani, and Friedman 2001). Determining how many hidden layers to include, and how to connect them, is informed by expert knowledge and well beyond the scope of this thesis; decades of research has been required to derive sensible new configurations.

Training Batches

ANNs can either be trained using complete data, in batches, or online. This decision is usually data-driven and will affect the maximum number of iterations used to train the algorithm; as such this will also often be chosen by expert-knowledge and not empirical methods such as cross-validation.

Neural Terminology

Neural network terminology often reflects the structures of the brain. Therefore ANN units are referred to as nodes or neurons and sometimes the connections between neurons are referred to as synapses. Neurons are said to be 'fired' if they are 'activated'. The simplest example of activating a neuron is with the Heaviside activation function with a threshold of 0: $\sigma(v) = \mathbb{I}(v \geq 0)$. Then a node is activated and passes its output to the next layer if its value is positive, otherwise it contributes no value to the next layer.

10.1.2. Neural Networks for Survival Analysis

Surveying neural networks is a non-trivial task as there has been a long history in machine learning of publishing very specific data-driven neural networks with limited applications; this is also true in survival analysis. This does mean however that where limited developments for survival were made in other machine learning classes, ANN survival adaptations have been around for several decades. A review in 2000 by Schwarzer *et al.* surveyed 43 ANNs for diagnosis and prognosis published in the first half of the 90s, however only up to ten of these are specifically for survival data.[1] Of those, Schwarzer *et al.* deemed three to be 'na"ive applications to survival data', and recommended for future research models developed by Liestøl *et al.* (1994) (Liestol, Andersen, and Andersen 1994), Faraggi and Simon (1995) (Faraggi and Simon 1995), and Biganzoli *et al.* (1998) (E. Biganzoli et al. 1998).

This survey will not be as comprehensive as the 2000 survey, and nor has any survey since, although there have been several ANN reviews (B. D. Ripley and Ripley 2001; Huang et al. 2020b; Ohno-Machado 1996; Yang 2010; W. Zhu et al. 2020). ANNs are considered to be a black-box model, with interpretability decreasing steeply as the number of hidden layers and nodes increases. In terms of accessibility there have been relatively few open-source packages developed for survival ANNs; where these are available the focus has historically been in Python, with no R implementations. The new **survivalmodels** (R. Sonabend 2020) package,[2] implements these Python models via **reticulate** (Ushey, Allaire, and Tang 2020). No recurrent neural netwoks are included in this survey though the survival models SRN (Oh et al. 2018) and RNN-Surv (Giunchiglia, Nemchenko, and Schaar 2018) are acknowledged. \\ This survey is made slightly more difficult as neural networks are often proposed for many different tasks, which are not necessarily clearly advertised in a paper's title or abstract. For example, many papers claim to use neural networks for survival analysis and make comparisons to Cox models, whereas the task tends to be death at a particular (usually 5-year) time-point (classification) (Han et al. 2018; Lundin et al. 1999; B. D. Ripley and Ripley 2001; R. M. Ripley, Harris, and Tarassenko 1998; Huseyin Seker et al. 2002), which is often not made clear until mid-way through the paper. Reviews and surveys have also conflated these different tasks, for example a very recent review concluded superior performance of ANNs over Cox models, when in fact this is only in classification (Huang et al. 2020a) (RM2) {sec:car_reduxstrats_mistakes}. To clarify, this form of classification task does fall into the general *field* of survival analysis, but not the survival *task* ((**box-task-surv?**)). Therefore this is not a comment on the classification task but a reason for omitting these models from this survey.

Using ANNs for feature selection (often in gene expression data) and computer vision is also very common in survival analysis, and indeed it is in this area that most success has been seen (Bello et al. 2019; Y.-C. Chen, Ke, and Chiu 2014; Cui et al. 2020; Lao et al. 2017; McKinney et al. 2020; Rietschel, Yoon, and Schaar 2018; H. Seker et al. 2002; Zhang et al. 2020; X. Zhu, Yao, and Huang 2016), but these are again beyond the scope of this survey. \\ The key difference between neural networks is in their output layer, required data transformations, the model prediction, and the loss function used to fit the model. Therefore the following are discussed for each of the surveyed models: the loss function for training, L, the model prediction type, $\hat{g}$, and any required data transformation. Notation is continued from the previous surveys with the addition of θ denoting model weights (which will be different for each model).

[1] Schwarzer conflates the prognosis and survival task, therefore it is not clear if all 10 of these are for time-to-event data (at least five definitely are).

[2] Created in order to run the experiments in [@Sonabend2021b].

10.1.2.1. Probabilistic Survival Models

Unlike other classes of machine learning models, the focus in ANNs has been on probabilistic models. The vast majority make these predictions via reduction to binary classification **??**. Whilst almost all of these networks implicitly reduce the problem to classification, most are not transparent in exactly how they do so and none provide clear or detailed interface points in implementation allowing for control over this reduction. Most importantly, the majority of these models do not detail how valid survival predictions are derived from the binary setting,[3] which is not just a theoretical problem as some implementations, such as the Logistic-Hazard model in **pycox** (Kvamme 2018), have been observed to make survival predictions outside the range $[0, 1]$. This is not a statement about the performance of models in this section but a remark about the lack of transparency across all probabilistic ANNs. \\ Many of these algorithms use an approach that formulate the Cox PH as a non-linear model and minimise the partial likelihood. These are referred to as 'neural-Cox' models and the earliest appears to have been developed by Faraggi and Simon (Faraggi and Simon 1995). All these models are technically composites that first predict a ranking, however they assume a PH form and in implementation they all appear to return a probabilistic prediction. \\ **ANN-COX** {#mod-anncox}\ Faraggi and Simon (Faraggi and Simon 1995) proposed a non-linear PH model

$$h(\tau|X_i, \theta) = h_0(\tau) \exp(\phi(X_i\beta)) \tag{10.7}$$

where ϕ is the sigmoid function and $\theta = \{\beta\}$ are model weights. This model, 'ANN-COX', estimates the prediction functional, $\hat{g}(X^*) = \phi(X^*\hat{\beta})$. The model is trained with the partial-likelihood function

$$L(\hat{g}, \theta|\mathcal{D}_0) = \prod_{i=1}^{n} \frac{\exp(\sum_{m=1}^{M} \alpha_m \hat{g}_m(X^*))}{\sum_{j \in \mathcal{R}_{t_i}} \exp(\sum_{m=1}^{M} \alpha_m \hat{g}_m(X^*))}$$

where $\mathcal{R}_{t_i}$ is the risk group alive at t_i; M is the number of hidden units; $\hat{g}_m(X^*) = (1 + \exp(-X^*\hat{\beta}_m))^{-1}$; and $\theta = \{\beta, \alpha\}$ are model weights.

The authors proposed a single hidden layer network, trained using back-propagation and weight optimisation with Newton-Raphson. This architecture did not outerperform a Cox PH (Faraggi and Simon 1995). Further adjustments including (now standard) pre-processing and hyper-parameter tuning did not improve the model performance (Mariani et al. 1997). Further independent studies demonstrated worse performance than the Cox model (Faraggi and Simon 1995; Xiang et al. 2000). \\ **COX-NNET** {#mod-coxnnet}\ COX-NNET (Ching, Zhu, and Garmire 2018) updates the ANN-COX by instead maximising the regularized partial log-likelihood

$$L(\hat{g}, \theta|\mathcal{D}_0, \lambda) = \sum_{i=1}^{n} \Delta_i \Big[\hat{g}(X_i) \; - \; \log\Big(\sum_{j \in \mathcal{R}_{t_i}} \exp(\hat{g}(X_j))\Big)\Big] + \lambda(\|\beta\|_2 + \|w\|_2)$$

with weights $\theta = (\beta, w)$ and where $\hat{g}(X_i) = \sigma(wX_i + b)^T\beta$ for bias term b, and activation function σ; σ is chosen to be the tanh function ((Equation 10.4)). In addition to weight decay, dropout (Srivastava et al. 2014) is employed to prevent overfitting. Dropout can be thought of as a similar concept to the variable selection in random forests, as each node is randomly deactivated with probability p, where p is a hyper-parameter to be tuned.

[3] One could assume they use procedures such as those described in Tutz and Schmid (2016) [@Tutz2016] but there is rarely transparent writing to confirm this.

Independent simulation studies suggest that COX-NNET does not outperform the Cox PH (Michael F. Gensheimer and Narasimhan 2019). \\ **DeepSurv** {#mod-deepsurv}\ DeepSurv (J. L. Katzman et al. 2018) extends these models to deep learning with multiple hidden layers. The chosen error function is the average negative log-partial-likelihood with weight decay

$$L(\hat{g}, \theta | \mathcal{D}_0, \lambda) = -\frac{1}{n^*} \sum_{i=1}^{n} \Delta_i \Big[\Big(\hat{g}(X_i) - \log \sum_{j \in \mathcal{R}_{t_i})} \exp(\hat{g}(X_j) \Big) \Big] + \lambda \|\theta\|_2^2$$

where $n^* := \sum_{i=1}^{n} \mathbb{I}(\Delta_i = 1)$ is the number of uncensored observations and $\hat{g}(X_i) = \phi(X_i|\theta)$ is the same prediction object as the ANN-COX. State-of-the-art methods are used for data pre-processing and model training. The model architecture uses a combination of fully-connected and dropout layers. Benchmark experiments by the authors indicate that DeepSurv can outperform the Cox PH in ranking tasks (J. Katzman et al. 2016; J. L. Katzman et al. 2018) although independent experiments do not confirm this (Zhao and Feng 2020).

Cox-Time {#mod-coxtime}\ Kvamme *et al.* (Kvamme, Borgan, and Scheel 2019) build on these models by allowing time-varying effects. The loss function to minimise, with regularization, is given by

$$L(\hat{g}, \theta | \mathcal{D}_0, \lambda) = \frac{1}{n} \sum_{i:\Delta_i=1} \log \Big(\sum_{j \in \mathcal{R}_{t_i}} \exp[\hat{g}(X_j, T_i) - \hat{g}(X_i, T_i)] \Big) + \lambda \sum_{i:\Delta_i=1} \sum_{j \in \mathcal{R}_{t_i}} |\hat{g}(X_j, T_i)|$$

where $\hat{g} = \hat{g}_1, ..., \hat{g}_n$ is the same non-linear predictor but with a time interaction and λ is the regularization parameter. The model is trained with stochastic gradient descent and the risk set, $\mathcal{R}_{t_i}$, in the equation above is instead reduced to batches, as opposed to the complete dataset. ReLU activations (Nair and Hinton 2010) and dropout are employed in training. Benchmark experiments indicate good performance of Cox-Time, though no formal statistical comparisons are provided and hence no comment about general performance can be made. \\ **ANN-CDP** {#mod-anncdp}\ One of the earliest ANNs that was noted by Schwarzer *et al.*[@Schwarzer2000] was developed by Liestøl *et al.*[@Liestol1994] and predicts conditional death probabilities (hence 'ANN-CDP'). The model first partitions the continuous survival times into disjoint intervals $\mathcal{I}_k, k = 1, ..., m$ such that $\mathcal{I}_k$ is the interval $(t_{k-1}, t_k]$. The model then studies the logistic Cox model (proportional odds) (Cox 1972) given by

$$\frac{p_k(\mathbf{x})}{q_k(\mathbf{x})} = \exp(\eta + \theta_k)$$

where $p_k = 1 - q_k$, $\theta_k = \log(p_k(0)/q_k(0))$ for some baseline probability of survival, $q_k(0)$, to be estimated; η is the usual linear predictor, and $q_k = P(T \geq T_k | T \geq T_{k-1})$ is the conditional survival probability at time T_k given survival at time T_{k-1} for $k = 1, ..., K$ total time intervals. A logistic activation function is used to predict $\hat{g}(X^*) = \phi(\eta + \theta_k)$, which provides an estimate for $\hat{p}_k$.

The model is trained on discrete censoring indicators D_{ki} such that $D_{ki} = 1$ if individual i dies in interval $\mathcal{I}_k$ and 0 otherwise. Then with K output nodes and maximum likelihood estimation to find the model parameters, $\hat{\eta}$, the final prediction provides an estimate for the conditional death probabilities $\hat{p}_k$. The negative log-likelihood to optimise is given by

$$L(\hat{g}, \theta | \mathcal{D}_0) = \sum_{i=1}^{n} \sum_{k=1}^{m_i} [D_{ki} \log(\hat{p}_k(X_i)) + (1 - D_{ki}) \log(\hat{q}_k(X_i))]$$

where m_i is the number of intervals in which observation i is not censored.

Liestøl *et al.* discuss different weighting options and how they correspond to the PH assumption. In the most generalised case, a weight-decay type regularization is applied to the model weights given by

$$\alpha \sum_l \sum_k (w_{kl} - w_{k-1,l})^2$$

where w are weights, and α is a hyper-parameter to be tuned, which can be used alongside standard weight decay. This corresponds to penalizing deviations from proportionality thus creating a model with approximate proportionality. The authors also suggest the possibility of fixing the weights to be equal in some nodes and different in others; equal weights strictly enforces the proportionality assumption. Their simulations found that removing the proportionality assumption completely, or strictly enforcing it, gave inferior results. Comparing their model to a standard Cox PH resulted in a 'better' negative log-likelihood, however this is not a precise evaluation metric and an independent simulation would be preferred. Finally Listøl *et al.* included a warning "The flexibility is, however, obtained at unquestionable costs: many parameters, difficult interpretation of the parameters and

a slow numerical procedure" (Liestol, Andersen, and Andersen 1994). \\ **PLANN** {#mod-plann}\ Biganzoli *et al.* (1998) (E. Biganzoli et al. 1998) studied the same proportional-odds model as the ANN-CDP (Liestol, Andersen, and Andersen 1994). Their model utilises partial logistic regression (Efron 1988) with added hidden nodes, hence 'PLANN'. Unlike ANN-CDP, PLANN predicts a smoothed hazard function by using smoothing splines. The continuous time outcome is again discretised into disjoint intervals $t_m, m = 1, ..., M$. At each time-interval, t_m, the number of events, d_m, and number of subjects at risk, n_m, can be used to calculate the discrete hazard function,[4]

$$\hat{h}_m = \frac{d_m}{n_m}, m = 1, ..., M \tag{10.8}$$

This quantity is used as the target to train the neural network. The survival function is then estimated by the Kaplan-Meier type estimator,

$$\hat{S}(\tau) = \prod_{m:t_m \leq \tau} (1 - \hat{h}_m) \tag{10.9}$$

The model is fit by employing one of the more 'usual' survival reduction strategies in which an observation's survival time is treated as a covariate in the model (Tutz and Schmid 2016). As this model uses discrete time, the survival time is discretised into one of the M intervals. This approach removes the proportional odds constraint as interaction effects between time and covariates can be modelled (as time-updated covariates). Again the model makes predictions at a given time m, $\phi(\theta_m + \eta)$, where η is the usual linear predictor, θ is the baseline proportional odds hazard $\theta_m = \log(h_m(0)/(1 - h_m(0))$. The logistic activation provides estimates for the discrete hazard,

$$h_m(X_i) = \frac{\exp(\theta_m + \hat{\eta})}{1 + \exp(\theta_m + \hat{\eta})}$$

which is smoothed with cubic splines (Efron 1988) that require tuning.

A cross-entropy error function is used for training

$$L(\hat{h}, \theta | \mathcal{D}_0, a) = -\sum_{m=1}^{M} \Big[\hat{h}_m \log\Big(\frac{h_l(X_i, a_l)}{\hat{h}_m}\Big) + (1 - \hat{h}_m) \log\Big(\frac{1 - h_l(X_i, a_l)}{1 - \hat{h}_m}\Big)\Big] n_m$$

where $h_l(X_i, a_l)$ is the discrete hazard h_l with smoothing at mid-points a_l. Weight decay can be applied and the authors suggest $\lambda \approx 0.01 - 0.1$ (E. Biganzoli et al. 1998), though they make use of an AIC type criterion instead of cross-validation.

This model makes smoothed hazard predictions at a given time-point, τ, by including τ in the input covariates X_i. Therefore the model first requires transformation of the input data by replicating all observations and replacing the single survival indicator Δ_i, with a time-dependent indicator D_{ik}, the same approach as in ANN-CDP. Further developments have extended the PLANN to Bayesian modelling, and for competing risks (E. M. Biganzoli, Ambrogi, and Boracchi 2009).

No formal comparison is made to simpler model classes. The authors recommend ANNs primarily for exploration, feature selection, and understanding underlying patterns in the data (E. M. Biganzoli, Ambrogi, and Boracchi 2009). \\ **Nnet-survival** {#mod-nnetsurvival}\ Aspects of the PLANN algorithm have been generalised into discrete-time survival algorithms in several papers

[4]Derivation of this as a 'hazard' estimator follows trivially by comparison to the Nelson-Aalen estimator.

(Michael F. Gensheimer and Narasimhan 2019; **Kvamme2019?**; Mani et al. 1999; Street 1998). Various estimates have been derived for transforming the input data to a discrete hazard or survival function. Though only one is considered here as it is the most modern and has a natural interpretation as the 'usual' Kaplan-Meier estimator for the survival function. Others by Street (1998) (Street 1998) and Mani (1999) (Mani et al. 1999) are acknowledged. The discrete hazard estimator (Equation 10.8), $\widehat{h}$, is estimated and these values are used as the targets for the ANN. For the error function, the mean negative log-likelihood for discrete time (**Kvamme2019?**) is minimised to estimate $\widehat{h}$,

$$L(\widehat{h}, \theta|\mathcal{D}_0) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k(T_i)}(\mathbb{I}(T_i = \tau_j, \Delta_i = 1)\log[\widehat{h}_i(\tau_j)] + \\ (1 - \mathbb{I}(T_i = \tau_j, \Delta_i = 1))\log(1 - \widehat{h}_i(\tau_j)))$$

where $k(T_i)$ is the time-interval index in which observation i dies/is censored, τ_j is the jth discrete time-interval, and the prediction of $\widehat{h}$ is obtained via

$$\widehat{h}(\tau_j|\mathcal{D}_0) = [1 + \exp(-\widehat{g}_j(\mathcal{D}_0))]^{-1}$$

where $\widehat{g}_j$ is the jth output for $j = 1, ..., m$ discrete time intervals. The number of units in the output layer for these models corresponds to the number of discrete-time intervals. Deciding the width of the time-intervals is an additional hyper-parameter to consider.

Gensheimer and Narasimhan's 'Nnet-survival' (Michael F. Gensheimer and Narasimhan 2019) has two different implementations. The first assumes a PH form and predicts the linear predictor in the final layer, which can then be composed to a distribution. Their second 'flexible' approach instead predicts the log-odds of survival in each node, which are then converted to a conditional probability of survival, $1 - h_j$, in a given interval using the sigmoid activation function. The full survival function can be derived with (Equation 10.9). The model has been demonstrated not to outperform the Cox PH w.r.t. Harrell's C or the Graf (Brier) score (Michael F. Gensheimer and Narasimhan 2019). \\ **PC-Hazard** {#mod-pchazard}\ Kvamme and Borgan deviate from nnet-survival in their 'PC-Hazard' (**Kvamme2019?**) by first considering a discrete-time approach with a softmax activation function influenced by multi-class classification. They expand upon this by studying a piecewise constant hazard function in continuous time and defining the mean negative log-likelihood as

$$L(\widehat{g}, \theta|\mathcal{D}_0) = -\frac{1}{n}\sum_{i=1}^{n}\left(\Delta_i X_i \log\tilde{\eta}_{k(T_i)} - X_i\tilde{\eta}_{k(T_i)}\rho(T_i) - \sum_{j=1}^{k(T_i)-1}\tilde{\eta}_j X_i\right)$$

where $k(T_i)$ and τ_i is the same as defined above, $\rho(t) = \frac{t-\tau_{k(t)-1}}{\Delta\tau_{k(t)}}$, $\Delta\tau_j = \tau_j - \tau_{j-1}$, and $\tilde{\eta}_j := \log(1 + \exp(\widehat{g}_j(X_i))$ where again $\widehat{g}_j$ is the jth output for $j = 1, ..., m$ discrete time intervals. Once the weights have been estimated, the predicted survival function is given by

$$\widehat{S}(\tau, X^*|\mathcal{D}_0) = \exp(-X^*\tilde{\eta}_{k(\tau)}\rho(\tau))\prod_{j=1}^{k(\tau)-1}\exp(-\tilde{\eta}_j(X^*))$$

Benchmark experiments indicate similar performance to nnet-survival (**Kvamme2019?**), an unsurprising result given their implementations are identical with the exception of the loss function

(**Kvamme2019?**), which is also similar for both models. A key result found that varying values for interval width lead to significant differences and therefore should be carefully tuned. \\ **DNNSurv** {#mod-dnnsurv}\ A very recent (pre-print) approach (Zhao and Feng 2020) instead first computes 'pseudo-survival probabilities' and uses these to train a regression ANN with sigmoid activation and squared error loss. These pseudo-probabilities are computed using a jackknife-style estimator given by

$$\tilde{S}_{ij}(T_{j+1}, \mathcal{R}_{t_j}) = n_j \hat{S}(T_{j+1}|\mathcal{R}_{t_j}) - (n_j - 1)\hat{S}^{-i}(T_{j+1}|\mathcal{R}_{t_j})$$

where $\hat{S}$ is the IPCW weighted Kaplan-Meier estimator (defined below) for risk set $\mathcal{R}_{t_j}$, $\hat{S}^{-i}$ is the Kaplan-Meier estimator for all observations in $\mathcal{R}_{t_j}$ excluding observation i, and $n_j := |\mathcal{R}_{t_j}|$. The IPCW weighted KM estimate is found via the IPCW Nelson-Aalen estimator,

$$\hat{H}(\tau|\mathcal{D}_0) = \sum_{i=1}^{n} \int_0^{\tau} \frac{\mathbb{I}(T_i \leq u, \Delta_i = 1)\hat{W}_i(u)}{\sum_{j=1}^{n} \mathbb{I}(T_j \geq u)\hat{W}_j(u)} \, du$$

where $\hat{W}_i, \hat{W}_j$ are subject specific IPC weights.

In their simulation studies, they found no improvement over other proposed neural networks. Arguably the most interesting outcome of their paper are comparisons of multiple survival ANNs at specific time-points, evaluated with C-index and Brier score. Their results indicate identical performance from all models. They also provide further evidence of neural networks not outperforming a Cox PH when the PH assumption is valid. However, in their non-PH dataset, DNNSurv appears to outperform the Cox model (no formal tests are provided). Data is replicated similarly to previous models except that no special indicator separates censoring and death, this is assumed to be handled by the IPCW pseudo probabilities. \\ **DeepHit** {#mod-deephit}\ DeepHit (C. Lee et al. 2018) was originally built to accommodate competing risks, but only the non-competing case is discussed here (Kvamme, Borgan, and Scheel 2019). The model builds on previous approaches by discretising the continuous time outcome, and makes use of a composite loss. It has the advantage of making no parametric assumptions and directly predicts the probability of failure in each time-interval (which again correspond to different terminal nodes), i.e. $\hat{g}(\tau_k|\mathcal{D}_1) = \hat{P}(T^* = \tau_k|X^*)$ where again $\tau_k, k = 1, ..., K$ are the distinct time intervals. The estimated survival function is found with $\hat{S}(\tau_K|X^*) = 1 - \sum_{k=1}^{K} \hat{g}_i(\tau_k|X^*)$. ReLU activations were used in all fully connected layers and a softmax activation in the final layer. The losses in the composite error function are given by

$$L_1(\hat{g}, \theta|\mathcal{D}_0) = -\sum_{i=1}^{N} [\Delta_i \log(\hat{g}_i(T_i)) + (1 - \Delta_i) \log(\hat{S}_i(T_i))]$$

and

$$L_2(\hat{g}, \theta|\mathcal{D}_0, \sigma) = \sum_{i \neq j} \Delta_i \mathbb{I}(T_i < T_j)\sigma(\hat{S}_i(T_i), \hat{S}_j(T_i))$$

for some convex loss function σ and where $\hat{g}_i(t) = \hat{g}(t|X_i)$. Again these can be seen to be a cross-entropy loss and a ranking loss. Benchmark experiments demonstrate the model outperforming the Cox PH and RSFs (C. Lee et al. 2018) with respect to separation, and an independent experiment supports these findings (Kvamme, Borgan, and Scheel 2019). However, the same independent study demonstrated worse performance than a Cox PH w.r.t. the integrated Brier score (Graf et al. 1999).

10.1.2.2. Deterministic Survival Models

Whilst the vast majority of survival ANNs have focused on probabilistic predictions (often via ranking), a few have also tackled the deterministic or 'hybrid' problem. \\ **RankDeepSurv** {#mod-rankdeepsurv}\ Jing *et al.*[@Jing2019] observed the past two decades of research in survival ANNs and then published a completely novel solution, RankDeepSurv, which makes predictions for the survival time $\hat{T} = (\hat{T}_1, ..., \hat{T}_n)$. They proposed a composite loss function

$$L(\hat{T}, \theta | \mathcal{D}_0, \alpha, \gamma, \lambda) = \alpha L_1(\hat{T}, T, \Delta) + \gamma L_2(\hat{T}, T, \Delta) + \lambda \|\theta\|_2^2$$

where θ are the model weights, $\alpha, \gamma \in \mathbb{R}_{>0}$, λ is the shrinkage parameter, by a slight abuse of notation $T = (T_1, ..., T_n)$ and $\Delta = (\Delta_1, ..., \Delta_n)$, and

$$L_1(\hat{T}, \theta | \mathcal{D}_0) = \frac{1}{n} \sum_{\{i: I(i)=1\}} (\hat{T}_i - T_i)^2; \quad I(i) = \begin{cases} 1, & \Delta_i = 1 \cup (\Delta_i = 0 \cap \hat{T}_i \leq T_i) \\ 0, & \text{otherwise} \end{cases}$$

$$L_2(\hat{T}, \theta | \mathcal{D}_0) = \frac{1}{n} \sum_{\{i,j: I(i,j)=1\}}^{n} [(T_j - T_i) - (\hat{T}_j - \hat{T}_i)]^2; \quad I(i,j) = \begin{cases} 1, & T_j - T_i > \hat{T}_j - \hat{T}_i \\ 0, & \text{otherwise} \end{cases}$$

where $\hat{T}_i$ is the predicted survival time for observation i. A clear contrast can be made between these loss functions and the constraints used in SSVM-Hybrid (Vanya Van Belle, Pelckmans, Van Huffel, et al. 2011) (Section 8.1.2). L_1 is the squared second constraint in 8.1.2.1 and L_2 is the squared first constraint in 8.1.2.1. However L_1 in RankDeepSurv discards the squared error difference for all censored observations when the prediction is lower than the observed survival time; which is problematic as if someone is censored at time T_i then it is guaranteed that their true survival time is greater than T_i (this constraint may be more sensible if the inequality were reversed). An advantage to this loss is, like the SSVM-Hybrid, it enables a survival time interpretation for a ranking optimised model; however these 'survival times' should be interpreted with care.

The authors propose a model architecture with several fully connected layers with the ELU (Clevert, Unterthiner, and Hochreiter 2015) activation function and a single dropout layer. Determining the success of this model is not straightforward. The authors claim superiority of RankDeepSurv over Cox PH, DeepSurv, and RSFs however this is an unclear comparison (RM2) {sec:car_reduxstrats_mistakes} that requires independent study.

10.1.3. Novel Adaptations

In stark contrast to other model classes, the vast majority of survival ANNs have focused on optimising probabilistic predictions and not relative risks. There does not appear to a model that directly optimises separation via some concordance measure. One simple method could consider RankDeepSurv (Jing et al. 2019) without L_1. Interestingly, Jing *et al.*[@Jing2019] compare RankDeepSurv to a model using L_1 only but not to a model using L_2 only, which would be similar to SSVM-Rank (Vanya Van Belle, Pelckmans, Van Huffel, et al. 2011) (Section 8.1.2). RankDeepSurv also likely suffers from the same computational problems as RANKSVMC (Vanya Van Belle et al. 2007). However this could be resolved by employing the same nearest-neighbours methodology (Vanya Van Belle et al. 2008) as in SSVM-Rank. A disadvantage of RankDeepSurv is that the loss does not compare right-censored observations to possibly-correct predictions. Two possible methods to resolve this are either to include a hyper-parameter for 'mean time alive from censoring', ϵ, or an

MRL imputation method, similarly to SSVR-MRL (Goli, Mahjub, Faradmal, and Soltanian 2016) (Section 8.1.2). With these adaptations, L_1 is given by

$$L_1(\hat{T}, \theta|\mathcal{D}_0, \epsilon) = \frac{1}{n}\sum_{i=1}^{n} \Delta_i(\hat{T}_i - T_i)^2 + (1-\Delta_i)(\hat{T}_i - T_i - \epsilon_i)^2$$

where either ϵ_i are hyper-parameters for tuning (all could be set equal to prevent overfitting) or $\epsilon_i = MRL(T_i)$. L_2 is given by,

$$L_2(\hat{T}, \theta|\mathcal{D}_0) = \frac{1}{n}\sum_{\{i:I(i)=1\}}^{n} [(T_i - T_{j(i)}) - (\hat{T}_i - \hat{T}_{j(i)})]^2;$$

$$I(i) = \begin{cases} 1, & T_i - T_{j(i)} > \hat{T}_i - \hat{T}_{j(i)} \\ 0, & \text{otherwise} \end{cases}$$

where $T_{j(i)}$ is the survival time of the nearest non-censored neighbour with the largest survival time smaller than T_i, the same definition as given by Van Belle *et al.*[@VanBelle2011b]. The adaptation to L_1 prevents predictions from being discarded in the loss and simultaneously increases penalization applied to under-predictions. The adapted L_2 should also have a faster run-time.

10.1.4. Conclusions

There have been many advances in neural networks for survival analysis. It is not possible to review all proposed survival neural networks without diverting too far from the thesis scope. This survey of ANNs should demonstrate two points: firstly that the vast majority (if not all) of survival ANNs are reduction models that either find a way around censoring via imputation or discretisation of time-intervals, or by focusing on partial likelihoods only; secondly that no survival ANN is APT. \\ Despite ANNs being highly performant in other areas of supervised learning, there is strong evidence that the survival ANNs above are inferior to a Cox PH when the data follows the PH assumption or when variables are linearly related (Michael F. Gensheimer and Narasimhan 2018; Luxhoj and Shyur 1997; Ohno-Machado 1997; Puddu and Menotti 2012; Xiang et al. 2000; Yang 2010; Yasodhara, Bhat, and Goldenberg 2018; Zhao and Feng 2020). There are not enough experiments to make conclusions in the case when the data is non-PH. Experiments in (R. E. B. Sonabend 2021) support the finding that survival ANNs are not performant.

There is evidence that many papers introducing neural networks do not utilise proper methods of comparison or evaluation (Franz J. Király, Mateen, and Sonabend 2018) and in conducting this survey, these findings are further supported. Many papers made claims of being 'superior' to the Cox model based on unfair comparisons (RM2){sec:car_reduxstrats_mistakes} or miscommunicating (or misinterpreting) results (e.g. (Fotso 2018)). At this stage, it does not seem possible to make any conclusions about the effectiveness of neural networks in survival analysis. Moreover, even the authors of these models have pointed out problems with transparency (E. M. Biganzoli, Ambrogi, and Boracchi 2009; Liestol, Andersen, and Andersen 1994), which was further highlighted by Schwarzer *et al.*[@Schwarzer2000].

Finally, accessibility of neural networks is also problematic. Many papers do not release their code and instead just state their networks architecture and available packages. In theory, this is enough to build the models however this does not guarantee the reproducibility that is usually expected. For users with a technical background and good coding ability, many of the models above could

be implemented in one of the neural network packages in R, such as **nnet** (N. Venables and D. Ripley 2002) and **neuralnet** (Fritsch, Guenther, and N. Wright 2019); though in practice the only package that does contain these models, **survivalmodels**, does not directly implement the models in R (which is much slower than Python) but provides a method for interfacing the Python implementations in **pycox** (Kvamme 2018).

11. Evaluation

TODO (150-200 WORDS)

This chapter studies how to evaluate the predictions arising from the surveyed models in the previous chapter. 'Model evaluation' is as vague a phrase as 'human evaluation'. A human could be evaluated by a series of exams, physical or neurological tests, aesthetics, etc. Likewise a model could be evaluated according to how well it fits to training data, the quality of predictions on new data, the average prediction, and many more methods. This chapter aims to provide a nuanced approach to defining, understanding, and examining model evaluation. Evaluation is defined in further detail in Section 11.1 and throughout this chapter the definition will continue to be refined and specialised to specific sub-types of evaluation, including discrimination (Section 11.5), calibration (Section 11.6), and overall predictive performance (Section 11.7).

Evaluation is a surprising source of disagreement in the literature with some arguing that the process can often be ignored completely (Laan, Polley, and Hubbard 2007; Wolpert 1992). There is a larger divide in survival analysis as many believe that the primary (possibly only) goal is risk prediction (H. C. Chen et al. 2012; Newson 1983; Pencina, D'Agostino, and Song 2012) and thus other forms of evaluation are not required. These strict views can undermine an integral part of the model building and deployment process, and create more division than necessary. This thesis advocates for strict implementation of model evaluation as a critical part of the model building process as well as in continuous monitoring of deployed models. Without rigorous evaluation, a model cannot be 'trusted' to perform well and could be as useless as making random guesses for all predictions. This is critical in survival analysis, which has important applications in healthcare and finance, in these sectors models that have not been evaluated are potentially dangerous.

An infamous example of evaluation going wrong is the Google Flu Trends (GFT) model[1], which claimed to accurately predict future flu trends but was in fact deemed by many a complete failure as it significantly overestimated all predictions, in some cases doubling the true figures (Lazer et al. 2014). The GFT model was never utilised (at least openly) in policy and as such no lasting harm was created. However it is not hard to imagine the problems that would be caused by such a model if it was utilised and trusted during the time of COVID-19. On a more individual level, as machine learning is increasingly deployed in public sectors, major decisions for patients could become increasingly automated (or at least machine-assisted). Patients should expect their models to be as trained and tested as their doctors.

This chapter attempts to highlight the purpose and need of evaluation in survival analysis by first giving a high-level overview to evaluation as a concept, then providing a brief review of commonly-used survival measures and finally extensive treatment to scoring rules for evaluation of probabilistic predictions, including novel definitions and proofs for properness of scoring rules. The term *measure*

[1] https://www.google.org/flutrends/about/

will be used throughout this chapter to refer to functions or 'metrics' that quantify some aspect of model evaluation, this should not be confused with a mathematical measure.

The APT criteria will be utilised to survey these measures. For transparency and accessibility, these are straightforward to apply to measures with the same definitions as for models. For predictive performance this is more complicated as it depends on the model class. Therefore optimal measure performance definitions will be covered within each section.

Notation and Terminology

The notation introduced in Chapter 3 is recapped for use in this chapter. The generative template is given by (X, T, Δ, Y, C) *t.v.i.* $\mathcal{X} \times \mathcal{T} \times \{0,1\} \times \mathcal{T} \times \mathcal{T}$ where $\mathcal{X} \subseteq \mathbb{R}^p$ and $\mathcal{T} \subseteq \mathbb{R}_{\geq 0}$, where C, Y are unobservable, $T := \min\{Y, C\}$, and $\Delta = \mathbb{I}(Y = T)$. Specific survival data is given by training data, $\mathcal{D}_0 = \{(X_1, T_1, \Delta_1), ..., (X_n, T_n, \Delta_n)\}$ where $(X_i, T_i, \Delta_i) \overset{i.i.d.}{\sim} (X, T, \Delta)$, and test data, $\mathcal{D}_1 = \{(X_1^*, T_1^*, \Delta_1^*), ..., (X_m^*, T_m^*, \Delta_m^*)\}$ where $(X_i^*, T_i^*, \Delta_i^*) \overset{i.i.d.}{\sim} (X, T, \Delta)$.

11.1. Evaluation Overview

11.1.1. What is Evaluation?

Evaluation is the process of examining a model's relationship to data, which may refer to the model's relationship to training data, i.e. how well the model is 'fit' to this data, or the relationship to testing data, i.e. how 'good' are the predictions from the model. In this thesis, only three types of evaluation measure are considered and qualitative definitions of these are given here; more precise definitions appear later in the chapter.

- Discrimination – A model's discriminatory power refers to how well it separates observations that are at a higher or lower risk of event. Therefore discrimination is also sometimes referred to as *separation*. For example, a model with good discrimination will predict that (at a given time) a dead patient has a higher probability of being dead than an alive patient. These measures are the most common in survival and assess relative risk or rank predictions.
- Calibration – There is no single agreed upon definition of model calibration, with definitions varying from paper to paper (Collins et al. 2014; F. E. Harrell, Lee, and Mark 1996; Rahman et al. 2017; Van Houwelingen 2000). Generally, a model is said to be well-calibrated if the average predicted values from the model are in some 'agreement' (which is specified by the chosen measure) with the average true observed values.
- Predictive Performance – A model is said to have good predictive performance (or sometimes 'predictive accuracy') if its predictions for new data are 'close to' the truth.

These are referred to as measures of predictive ability as they draw conclusions about the ability of the model to make predictions.[2]

Using these definitions as a primary taxonomy for survival measures is problematic as without clear definitions there can be significant overlap between model 'classes'. Instead this thesis advocates

[2]Measures of predictive ability measure a model's *ability* to make any form of prediction. Measures of predictive performance measure the *performance* of the predictions. In this section a model's predictive ability refers to all three of discrimination, calibration, and predictive performance.

for the same taxonomy as in the previous chapter and categorises measures by the return type that they evaluate: survival time, ranking, or survival distribution.

Goodness-of-fit measures are very briefly discussed in Section 11.3 for completeness, however these are generally out of scope in this thesis as the vast majority (if any) cannot evaluate machine learning models.

11.2. Why are Models Evaluated?

A key element of the scientific method is experiments and validation. In the usual workflow of the scientific method:

- a hypothesis is proposed;
- predictions are made; and
- experiments are performed to test the hypothesis based on these predictions.

For statistical models the same principles are upheld:

i. a model is proposed (by manual or automated selection with possible tuning);
ii. predictions are made either internally (cross-validation) or externally (held-out data); and
iii. validation is performed on these predictions in order to infer something about the model's performance.

The model can then be considered 'good' or 'bad' and either deployed, adjusted, or discarded. As these are models that are run on a computer (as opposed to experiments in the real-world), the process from fitting to validating is relatively quick and as such multiple proposed models can be evaluated and compared at the same time. This provides two key use-cases for evaluation:

i. demonstrating model performance; and
ii. model comparison/selection.

Resistance to model evaluation can be found in the machine learning community. One such example are proponents of inhomogeneous ensemble methods, which combine predictions from multiple different models into a single prediction. The arguments for these models are that:

i. model evaluation can never be precise enough, or strong enough guarantees cannot be given (Jiao and Du 2016); and
ii. ensemble methods can guarantee a better performance than the individual component models and therefore evaluation of the components is not required.

For example, 'super learners' (Laan, Polley, and Hubbard 2007) are a class of such model and claim[3] to guarantee that a super learner will always perform as well as, if not better, than its component models: ''...the super learner framework allows a researcher to try many prediction algorithms...knowing that the final combined super learner fit will either be the best fit or near the best fit" (Polley and Van Der Laan 2010). This has three problems, it:

i. assumes that researchers will only fit sensible prediction algorithms;
ii. advocates for complex ensemble models instead of transparent and parsimonious ones; and

[3]Testing this claim is tangential so for now will be assumed true.

iii. assumes that a super learner is guaranteed to be the (near) 'best fit', which actively discourages simpler models being tested separately.

Each of these problems can be resolved by researchers only fitting sensible models and opting for an Occam's Razor approach where inhomogeneous ensemble methods are used only if they outperform simpler models, thus requiring validation to test this.

By the parsimony principle, if two models have the same predictive performance (within some degree of confidence), then the simpler and more transparent model is preferred. Even a very slight gain in predictive performance could be outweighed by a large increase to complexity. All models, whether simple or complex, should be critically compared to many alternatives. At the very least a model should be compared to a baseline (Section 11.7.5.1) as many performance measures are uninterpretable without a point of comparison (Gressmann et al. 2018).

11.2.1. How are Models Evaluated?

The process of evaluation in machine learning is briefly given as a key method in Section 2.1 and relevant parts are repeated here. The evaluation process itself is a simple application of a suitable mathematical function to predictions and true data. Let L be some evaluation measure and for now assume L is a measure evaluating deterministic predictions (the following generalises to other types trivially). A model will either be evaluated on each prediction separately, in which case $L : \mathbb{R} \times \mathbb{R} \to \bar{\mathbb{R}}$ or the measure is calculated for all predictions simultaneously, in which case $L : \mathbb{R}^m \times \mathbb{R}^m \to \bar{\mathbb{R}}$. Specifically the loss parameters are observed (true) outcomes, Y, and predictions of this outcome, $\hat{Y}$. L is usually referred to as a *loss* when L should be minimised for optimal prediction, whereas a *score* is the term given when L should be maximised.

All evaluation measures discussed in this thesis are out-of-sample measures and therefore evaluation takes place after the model makes predictions on held-out test data.

Specific choices for L are now reviewed.

11.3. In-Sample Measures

In-sample measures are not examined in this thesis as no in-sample measures could be found that are applicable to all machine learning methods and therefore are out of scope for this thesis. Instead, the interested reader is referred to the papers and references listed below:

Residuals

For discussion about model residuals, refer to texts on survival modelling fitting and goodness-of-fit such as:

- Collett (2014)
- Hosmer Jr, Lemeshow, and May (2011)

Both provide a comprehensive overview to model residuals for semi- and fully-parametric low-complexity survival models.

R^2 measures

R^2 type measures have been the focus of several reviews and surveys, in particular the following are recommended:

- Choodari-Oskooei, Royston, and Parmar (2012a) — For a comprehensive review and simulation study of R^2 type measures
- Kent and O'Quigley (1988) — Defines the commonly utilised Kent and O'Quigley R^2 measure
- Patrick Royston and Sauerbrei (2004) — Defines the commonly utilised Royston and Sauerbrei R^2 measure

Likelihood and Information Criteria

Measures of likelihood and information criteria (e.g. AIC, BIC) are commonly utilised in in-sample model comparison of low-complexity survival models though in general are harder (if not impossible) to compute on ML alternatives.

These criterion are originally defined in:

- Akaike (1974) — For the introduction of the AIC
- Schwarz (1978) — For the introduction of the BIC

These are discussed for survival analysis in:

- Volinsky and Raftery (2000) — For discussion on the BIC for survival models.
- HURVICH and TSAI (1989) — Definition of corrected AIC for survival models, AIC_C
- Liang and Zou (2008) — 'Improved' AIC for survival models.

11.4. Evaluating Survival Time

There appears to be little research into measures for evaluating survival time predictions, which is likely due to this task usually being of less interest than the others (Section 3.3). Common measures in survival analysis for survival time predictions are the same as regression measures but with an additional indicator variable to remove censoring. Three common regression measures are the mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE). These are respectively defined for survival analysis as

Definition 11.1 (Survival time measures)**.** Let $\mathcal{T}^m \subseteq \mathbb{R}^m_{>0}$, $\hat{t} = \hat{t}_1, ..., \hat{t}_m, t = t_1, ..., t_m$, $\delta = \delta_1, ..., \delta_m$, and $d \coloneqq \sum_{i=1}^m \delta_i$, then

i. The *censoring-adjusted mean absolute error*, MAE_C is defined by

$$MAE_C : \mathcal{T}^m \times \mathcal{T}^m \times \{0,1\}^m \to \mathbb{R}_{\geq 0}; (\hat{t}, t, \delta) \mapsto \frac{1}{d} \sum_{i=1}^{m} \delta_i |t_i - \hat{t}_i|$$

i. The *censoring-adjusted mean squared error*, MSE_C is defined by

$$MSE_C : \mathcal{T}^m \times \mathcal{T}^m \times \{0,1\}^m \to \mathbb{R}_{\geq 0}; (\hat{t}, t, \delta) \mapsto \frac{1}{d} \sum_{i=1}^{m} \delta_i (t_i - \hat{t}_i)^2$$

i. The *censoring-adjusted root mean squared error*, $RMSE_C$ is defined by

$$RMSE_C : \mathcal{T}^m \times \mathcal{T}^m \times \{0,1\}^m \to \mathbb{R}_{\geq 0}; (\hat{t}, t, \delta) \mapsto \sqrt{MSE_C(t, \hat{t}, \delta)}$$

These are referred to as 'distance' measures as they measure the distance between the true, (t, δ), and predicted, $\hat{t}$, values. This approach is not ideal as the removal of censored observations results in increased bias as the proportion of censoring increases (Section 11.5.1). Furthermore these measures make some assumptions that are likely not valid in a survival setting. For example these metrics assume that an over-prediction should be penalised equally as much as an under-prediction, whereas in survival data it is likely that a model should be overly-cautious and under-predict survival times, i.e. it is safer to predict a patient is more at risk and will die sooner rather than less risk and die later.

These measures are clearly transparent and accessible as off-shelf implementation is straightforward, though **mlr3proba** (R. Sonabend et al. 2021) was the only R package found to implement these. For performance, no conclusions can be drawn as no research could be found into the theoretical properties of these losses; despite this there is evidence of them being utilised in the literature (P. Wang, Li, and Reddy 2019).

11.5. Evaluating Continuous Rankings

The next category of survival measures assess predictive performance via discrimination for the evaluation of continuous ranking predictions. Assessment of continuous rankings are also possible by measures of calibration however few methods could be found that generalised to all (not just PH) model forms. Therefore this section exclusively discusses measures of discrimination. First time-independent concordance indices (Section 11.5.1) are discussed and then time-dependent AUCs (Section 11.5.2).

Measures of discrimination identify how well a model can separate patients into different risk groups. A model has perfect discrimination if it correctly predicts that patient i is at higher risk of death than patient j if patient i dies first. This risk of death is derived from the ranking prediction type. All discrimination measures are ranking measures, which means that the exact predicted value is irrelevant, only its relative ordering is required. For example given predictions $\{100, 2, 299.3\}$, only their rankings, $\{2, 1, 3\}$, are used by measures of discrimination.

11.5.1. Concordance Indices

The simplest form of discrimination measures are concordance indices, which in general measure the proportion of cases in which the model correctly separates a pair of observations into 'low' and 'high' risk.

Definition 11.2 (Concordance)**.** Let (i, j) be a pair of observations with outcomes$\backslash \{(t_i, \delta_i), (t_j, \delta_j)\} \overset{i.i.d.}{\sim} (T, \Delta)$ and let $y_i, y_j \in \mathbb{R}$ be their respective risk predictions. Then (i, j) are called (F. E. J. Harrell et al. 1984; F. E. Harrell, Califf, and Pryor 1982):

- *Comparable* if $t_i < t_j$ and $\delta_i = 1$;
- *Concordant* if $y_i > y_j$.[4]

A concordance index (C-index) is a weighted proportion of the number of concordant pairs over the number of comparable pairs. As such, a C-index value is between $[0, 1]$ with 1 indicating perfect separation, 0.5 indicating no separation, and 0 being separation in the 'wrong direction', i.e. all high risk patients being ranked lower than all low risk patients. Concordance measures may either be reported as a value in $[0, 1]$, a percentage, or as 'discriminatory power'. Discriminatory power refers to the percentage improvement of a model's discrimination above the baseline value of 0.5. For example if a model has a concordance of 0.8 then its discriminatory power is $(0.8 - 0.5)/0.5 = 60$. This representation of discrimination provides more information by encoding the model's improvement over some baseline although is often confused with reporting concordance as a percentage (e.g. reporting a concordance of 0.8 as 80%).

The most common concordance indices can be expressed as a general measure.

Definition 11.3 (C-index)**.** Let $\mathcal{T}^m \subseteq \mathbb{R}^m_{>0}$, $y = y_1, ..., y_m, t = t_1, ..., t_m$, $\delta = \delta_1, ..., \delta_m$, and let W be a weighting function. Then, the *survival concordance index* is defined by,

$$C : \mathbb{R}^m \times \mathcal{T}^m \times \{0, 1\}^m \times \mathbb{R}_{\geq 0} \to [0, 1];$$
$$(y, t, \delta|\tau) \mapsto \frac{\sum_{i \neq j} W(t_i)\mathbb{I}(t_i < t_j, y_i > y_j, t_i < \tau)\delta_i}{\sum_{i \neq j} W(t_i)\mathbb{I}(t_i < t_j, t_i < \tau)\delta_i}$$

for some cut-off time τ.

The choice of W specifies a particular evaluation measure (see below). To evaluate the discrimination of a prediction functional, $\hat{g}$, with predicted rankings from the model, $r = r_1, ..., r_m$, the concordance is calculated as $\backslash C(r, (T^*_1, ..., T^*_m), (\Delta^*_1, ..., \Delta^*_m)|\tau)$ for some choice of $\tau \in \mathbb{R}_{\geq 0}$. The use of the cut-off τ mitigates against decreased sample size over time due to the removal of censored observations. There are multiple methods for dealing with tied times but in practice a value of 0.5 is usually taken when $t_i = t_j$ (Therneau and Atkinson 2020). The following weights have been proposed for the concordance index (Therneau and Atkinson 2020):

[4] Recall (Section 3.3) this thesis defines the risk ranking such that a higher value implies higher risk of death and so a pair is concordant if $\mathbb{I}(t_i < t_j, y_i > y_j)$, whereas this would be $\mathbb{I}(t_i < t_j, y_i < y_j)$ if a higher value implied a lower risk of death.

- $W(t_i) = 1$ – This is Harrell's concordance index, C_H (F. E. J. Harrell et al. 1984; F. E. Harrell, Califf, and Pryor 1982), which is widely accepted to be the most common survival measure (Collins et al. 2014; **GonenHeller2005?**; Rahman et al. 2017). There is no cut-off in the original definition of C_H ($\tau = \infty$).
- $W(t_i) = [\hat{G}_{KM}(t_i)]^{-2}$ – This is Uno's C, C_U (Uno et al. 2011). $\hat{G}_{KM}$ is the Kaplan-Meier estimate of the survival function of the censoring distribution fit on training data. This is referred to as an Inverse Probability of Censoring Weighted (IPCW) measure as the estimated censoring distribution is utilised to weight the measure in order to compensate for removed censored observations.
- $W(t_i) = [\hat{G}_{KM}(t_i)]^{-1}$
- $W(t_i) = \hat{S}_{KM}(t_i)$. $\hat{S}_{KM}$ is the Kaplan-Meier estimator of the survival distribution.
- $W(t_i) = \hat{S}_{KM}(t_i)/\hat{G}_{KM}(t_i)$

All methods assume that censoring is conditionally-independent of the event given the features (Section 3.1.2), otherwise weighting by $\hat{S}_{KM}$ or $\hat{G}_{KM}$ would not be applicable. It is assumed here that $\hat{S}_{KM}$ and $\hat{G}_{KM}$ are estimated on the training data and not the testing data (though the latter is often seen in implementation (Therneau 2015)).

With respect to being APT, all concordance indices are highly transparent and accessible, with many off-shelf implementations. With respect to performance, Choodari-Oskooei *et al.* (2012) (Choodari-Oskooei, Royston, and Parmar 2012a) define a measure as performant if it is:[5]

i. independent of censoring;
ii. interpretable; and
iii. robust against outliers.

This second property is already covered by 'transparency'. The third property is guaranteed for all measures of concordance, which are ranking measures; all outliers are removed once ranks are applied to predictions. Therefore the first property, ''a measure that is the least affected by the amount of censoring is generally preferred" (Choodari-Oskooei, Royston, and Parmar 2012a), is now considered.

Several papers have shown that C_H is affected by the precense of censoring (Koziol and Jia 2009; Pencina, D'Agostino, and Song 2012; Patrick Royston and Altman 2013; Uno et al. 2011) as the measure ignores pairs in which the shorter survival time is censored. Despite this, C_H is still the most widely utilised measure and moreover if a suitable cut-of τ is chosen, then all these weightings perform very similarly (Rahman et al. 2017; Schmid and Potapov 2012).

Measures that utilise other weightings have been demonstrated to be less affected by censoring than C_H (Rahman et al. 2017). However if a poor choice is selected for τ then IPCW measures (which include $\hat{G}_{KM}$ in the weighting) can be highly unstable (Rahman et al. 2017). For example, the variance of C_U has been shown to drastically increase more than other measures with increased censoring (Schmid and Potapov 2012).

None of these measures are perfect and all have been shown to be affected to some extent by censoring (Schmid and Potapov 2012), which can lead to both under-confidence and over-confidence in the model's discriminatory ability. For example, C_U has been observed to report values as low as 0.2 when the 'true estimate' was 0.6 (Schmid and Potapov 2012). Therefore interpreting a value

[5]This paper refers specifically to measures of explained variation and therefore only the properties that generalise to all measures are included here.

from these measures can be very difficult, for example naively reporting a concordance of 60% when $C_H = 0.6$ would be incorrect as this value may mean very different things for different amounts of censoring. Whilst intepreting these measures may be difficult, it is not impossible as all these estimators tend to produce values around a similar range (Rahman et al. 2017; Schmid and Potapov 2012). Therefore this thesis advocates for multiple concordance indices being reported alongside expert interpretation that takes into account sample size and censoring proportions (Schmid and Potapov 2012) as well as 'risk profiles' (how at risk patients are) (Rahman et al. 2017).

For within-study model comparison, instability from censoring is not of concern as the measure will be affected equally across all models; though interpretation remains difficult. However a concordance from one study cannot be compared to that from another if the datasets differ greatly in the proportion of censoring. Future research could consider more robust concordance indices that can provide greater ease of interpretation.

As well as the concordance indices discussed here, another promiment alterntive was derived by G"onen and Heller (2005) (**GonenHeller2005?**). However as this is only applicable to the Cox PH it is out of scope for this thesis, which is primarily concerned with generalisable measures for model comparison.

In simulation experiments, the concordance indices that tended to perform 'better' were those based on AUC-type measures, these are now discussed.

11.5.2. AUC Measures

AUC, or AUROC, measures calculate the Area Under the Receiver Operating Characteristic (ROC) Curve, which is a plot of the *sensitivity* (or true positive rate (TPR)) against \$1 - \$*specificity* (or true negative rate (TNR)) at varying thresholds (described below) for the predicted probability (or risk) of event. Figure 11.1 visualises ROC curves for two classification models. The blue line is a featureless baseline that has no discrimination. The red line is a decision tree with better discrimination as it comes closer to the top-left corner.

In a classification setting with no censoring, the AUC has the same interpretation as Harrell's C (Uno et al. 2011). AUC measures for survival analysis have been developed in order to provide a time-dependent measure of discriminatory ability (Patrick J. Heagerty, Lumley, and Pepe 2000). The proposed concordance indices described above are time-independent, which is useful for producing a single statistic. However, in a survival setting it can reasonably be expected for a model to perform differently over time and therefore time-dependent measures are advantageous. First discussion around computation of TPR and TNR are provided and then how these are incorporated into the AUC equation.

The AUC, TPR, and TNR are derived from the *confusion matrix* in a binary classification setting. Let $b, \widehat{b} \in \{0, 1\}$ be the true and predicted binary outcomes respectively. The confusion matrix is

	$b = 1$	$b = 0$
$\widehat{b} = 1$	TP	FP
$\widehat{b} = 0$	FN	TN

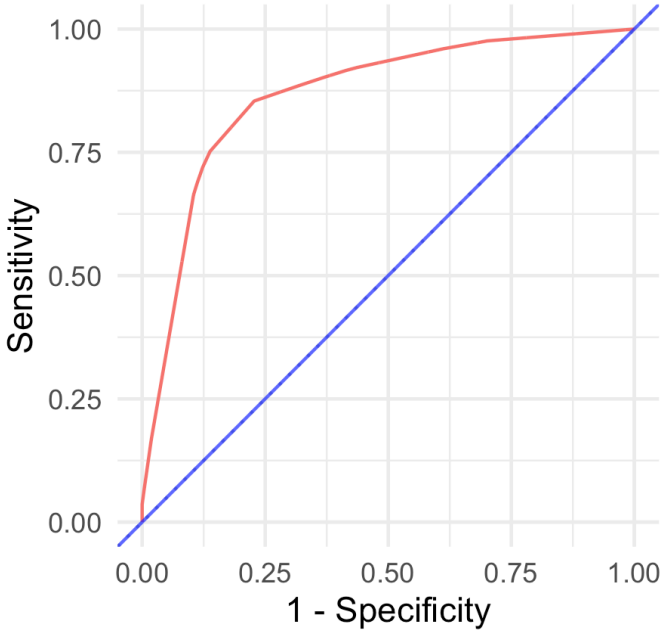


Figure 11.1.: ROC Curves for a classification example. Red is a decision tree with good discrimination as it 'hugs' the top-left corner. Blue is a featureless baseline with no discrimination as it sits on $y = x$.

where $TN := \sum_i \mathbb{I}(b = 0, \hat{b} = 0)$ is the number of (#) true negatives, $TP := \sum_i \mathbb{I}(b = 1, \hat{b} = 1)$ is # true positives, $FP := \sum_i \mathbb{I}(b = 0, \hat{b} = 1)$ is # false positives, and $FN := \sum_i \mathbb{I}(b = 1, \hat{b} = 0)$ is # false negatives. From these are derived

$$TPR := \frac{TP}{TP + FN} \tag{11.1}$$

$$TNR := \frac{TN}{TN + FP} \tag{11.2}$$

In classification, a probabilistic prediction of an event can simply be *thresholded* (or 'binarised') to obtain a deterministic prediction. For a predicted $\hat{p} := \hat{P}(b = 1)$, and threshold α, the thresholded binary prediction is given by $\hat{b} := \mathbb{I}(\hat{p} > \alpha)$. In survival analysis, this is complicated as either models only predict a continuous ranking (and not a probability of death), or a full survival distribution, which implies that the probability of death changes over time; it is the first of these that is utilised in AUC measures. Two primary methods for doing so have emerged, the first is to use an IPCW method to weight the thresholded linear predictor by an estimated censoring distribution at a given time, the second is to first classify cases and controls then compute estimators based on these classes. All measures of TPR, TNR and AUC are in the range $[0, 1]$ with larger values preferred.

Weighting the linear predictor was proposed by Uno *et al.* (2007) (Uno et al. 2007) and provides a method for estimating TPR and TNR via

$$TPR_U : \mathbb{R}^m \times \mathbb{R}^m_{\geq 0} \times \{0,1\}^m \times \mathbb{R}_{\geq 0} \times \mathbb{R} \to [0,1];$$
$$(\hat{\eta}, t, \delta | \tau, \alpha) \mapsto \frac{\sum_{i=1}^m \delta_i \mathbb{I}(k(\hat{\eta}_i) > \alpha, t_i \leq \tau)[\hat{G}_{KM}(t_i)]^{-1}}{\sum_{i=1}^m \delta_i \mathbb{I}(t_i \leq \tau)[\hat{G}_{KM}(t_i)]^{-1}}$$

and

$$TNR_U : \mathbb{R}^m \times \mathbb{R}^m_{\geq 0} \times \mathbb{R}_{\geq 0} \times \mathbb{R} \to [0,1];$$
$$(\hat{\eta}, t | \tau, \alpha) \mapsto \frac{\sum_{i=1}^m \mathbb{I}(k(\hat{\eta}_i) \leq \alpha, t_i > \tau)}{\sum_{i=1}^m \mathbb{I}(t_i > \tau)}$$

where τ is the time at which to evaluate the measure, α is a cut-off for the linear predictor, and k is a known, strictly increasing, differentiable function. k is chosen depending on the model choice, for example if the fitted model is PH then $k(x) = 1 - \exp(-\exp(x))$ (Uno et al. 2007). Similarities can be drawn between these equations and Uno's concordance index, in particular the use of IPCW. Censoring is again assumed to be at least random once conditioned on features. Plotting TPR_U against $1 - TNR_U$ for varying values of α provides the ROC.

The second method, which appears to be more prominent in the literature, is derived from Heagerty and Zheng (2005) (Patrick J. Heagerty and Zheng 2005). They define four distinct classes, in which observations are split into controls and cases.

An observation is a *case* at a given time-point if they are dead, otherwise they are a *control*. These definitions imply that all observations begin as controls and (hypothetically) become cases over time. Cases are then split into *incident* or *cumulative* and controls are split into *static* or *dynamic*. The choice between modelling static or dynamic controls is dependent on the question of interest. Modelling static controls implies that a 'subject does not change disease status' (Patrick J. Heagerty and Zheng 2005), and few methods have been developed for this setting (Kamarudin, Cox, and Kolamunnage-Dona 2017), as such the focus here is on *dynamic* controls. The incident/cumulative cases choice is discussed in more detail below.[6]

The TNR for dynamic cases is defined as

$$TNR_D(y, N | \alpha, \tau) = P(y_i \leq \alpha | N_i(\tau) = 0)$$

where $y = (y_1, ..., y_n)$ is some deterministic prediction and $N(\tau)$ is a count of the number of events in $[0, \tau)$. Heagerty and Zheng further specify y to be the predicted linear predictor $\hat{\eta}$. Cumulative/dynamic and incident/dynamic measures are available in software packages 'off-shelf', these are respectively defined by

$$TPR_C(y, N | \alpha, \tau) = P(y_i > \alpha | N_i(\tau) = 1)$$

and

[6] All measures discussed in this section evaluate model discrimination from 'markers', which may be a *predictive* marker (model predictions) or a *prognostic* marker (a single covariate). This section always defines a marker as a ranking prediction, which is valid for all measures discussed here with the exception of one given at the end.

$$TPR_I(y, N|\alpha, \tau) = P(y_i > \alpha | dN_i(\tau) = 1)$$

where $dN_i(\tau) = N_i(\tau) - N_i(\tau-)$. Practical estimation of these quantities is not discussed here.

The choice between the incident/dynamic (I/D) and cumulative/dynamic (C/D) measures primarily relates to the use-case. The C/D measures are preferred if a specific time-point is of interest (Patrick J. Heagerty and Zheng 2005) and is implemented in several applications for this purpose (Kamarudin, Cox, and Kolamunnage-Dona 2017). The I/D measures are preferred when the true survival time is known and discrimination is desired at the given event time (Patrick J. Heagerty and Zheng 2005).

Defining a time-specific AUC is now possible with

$$AUC(y, N|\tau) = \int_0^1 TPR(y, N|1 - TNR^{-1}(p|\tau), \tau)\ dp$$

Finally, integrating over all time-points produces a time-dependent AUC and as usual a cut-off is applied for the upper limit,

$$AUC^*(y, N|\tau^*) = \int_0^{\tau^*} AUC(y, N|\tau) \frac{2\hat{p}_{KM}(\tau)\hat{S}_{KM}(\tau)}{1 - \hat{S}^2_{KM}(\tau^*)}\ d\tau$$

where $\hat{S}_{KM}, \hat{p}_{KM}$ are survival and mass functions estimated with a Kaplan-Meier model on training data.

Since Heagerty and Zheng's paper, other methods for calculating the time-dependent AUC have been devised, including by Chambless and Diao (Chambless and Diao 2006), Song and Zhou (Song and Zhou 2008), and Hung and Chiang (Hung and Chiang 2010). These either stem from the Heagerty and Zheng paper or ignore the case/control distinction and derive the AUC via different estimation methods of TPR and TNR. Blanche *et al.* (2012) (Blanche, Latouche, and Viallon 2012) surveyed these and concluded ''regarding the choice of the retained definition for cases and controls, no clear guidance has really emerged in the literature'', but agree with Heagerty and Zeng on the use of C/D for clinical trials and I/D for 'pure' evaluation of the marker. Blanche *et al.* (2013) (Blanche, Dartigues, and Jacqmin-Gadda 2013) published a survey of C/D AUC measures with an emphasis on non-parametric estimators with marker-dependent censoring, including their own Conditional IPCW (CIPCW) AUC,

$$AUC_B(y, t, \delta, \hat{G}|\tau) = \frac{\sum_{i \neq j} \mathbb{I}(y_i > y_j)\mathbb{I}(t_i \leq \tau, t_j > \tau)\frac{\delta_i}{m^2 \hat{G}(t_i|y_i)\hat{G}(\tau|y_j)}}{\left(\sum_{i=1}^m \mathbb{I}(t_i \leq \tau)\frac{\delta_i}{m\hat{G}(t_i|y_i)}\right)\left(\sum_{j=1}^m \mathbb{I}(t_j > \tau)\frac{1}{m\hat{G}(\tau|y_j)}\right)}$$

where $t = (t_1, ..., t_m)$, and $\hat{G}$ is the Akritas (Akritas 1994) estimator of the censoring distribution (Section 5.1.1). It can be shown that setting the λ parameter of the Akritas estimator to 1 results in the IPCW estimators (Blanche, Dartigues, and Jacqmin-Gadda 2013). However unlike the previous measures in which a deterministic prediction can be substituted for the marker, this is not valid for this estimator and as such this cannot be used for predictions. This is clear from the weights, $\hat{G}(t|y)$, in the equation which are dependent on the prediction itself. The purpose of the CIPCW

method is to adapt the IPCW weights to be conditioned on the data covariates, which is not the case when y is a predictive marker. Hence the following adaptation is considered instead,

$$AUC^*_B(y, x, t, \delta, \hat{G}|\tau) = \frac{\sum_{i \neq j} \mathbb{I}(y_i > y_j)\mathbb{I}(t_i \leq \tau, t_j > \tau)\frac{\delta_i}{m^2 \hat{G}(t_i|x_i)\hat{G}(\tau|x_j)}}{\left(\sum_{i=1}^m \mathbb{I}(t_i \leq \tau)\frac{\delta_i}{m\hat{G}(t_i|x_i)}\right)\left(\sum_{j=1}^m \mathbb{I}(t_j > \tau)\frac{1}{m\hat{G}(\tau|x_j)}\right)}$$

where x are random covariates (possibly from a separate training dataset).

AUC measures are less transparent and less accessible than the simpler time-independent concordance indices, only the **survAUC** (Potapov, Adler, and Schmid 2012) package could be found that implements these measures. For performance, reviews of these measures have produced (sometimes markedly) different results (Blanche, Latouche, and Viallon 2012; Li, Greene, and Hu 2018; Kamarudin, Cox, and Kolamunnage-Dona 2017) with no clear consensus on how and when these measures should be used. The primary advantage of these measures is to extend discrimination metrics to be time-dependent. However it is unclear how to interpret a threshold of a linear predictor and moreover if this is even the 'correct' quantity to threshold, especially when survival distribution predictions are the more natural object to evaluate over time. Methods for evaluating these distribution predictions are now discussed.

11.6. Evaluating Distributions by Calibration

The final discussed measures are for evaluating survival distributions. First measures of calibration are briefly discussed in this section and then extensive treatment is given to scoring rules (Section 11.7).

Random Variable and Distribution Notation

Throughout these next two sections, two different notations are utilised for random variables and distributions. The first is the 'standard' notation, for example if ζ is a continuous probability distribution and $X \sim \zeta$ is a random variable, then f_X is the probability density function of X. The second notation associates distribution functions directly with the distribution and not the variable. For example if ζ is a continuous probability distribution then $\zeta.f$ is the probability density function of ζ. Analogously for the probability mass, cumulative distribution, hazard, cumulative hazard, and survival functions of $X \sim \zeta$, $p_X/\zeta.p, F_X/\zeta.F, h_X/\zeta.h, H_X/\zeta.H, S_X/\zeta.S$. This notation provides a clearer separation of probability distributions and random variables, which in turn allows for cleaner proofs involving probability distributions.

Measures of Calibration

Few measures of calibration exist in survival analysis (Rahman et al. 2017) and this is likely due to the meaning of calibration being unclear in this context (Van Houwelingen 2000). This is compounded by the fact that calibration is often evaluated graphically, which can leave room for high subjectivity and thus may be restricted to expert interpretation. For these reasons, measures of calibration are only considered in this thesis with respect to accessibility and transparency as there is no clear meaning for what makes a calibration measure performant. Many methods of

calibration are restricted to calibration and re-calibration of PH models (Demler, Paynter, and Cook 2015; Van Houwelingen 2000), none of these are considered here as they do not generalise to all (or at least many) survival models.

Point and Probabilistic Calibration

Andres *et al.* (2018) (Andres et al. 2018) derived a taxonomy for calibration measures to separate measures that only evaluate distributions at a single time-point ('1-Calibration') and measures that evaluate distributions at all time-points ('distributional-calibration'). This section will use the same taxonomy but in keeping with machine learning terminology will refer to '1-Calibration' as 'Point Calibration' and 'distributional-calibration' as 'Probabilistic Calibration'.

All measures considered previously can be viewed as 'point' measures as they evaluate predictions at a single point, specifically comparing the predicted linear predictor (more generally relative risk) or survival time to the true time of death. However calibration measures and scoring rules instead evaluate predicted distributions and specifically functions that vary over time, hence it is often of more interest to evaluate these functions at multiple (all if discrete) time-points in order to derive a metric that captures changes over time. For example one may expect probabilistic predictions to be more accurate in the near-future and to steadily worsen as uncertainty increases over time (both mathematical (censoring) and real-world uncertainty), and therefore a measure that only evaluates distributions at a single (possibly early) time-point cannot assess the true variation in the prediction.

Mathematically this difference in measures may be considered as follows: Let $\mathcal{P}$ be a set of distributions over $\mathcal{T} \subseteq \mathbb{R}_{>0}$, then a point measure for evaluating distributions is given by,

$$L_1 : \mathcal{P} \times \mathcal{T} \times \{0,1\} \times \mathcal{T} \to \bar{\mathbb{R}}; \quad (\zeta, t, \delta|\tau) \mapsto g_1(\zeta.\rho(\tau), t, \delta)$$

and a probabilistic measure is given by,

$$L_P : \mathcal{P} \times \mathcal{T} \times \{0,1\} \times \mathbb{R}_{>0} \to \bar{\mathbb{R}}; \quad (\zeta, t, \delta|\tau^*) \mapsto \int_0^{\tau^*} g_P(\zeta.\rho(\tau), t, \delta)\ d\tau$$

or

$$L_P : \mathcal{P} \times \mathcal{T} \times \{0,1\} \times \mathbb{R}_{>0} \to \bar{\mathbb{R}}; \quad (\zeta, t, \delta|\tau^*) \mapsto \sum_{\tau=0}^{\tau^*} g_P(\zeta.\rho(\tau), t, \delta)$$

where τ^* is some cut-off for the measure to control uncertainty increasing over time, ρ is usually the survival function but may be any distribution-defining function, and g_1, g_P are functions corresponding to specific measures (some examples in next two sections). Note that τ is an argument (not a free variable) of L_1 as the fixed choice of τ is measure-dependent; usually $\tau = t$.

Less abstractly, a point-calibration measure will evaluate a function of the predicted distribution at a single time-point whereas a probabilistic measure evaluates the distribution over a range of time-points; in both cases the evaluated quantity is compared to the observed outcome, (T^*, Δ^*).

11.6.1. Point Calibration

Point calibration measures can be further divided into metrics that evaluate calibration at a single time-point (by reduction) and measures that evaluate an entire distribution by only considering the event time. The subtle difference significantly affects conclusions that can be drawn. In the first case, a calibration measure can only draw conclusions at that one time-point, whereas the second case can draw conclusions about the calibration of the entire distribution.

11.6.1.1. Calibration by Reduction

Point calibration measures are implicitly reduction methods as they attempt to evaluate a full distribution based on a single point only. For example given a predicted survival function $\zeta.S$, then one could select a time-point τ^* and calculate the survival function at this time, $\zeta.S(\tau^*)$, probabilistic classification calibration measures can then be utilised. Using this approach one may employ common calibration methods such as the Hosmer–Lemeshow test (Hosmer and Lemeshow 1980). Calibration at a single point in this manner is not particularly useful as a model may be well-calibrated at one time-point and then poorly calibrated at all others (Haider et al. 2020). To overcome this one could perform the Hosmer–Lemeshow test (or any other applicable test) multiple times at different values of $\tau^* \in \mathbb{R}_{\geq 0}$. However doing so is inefficient and can lead to problems with 'multiple testing'; hence these single-point methods are not considered further.

11.6.1.2. Houwelingen's α

Methods that evaluate entire distributions based on a single point may be more useful as conclusions can be drawn at the distribution level. One such method is termed here 'Houwelingen's α'. van Houwelingen proposed several measures (Van Houwelingen 2000) for calibration but only one generalises to all probabilistic survival models. This method evaluates the predicted cumulative hazard function, $\zeta_i.H$ (for some predicted distribution ζ_i), by comparing $\zeta_i.H$ to the 'true' hypothetical cumulative hazard, H. The test statistic, H_α, is defined by

$$H_\alpha := \frac{\sum_i H_i(T_i^*)}{\sum_i \zeta_i.H(T_i^*)} \approx \frac{\sum_i \Delta_i^*}{\sum_i \zeta_i.H(T_i^*)}$$

where $\zeta = (\zeta_1, ..., \zeta_m)$ are predicted distributions and $\{(T_1^*, \Delta_1^*), ..., (T_m^*, \Delta_m^*)\} \overset{i.i.d.}{\sim} (T, \Delta)$ is some test data. The model is therefore well-calibrated if $H_\alpha = 1$. This has standard error $SE(H_\alpha) = \exp(1/\sqrt{(\sum_i \Delta_i^*)})$.

The approximate equality is motivated by formulating survival data as a counting process and noting that in this setting the cumulative hazard function can estimate the number of events in a time-period (Hosmer Jr, Lemeshow, and May 2011). No study could be found that utilised H_α for model comparison, possibly because graphical methods are favoured. This method can infer results about the calibration of an entire model and not just at a single point because the measure is calculated at a meaningful time (the event time) and utilises known results from counting processes to verify if the expected number of deaths equals the observed number of deaths.

However, as with the reduction method, the statistic is derived from a single point (the observed event time) for each individual and thus it is possible that the model is well-calibrated only for making predictions at the event time, but not over the full $\mathbb{R}_{>0}$ range.

11.6.2. Probabilistic Calibration

Unlike other areas of evaluation, graphical methods are favoured in calibration and possibly more so than numerical ones. Graphical methods compare the average predicted distribution to the expected distribution. As the expected distribution is itself unknown, this is often estimated with the Kaplan-Meier curve.

11.6.2.1. Kaplan-Meier Comparison

The simplest graphical comparison compares the average predicted survival curve to the Kaplan-Meier curve estimated on the testing data. Formally, let $\setminus\ \zeta_1.S, ..., \zeta_m.S$ be predicted survival functions, then the average predicted survival function is a mixture of these distributions, $\frac{1}{m}\sum_{i=1}^{m} \zeta_i.S(\tau)$. Plotting this mixture and the Kaplan-Meier on τ vs $S(\tau)$ allows a visual comparison of how closely these curves align. An example is given in Figure 11.2, the Cox model (CPH) is well-calibrated as it almost perfectly overlaps the Kaplan-Meier estimator, whereas predictions from the poorly-calibrated support vector machine (SVM) are far from this line.

This approach is both simple and interpretable. In the example above one can conclude: on average, the trained Cox PH predicts a distribution just as well as (or very close to) an unconditional estimator using the real test data. A major caveat is that conclusions are at an average *population* level with no individual-level measurement.

In order to capture finer information on a level closer to inidivduals, calibration can be applied to the predicted relative risks or linear predictor. One such approach is to bin the predictions to create different 'risk groups' from low-to-high risk (Patrick Royston and Altman 2013). These groups are then plotted against a stratified Kaplan-Meier estimator. This allows for a more nuanced approach to calibration and can simultaneously visualise a model's discrimination. However this method is far less transparent as it adds even more subjectivity around how many risk groups to create and how to create them (Patrick Royston and Altman 2013).

11.6.2.2. D-Calibration

D-Calibration (Andres et al. 2018; Haider et al. 2020) is a very recent method that aims to evaluate a model's calibration at all time-points in a predicted survival distribution. The D-calibration measure is identical to the χ^2 test-statistic, which is usually written as follows

$$\chi^2 := \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

where $O_1, ..., O_n$ is the observed number of events in n groups and $E_1, ..., E_n$ is the expected number of events. The statistic is utilised to determine if the underlying distribution of the observed events follows a theoretical/expected distribution.

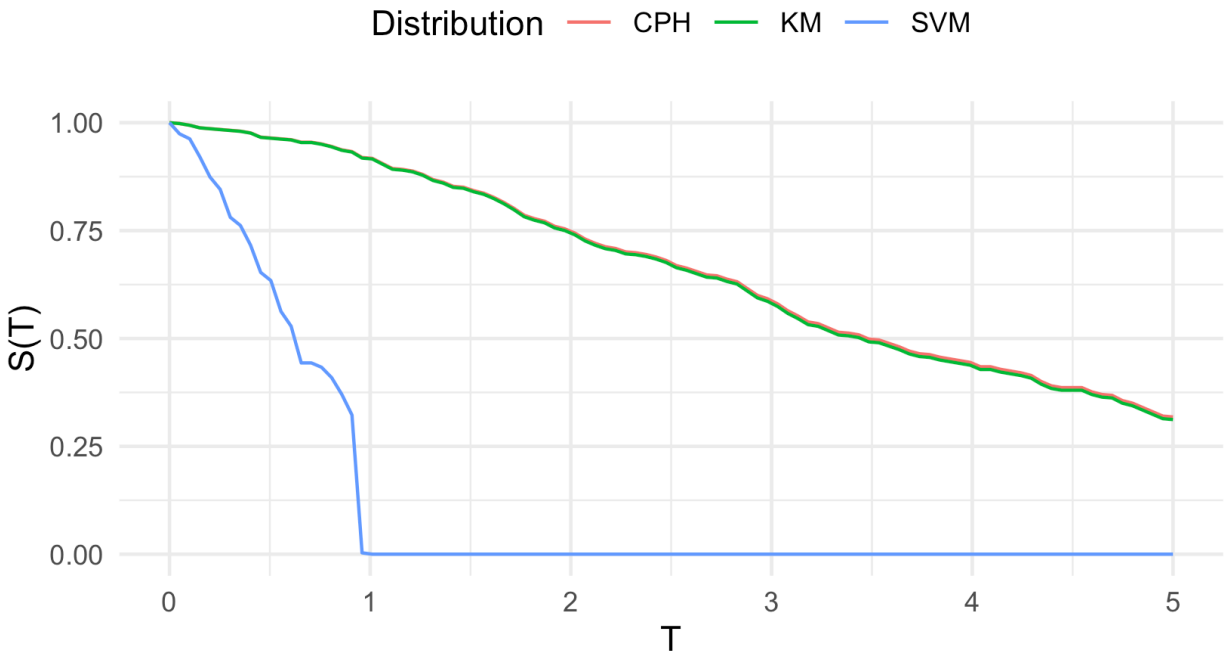


Figure 11.2.: Assessing the calibration of a Cox PH (CPH) and SVM (with distribution composition by PH form and Kaplan-Meier (Chapter 12)) by comparing the average survival prediction to a Kaplan-Meier (KM) estimate on the testing dataset. x-axis is time and y-axis is the predicted survival functions evaluated over time. The CPH (red line) is said to be well-calibrated as it almost perfectly overlaps the Kaplan-Meier (green line), whereas the SVM (blue line) is far from this line. Models trained and tested on randomly simulated data from the `simsurv` (Brilleman 2019) package in **mlr3proba** (R. Sonabend et al. 2021).

The D-Calibration measure tests if predictions (observations) from the survival functions of predicted distributions, $\zeta_1.S, ..., \zeta_m.S$, follow the uniform distribution as expected. The following lemma motivates this test.

Lemma 11.1. *Let ζ be a continuous probability distribution and let $X \sim \zeta$ be a random variable. Let S_X be the survival function of X. Then $S_X(X) \sim \mathcal{U}(0, 1)$.*

In order to utilise the χ^2 test (for categorical variables), the $[0, 1]$ codomain of $\zeta_i.S$ is cut into B disjoint contiguous intervals ('bins') over the full range $[0, 1]$. Let m be the total number of observations in the test data. Then assuming a discrete uniform distribution as the theoretical distribution, the expected number of events is m/B.

The observed number of events in bin i, O_i, is defined as follows: Define b_i as the set of observations that die in the ith bin, formally defined by $b_i := \{j \in 1, ..., m : \lceil \zeta_j.S(T_j^*)B \rceil = i\}$, where $j = 1, ..., m$ are the indices of the test observations and $\zeta = (\zeta_1, ..., \zeta_m)$ are predicted distributions.[7] Then, $O_i = |b_i|, \forall i \in 1, ..., B$.

The D-Calibration measure, or χ^2 statistic, is now defined by,

$$D_{\chi^2}(\zeta, T^*) := \frac{\sum_{i=1}^{B} (O_i - \frac{m}{B})^2}{m/B}$$

This measure has several useful properties. Firstly, a p-value can be derived from χ^2_{B-1} to hypothesis test if a single model is 'D-calibrated'. Secondly, as a model is increasingly well-calibrated it holds that $D_{\chi^2} \to 0$ (as the number of observed events approach expected events), which motivates utilising the test for model comparison. Thirdly, the theory lends itself very nicely to an intuitive graphical calibration method:

If a model is D-calibrated, i.e. predicted distributions from the model result in a low D-calibration, then one expects,

$$p = \frac{\sum_i \mathbb{I}(T_i^* \leq \zeta_i.F^{-1}(p))}{|T^*|} \tag{11.3}$$

where $p \in [0, 1]$ and $\zeta_i.F^{-1}$ is the inverse cumulative distribution function of the ith predicted distribution. In words, if a model is D-calibrated then the number of deaths occurring at or before each quantile should be equal to the quantile itself, for example 50% of deaths should occur before their predicted median survival time. Therefore one can graphically test for D-calibration by plotting p on the x-axis and the RHS of Equation 11.3 on the y-axis. A D-calibrated model should result in a straight line on $x = y$. This is visualised in Figure 11.3 for the same models as in Figure 11.2. Again the SVM is terribly-calibrated but the CPH is better calibrated. In this case it is clearer that the D-calibration of the CPH is not perfect, especially at higher quantiles. Comparison to χ^2_9 indicates the CPH is D-calibrated whereas the SVM is not.

[7] This is a slightly simplified procedure which omits handling of censoring, but this is easily extended in the full algorithm, see Algorithm 2 of Haider *et al.* (2020) (Haider et al. 2020).

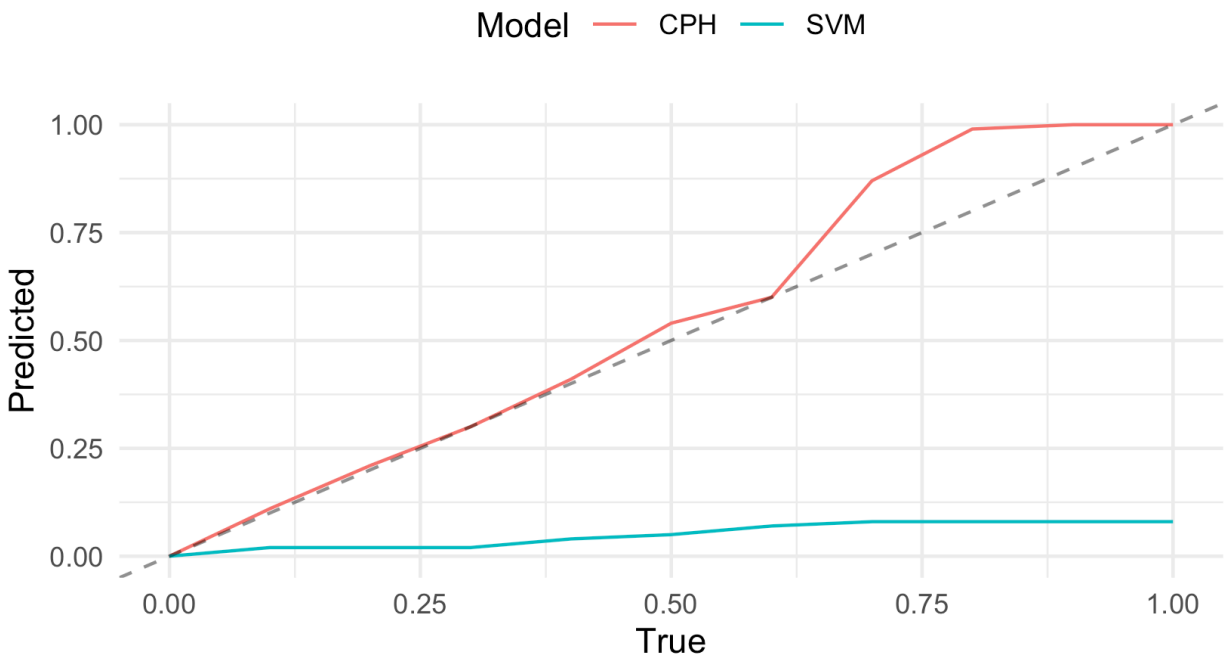


Figure 11.3.: Assessing the D-calibration of the Cox PH (CPH) and SVM from the same data as Figure 11.2: models trained and tested on randomly simulated data from the `simsurv` (Brilleman 2019) package in **mlr3proba** (R. Sonabend et al. 2021). x-axis are quantiles in $[0, 1]$ and y-axis are predicted quantiles from the models. The dashed line is $y = x$. Again the SVM is terribly calibrated and the CPH is better calibrated as it is closer to $y = x$.

11.6.2.3. Transparency and Accessibility

It has already been stated that performance cannot be considered for calibration measures however it is unclear if any of these measures are even accessible or transparent as they often require expert interpretation to prevent erroneous conclusions. This is demonstrated by example using the same data and models as in Figure 11.3. The predictions from these models are evaluated with Harrell's C (Section 11.5.1), the Integrated Graf Score (Section 11.7.3), D-Calibration, and Houwelingen's α (Table 11.2). All measures agree that the SVM performs poorly. In contrast, whilst the Cox PH (CPH) is well-calibrated according to both measures, its concordance is quite bad (barely above baseline). Haider *et al.*[@Haider2020] claimed that if a model is D-Calibrated then a 'patient should believe the prediction from the survival curve', these results clearly demonstrate otherwise. Measures of calibration alone are clearly not sufficient to determine if a survival curve prediction should be 'believed' and should therefore be computed alongside measures of discrimination or scoring rules, discussed next.

Model	KM	CPH	SVM
C_H^1	0.5	0.52	0.45
L_{IGS}^2	0.18	0.18	0.52
H_α^3	0.99	1.00	15.42
$D_{\chi^2}^4$	2.23^*	7.03^*	1.02×10^{10}

Table 11.2.: Comparison of numerical calibration metrics. Same models and data as in Figure 11.2: models trained and tested on randomly simulated data from the `simsurv` (Brilleman 2019) package in **mlr3proba**.

1. Harrell's C (Section 11.5.1). 2. Integrated Graf Score (Section 11.7.3). 3. Houwelingen's α (Section 11.6.1). 4. D-Calibration statistic. A $'*'$ indicates the model is D-Calibrated according to a χ_9^2 test.

11.7. Evaluating Distributions by Scoring Rules

Scoring rules evaluate probabilistic predictions and (attempt to) measure the overall predictive ability of a model, i.e. both calibration and discrimination (Gneiting and Raftery 2007; Murphy 1973). Scoring rules have been gaining in popularity for the past couple of decades since probabilistic forecasts were recognised to be superior than deterministic predictions for capturing uncertainity in predictions (A. P. Dawid 1984; A. Philip Dawid 1986). Formalisation and development of scoring rules has primarily been due to Dawid (A. P. Dawid 1984; A. Philip Dawid 1986; A. Philip Dawid and Musio 2014) and Gneiting and Raftery (Gneiting and Raftery 2007); though the earliest measures promoting ''rational" and ''honest" decision making date back to the 1950s (Brier 1950; Good 1952). Whilst several scoring rules have been proposed for classification problems, fewer exist for probabilistic regression predictions (Gneiting and Raftery 2007) and even fewer for survival analysis. In practice, only three continuous scoring rules for regression are employed (though the last two of these are often conflated), the integrated Brier score (Brier 1950), the log loss (Good

1952), and the integrated log loss.[8] In survival analysis only one scoring rule was found to be routinely employed. In fact, there is no recognised definition of a scoring rule in survival analysis, nor definitions for the fundamental scoring rule properties of (strict) properness. This section attempts to fill these gaps and to explore the proposed scoring rules for survival analysis.[9]

This survey of survival scoring rules covers:

i. basic definitions for scoring rules and properties;
ii. proposed scoring rules for survival analysis;
iii. proofs for (strict) properness; and
iv. baselines and standard errors for scoring rules.

Key contributions include demonstrating that no commonly-utilised survival scoring rule is proper and deriving a class of strictly proper outcome-independent scoring rules with strict assumptions (see Section 11.7.2 for definitions and Section 11.7.4 for proofs).

Each of these subsections is built up in complexity, starting with binary classification, then probabilistic regression, and finally survival. This is required to demonstrate how the survival setting makes use of the other two for scoring rules.

To recap the notation from Chapter 3, the three mathematical settings are defined by the generative processes:

- Regression: (X, Y) *t.v.i.* $\mathcal{X} \times \mathcal{Y}$ where $\mathcal{X} \subseteq \mathbb{R}^p$ and $\mathcal{Y} \subseteq \mathbb{R}$.
- Classification: (X, Y) *t.v.i.* $\mathcal{X} \times \mathcal{Y}$ where $\mathcal{X} \subseteq \mathbb{R}^p$ and $\mathcal{Y} = \{0, 1\}$.
- Survival: (X, T, Δ, Y, C) *t.v.i.* $\mathcal{X} \times \mathcal{T} \times \{0, 1\} \times \mathcal{T} \times \mathcal{T}$ where $X \subseteq \mathbb{R}^p$ and $\mathcal{T} \subseteq \mathbb{R}_{\geq 0}$, where C, Y are unobservable, $T := \min\{Y, C\}$, and $\Delta = \mathbb{I}(Y = T)$.

As the sections are clearly separated, the overloaded notation will be clear from context.

11.7.1. Classification and Regression Scoring Rules

Definitions and losses in the classification setting are first discussed and then the same in the regression setting.

11.7.1.1. Classification

All scoring rules were initially derived from the binary classification setting, in this case scoring rules are considered to have the form in Box 11.1.

[8] These often appear under many different names. The Brier score is often referred to as the 'squared-error loss', or 'quadratic score', and the log loss often appears as the 'log score', 'logarithmic loss', 'cross-entropy loss', or 'negative log-likelihood'.

[9] In this section a 'scoring rule' refers to the general class of measures that evaluate a probabilistic prediction and a 'loss' refers to the specific function to be minimised. As all scoring rules are optimally minimised in this survey, the terms are used interchangeably.

Binary classification loss

Box 11.1. *Let $\mathcal{P}$ be some family of distributions over $\mathcal{Y} = \{0, 1\}$ containing at least two elements. Then for a predicted distribution in $\mathcal{P}$, any real-valued function with the signature $L : \mathcal{P} \times \mathcal{Y} \to \bar{\mathbb{R}}$ will be considered as a* binary classification loss.

Any arbitrary function can be a binary classification loss as long as it satisfies the conditions in Box 11.1, for example $L(\zeta, y) = 0$ is a valid loss for all $\zeta \in \mathcal{P}$ and all $y \in \mathcal{Y}$. Therefore a scoring rule is generally only considered useful if it satisfies the properties below (Gneiting and Raftery 2007).

Definition 11.4 (Classification loss properness)**.** A classification loss $L : \mathcal{P} \times \mathcal{Y} \to \bar{\mathbb{R}}$ is called:

i. *Proper* if: for any distributions p_Y, p in $\mathcal{P}$ and for any random variables $Y \sim p_Y$, it holds that

$$\mathbb{E}[L(p_Y, Y)] \leq \mathbb{E}[L(p, Y)]$$

i. *Strictly proper* if in addition to being proper it holds, for the same quantification of variables, that

$$\mathbb{E}[L(p_Y, Y)] = \mathbb{E}[L(p, Y)] \Leftrightarrow p = p_Y$$

Proper scoring rules provide a method of model comparison as, by definition, predictions closest to the true distribution will result in lower expected losses.[10] On the other hand, if a scoring rule is not proper ('improper' (Gneiting and Raftery 2007)) then it has no meaningful comparison as it is unknown if the optimal model would have a lower or higher loss than any sub-optimal one. A strictly proper scoring rule has additional important uses such as in model optimisation, i.e. if a loss is strictly proper then minimisation of the loss will result in the 'optimum score estimator based on the scoring rule' (Gneiting and Raftery 2007). Whilst properness is usually a minimal acceptable property for a scoring rule, it is generally not sufficient on its own. For example, take the following classification loss,

$$L : \mathcal{P} \times \mathcal{Y} \to \bar{\mathbb{R}}; \quad (\zeta, y) \mapsto 42$$

This is proper as the loss, L, is always equal to 42 and therefore is minimised by the true distribution of Y but the loss is clearly useless. Properness and strict properness properties are utilised to determine if a scoring rule is performant and will be stated (if previously proved/disproved) or proved/disproved for all losses going forward.

[10] Further details for model comparison are not provided here as the topic is complex and with many open questions, see e.g. (Demšar 2006; Dietterich 1998; Nadeau and Bengio 2003).

Losses

The two most widely used scoring rules for classification are the Brier score (Brier 1950) and log loss (Good 1952).[11]

The (binary classification) log loss is defined by

$$L_{LL} : \mathcal{P} \times \mathcal{Y} \to \mathbb{R}_{\geq 0};$$
$$(\zeta, y) \mapsto -\mathbb{I}(y = 1) \log(\zeta.p(1)) - \mathbb{I}(y = 0) \log(\zeta.p(0))$$

or more simply

$$(\zeta, y) \mapsto -\log \zeta.p(y)$$

The (binary classification) Brier score is defined by

$$L_{BS} : \mathcal{P} \times \mathcal{Y} \to [0, 1]; \quad (\zeta, y) \mapsto (y - \zeta.p(y))^2$$

These are both strictly proper scoring rules (A. Philip Dawid and Musio 2014) and are visualised in Figure 11.4 to demonstrate their properties. The figure highlights the 'honesty' property of the scoring rules (i.e. their strict properness) as both losses are shown to be minimised when the true prediction is made. The plot also demonstrates baselines for interpretability (Section 11.7.5.1). For the Brier score and log loss, any result below 0.25 and 0.693 respectively indicates a prediction better than a constant uninformed prediction of $\zeta.p(1) = 0.5$. Therefore classification scoring rules provide a method to simultaneously encourage honest predictions and have in-built informative baselines for external reference.

11.7.1.2. Regression

The definition of a probabilistic regression scoring rule follows similarly to the classification setting after a re-specification of the target domain.

Probabilistic regression loss

Box 11.2. *Let $\mathcal{P}$ be some family of distributions over $\mathcal{Y} \subseteq \mathbb{R}$ containing at least two elements. Then for a predicted distribution in $\mathcal{P}$, any real-valued function with the signature $L : \mathcal{P} \times \mathcal{Y} \to \bar{\mathbb{R}}$ will be considered as a* probabilistic regression loss.

Definition 11.5 (Regression loss properness)**.** A probabilistic regression loss $L : \mathcal{P} \times \mathcal{Y} \to \bar{\mathbb{R}}$ is called:

i. *Proper* if: for any distributions p_Y, p in $\mathcal{P}$ and for any random variables $Y \sim p_Y$, it holds that

[11] Despite being called a 'score', the Brier score is in fact a loss to be minimised.

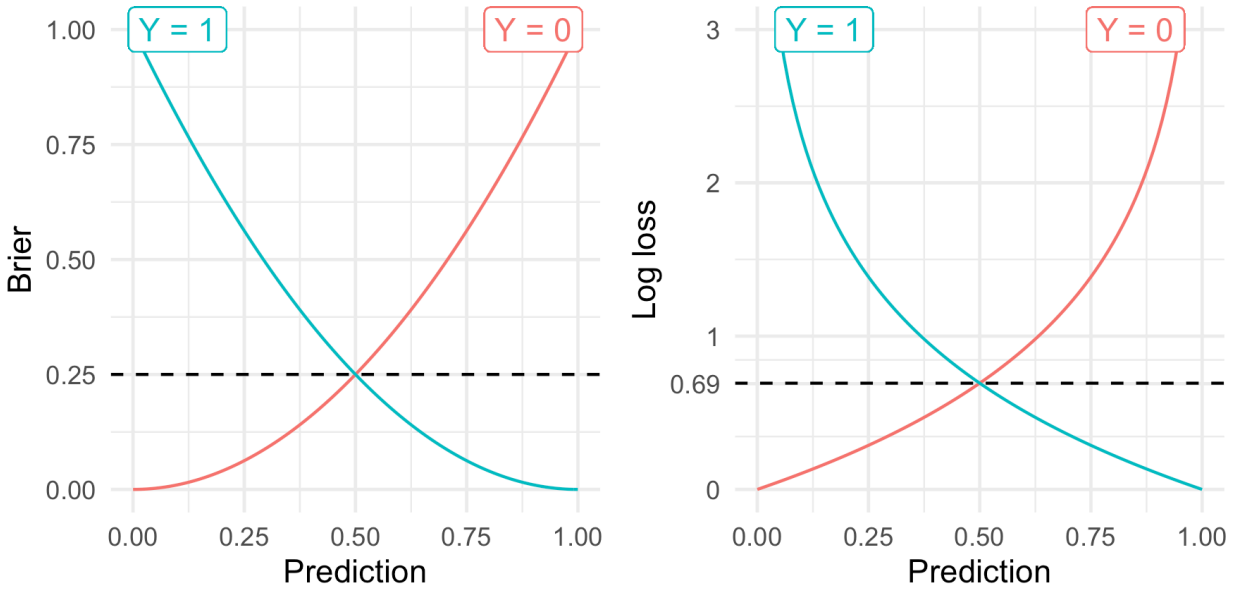


Figure 11.4.: Brier and log loss scoring rules for a binary outcome and varying probabilistic predictions. x-axis is a probabilistic prediction in $[0, 1]$, y-axis is Brier score (left) and log loss (right). Blue lines are varying Brier score/log loss over different predicted probabilities when the true outcome is 1. Red lines are varying Brier score/log loss over different predicted probabilities when the true outcome is 0. Both losses are minimised with the correct prediction, i.e. if $\zeta.p(1) = 1$ when $y = 1$ and $\zeta.p(1) = 0$ when $y = 0$ for a predicted discrete distribution ζ.

$$\mathbb{E}[L(p_Y, Y)] \leq \mathbb{E}[L(p, Y)]$$

i. *Strictly proper* if in addition to being proper it holds, for the same quantification of variables, that

$$\mathbb{E}[L(p_Y, Y)] = \mathbb{E}[L(p, Y)] \Leftrightarrow p = p_Y$$

Losses

In the regression setting, classification scoring rules are extended by instead considering distribution functions and integrating these over $\mathcal{Y} \subseteq \mathbb{R}$.

The Integrated Brier Score (IBS) is defined by,[12]

$$L_{IBS} : \mathcal{P} \times \mathcal{Y} \to [0, 1]; \quad (\zeta, y) \mapsto \int_{\mathcal{Y}} (\mathbb{I}(y \leq \tau) - \zeta.F(\tau))^2 \, d\tau \tag{11.4}$$

The extension from the classification Brier score is intuitive, instead of evaluating if the predicted pmf is 'correct' at a single point, the predicted cumulative distribution function is compared with the true event status over the entire distribution.

The log loss has two adaptations for continuous predictions. The first is analogous to the IBS and is termed the Integrated Log Loss (ILL)

$$\begin{aligned} &L_{ILL} : \mathcal{P} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}; \\ &(\zeta, y) \mapsto -\int_{\mathcal{Y}} \mathbb{I}(y \leq \tau) \log[\zeta.F(\tau)] + \mathbb{I}(y > \tau) \log[\zeta.S(\tau)] \, d\tau \end{aligned}$$

This follows the 'longer' form of the binary classification log loss and considers the cumulative probability of events over all time-points. A second adaptation to the log loss instead considers the 'simpler' form and replaces the probability mass function with the probability density function. Again this measure is intuitive as a perfect distributional prediction will assign the highest point of density to the point at which the event occurs. This variant of the log loss does not have a specific name but it is termed here the 'density log loss', L_{DLL}, and is formally defined by,

$$L_{DLL} : \mathcal{P} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}; \quad (\zeta, y) \mapsto -\log[\zeta.f(y)] \tag{11.5}$$

where $\mathcal{P}$ is a family of absolutely continuous distributions over $\mathcal{Y}$ with defined density functions.

All three of these losses are strictly proper (Gneiting and Raftery 2007; Gressmann et al. 2018).

[12] also known as the Continuous Ranked Probability Score (CRPS).

11.7.2. Survival Scoring Rule Definitions

Losses in the survival setting compare predicted survival distributions to the observed outcome tuple (time and censoring). A large class of survival losses additionally incorporate an estimator of the unknown censoring distribution, in order to attempt meaningful comparison. This second group of losses are termed here as 'approximate' losses as the true censoring distribution is never known and hence an estimate of the loss is approximate at best.

Survival loss

Box 11.3. *Let $\mathcal{T} \subseteq \mathbb{R}_{\geq 0}$ and let $\mathcal{C}, \mathcal{P}$ be any two distinct families of distributions over $\mathcal{T}$, containing at least two elements. Then,*

- *Any real-valued function with the signature $L : \mathcal{P} \times \mathcal{T} \times \{0, 1\} \to \bar{\mathbb{R}}$ will be considered as a* survival loss.
- *Any real-valued function with the signature $L : \mathcal{P} \times \mathcal{T} \times \{0, 1\} \times \mathcal{C} \to \bar{\mathbb{R}}$ will be considered as an* approximate survival loss.

Two separate novel definitions for (strict) properness are provided: the first captures the general case in which no assumptions are made about the censoring distribution; the second assumes that censoring is conditionally event-independent.

Definition 11.6 (Survival loss properness)**.** A survival loss $L : \mathcal{P} \times \mathcal{T} \times \{0, 1\} \to \bar{\mathbb{R}}$ is called:

i. *Proper* if: for any distributions p_Y, p in $\mathcal{P}$; and for any random variables $Y \sim p_Y$, and C t.v.i. $\mathcal{T}$; with $T := \min\{Y, C\}$ and $\Delta := \mathbb{I}(T = Y)$; it holds that,

$$\mathbb{E}[L(p_Y, T, \Delta)] \leq \mathbb{E}[L(p, T, \Delta)]$$

i. *Strictly proper* if in addition to being proper it holds, for the same quantification of variables, that

$$\mathbb{E}[L(p_Y, T, \Delta)] = \mathbb{E}[L(p, T, \Delta)] \Leftrightarrow p = p_Y$$

i. *Outcome-independent proper* if: for any distributions p_Y, p in $\mathcal{P}$; and for any random variables $Y \sim p_Y$, and C t.v.i. $\mathcal{T}$, where $C \perp\!\!\!\perp Y$; with $T := \min\{Y, C\}$ and $\Delta := \mathbb{I}(T = Y)$; it holds that,

$$\mathbb{E}[L(p_Y, T, \Delta)] \leq \mathbb{E}[L(p, T, \Delta)]$$

i. *Outcome-independent strictly proper* if in addition to being outcome-independent proper it holds, for the same quantification of variables, that

$$\mathbb{E}[L(p_Y, T, \Delta)] = \mathbb{E}[L(p, T, \Delta)] \Leftrightarrow p = p_Y$$

These final two definitions are 'weaker' but provide a term for losses that are improper in general but are (strictly) proper under common (though possibly strict) assumptions about the censoring distribution. Note by definition that if a loss is:

i. (strictly) proper then it is also outcome-independent (strictly) proper;
ii. (outcome-independent) strictly proper then it is also (outcome-independent) proper

Analogous definitions are now provided for approximate survival losses.

Definition 11.7 (Survival approximate loss properness)**.** An approximate survival loss $L : \mathcal{P} \times \mathcal{T} \times \{0, 1\} \times \mathcal{C} \to \bar{\mathbb{R}}$ is called:

i. *Proper* if: for any distributions p_Y, p in $\mathcal{P}$ and $c \in \mathcal{C}$; and for any random variables $Y \sim p_Y$ and $C \sim c$; with $T := \min\{Y, C\}$ and $\Delta := \mathbb{I}(T = Y)$; it holds that,

$$\mathbb{E}[L(p_Y, T, \Delta|c)] \leq \mathbb{E}[L(p, T, \Delta|c)]$$

i. *Strictly proper* if in addition to being proper it holds, for the same quantification of variables, that

$$\mathbb{E}[L(p_Y, T, \Delta|c)] = \mathbb{E}[L(p, T, \Delta|c)] \Leftrightarrow p = p_Y$$

i. *Outcome-independent proper* if: for any distributions p_Y, p in $\mathcal{P}$ and $c \in \mathcal{C}$; and for any random variables $Y \sim p_Y$ and $C \sim c$, where $C \perp\!\!\!\perp Y$; with $T := \min\{Y, C\}$ and $\Delta := \mathbb{I}(T = Y)$; it holds that,

$$\mathbb{E}[L(p_Y, T, \Delta|c)] \leq \mathbb{E}[L(p, T, \Delta|c)]$$

i. *Outcome-independent strictly proper* if in addition to being outcome-independent proper it holds, for the same quantification of variables, that

$$\mathbb{E}[L(p_Y, T, \Delta|c)] = \mathbb{E}[L(p, T, \Delta|c)] \Leftrightarrow p = p_Y$$

As the true censoring distribution, c, can never be known exactly, this definition allows for approximate losses to be proper in the asymptotic (with infinite training data) if they include estimators of c that are convergent in distribution. Proper approximate losses are therefore useful in modern predictive settings in which 'big data' is very common and thus estimators, such as the Kaplan-Meier, can converge to the true censoring distribution. However approximate losses may provide misleading results when the sample size is small; future research should ascertain what 'small' means for individual losses.

11.7.3. Common Survival Scoring Rules

The IBS, ILL, and DLL are now extended to the survival setting by suitably incorporating censoring and their properness properties are then discussed in Section 11.7.4. Measures are split into 'classes', which represent the basic form of the measure.

11.7.3.1. Squared Survival Losses

The analogue to the IBS for survival analysis is termed here as the Integrated Graf Score (IGS) as it was extensively discussed and promoted by Graf (Graf and Schumacher 1995; Graf et al. 1999).

Definition 11.8 (Integrated Graf score (IGS))**.**

$$\begin{aligned} &L_{IGS} : \mathcal{P} \times \mathcal{T} \times \{0,1\} \times \mathcal{C} \to [0,1]; \\ &(\zeta, t, \delta | \hat{G}_{KM}) \mapsto \int_0^{\tau^*} \frac{\zeta.S^2(\tau)\mathbb{I}(t \leq \tau, \delta = 1)}{\hat{G}_{KM}(t)} + \frac{\zeta.F^2(\tau)\mathbb{I}(t > \tau)}{\hat{G}_{KM}(\tau)} \, d\tau \end{aligned} \tag{11.6}$$

where $\zeta.S^2(\tau) = (\zeta.S(\tau))^2$, analogously for $\zeta.F^2$, and $\tau^* \in \mathcal{T}$ is an upper threshold to compute the loss up to.

The IGS consistently estimates the mean square error $L(t, S|\tau^*) = \int_0^{\tau^*} [\mathbb{I}(t > \tau) - S(\tau)]^2 d\tau$, where S is the correctly specified survival function, when censoring is uninformative only (Gerds and Schumacher 2006). This is intuitive as the IGS utilises the marginal Kaplan-Meier estimator to estimate the censoring distribution. Therefore CIPCW estimates such as the Cox model or Akritas estimator could instead be considered for $\hat{G}_{KM}$ and these have been demonstrated to have less bias when censoring is informative (Gerds and Schumacher 2006). However this raises concerns as now separate models have to be trained and predicted, which could need validation themselves, and therefore the final measure is even more difficult to interpret. Graf claimed that the IGS is strictly proper (Graf et al. 1999) however as no definition of properness was provided this claim cannot be validated. With the definition of properness provided in this thesis (Definition 11.7), the IGS is not even proper (Section 11.7.4.4).

One could instead consider extending the IBS by weighting by $\hat{G}_{KM}(t)$ only, giving the following loss.

Definition 11.9 (Reweighted Integrated Graf score (IGS*))**.** Let $\mathcal{P}$ be a family of absolutely continuous distributions over $\mathcal{T}$ with defined density functions. Then the *reweighted Integrated Graf score* (IGS*) is defined by

$$\begin{aligned} &L_{IGS^*} : \mathcal{P} \times \mathcal{T} \times \{0,1\} \times \mathcal{C} \to \mathbb{R}_{\geq 0}; \\ &(\zeta, t, \delta | \hat{G}_{KM}) \mapsto \frac{\delta \int_{\mathcal{T}} (\mathbb{I}(t \leq \tau) - \zeta.F(\tau))^2 \, d\tau}{\hat{G}_{KM}(t)} \end{aligned} \tag{11.7}$$

IGS* is outcome-independent strictly proper (Section 11.7.4.3).

11.7.3.2. Log Survival Losses

The ILL is similarly extended to the Integrated Survival Log Loss (ISLL) (Graf et al. 1999).

Definition 11.10 (Integrated survival log loss (ISLL))**.** The *integrated survival log loss* (ISLL) is defined by

$$\begin{aligned} &L_{ISLL} : \mathcal{P} \times \mathcal{T} \times \{0,1\} \times \mathcal{C} \to \mathbb{R}_{\geq 0}; \\ &(\zeta, t, \delta | \hat{G}_{KM}) \mapsto -\int_0^{\tau^*} \frac{\log[\zeta.F(\tau)]\mathbb{I}(t \leq \tau, \delta = 1)}{\hat{G}_{KM}(t)} + \frac{\log[\zeta.S(\tau)]\mathbb{I}(t > \tau)}{\hat{G}_{KM}(\tau)} \, d\tau \end{aligned} \tag{11.8}$$

where $\tau^* \in \mathcal{T}$ is an upper threshold to compute the loss up to.

The ISLL is not a proper approximate survival loss (Section 11.7.4.4). Again one could instead a different weighting in the denominator of the measure to give the following loss.

Definition 11.11 (Reweighted integrated survival log loss (ISLL*))**.** Let $\mathcal{P}$ be a family of absolutely continuous distributions over $\mathcal{T}$ with defined density functions. Then the *reweighted integrated survival log loss* (ISLL*) is defined by

$$\begin{aligned} &L_{ISLL^*} : \mathcal{P} \times \mathcal{T} \times \{0,1\} \times \mathcal{C} \to \mathbb{R}_{\geq 0}; \\ &(\zeta, t, \delta | \hat{G}_{KM}) \mapsto -\frac{\delta \int_{\mathcal{T}} \mathbb{I}(t \leq \tau) \log[\zeta.F(\tau)] + \mathbb{I}(t > \tau) \log[\zeta.S(\tau)] \, d\tau}{\hat{G}_{KM}(t)} \end{aligned} \tag{11.9}$$

ISLL* is an outcome-independent strictly proper scoring rule (Section 11.7.4.3).

The DLL can be extended in one of two ways, the first simply removes all censored observations.

Definition 11.12 (Survival density log loss (SDLL))**.** Let $\mathcal{P}$ be a family of absolutely continuous distributions over $\mathcal{T}$ with defined density functions. Then the *survival density log loss* (SDLL) is defined by

$$L_{SDLL} : \mathcal{P} \times \mathcal{T} \times \{0,1\} \to \mathbb{R}_{\geq 0}; \quad (\zeta, t, \delta) \mapsto -\delta \log[\zeta.f(t)] \tag{11.10}$$

The SDLL is not a proper scoring rule (Section 11.7.4.2). The second extension to DLL adds the same IPC weighting as IGS* and ISLL*.

Definition 11.13 (Weighted survival density log loss (SDLL*))**.** Let $\mathcal{P}$ be a family of absolutely continuous distributions over $\mathcal{T}$ with defined density functions. Then the *weighted survival density log loss* (SDLL*) is defined by

$$L_{SDLL^*} : \mathcal{P} \times \mathcal{T} \times \{0,1\} \times \mathcal{C} \to \mathbb{R}_{\geq 0}; \quad (\zeta, t, \delta | \widehat{G}_{KM}) \mapsto -\frac{\delta \log[\zeta.f(t)]}{\widehat{G}_{KM}(t)} \tag{11.11}$$

SDLL* is outcome-independent strictly proper (Section 11.7.4.3).

11.7.3.3. Absolute Survival Losses

Whilst the IGS and ISLL appear to be the most common losses in the literature, there is one other class to briefly mention that is based on absolute error functions. For example, the 'absolute Brier score' proposed by Schemper and Henderson (Schemper and Henderson 2000) which is based on the mean absolute error. This takes a similar approach to the IGS and weights the loss at different time-points according to whether an observation is censored. Studies of this loss have demonstrated that it depends heavily on correct model specification and is biased when this is not the case (Choodari-Oskooei, Royston, and Parmar 2012b; Schmid et al. 2011). To prevent this bias, Schmid *et al.*[@Schmid2011] proposed the following robust approximate loss, termed here the 'Schmid score',

$$L(\zeta, t, \delta | \widehat{G}_{KM}) = \int_0^{\tau^*} \frac{\zeta.S(\tau)\mathbb{I}(t \leq \tau, \delta = 1)}{\widehat{G}_{KM}(t)} + \frac{\zeta.F(\tau)\mathbb{I}(t > \tau)}{\widehat{G}_{KM}(\tau)} \, d\tau$$

where $\widehat{G}_{KM}$ and τ^* are as defined above. Analogously to the IGS, the Schmid score consistently estimates the mean absolute error when censoring is uninformative (Schmid et al. 2011). Both scores tend to yield similar results (Schmid et al. 2011).

11.7.3.4. Comparing Weighting Methods

The IGS and ISLL are well-established survival losses however no discussion about IGS* and ISLL* could be found in the literature. On the surface these measures may look very similar but there are two important differences, which are illustrated below with the ISLL and ISLL*, recall these are defined as:

$$L_{ISLL^*}(\zeta, t, \delta | \widehat{G}_{KM}) = -\int_0^{\tau^*} \frac{\log[\zeta.F(\tau)]\mathbb{I}(t \leq \tau, \delta = 1)}{\widehat{G}_{KM}(t)} + \frac{\log[\zeta.S(\tau)]\mathbb{I}(t > \tau, \delta = 1)}{\widehat{G}_{KM}(t)} \, d\tau$$

$$L_{ISLL}(\zeta, t, \delta | \widehat{G}_{KM}) = -\int_0^{\tau^*} \frac{\log[\zeta.F(\tau)]\mathbb{I}(t \leq \tau, \delta = 1)}{\widehat{G}_{KM}(t)} + \frac{\log[\zeta.S(\tau)]\mathbb{I}(t > \tau)}{\widehat{G}_{KM}(\tau)} \, d\tau$$

The primary differences are (RHS of equations):

i. Always removing censored observations from L_{ISLL^*} (even when alive) whereas L_{ISLL} includes all observations when alive.

ii. L_{ISLL^*} weights alive and dead observations by $\hat{G}_{KM}(t)$ whereas L_{ISLL} weights dead observations by $\hat{G}_{KM}(t)$ and alive observations by $\hat{G}_{KM}(\tau)$

Analytically the difference between these weighting results has major implications as L_{ISLL^*} (and L_{IGS^*}) is outcome-independent strictly proper (Section 11.7.4.3) whereas L_{ISLL} (and L_{IGS}) is not even proper (Section 11.7.4.4). However whilst it has been demonstrated that the IGS consistently estimates the mean squared error (Gerds and Schumacher 2006), no theory exists for IGS*. Similarly no study has been made on ISLL* and SDLL*.

11.7.3.5. PECs

As well as evaluating probabilistic outcomes with integrated scoring rules, non-integrated scoring rules can also be utilised for evaluating distributions at a single point. For example, instead of evaluating a probabilistic prediction with the IGS over $\mathbb{R}_{\geq 0}$, instead one could compute the IGS at a single time-point, $\tau \in \mathbb{R}_{\geq 0}$, only. Plotting these for varying values of τ results in 'prediction error curves' (PECs), which provide a simple visualisation for how predictions vary over the outcome. PECs are especially useful for survival predictions as they can visualise the prediction 'over time'. PECs should only be used as a graphical guide and never for model comparison as they only provide information at a limited number of points. An example is provided in **?@fig-eval-pecs** for the IGS; the CPH is consistently better performing than the SVM.

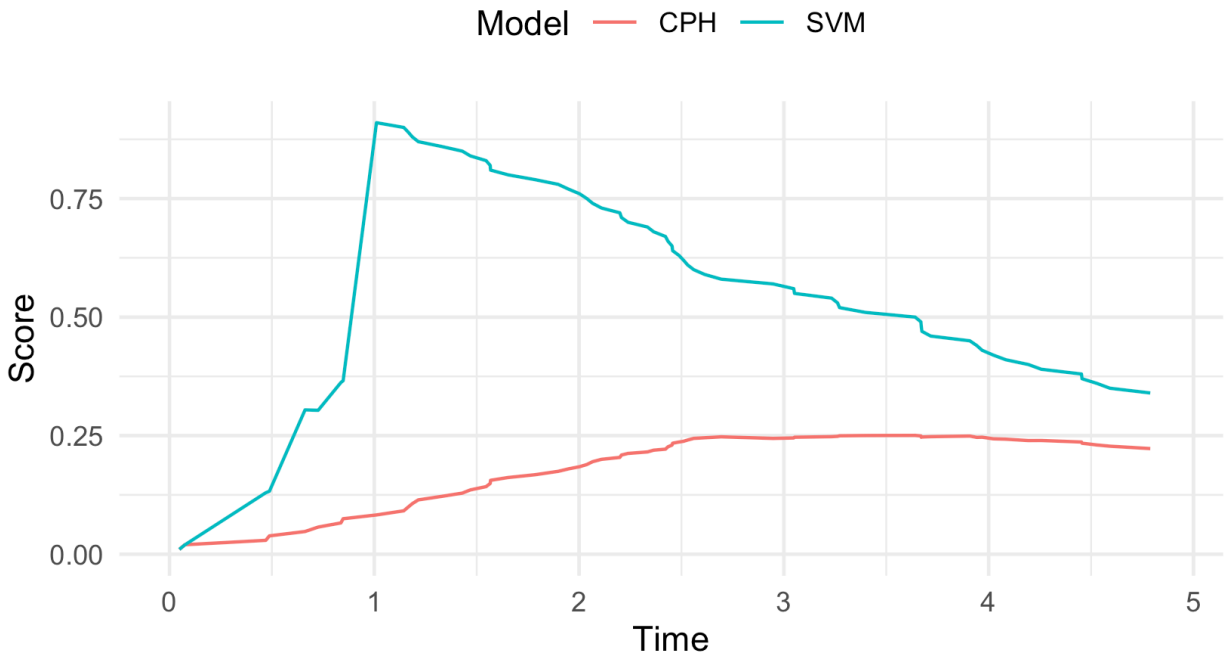


Figure 11.5.: Prediction error curves for the CPH and SVM models from Section 11.6. x-axis is time and y-axis is the IGS computed at different time-points. The CPH (red) performs better than the SVM (blue) as it scores consistently lower. Trained and tested on randomly simulated data from **mlr3proba**.

11.7.4. Properness of Survival Scoring Rules

As the IBS, ILL, and DLL are all strictly proper regression losses, one may assume the analogous survival losses are also strictly proper. No arguments could be found proving/disproving properness of the survival losses, which may be due to researchers assuming properness followed from the regression setting. Despite these estimators being demonstrated to have useful properties and to 'perform well' in simulation experiments (Choodari-Oskooei, Royston, and Parmar 2012a, 2012b; Gerds and Schumacher 2006), it transpires that none are proper. Key results in this section are collected in the following summary theorem.

Let $\mathcal{T} \subseteq \mathbb{R}_{>0}$ and let $\mathcal{C}, \mathcal{P}$ be two distinct families of distributions over $\mathcal{T}$ containing at least two elements and let $L_R : \mathcal{P} \times \mathcal{T} \to \bar{\mathbb{R}}$ be a regression scoring rule. Then the following statements are true:

i. L_{SDLL} is not: a) outcome-independent proper; b) outcome-independent strictly proper; c) proper; d) strictly proper ((**prop-sdll-proper?**)).
ii. Define the approximate survival loss,

$$L_S : \mathcal{P} \times \mathcal{T} \times \{0,1\} \times \mathcal{C} \to \bar{\mathbb{R}}; \quad (\zeta, t, \delta | \hat{G}_{KM}) \mapsto \frac{\delta L_R(\zeta, t)}{\hat{G}_{KM}(t)}$$

Then L_S is outcome-independent strictly proper if and only if L_R is strictly proper (Theorem 11.1). i. $L_{SDLL^*}, L_{IGS^*}, L_{ISLL^*}$ are all outcome-independent strictly proper ((**prop-approx-proper-losses?**)). i. L_{IGS} is not: a) outcome-independent proper; b) outcome-independent strictly proper; c) proper; d) strictly proper ((**prop-eval-igs?**)). i. L_{ISLL} is not: a) outcome-independent proper; b) outcome-independent strictly proper; c) proper; d) strictly proper ((**prop-eval-isll?**)).

The following conjectures are also made:

i. No survival loss, $L : \mathcal{P} \times \mathcal{T} \times \{0,1\} \to \bar{\mathbb{R}}$, is: a) outcome-independent strictly proper; b) strictly proper ((**conj-no-proper-loss?**)).
ii. No approximate survival loss, $L : \mathcal{P} \times \mathcal{T} \times \{0,1\} \times \mathcal{C} \to \bar{\mathbb{R}}$, is strictly proper ((**conj-approx-strictly?**)).

11.7.4.1. Definitions and Lemmas

Important proofs in this subsection follow after these definitions and lemmas.

Lemma 11.2. *Let $L : \mathcal{P} \times \mathcal{T} \times \{0,1\} \to \bar{\mathbb{R}}$ be a survival loss. Let $p_Y \in \mathcal{P}$, let $Y \sim p_Y$ and C t.v.i. $\mathcal{T}$ be random variables where $C \perp\!\!\!\perp Y$. Let $T := \min\{Y, C\}$ and $\Delta := \mathbb{I}(T = Y)$. Then if $\exists p \in \mathcal{P}, p \neq p_Y$, such that*

$$\mathbb{E}[L(p_Y, T, \Delta)] > \mathbb{E}[L(p, T, \Delta)]$$

Then, L is not:

i. outcome-independent proper;

ii. outcome-independent strictly proper;
iii. proper;
iv. strictly proper.

Lemma 11.3. *Let $L : \mathcal{P} \times \mathcal{T} \times \{0,1\} \times \mathcal{C} \to \bar{\mathbb{R}}$ be an approximate survival loss. Let $p_Y \in \mathcal{P}$ and let $c \in \mathcal{C}$. Let $Y \sim p_Y$ and $Ct.v.i.\mathcal{T}$ be random variables. Let $T := min\{Y, C\}$ and $\Delta := \mathbb{I}(T = Y)$. Then if $\exists p \in \mathcal{P}, p \neq p_Y$, such that*

$$\mathbb{E}[L(p_Y, T, \Delta|c)] > \mathbb{E}[L(p, T, \Delta|c)]$$

Then: L is not,

i. proper;
ii. strictly proper.

Definition 11.14 (Properness terminology)**.** Let $L : \mathcal{P} \times \mathcal{T} \times \{0,1\} \to \bar{\mathbb{R}}$ be a proper scoring rule and let p, p_Y be distributions in $\mathcal{P}$. Let $Y \sim p_Y$ and C t.v.i. $\mathcal{T}$ be random variables and let $T := \min\{Y, C\}$ and $\Delta := \mathbb{I}(T = Y)$. Then, (Gneiting and Raftery 2007)

i. $S_L(p_Y, p) := \mathbb{E}[L(p, T, \Delta)]$ is defined as the *expected penalty.*
ii. $H_L(p_Y) := S_L(p_Y, p_Y)$ is defined as the *(generalised) entropy of $p_Y \in \mathcal{P}$.*
iii. $D_L(p_Y, p) := S_L(p_Y, p) - H_L(p_Y)$ is defined as the *discrepancy* or *divergence* of $p \in \mathcal{P}$ from $p_Y \in \mathcal{P}$.

Similar definitions follow for the expected penalty, entropy, and divergence for an approximate survival loss $L : \mathcal{P} \times \mathcal{T} \times \{0,1\} \times \mathcal{C} \to \bar{\mathbb{R}}$.

Lemma 11.4. *Let $L : \mathcal{P} \times \mathcal{T} \times \{0,1\} \to \bar{\mathbb{R}}$ be a survival loss and let p_Y be a distribution in $\mathcal{P}$. Let $Y \sim p_Y$ and C t.v.i. $\mathcal{T}$ be random variables and let $T := \min\{Y, C\}$ and $\Delta := \mathbb{I}(T = Y)$. Then,*

- *$D_L(p_Y, p) \geq 0$ for all $p \in \mathcal{P}$ if L is proper*
- *$D_L(p_Y, p) > 0$ iff L is strictly proper and $p \neq p_Y$*

Definition 11.15 (Joint density)**.** Let X be an absolutely continuous random variable and let Y be a discrete random variable. Then,

i. The *mixed joint density* of (X, Y) is defined by

$$f_{X,Y}(x,y) = f_{X|Y}(x|y)P(Y=y)$$

where $f_{X|Y}(x|y)$ is the conditional probability density function of X given $Y = y$. i. The *mixed joint cumulative distribution function* of (X,Y) is given by

$$F_{X,Y}(x,y) = \sum_{z \le y} \int_{u=-\infty}^{x} f_{X,Y}(u,z)\ du$$

Lemma 11.5. *Let X, Y be jointly absolutely continuous random variables supported on the Reals with joint density function $f_{X,Y}(x,y)$ and let $Z = \mathbb{I}(X \le Y)$, then the mixed joint density of (X,Z) is given by*

$$f_{X,Z}(x,z) = \begin{cases} \int_x^\infty f_{X,Y}(x,y)\ dy, & z = 1 \\ \int_{-\infty}^x f_{X,Y}(x,y)\ dy, & z = 0 \end{cases}$$

Corollary 11.1. *Let X, Y be jointly absolutely continuous random variables supported on the Reals with joint density function $f_{X,Y}(x,y)$ and let $Z = \mathbb{I}(X \le Y)$. As a direct corollary to Lemma 11.5, if X and Y are independent then the mixed joint density of (X,Z) is given by*

$$f_{X,Z}(x,z) = \begin{cases} f_X(x)S_Y(x), & z = 1 \\ f_X(x)F_Y(x), & z = 0 \end{cases}$$

Lemma 11.6. *Let X, Y be jointly absolutely continuous random variables supported on the Reals with joint density function $f_{X,Y}(x,y)$ and let $Z = \mathbb{I}(X \le Y)$, then the mixed joint density of (Y,Z) is given by*

$$f_{Y,Z}(y,z) = \begin{cases} \int_{-\infty}^y f_{X,Y}(x,y)\ dx, & z = 1 \\ \int_y^\infty f_{X,Y}(x,y)\ dx, & z = 0 \end{cases}$$

In addition if $X \perp\!\!\!\perp Y$, then

$$f_{Y,Z}(y,z) = \begin{cases} f_Y(y)F_X(y), & z = 1 \\ f_Y(y)S_X(y), & z = 0 \end{cases}$$

11.7.4.2. No Strictly Proper Survival Loss

First it is proved that the survival density log loss is not outcome-independent proper and then a conjecture is made on the strict properness of all non-approximate losses.

The survival density log loss is not:

i. outcome-independent proper
ii. outcome-independent strictly proper
iii. proper
iv. strictly proper

Not only is the L_{SDLL} not outcome-independent proper but the counter-example in the proof is not even a rare edge case. Accounting for the censoring distribution is attempted by approximate losses, which are explored after the following conjecture.

Let $L : \mathcal{P} \times \mathcal{T} \times \{0, 1\} \to \bar{\mathbb{R}}$ be a survival loss, then L is not:

i. outcome-independent strictly proper;
ii. strictly proper;

This conjecture is motivated by identifying that as the true censoring distribution is always unknown, a counter-example can likely always be identified to contradict the loss being strictly proper.[13]

11.7.4.3. Strictly Proper Approximate Survival Losses

By making strict assumptions about the data, some survival scoring rules can still be useful, these assumptions are:

i. survival times and censoring times are independent;
ii. the training dataset is large enough to approximate the censoring distribution

With these assumptions, a large class of approximate losses can be outcome-independent strictly proper.

Theorem 11.1. *Let $L_R : \mathcal{P} \times \mathcal{T} \to \bar{\mathbb{R}}$ be a regression loss and define the approximate survival loss*

$$L_S : \mathcal{P} \times \mathcal{T} \times \{0, 1\} \times \mathcal{C} \to \bar{\mathbb{R}}; \quad (\zeta, t, \delta | \widehat{G}_{KM}) \mapsto \frac{\delta L_R(\zeta, t)}{\widehat{G}_{KM}(t)}$$

Then L_S is outcome-independent strictly proper if and only if L_R is strictly proper.

The following approximate survival losses are outcome-independent strictly proper:

i. L_{SDLL^*} – Equation 11.11
ii. L_{IGS^*} – Equation 11.7
iii. L_{ISLL^*} – Equation 11.9

[13]This conjecture is being explored as part of a theorem in a paper with external collaborators.

11.7.4.4. Non-Proper Approximate Survival Losses

From the previous proofs, it would be natural to assume that L_{IGS} and L_{ISLL} are also outcome-independent strictly proper, however this is not the case.

The integrated Graf score, L_{IGS}, is not:

i. outcome-independent proper
ii. outcome-independent strictly proper
iii. proper
iv. strictly proper

Whilst in this counter-example the value of $D_{IGS}(\xi, \zeta)$ is very close to zero, there will be other counter-examples with a more pronounced difference, though this is not required for the proof. Also note that again this is not a rare edge case, practically this example is reflected in any real-world scenario in which the prediction is close to the truth and when the censoring and survival times follow the same distribution.

Proposition 11.1. *The integrated survival log-loss, L_{ISLL}, is not:*

i. outcome-independent proper
ii. outcome-independent strictly proper
iii. proper
iv. strictly proper

Proof is not provided but follows with the same argumentation as the previous proposition and noting that a counter-example can always be found as C is unknown and cannot be removed from the equation.

Box 11.4. *Let $L : \mathcal{P} \times \mathcal{T} \times \{0, 1\} \times \mathcal{C} \to \bar{\mathbb{R}}$ be an approximate survival loss, then L is not strictly proper.*

This conjecture is motivated by noting that the joint distribution of (Y, C) is always unknown and thus a suitable counter-example to strict-properness can likely always be derived.[14]

11.7.5. Baselines and ERV

A common criticism of scoring rules is a lack of interpretability, e.g. without context an IGS of 0.5 or 0.0005 have no meaning. The final part of this section very briefly looks at two methods that help increase the interpretability of scoring rules. Scoring rules may already be considered less transparent than, say, concordance indices, as the underlying mathematics is more abstract, and therefore interpretability of the measure can play a large role in increasing transparency.

[14]This conjecture is being explored as part of a theorem in a paper with external collaborators.

11.7.5.1. Baselines

A baseline is either a model or a value that can be utilised to provide a reference value for a scoring rule, they provide a universal method to judge all models of the same class by (Gressmann et al. 2018).

In classification, an analytical baseline value can be derived for measures, i.e. a baseline model does not actually need to be fit to know what a 'good' value for the loss is. For example it is generally known that a Brier score is considered 'good' if it is below 0.25 or a log loss if it is below 0.693 (Section 11.7.1). Unfortunately simple analytical expressions are not possible in survival analysis as the losses are dependent on the distributions of both the survival and censoring time. Therefore all experiments in survival analysis must include a baseline model that can produce a reference value in order to derive meaningful results.

There is a clear consensus that the Kaplan-Meier estimator is the most sensible baseline model for survival modelling (Graf and Schumacher 1995; Lawless and Yuan 2010; Patrick Royston and Altman 2013) as it is the simplest model that can consistently estimate the true survival function. One could also consider the Akritas estimator as a tunable conditional baseline (Section 5.1.1).

Baseline models are often ignored in experiments when there is overconfidence in a particular model class, this is frequently the case in survival analysis in which a novel model class may only be compared to a Cox PH. This has practical and ethical implications. The calibration example in Section 11.6.1 demonstrates how one sophisticated model (CPH) may outperform another (SVM) and still perform worse than the Kaplan-Meier. Not including Kaplan-Meier in every experiment could lead to over-confidence in a novel model that is no better than an unconditional estimator (with no individual predictive ability).

11.7.5.2. Explained Residual Variation

Baseline models can also be utilised to derive a potentially more useful representation of scoring rules. Any scoring rule can be utilised to derive a measure of explained residual variation (ERV) (Edward L. Korn and Simon 1990; Edward L. Korn and Simon 1991) by standardising the loss with respect to a baseline, say Kaplan-Meier. For any survival loss L (analogously for an approximate survival loss), the ERV is,

$$\begin{aligned} &R_L : \mathcal{P} \times \mathcal{P} \times \mathbb{R}^m_{\geq 0} \times \{0,1\}^m \to [0,1]; \\ &(\zeta, \xi_0, t, \delta) \mapsto 1 - \frac{\frac{1}{m}\sum_{i=1}^m L(\zeta, t_i, \delta_i)}{\frac{1}{m}\sum_{i=1}^m L(\xi_0, t_i, \delta_i)} \end{aligned} \tag{11.12}$$

where $t = t_1, ..., t_m, \delta = \delta_1, ..., \delta_m$ and ζ should be a predicted distribution from a sophisticated (non-baseline) model and ξ_0 is a prediction from the Kaplan-Meier estimator.[15]

Representing a scoring rule in this manner improves interpretability by allowing for model comparison whilst simultaneously capturing the improvement from a baseline. Therefore instead of reporting some arbitrary loss value, say $L = 0.1$, one can instead report $R_L = 70$ which demonstrates a clear improvement (of 70%) over the baseline.

[15] Equation 11.12 assumes the numerator is always less than the denominator or more specifically that the sophisticated model is 'better' than the baseline; if this is not the case then $R^2_L < 0$. Therefore this representation should only be utilised when the model outperforms the baseline.

11.8. Conclusions

This chapter briefly reviewed different classes of survival measures before focusing on the application of scoring rules to survival analysis.

One finding of note from the review of survival measures is the possibility that research and debate has become too focused on measures of discrimination. For example, many papers state the flaws of Harrell's C index (**GonenHeller2005?**; Rahman et al. 2017; Schmid and Potapov 2012; Uno et al. 2007) however few acknowledge that simulation experiments have demonstrated that common alternatives yield very similar results to Harrell's C (Rahman et al. 2017; Therneau and Atkinson 2020) and moreover some promimnent alternatives, such as Uno's C (Uno et al. 2007), are actually harder to interpret due to very high variance (Rahman et al. 2017; Schmid and Potapov 2012). Whilst all concordance indices may be considered accessible and transparent, there is considerable doubt over their performance due to influence from censoring.

Focus on discrimination could be the reason for less development in survival time and calibration measures. There is evidence (P. Wang, Li, and Reddy 2019) of the censoring-adjusted RMSE, MAE, and MSE (Section 11.4) being used in evaluation but without any theoretical justification, which may lead to questionable results. Less development in calibration measures is likely due to these measures being more widely utilised for re-calibration of models and not in model comparison. The new D-Calibration measure (Andres et al. 2018; Haider et al. 2020) could prove useful for model comparison however independent simulation experiments and theoretical studies of the measure's properties would first be required. No calibration measures can be considered performant due to a lack of clear definition of a calibration measure for survival, moreover the reviewed measures may not even be transparent and accessible due to requiring expert interpretation.

The most problematic findings in this chapter lie in the survival scoring rules. Section 11.7.4 proved that no commonly used scoring rule is proper, which means that any results regarding model comparison based on these measures are thrown into question. It is also conjectured that no approximate survival loss can be strictly proper (in general), which is due to the joint distribution of the censoring and survival distribution always being unknown and impossible to estimate (though the marginal censoring distribution can be estimated). As demonstrated in Section 11.7.1, a proper scoring rule is not necessarily a useful one and therefore is not enough for robust model validation.

As an important caveat to the findings in this chapter, this thesis presents one particular definition of properness for survival scoring rules. This definition is partially subjective and other definitions could instead be considered. Therefore these losses should not be immediately dismissed outright. As well as deriving new losses that are (strictly) proper with respect to the definitions provided here, research may also be directed towards finding other sensible definitions of properness, or in confirming that the definition here is the only sensible option. As these are open research questions, the scoring rules discussed in this chapter are still utilised in evaluation for the benchmark experiment in **?@sec-bench**.

This chapter demonstrates that no survival measure on its own can capture enough information to fully evaluate a survival prediction. No measure is satisfactorily APT. This is a serious problem that will either lead (or already is leading) to less interest and uptake in survival modelling, or misunderstanding and deployment of sub-optimal models. Evaluation of survival models is still possible but currently requires expert interpretation to prevent misleading results. If the aim of a study is solely in assessing a model's discriminatory power, then measures of discrimination

alone are sufficient, otherwise a range of classes should be included to capture all aspects of model performance. This thesis advocates reporting *all* of the below to evaluate model performance:

- **Calibration**: Houwelingen's α and (Van Houwelingen 2007) *and* D-calibration (Haider et al. 2020).
- **Discrimination**: Harrell's (F. E. J. Harrell et al. 1984) *and* Uno's (Uno et al. 2011) C. By including two (or even more) measures of concordance, one can determine a feasible range for the 'true' discriminatory ability of the model instead of basing results on a single measure. Time-dependent AUCs can also be considered but these may require expert-interpretation and may only be advisable for discrimination-specific studies.
- **Scoring Rules**: When censoring is outcome-independent and a large enough training dataset is available, then the re-weighted integrated Graf score and re-weighted integrated survival log-loss (Section 11.7.3). Otherwise the IGS *and* ISLL (Graf et al. 1999) which should be interpreted together to ensure consistency in results.

If survival time prediction is the primary goal then RMSE_C and MAE_C can be included in the analysis however these should not form the primary conclusions due to a lack of theoretical justification. Instead, scoring rules should be utilised as a distributional prediction can always be composed into a survival time prediction (Chapter 12).

All measures discussed in this chapter, with the exception of the Blanche AUC, have been implemented in **mlr3proba** (R. Sonabend et al. 2021). The listed measures above are utilised in the benchmark experiment in **?@sec-bench**.

12. Pipelines - Composition and Reduction

TODO (150-200 WORDS)

In this chapter, composition and reduction are formally introduced, defined and demonstrated within survival analysis. Neither of these are novel concepts in general or in survival, with several applications already seen earlier when reviewing models (particularly in neural networks), however a lack of formalisation has led to much repeated work and at times questionable applications (Section 10.1). The primary purpose of this chapter is to formalise composition and reduction for survival and to unify references and strategies for future use. These strategies are introduced in the context of minimal 'workflows' and graphical 'pipelines' in order to maximise their generalisability. The pipelines discussed in this chapter are implemented in `mlr3proba`.

A *workflow* is a generic term given to a series of sequential operations. For example a standard ML workflow is fit/predict/evaluate, which means a model is fit, predictions are made, and these are evaluated. In this thesis, a *pipeline* is the name given to a concrete workflow. Section 12.1 demonstrates how pipelines are represented in this thesis.

Composition (Section 12.2) is a general process in which an object is built (or composed) from other objects and parameters. Reduction (Section 12.3) is a closely related concept that utilises composition in order to transform one problem into another. Concrete strategies for composition and reduction are detailed in sections Section 12.4 and Section 12.5.

Notation and Terminology

The notation introduced in Chapter 3 is recapped for use in this chapter: the generative survival template for the survival setting is given by (X, T, Δ, Y, C) *t.v.i.* $\mathcal{X} \times \mathcal{T} \times \{0, 1\} \times \mathcal{T} \times \mathcal{T}$ where $\mathcal{X} \subseteq \mathbb{R}^p$ and $\mathcal{T} \subseteq \mathbb{R}_{\geq 0}$, where C, Y are unobservable, $T := \min\{Y, C\}$, and $\Delta = \mathbb{I}(Y = T)$. Random survival data is given by $(X_i, T_i, \Delta_i, Y_i, C_i) \overset{i.i.d.}{\sim} (X, T, \Delta, Y, C)$. Usually data will instead be presented as a training dataset, $\mathcal{D}_0 = \{(X_1, T_1, \Delta_1), ..., (X_n, T_n, \Delta_n)\}$ where $(X_i, T_i, \Delta_i) \overset{i.i.d.}{\sim} (X, T, \Delta)$, and some test data $\mathcal{D}_1 = (X^*, T^*, \Delta^*) \sim (X, T, \Delta)$.

For regression models the generative template is given by (X, Y) t.v.i. $\mathcal{X} \subseteq \mathbb{R}^p$ and $Y \subseteq \mathbb{R}$. As with the survival setting, a regression training set is given by $\{(X_1, Y_1), ..., (X_n, Y_n)\}$ where $(X_i, Y_i) \overset{i.i.d.}{\sim} (X, Y)$ and some test data $(X^*, Y^*) \sim (X, Y)$.

12.1. Representing Pipelines

Before introducing concrete composition and reduction algorithms, this section briefly demonstrates how these pipelines will be represented in this thesis.

Pipelines are represented by graphs designed in the following way: all are drawn with operations progressing sequentially from left to right; graphs are comprised of nodes (or 'vertices') and arrows (or 'directed edges'); a rounded rectangular node represents a process such as a function or model fitting/predicting; a (regular) rectangular node represents objects such as data or hyper-parameters. Output from rounded nodes are sometimes explicitly drawn but when omitted the output from the node is the input to the next.

These features are demonstrated in **?@fig-car-example**. Say $y = 2$ and $a = 2$, then: data is provided ($y = 2$) and passed to the shift function ($f(x) = x + 2$), the output of this function ($y = 4$) is passed directly to the next ($h(x|a) = x^a$), this function requires a parameter which is also input ($a = 2$), finally the resulting output is returned ($y^* = 16$). Programmatically, $a = 2$ would be a hyper-parameter that is stored and passed to the required function when the function is called.

This pipeline is represented as a pseudo-algorithm in (**alg-car-ex?**), though of course is overly complicated and in practice one would just code $(y + 2)^\wedge a$.

Algorithm 6 Example pipeline.
Input Data, $y \in \mathbb{R}$. Parameter, $a \in \mathbb{R}$.
Output Transformed data, $x \in \mathbb{R}$.

$x \leftarrow y$
$x \leftarrow x + 2$
$x \leftarrow x^\wedge a$ **return** x

12.2. Introduction to Composition

This section introduces composition, defines a taxonomy for describing compositors (Section 12.2.1), and provides some motivating examples of composition in survival analysis (Section 12.2.2).

In the simplest definition, a model (be it mathematical, computational, machine learning, etc.) is called a *composite model* if it is built of two or more constituent parts. This can be simplest defined in terms of objects. Just as objects in the real-world can be combined in some way, so can mathematical objects. The exact 'combining' process (or 'compositor') depends on the specific composition, so too do the inputs and outputs. By example, a wooden table can be thought of as a composite object (Figure 12.1). The inputs are wood and nails, the combining process is hammering (assuming the wood is pre-chopped), and the output is a surface for eating. In mathematics, this process is mirrored. Take the example of a shifted linear regression model. This is defined by a linear regression model, $f(x) = \beta_0 + x\beta_1$, a shifting parameter, α, and a compositor $g(x|\alpha) = f(x) + \alpha$. Mathematically this example is overly trivial as this could be directly modelled with $f(x) = \alpha + \beta_0 + x\beta_1$, but algorithmically there is a difference. The composite model g, is defined by first fitting the linear regression model, f, and then applying a shift, α; as opposed to fitting a directly shifted model.

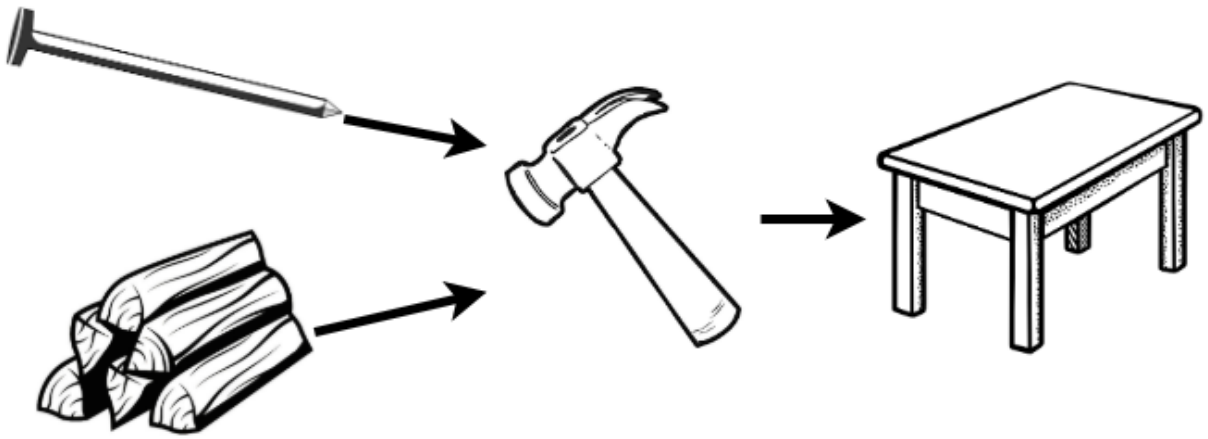


Figure 12.1.: Visualising composition in the real-world. A table is a composite object built from nails and wood, which are combined with a hammer 'compositor'. Figure not to scale.

Why Composition?

Tables tend to be better surfaces for eating your dinner than bundles of wood. Or in modelling terms, it is well-known that ensemble methods (e.g. random forests) will generally outperform their components (e.g. decision trees). All ensemble methods are composite models and this demonstrates one of the key use-cases of composition: improved predictive performance. The second key use-case is reduction, which is fully discussed in Section 12.3. Section 12.2.2 motivates composition in survival analysis by demonstrating how it is already prevalent but requires formalisation to make compositions more transparent and accessible.

Composite Model vs. Sub-models

A bundle of wood and nails is not a table and $1,000$ decision trees are not a random forest, both require a compositor. The compositor in a composite model combines the components into a single model. Considering a composite model as a single model enables the hyper-parameters of the compositor and the component model(s) to be efficiently tuned whilst being evaluated as a single model. This further allows the composite to be compared to other models, including its own components, which is required to justify complexity instead of parsimony in model building (Section 11.2).

12.2.1. Taxonomy of Compositors

Just as there are an infinite number of ways to make a table, composition can come in infinite forms. However there are relatively few categories that these can be grouped into. Two primary

taxonomies are identified here. The first is the 'composition type' and relates to the number of objects composed:

[i)] i. Single-Object Composition (SOC) – This form of composition either makes use of parameters or a transformation to alter a single object. The shifted linear regression model above is one example of this, another is given in Section 12.4.3. i. Multi-Object Composition (MOC) – In contrast, this form of composition combines multiple objects into a single one. Both examples in Section 12.2.2 are multi-object compositions.

The second grouping is the 'composition level' and determines at what 'level' the composition takes place:

[i)] i. Prediction Composition – This applies at the level of predictions; the component models could be forgotten at this point. Predictions may be combined from multiple models (MOC) or transformed from a single model (SOC). Both examples in Section 12.2.2 are prediction compositions. i. Task Composition – This occurs when one task (e.g. regression) is transformed to one or more others (e.g. classification), therefore always SOC. This is seen mainly in the context of reduction (Section 12.3). i. Model Composition – This is commonly seen in the context of wrappers (Section 12.5.6.4), in which one model is contained within another. i. Data Composition – This is transformation of training/testing data types, which occurs at the first stage of every pipeline.

12.2.2. Motivation for Composition

Two examples are provided below to demonstrate common uses of composition in survival analysis and to motivate the compositions introduced in Section 12.4.

Example 1: Cox Proportional Hazards

Common implementations of well-known models can themselves be viewed as composite models, the Cox PH is the most prominent example in survival analysis. Recall the model defined by

$$h(\tau|X_i) = h_0(\tau)\exp(\beta X_i)$$

where h_0 is the baseline hazard and β are the model coefficients.

This can be seen as a composite model as Cox defines the model in two stages (Cox 1972): first fitting the β-coefficients using the partial likelihood and then by suggesting an estimate for the baseline distribution. This first stage produces a linear predictor return type (Section 3.3) and the second stage returns a survival distribution prediction. Therefore the Cox model for linear predictions is a single (non-composite) model, however when used to make distribution predictions then it is a composite. Cox implicitly describes the model as a composite by writing ''alternative simpler procedures would be worth having" (Cox 1972), which implies a decision in fitting (a key feature of composition). This composition is formalised in Section 12.4.1 as a general pipeline (C1). The Cox model utilises the (C1) pipeline with a PH form and Kaplan-Meier baseline.

Example 2: Random Survival Forests

Fully discussed in Section 7.1, random survival forests are composed from many individual decision trees via a prediction composition algorithm ((**alg-rsf-pred?**)). In general, random forests perform better than their component decision trees, which tends to be true of all ensemble methods. Aggregation of predictions in survival analysis requires slightly more care than other fields due to the multiple prediction types, however this is still possible and is formalised in Section 12.4.4.

12.3. Introduction to Reduction

This section introduces reduction, motivates its use in survival analysis (Section 12.3.1), details an abstract reduction pipeline and defines the difference between a complete/incomplete reduction (Section 12.3.2), and outlines some common mistakes that have been observed in the literature when applying reduction (Section 12.3.3).

Reduction is a concept found across disciplines with varying definitions. This report uses the Langford definition: reduction is ''a complex problem decomposed into simpler subproblems so that a solution to the subproblems gives a solution to the complex problem" (Langford et al. 2016). Generalisation (or induction) is a common real-world use of reduction, for example sampling a subset of a population in order to estimate population-level results. The true answer (population-level values) may not always be found in this way but very good approximations can be made with simpler sub-problems (sub-sampling).

Reductions are workflows that utilise composition. By including hyper-parameters, even complex reduction strategies can remain relatively flexible. To illustrate reduction by example, recall the table-building example (Section 12.2) in which the task of interest is to acquire a table. The most direct but complex solution is to fell a tree and directly saw it into a table (Figure 12.2, top), clearly this is not a sensible process. Instead the problem can be reduced into simpler sub-problems: saw the tree into bundles of wood, acquire nails, and then use the 'hammer compositor' (Figure 12.1) to create a table (Figure 12.2, bottom).

In a modelling example, predicting a survival distribution with the Cox model can be viewed as a reduction in which two sub-problems are solved and composed:

i. predict continuous ranking;
ii. estimate baseline hazard; and
iii. compose with (C1) (Section 12.4.1).

This is visualised as a reduction strategy in **?@fig-car-cargraph**. The entire process from defining the original problem, to combining the simpler sub-solutions (in green), is the reduction (in red).

12.3.1. Reduction Motivation

Formalisation of reduction positively impacts upon accessibility, transparency, and predictive performance. Improvements to predictive performance have already been demonstrated when comparing random forests to decision trees. In addition, a reduction with multiple stages and many hyper-parameters allows for fine tuning for improved transparency and model performance (usual overfitting caveat applies, as does the trade-off described in Section 12.6).

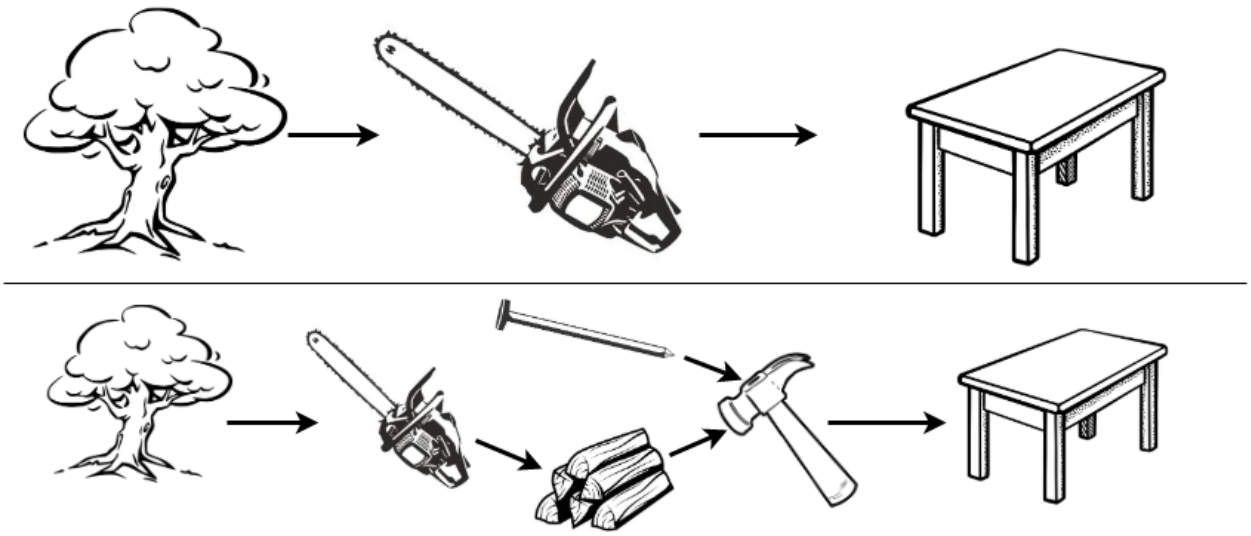


Figure 12.2.: Visualising reduction in the real-world. The complex process (top) of directly sawing a tree into a table is inefficient and unnecessarily complex. The reduction (bottom) that involves first creating bundles of wood is simpler, more efficient, and yields the same result, though technically requiring more steps.

The survey of ANNs (Section 10.1) demonstrated how reduction is currently utilised without transparency. Many of these ANNs are implicitly reductions to probabilistic classification (Section 12.5.6.6) however none include details about how the reduction is performed. Furthermore in implementation, none provide interface points to the reduction hyper-parameters. Formalisation encourages consistent terminology, methodology and transparent implementation, which can only improve model performance by exposing further hyper-parameters.

Accessibility is improved by formalising specific reduction workflows that previously demanded expert knowledge in deriving, building, and running these pipelines. All regression reductions in this chapter, are implemented in \ `mlr3proba` (R. Sonabend et al. 2021) and can be utilised with any possible survival model.

Finally there is an economic and efficiency advantage to reduction. A reduction model is relatively 'cheap' to explore as they utilise pre-established models and components to solve a new problem. Therefore if a certain degree of predictive ability can be demonstrated from reduction models, it may not be worth the expense of pursuing more novel ideas and hence reduction can help direct future research.

12.3.2. Task, Loss, and Data Reduction

Reduction can be categorised into task, loss, and data reduction, often these must be used in conjunction with each other. The direction of the reductions may be one- or two-way; this is visualised in **?@fig-car-reduxdiag**. This diagram should not be viewed as a strict fit/predict/evaluation

workflow but instead as a guidance for which tasks, T, data, D, models, M, and losses, L, are required for each other. The subscript O refers to the original object 'level' before reduction, whereas the subscript R is in reference to the reduced object.

The individual task, model, and data compositions in the diagram are listed below, the reduction from survival to classification (Section 12.5.6) is utilised as a running example to help exposition.

- $T_O \to T_R$: By definition of a machine learning reduction, task reduction will always be one way. A more complex task, T_O, is reduced to a simpler one, T_R, for solving. T_R could also be multiple simpler tasks. For example, solving a survival task, T_O, by classification, T_R (Section 12.5.6).
- $T_R \to M_R$: All machine learning tasks have models that are designed to solve them. For example logistic regression, M_R, for classification tasks, T_R.
- $M_R \to M_O$: The simpler models, M_R, are used for the express purpose to solve the original task, T_O, via solving the simpler ones. To solve T_O, a compositor must be applied, which may transform one (SOC) or multiple models (MOC) at a model- or prediction-level, thus creating M_O. For example predicting survival probabilities with logistic regression, M_R, at times $1, ..., \tau^*$ for some $\tau^* \in \mathbb{N}_{>0}$ (Section 12.5.6.4).
- $M_O \to T_O$: The original task should be solvable by the composite model. For example predicting a discrete survival distribution by concatenating probabilistic predictions at the times $1, ..., \tau^*$ (Section 12.5.6.6).
- $D_O \to D_R$: Just as the tasks and models are reduced, the data required to fit these must likewise be reduced. Similarly to task reduction, data reduction can usually only take place in one direction, to see why this is the case take an example of data reduction by summaries. If presented with 10 data-points $\{1, 1, 1, 5, 7, 3, 5, 4, 3, 3\}$ then these could be reduced to a single point by calculating the sample mean, 3.3. Clearly given only the number 3.3 there is no strategy to recover the original data. There are very few (if any) data reduction strategies that allow recovery of the original data. Continuing the running example, survival data, D_O, can be binned (Section 12.5.6.1) to classification data, D_R.

There is no arrow between D_O and M_O as the composite model is never fit directly, only via composition from $M_R \to M_O$. However, the original data, D_O, is required when evaluating the composite model against the respective loss, L_O.[1] Reduction should be directly comparable to non-reduction models, hence this diagram does not include loss reduction and instead insists that all models are compared against the same loss L_O.

A reduction is said to be *complete* if there is a full pipeline from $T_O \to M_O$ and the original task is solved, otherwise it is *incomplete*. The simplest complete reduction is comprised of the pipeline $T_O \to T_R \to M_R \to M_O$. Usually this is not sufficient on its own as the reduced models are fit on the reduced data, $D_R \to M_R$.

A complete reduction can be specified by detailing:

i. the original task and the sub-task(s) to be solved, $T_O \to T_R$;
ii. the original dataset and the transformation to the reduced one, $D_O \to D_R$ (if required); and
iii. the composition from the simpler model to the complex one, $M_R \to M_O$.

[1] A complete diagram would indicate that D_O is split into training data, which is subsequently reduced, and test data, which is passed to L_O. All reductions in this section can be applied to any data splitting process.

12.3.3. Common Mistakes in Implementation of Reduction

In surveying models and measures, several common mistakes in the implementation of reduction and composition were found to be particularly prevalent and problematic throughout the literature. It is assumed that these are indeed mistakes (not deliberate) and result from a lack of prior formalisation. These mistakes were even identified 20 years ago (Schwarzer, Vach, and Schumacher 2000) but are provided in more detail in order to highlight their current prevalence and why they cannot be ignored.

RM1. Incomplete reduction. This occurs when a reduction workflow is presented as if it solves the original task but fails to do so and only the reduction strategy is solved. A common example is claiming to solve the survival task by using binary classification, e.g. erroneously claiming that a model predicts survival probabilities (which implies distribution) when it actually predicts a five year probability of death ((**box-task-classif?**)). This is a mistake as it misleads readers into believing that the model solves a survival task ((**box-task-surv?**)) when it does not. This is usually a semantic not mathematical error and results from misuse of terminology. It is important to be clear about model predict types (Section 3.3) and general terms such as 'survival predictions' should be avoided unless they refer to one of the three prediction tasks. RM2. Inappropriate comparisons. This is a direct consequence of (RM1) and the two are often seen together. (RM2) occurs when an incomplete reduction is directly compared to a survival model (or complete reduction model) using a measure appropriate for the reduction. This may lead to a reduction model appearing erroneously superior. For example, comparing a logistic regression to an RSF (Section 7.1) for predicting survival probabilities at a single time using the accuracy measure is an unfair comparison as the RSF is optimised for distribution predictions. This would be non-problematic if a suitable composition is clearly utilised. For example a regression SSVM predicting survival time cannot be directly compared to a Cox PH. However the SSVM can be compared to a CPH composed with the probabilistic to deterministic compositor (C3), then conclusions can be drawn about comparison to the composite survival time Cox model (and not simply a Cox PH). RM3. Na"ive censoring deletion. This common mistake occurs when trying to reduce survival to regression or classification by simply deleting all censored observations, even if censoring is informative. This is a mistake as it creates bias in the dataset, which can be substantial if the proportion of censoring is high and informative. More robust deletion methods are described in Section 12.4.5. RM4. Oversampling uncensored observations. This is often seen when trying to reduce survival to regression or classification, and often alongside (RM3). Oversampling is the process of replicating observations to artificially inflate the sample size of the data. Whilst this process does not create any new information, it can help a model detect important features in the data. However, by only oversampling uncensored observations, this creates a source of bias in the data and ignores the potentially informative information provided by the proportion of censoring.

12.4. Composition Strategies for Survival Analysis

Though composition is common practice in survival analysis, with the Cox model being a prominent example, a lack of formalisation means a lack of consensus in simple operations. For example, it is often asked in survival analysis how a model predicting a survival distribution can be used to return a survival time prediction. A common strategy is to define the survival time prediction as the median of the predicted survival curve however there is no clear reason why this should be more sensible than returning the distribution mean, mode, or some random quantile. Formalisation allow

these choices to be analytically compared both theoretically and practically as hyper-parameters in a workflow. Four prediction compositions are discussed in this section ((**tab-car-taxredcar?**)), three are utilised to convert prediction types between one another, the fourth is for aggregating multiple predictions. One data composition is discussed for converting survival to regression data. Each is first graphically represented and then the components are discussed in detail. As with losses in the previous chapter, compositions are discussed at an individual observation level but extend trivially to multiple observations.

ID[1]	Composition	Type[2]	Level[3]
C1)	Linear predictor to distribution	MOC	Prediction
C2)	Survival time to distribution	MOC	Prediction
C3)	Distribution to survival time	SOC	Prediction
C4)	Survival model averaging	MOC	Prediction
C5)	Survival to regression	SOC	Data

Table 12.1.: Compositions formalised in Section 12.4. {tbl-car-taxredcar}

1. ID for reference throughout this thesis. 2. Composition type. Multi-object composition (MOC) or single-object composition (SOC). 3. Composition level.

12.4.1. C1) Linear Predictor → Distribution

This is a prediction-level MOC that composes a survival distribution from a predicted linear predictor and estimated baseline survival distribution. The composition (**?@fig-car-comp-distr**) requires:

- $\hat{\eta}$: Predicted linear predictor. $\hat{\eta}$ can be tuned by including this composition multiple times in a benchmark experiment with different models predicting $\hat{\eta}$. In theory any continuous ranking could be utilised instead of a linear predictor though results may be less sensible (Section 12.6).
- $\hat{S}_0$: Estimated baseline survival function. This is usually estimated by the Kaplan-Meier estimator fit on training data, $\hat{S}_{KM}$. However any model that can predict a survival distribution can estimate the baseline distribution (caveat: see Section 12.6) by taking a uniform mixture of the predicted individual distributions: say $\xi_1, ..., \xi_m$ are m predicted distributions, then $\hat{S}_0(\tau) = \frac{1}{m}\sum_{i=1}^{m} \xi_i.S(\tau)$. The mixture is required as the baseline must be the same for all observations. Alternatively, parametric distributions can be assumed for the baseline, e.g. $\xi = \text{Exp}(2)$ and $\xi.S(t) = \exp(-2t)$. As with $\hat{\eta}$, this parameter is also tunable.
- M: Chosen model form, which theoretically can be any non-increasing right-continuous function but is usually one of:
- Proportional Hazards (PH): $S_{PH}(\tau|\eta, S_0) = S_0(\tau)^{\exp(\eta)}$
- Accelerated Failure Time (AFT): $S_{AFT}(\tau|\eta, S_0) = S_0(\frac{\tau}{\exp(\eta)})$
- Proportional Odds (PO): $S_{PO}(\tau|\eta, S_0) = \frac{S_0(\tau)}{\exp(-\eta)+(1-\exp(-\eta))S_0(\tau)}$

Models that predict linear predictors will make assumptions about the model form and therefore dictate sensible choices of M, for example the Cox model assumes a PH form. This does not mean other choices of M cannot be specified but that interpretation may be more difficult (Section 12.6). The model form can be treated as a hyper-parameter to tune. * C: Compositor returning the composed distribution, $\zeta := C(M, \hat{\eta}, \hat{S}_0)$ where ζ has survival function $\zeta.S(\tau) = M(\tau|\hat{\eta}, \hat{S}_0)$.

Pseudo-code for training ((**alg-car-comp-distr-fit?**)) and predicting ((**alg-car-comp-distr-pred?**)) this composition as a model 'wrapper' with sensible parameter choices (Section 12.6) is provided in appendix (**app-car?**).

12.4.2. C2) Survival Time → Distribution

This is a prediction-level MOC that composes a distribution from a predicted survival time and assumed location-scale distribution. The composition (**?@fig-car-comp-response**) requires:

- $\hat{T}$: A predicted survival time. As with the previous composition, this is tunable. In theory any continuous ranking could replace $\hat{T}$, though the resulting distribution may not be sensible (Section 12.6).
- ξ: A specified location-scale distribution, $\xi(\mu, \sigma)$, e.g. Normal distribution.
- $\hat{\sigma}$: Estimated scale parameter for the distribution. This can be treated as a hyper-parameter or predicted by another model.
- C: Compositor returning the composed distribution $\zeta := C(\xi, \hat{T}, \hat{\sigma}) = \xi(\hat{T}, \hat{\sigma})$.

Pseudo-code for training ((**alg-car-comp-response-fit?**)) and predicting ((**alg-car-comp-response-pred?**)) this composition as a model 'wrapper' with sensible parameter choices (Section 12.6) is provided in appendix (**app-car?**).

12.4.3. C3) Distribution → Survival Time Composition

This is a prediction-level SOC that composes a survival time from a predicted distribution. Any paper that evaluates a distribution on concordance is implicitly using this composition in some manner. Not acknowledging the composition leads to unfair model comparison (Section 12.3.3). The composition (**?@fig-car-comp-crank**) requires:

- ζ: A predicted survival distribution, which again is 'tunable'.
- ϕ: A distribution summary method. Common examples include the mean, median and mode. Other alternatives include distribution quantiles, $\zeta.F^{-1}(\alpha), \backslash\alpha \in [0, 1]$; α could be tuned as a hyper-parameter.
- C: Compositor returning composed survival time predictions, $\hat{T} := C(\phi, \zeta) = \phi(\zeta)$.

Pseudo-code for training ((**alg-car-comp-crank-fit?**)) and predicting ((**alg-car-comp-crank-pred?**)) this composition as a model 'wrapper' with sensible parameter choices (Section 12.6) is provided in appendix (**app-car?**).

12.4.4. C4) Survival Model Averaging

Ensembling is likely the most common composition in machine learning. In survival it is complicated slightly as multiple prediction types means one of two possible compositions is utilised to average predictions. The (**?@fig-car-comp-avg**) composition requires:

- $\rho = \rho_1, ..., \rho_B$: B predictions (not necessarily from the same model) of the same type: ranking, survival time or distribution; again 'tunable'.
- $w = w_1, ..., w_B$: Weights that sum to one.
- C: Compositor returning combined predictions, $\hat{\rho} := C(\rho, w)$ where $C(\rho, w) = \frac{1}{B}\sum_{i=1}^{B} w_i\rho_i$, if ρ are ranking of survival time predictions; or $C(\rho, w) = \zeta$ where ζ is the distribution defined by the survival function $\zeta.S(\tau) = \frac{1}{B}\sum_{i=1}^{B} w_i\rho_i.S(\tau)$, if ρ are distribution predictions.

Pseudo-code for training ((**alg-car-comp-avg-fit?**)) and predicting ((**alg-car-comp-avg-pred?**)) this composition as a model 'wrapper' with sensible parameter choices (Section 12.6) is provided in appendix (**app-car?**).

12.4.5. C5) Survival to Regression Data

This is a data-level SOC that transforms survival data to regression data by either removing censored observations or 'imputing' survival times. This composition is frequently incorrectly utilised (Section 12.3.3) and therefore more detail is provided here than previous compositions. Note that the previous compositions were prediction-level transformations that occur after a survival model makes a prediction, whereas this composition is on a data-level and can take place before model training or predicting.

In Statistics, there are only two methods for removing 'missing' values: deletion and imputation; both of these have been attempted for censoring.

Censoring can be beneficial, harmful, or neutral; each will affect the data differently if deleted or imputed. Harmful censoring occurs if the reason for censoring is negative, for example drop-out due to disease progression. Harmful censoring indicates that the true survival time is likely soon after the censoring time. Beneficial censoring occurs if censoring is positive, for example drop-out due to recovery. This indicates that the true survival time is likely far from the censoring time. Finally neutral censoring occurs when no information can be gained about the true survival time from the censoring time. Whilst the first two of these can be considered to be dependent on the outcome, neutral censoring is often the case when censoring is independent of the outcome conditional on the data, which is a standard assumption for the majority of survival models and measures.

12.4.5.1. Deletion #{sec-car-pipelines-survreg-del}

Deletion is the process of removing observations from a dataset. This is usually seen in 'complete case analysis' in which observations with 'missingness', covariates with missing values, are removed from the dataset. In survival analysis this method is somewhat riskier as the subjects to delete depend on the outcome and not the features. Three methods are considered, the first two are a more brute-force approach whereas the third allows for some flexibility and tuning.

Complete Deletion

Deleting all censored observations is simple to implement with no computational overhead. Complete deletion results in a smaller regression dataset, which may be significantly smaller if the proportion of censoring is high. If censoring is uninformative, the dataset is suitably large and the proportion of censoring suitably low, then this method can be applied without further consideration. However if censoring is informative then deletion will add bias to the dataset, although the 'direction' of bias cannot be known in advance. If censoring is harmful then censored observations will likely have a similar profile to those that died, thus removing censoring will artificially inflate the proportion of those who survive. Conversely if censoring is beneficial then censored observations may be more similar to those who survive, thus removal will artificially inflate the proportion of those who die.

Omission

Omission is the process of omitting the censoring indicator from the dataset, thus resulting in a regression dataset that assumes all observations experienced the event. Complete deletion results in a smaller dataset of dead patients, omission results in no sample size reduction but the outcome may be incorrect. This reduction strategy is likely only justified for harmful censoring. In this case the true survival time is likely close to the censoring time and therefore treating censored observations as dead may be a fair assumption.

IPCW

If censoring is conditionally-outcome independent then deletion of censored events is possible by using Inverse Probability of Censoring Weights (IPCW). This method has been seen several times throughout this thesis in the context of models and measures. It has been formalised as a composition technique by Vock *et al.* (2016) (Vock et al. 2016) although their method is limited to binary classification. Their method weights the survival time of uncensored observations by $w_i = 1/\hat{G}_{KM}(T_i)$ and deletes censored observations, where $\hat{G}_{KM}$ is the Kaplan-Meier estimate of the censoring distribution fit on training data. As previously discussed, one could instead consider the Akritas (or any other) estimator for $\hat{G}_{KM}$.

Whilst this method does provide a 'safer' way to delete censored observations, there is not a necessity to do so. Instead consider the following weights

$$w_i = \frac{\Delta_i + \alpha(1 - \Delta_i)}{\hat{G}_{KM}(T_i)} \tag{12.1}$$

where $\alpha \in [0, 1]$ is a hyper-parameter to tune. Setting $\alpha = 1$ equally weights censored and uncensored observations and setting $\alpha = 0$ recovers the setting in which censored observations are deleted. It is assumed $\hat{G}_{KM}$ is set to some very small ϵ when $\hat{G}_{KM}(T_i) = 0$. When $\alpha \neq 0$ this becomes an imputation method, other imputation methods are now discussed.

12.4.5.2. Imputation

Imputation methods estimate the values of missing data conditional on non-missing data and other covariates. Whilst the true value of the missing data can never be known, by carefully conditioning on the 'correct' covariates, good estimates for the missing value can be obtained to help prevent a loss of data. Imputing outcome data is more difficult than imputing covariate data as models are then trained on 'fake' data. However a poor imputation should still be clear when evaluating a model as testing data remains un-imputed. By imputing censoring times with estimated survival times, the censoring indicator can be removed and the dataset becomes a regression dataset.

Gamma Imputation

Gamma imputation (D. Jackson et al. 2014) incorporates information about whether censoring is harmful, beneficial, or neutral. The method imputes survival times by generating times from a shifted proportional hazards model

$$h(\tau) = h_0(\tau)\exp(\eta + \gamma)$$

where η is the usual linear predictor and $\gamma \in \mathbb{R}$ is a hyper-parameter determining the 'type' of censoring such that $\gamma > 0$ indicates harmful censoring, $\gamma < 0$ indicates beneficial censoring, and $\gamma = 0$ is neutral censoring. This imputation method has the benefit of being tunable as γ is a hyper-parameter and there is a choice of variables to condition the imputation. No independent experiments exist studying how well this method performs, nor discussing the theoretical properties of the method.

MRL

The Mean Residual Lifetime (MRL) estimator has been previously discussed in the context of SVMs (Section 8.1.2). Here the estimator is extended to serve as an imputation method. Recall the MRL function, $MRL(\tau|\hat{S}) = \int_\tau^\infty \hat{S}(u)\ du/\hat{S}(\tau)$, where $\hat{S}$ is an estimate of the survival function of the underlying survival distribution (e.g. $\hat{S}_{KM}$). The MRL is interpreted as the expected remaining survival time after the time-point τ. This serves as a natural imputation strategy where given the survival outcome (T_i, Δ_i), the new imputed time T_i' is given by

$$T_i' = T_i + (1 - \Delta_i)MRL(T_i|\hat{S})$$

where $\hat{S}$ would be fit on the training data and could be an unconditional estimator, such as Kaplan-Meier, or conditional, such as Akritas. The resulting survival times are interpreted as the true times for those who died and the expected survival times for those who were censored.

Buckley-James

Buckley-James (Buckley and James 1979) is another imputation method discussed earlier (Section 9.1). The Buckley-James method uses an iterative procedure to impute censored survival times by the conditional expectation given censoring times and covariates (Z. Wang and Wang 2010). Given the survival tuple for an outcome (T_i, Δ_i), the new imputed time T_i' is

$$T_i' = \begin{cases} T_i, & \Delta_i = 1 \\ X_i\hat{\beta} + \frac{1}{\hat{S}_{KM}(e_i)} \sum_{e_i < e_k} \hat{p}_{KM}(e_k) e_k & \Delta_i = 0 \end{cases}$$

where $\hat{S}_{KM}$ is the Kaplan-Meier estimator of the survival distribution estimated on training data and with associated pmf $\hat{p}_{KM}$ and $e_i = T_i - X_i\hat{\beta}$ where $\hat{\beta}$ are estimated coefficients of a linear regression model fit on (X_i, T_i). Given the least squares approach, more parametric assumptions are made than other imputation methods and it is more complex to separate model fitting from imputation. Hence, this imputation may only be appropriate on a limited number of data types.

Alternative Methods

Other methods have been proposed for 'imputing' censored survival times though with either less clear discussion or to no benefit. Multiple imputation by chained equations (MICE) has been demonstrated to perform well with covariate data and even outcome data (in a non-survival setting). However no adaptations have been developed to incorporate censoring times into the imputation and therefore is less informative than Gamma imputation.

Re-calibration of censored survival times (Vinzamuri, Li, and Reddy 2017) uses an iterative update procedure to 're-calibrate' censoring times however the motivation behind the method is not sufficiently clear to be of interest in general survival modelling tasks outside of the authors' specific pipelines.

Finally parametric imputation is defined by making random draws from truncated probability distributions and adding these to the censoring time (P. Royston 2001; Patrick Royston, Parmar, and Altman 2008). Whilst this method is arguably the simplest method and will lead to a sufficiently random sample, i.e. not one skewed by the imputation process, in practice the randomness leads to unrealistic results, with some imputed times being very far from the original censoring times and some being very close.

12.4.5.3. The Decision to Impute or Delete

Deletion methods are simple to implement and fast to compute however they can lead to biasing the data or a significant sample reduction if used incorrectly. Imputation methods can incorporate tuning and have more relaxed assumptions about the censoring mechanism, though they may lead to over-confidence in the resulting outcome and therefore add bias into the dataset. In some cases, the decision to impute or delete is straightforward, for example if censoring is uninformative and only few observations are censored then complete deletion is appropriate. If it is unknown if censoring is informative then this can crudely be estimated by a benchmark experiment. Classification models can be fit on $\{(X_1, \Delta_1), ..., (X_n, \Delta_n)\}$ where $(X_i, \Delta_i) \in \mathcal{D}_0$. Whilst not an exact test, if any model significantly outperforms a baseline, then this may indicate censoring is informative. This

is demonstrated in (**tab-car-predcens?**), in which a logistic regression outperforms a featureless baseline in correctly predicting if an observation is censored when censoring is informative, but is no better than the baseline when censoring is uninformative.

Data	Baseline	Logistic Regression
`Sim1`	0.20 (0.14, 0.26)	0.02 (0.01, 0.03)
`Sim7`	0.19 (0.14, 0.24)	0.16 (0.13, 0.19)

Table 12.2.: Estimating censoring dependence by prediction. `Sim1` is informative censoring and `Sim7` is uninformative. Logistic regression is compared to a featureless baseline with the Brier score with standard errors. Censoring can be significantly predicted to 95% confidence when informative (`Sim1`) but not when uninformative (`Sim7`).

12.5. Novel Survival Reductions

This section collects the various strategies and settings discussed previously into complete reduction workflows. (**tab-car-reduxes?**) lists the reductions discussed in this section with IDs for future reference. All strategies are described by visualising a graphical pipeline and then listing the composition steps required in fitting and predicting.

This section only includes novel reduction strategies and does not provide a survey of pre-existing strategies. This limitation is primarily due to time (and page) constraints as every method has very distinct workflows that require complex exposition. Well-established strategies are briefly mentioned below and future research is planned to survey and compare all strategies with respect to empirical performance (i.e. in benchmark experiments).

Two prominent reductions are 'landmarking' (Van Houwelingen 2007) and piecewise exponential models (M. Friedman 1982). Both are reductions for time-varying covariates and hence outside the scope of this thesis. Relevant to this thesis scope is a large class of strategies that utilise 'discrete time survival analysis' (Tutz and Schmid 2016); these strategies include reductions (R7) and (R8). Methodology for discrete time survival analysis has been seen in the literature for the past three decades (Liestol, Andersen, and Andersen 1994). The primary reduction strategy for discrete time survival analysis is implemented in the R package `discSurv` (Welchowski and Schmid 2019); this is very similar to (R7) except that it enforces stricter constraints in the composition procedures and forces a 'discrete-hazard' instead of 'discrete-survival' representation (Section 12.5.6.2).

ID	Original Survival Task	Reduced Task
R1)	Probabilistic	Probabilistic Regression
R2)	Probabilistic	Deterministic Regression
R3)	Deterministic	Deterministic Regression
R4)	Deterministic	Probabilistic Distribution
R5)	Probabilistic	Deterministic Regression
R6)	Ranking	Deterministic Regression
R7)	Probabilistic	Probabilistic Classification
R8)	Deterministic	Probabilistic Classification

ID	Original Survival Task	Reduced Task

Table 12.3.: Survival reductions in Section 12.5. First column is a unique identifier for the strategy, second column is the original suvival task of interest, third column is the reduced task that will be solved as a surrogate in the workflow. {#tbl:car-reduxes}

12.5.1. R1) Probabilistic Survival → Probabilistic Regression

This is perhaps the most natural reduction strategy as the survival task can be thought of as probabilistic regression with censoring. Steps and compositions of the reduction (**?@fig-car-R1**):

Fit F1) A survival dataset, $\mathcal{D}_0$, is composed with (C5) to a regression dataset, $\mathcal{D}_R$. F2) A *probabilistic* regression model, g, with hyper-parameters, ϕ, is fit on the composed regression data. It is important to select a model that will only predict distributions supported over $\mathbb{R}_{\geq 0}$ in order to reflect the survival setting. **Predict** P1) Testing survival data, $\mathcal{D}_1$, is passed to the trained regression model, $\hat{g}$, without further data composition, and distributions are predicted, $\zeta = \zeta_1, ..., \zeta_m$.

12.5.2. R2) Probabilistic Survival → Deterministic Regression

This is almost identical to the previous reduction but utilises deterministic regression models and composition to distribution predictions. Steps and compositions of the reduction (**?@fig-car-R2**):

Fit F1) A survival dataset, $\mathcal{D}_0$, is composed with (C5) to a regression dataset, $\mathcal{D}_R$. F2) A *deterministic* regression model, g, with hyper-parameters, ϕ, is fit on the composed regression data. It is important to select a model that will only predict positive values in order to reflect the survival setting. **Predict** P1) Testing survival data, $\mathcal{D}_1$, is passed to the trained regression model, $\hat{g}$, without further data composition, and survival times are predicted, $\hat{T} = \hat{T}_1, ..., \hat{T}_m$. P2) Survival times are composed with (C2) to distribution predictions $\zeta = \zeta_1, ..., \zeta_m$.

12.5.3. R3) Deterministic Survival → Deterministic Regression

This reduction is identical to (R2) except (P2) is omitted.

12.5.4. R4) Deterministic Survival → Probabilistic Regression

This is identical to (R1) with an additional composition to survival time. Steps and compositions of the reduction (**?@fig-car-R4**):

Fit F) Same as (R1). **Predict** P1) Testing survival data, $\mathcal{D}_1$, is passed to the trained regression model, $\hat{g}$, without further data composition, and distributions are predicted, $\zeta = \zeta_1, ..., \zeta_m$. P2) Distributions are composed with (C3) to survival times $\hat{T} = \hat{T}_1, ..., \hat{T}_m$.

12.5.5. R5) Probabilistic Survival → Deterministic Regression (II)

These next two reductions utilise deterministic regression to predict linear predictors. This first reduction additionally composes the linear predictor to a distribution prediction. Steps and compositions of the reduction (**?@fig-car-R5**):

Fit F1) A survival model, g_0, with a linear predictor prediction type is fit on a survival dataset, $\mathcal{D}_0$. F2) The model is inspected and the fitted linear predictors, η, are returned. F3) A deterministic regression model, g, is fit on (X_i, η_i) with η_i as the target. **Predict** P1) Testing survival data, $\mathcal{D}_1$, is passed to the trained regression model, $\hat{g}$, and linear predictors are predicted, $\hat{\eta} = \hat{\eta}_1, ..., \hat{\eta}_m$. P2) Linear predictors are composed with (C1) to survival distributions $\zeta = \zeta_1, ..., \zeta_m$. The most sensible choice of model form for the (C1) composition will be dictated by g_0, e.g. does it have an underlying PH form?

R6) Ranking Survival → Deterministic Regression {#sec-car-reduxes-r6}

This reduction is identical to (R5) except (P2) is omitted. Whilst this is categorised as solving a ranking task, the predicted quantities can be interpreted as linear predictors (given the model form specified by g_0).

12.5.6. R7-R8) Survival → Probabilistic Classification

Two separate reductions are presented in **?@fig-car-R7R8** however as both are reductions to probabilistic classification and are only different in the very last step, both are presented in this section. Steps and compositions of the reduction (**?@fig-car-R7R8**):

Fit F1) A survival dataset, $\mathcal{D}_0$, is binned, B, with a continuous to discrete data composition (Section 12.5.6.1). F2) A multi-label classification model, with adaptations for censoring, $g_L(D_B|\theta)$, is fit on the transformed dataset, D_B. Optionally, g_L could be further reduced to binary, g_B, or multi-class classification, g_c, (Section 12.5.6.4). **Predict** P1) Testing survival data, $\mathcal{D}_1$, is passed to the trained classification model, $\hat{g}$, to predict pseudo-survival probabilities $\tilde{S}$ (or optionally hazards (Section 12.5.6.2)). P2a) Predictions can be composed, T_1, into a survival distribution prediction, $\zeta = \zeta_1, ..., \zeta_m$ (Section 12.5.6.6); or, P2b) Predictions can be composed, T_2, to survival time predictions, $\hat{T} = \hat{T}_1, ..., \hat{T}_m$ (Section 12.5.6.7).

Further details for binning, multi-label classification, and transformation of pseudo-survival probabilities are now provided.

12.5.6.1. Composition: Binning Survival Times

An essential part of the reduction is the transformation from a survival dataset to a classification dataset, which requires two separate compositions. The first (discussed here) is to discretise the survival times ($B(\mathcal{D}_0|w)$ in **?@fig-car-R7R8**) and the second is to merge the survival time and censoring indicator into a single outcome (Section 12.5.6.2).

Discretising survival times is achieved by the common 'binning' composition, in which a continuous outcome is discretised into 'bins' according to specified thresholds. These thresholds are usually determined by specifying the width of the bins as a hyper-parameter w.[2] This is a common

[2]Binning is described here with equal widths but generalises to unequal widths trivially.

transformation and therefore further discussion is not provided here. An example is given below with the original survival data on the left and the binned data on the right ($w = 1$).

X	Time (Cont.)	Died
1	1.56	0
2	2	1
3	3.3	1
4	3.6	0
5	4	0

X	Time (Disc.)	Died
1	[1, 2)	0
2	[2, 3)	1
3	[3, 4)	1
4	[3, 4)	0
5	[4, 5)	0

12.5.6.2. Composition: Survival to Classification Outcome

The binned dataset still has the unique survival data format of utilising two outcomes for training (time and status) but only making a prediction for one outcome (distribution). In order for this to be compatible with classification, the two outcome variables are composed into a single variable.[3] This is achieved by casting the survival times into a 'wide' format and creating a new outcome indicator.[4] Two outcome transformations are possible, the first represents a discrete survival function and the second represents a discrete hazard function.[5]

Discrete Survival Function Composition

In this composition, the data in the transformed dataset represents the discrete survival function. The new indicator is defined as follows,

$$Y_{i;\tau} := \begin{cases} 1, & T_i > \tau \\ 0, & T_i \leq \tau \cap \Delta_i = 1 \\ -1, & T_i \leq \tau \cap \Delta_i = 0 \end{cases}$$

At a given discrete time τ, an observation, i, is either alive ($Y_{i;\tau} = 1$), dead ($Y_{i;\tau} = 0$), or censored ($Y_{i;\tau} = -1$). Therefore $\hat{P}(Y_{i;\tau} = 1) = \hat{S}_i(\tau)$, motivating this particular choice of representation.

[3] This is the first key divergence from other discrete-time classification strategies, which use the censoring indicator as the outcome and the time outcome as a feature.

[4] This is the second key divergence from other discrete-time classification strategies, which keep the data in a 'long' format.

[5] This is the final key divergence from other discrete-time classification strategies, which enforce the discrete hazard representation.

This composition is demonstrated below with the binned data (left) and the composed classification data (right).

X	Time (Disc.)	Died
1	[1, 2)	0
2	[2, 3)	1
3	[3, 4)	1
4	[3, 4)	0
5	[4, 5)	0

X	[1,2)	[2,3)	[3,4)	[4,5)
1	-1	-1	-1	-1
2	1	0	0	0
3	1	1	0	0
4	1	1	-1	-1
5	1	1	-1	-1

Discrete Hazard Function Composition

In this composition, the data in the transformed dataset represents the discrete hazard function. The new indicator is defined as follows,

$$Y^*_{i;\tau} := \begin{cases} 1, & T_i = \tau \cap \Delta_i = 1 \\ -1, & T_i = \tau \cap \Delta_i = 0 \\ 0, & \text{otherwise} \end{cases}$$

At a given discrete time τ, an observation, i, either experiences the event ($Y^*_{i;\tau} = 1$), experiences censoring ($Y_{i;\tau} = -1$), or neither ($Y_{i;\tau} = 0$). Utilising sequential multi-label classification problem transformation methods (Section 12.5.6.4) results in $\widehat{P}(Y^*_{i;\tau} = 1) = \widehat{h}_i(\tau)$. If methods are utilised that do not 'look back' at predictions then $\widehat{P}(Y^*_{i;\tau} = 1) = \widehat{p}_i(\tau)$ (Section 12.5.6.4).[6]

This composition is demonstrated below with the binned data (left) and the composed classification data (right).

X	Time (Disc.)	Died
1	[1, 2)	0
2	[2, 3)	1
3	[3, 4)	1
4	[3, 4)	0
5	[4, 5)	0

[6]This important distinction is not required in other discrete-time reduction strategies that automatically condition the prediction by including time as a feature.

X	[1,2)	[2,3)	[3,4)	[4,5)
1	-1	0	0	0
2	0	1	0	0
3	0	0	1	0
4	0	0	-1	0
5	0	0	0	-1

Multi-Label Classification Data

In both compositions, survival data t.v.i. $\mathbb{R}^p \times \mathbb{R}_{\geq 0} \times \{0, 1\}$ is transformed to multi-label classification data t.v.i. $\mathbb{R}^p \times \{-1, 0, 1\}^K$ for K binned time-intervals. The multi-label classification task is defined in Section 12.5.6.4 with possible algorithms.

The discrete survival representation has a slightly more natural interpretation and is 'easier' for classifiers to use for training as there are more positive events (i.e. more observations alive) to train on, whereas the discrete hazard representation will have relatively few events in each time-point. However the hazard representation leads to more natural predictions (Section 12.5.6.6).

A particular bias that may easily result from the composition of survival to classification data is now discussed.

12.5.6.3. Reduction to Classification Bias

The reduction to classification bias is commonly known (Zhou et al. 2005) but is reiterated briefly here as it must be accounted for in any automated reduction to classification workflow. This bias occurs when making classification predictions about survival at a given time and incorrectly censoring patients who have not been observed long enough, instead of removing them.

By example, say the prediction of interest is five-year survival probabilities after a particular diagnosis, clearly a patient who has only been diagnosed for three years cannot inform this prediction. The bias is introduced if this patient is censored at five-years instead of being removed from the dataset. The result of this bias is to artificially inflate the probability of survival at each time-point as an unknown outcome is treated as censored and therefore alive.

This bias is simply dealt with by removing patients who have not been alive 'long enough'.[7] Paradoxically, even if a patient is observed to die before the time-point of interest, they should still be removed if they have not been in the dataset 'long enough' as failing to do so will result in a bias in the opposite direction, thus over-inflating the proportion of dead observations.

Accounting for this bias is particularly important in the multi-label reduction as the number of observable patients will decrease over time due to censoring.

[7] Accounting for this bias is only possible if the study start and end dates are known, as well as the date the patient entered the study.

12.5.6.4. Multi-Label Classification Algorithms

As the work in this section is completely out of the thesis scope, the full text is in appendix (**app-mlc?**). The most important contributions from this section are:

- Reviewing problem transformation methods (Tsoumakas and Katakis 2007) for multi-label classification;
- Identifying that only binary relevance, nested stacking, and classifier chains are appropriate in this reduction; and
- Generalising these methods into a single wrapper for any binary classifier, the 'LWrapper'.

12.5.6.5. Censoring in Classification

Classification algorithms cannot natively handle the censoring that is included in the survival reduction, but this can be incorporated using one of two approaches.

Multi-Class Classification

All multi-label datasets can also handle multi-class data, hence the simplest way in which to handle censoring is to make multi-class predictions in each label for the outcome Y_τ $t.v.i.\{-1, 0, 1\}$. Many off-shelf classification learners can make multi-class predictions natively and simple reductions exist for those that cannot. As a disadvantage to this method, classifiers would then predict if an individual is dead or alive or censored (each mutually exclusive), and not simply alive or dead. Though this could be perceived as an advantage when censoring is informative as this will accurately reflect a real-world competing-risks set-up.

Subsetting/Hurdle Models

For this approach, the multi-class task is reduced to two binary class tasks: first predict if a subject is censored or not (dead or alive) and only if the prediction for censoring is below some threshold, $\alpha \in [0, 1]$, then predict if the subject is alive or not (dead or censored). If the probability of censoring is high in the first task then the probability of being alive is automatically set to zero in the final prediction, otherwise the prediction from the second task is used. Any classifier can utilise this approach and it has a meaningful interpretation, additionally α is a tunable hyper-parameter. The main disadvantage is increases to storage and run-time requirements as double the number of models may be fit.

Once the datasets have been composed to classification datasets and censoring is suitably incorporated by either approach, then any probabilistic classification model can be fit on the data. Predictions from these models can either be composed to a distribution prediction (R7) or a survival time prediction (R8).

12.5.6.6. R7) Probabilistic Survival → Probabilistic Classification

This final part of the (R7) reduction is described separately for discrete hazard and survival representations of the data (Section 12.5.6.2).

Discrete Hazard Representation

In this representation recall that predictions of the positive class, $P(Y_\tau = 1)$, are estimating the quantity $h(\tau)$. These predictions provide a natural and efficient transformation from predicted hazards to survival probabilities. Let $\hat{h}_i$ be a predicted hazard function for some observation i, then the survival function for that observation can be found with a Kaplan-Meier type estimator,

$$\tilde{S}_i(\tau^*) = \prod_\tau 1 - \hat{h}_i(\tau)$$

Now predictions are for a pseudo-survival function, which is 'pseudo' as it is not right-continuous. Resolving this is discussed below.

Discrete Survival Representation

In this representation, $P(Y_\tau = 1)$ is estimating $S(\tau)$, which means that predictions from a classification model result in discrete point predictions and not a right-continuous function. More importantly, there is no guarantee that a non-increasing function will be predicted, i.e. there is no guarantee that $P(Y_j = 1) < P(Y_i = 1)$, for time-points $j > i$.

Unfortunately there is no optimal way of dealing with predictions of this sort and 'mistakes' of this kind have been observed in some software implementation. One point to note is that in practice these are quite rare as the probability of survival will always decrease over time. Therefore the 'usual' approach is quite 'hacky' and involves imputing increasing predictions with the previous prediction, formally,

$$\tilde{S}(i+1) := \min\{P(Y_{i+1} = 1), P(Y_i = 1)\}, \forall i = \mathbb{R}_{\geq 0}$$

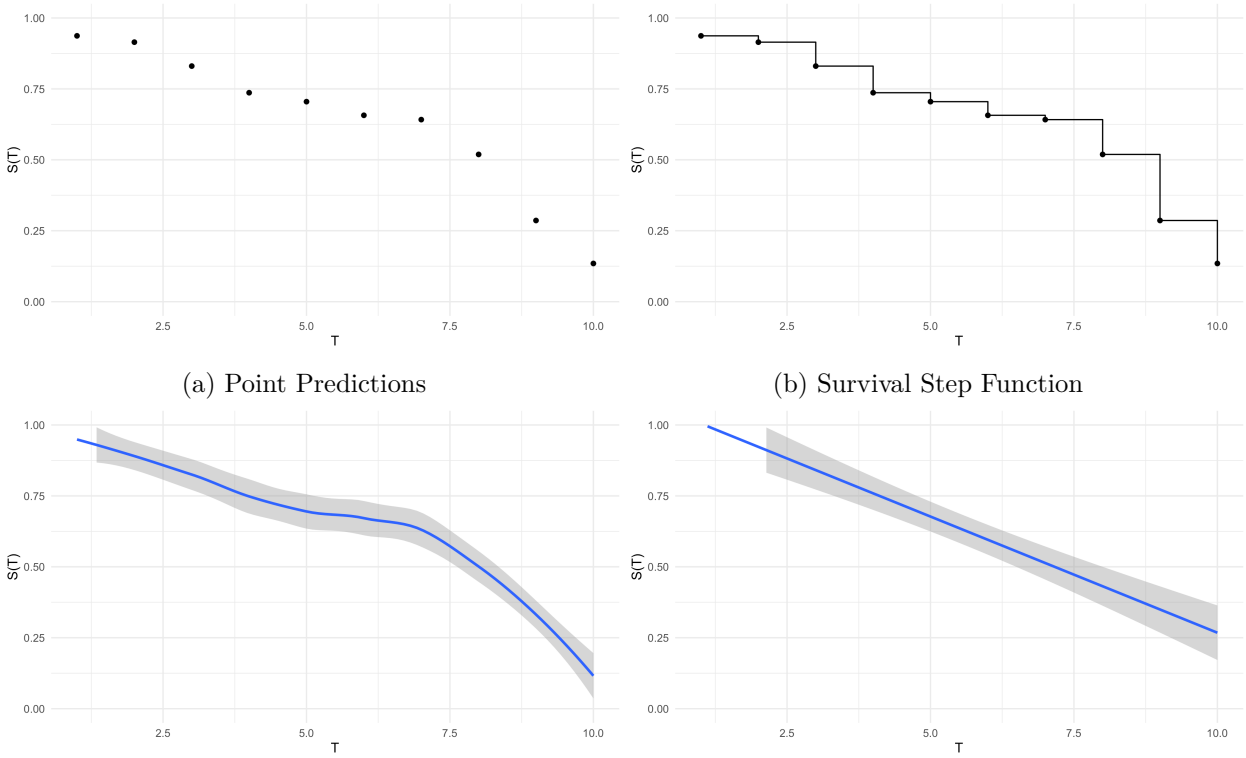
assuming $\tilde{S}(0) = 1$. Future research should seek more robust alternatives.

Right-Continuous Survival Function

From either representation, a \ non-increasing but non-continuous pseudo-survival function, $\tilde{S}$, is now predicted. Creating a right-continuous function ('$T_1(\tilde{S})$' in **?@fig-car-R7**) from these point predictions (Figure 12.3 (a)) is relatively simple and well-known with accessible off-shelf software. At the very least, one can assume a constant hazard rate between predictions and cast them into a step function (Figure 12.3 (b)). This is a fairly common assumption and is usually valid as bin-width decreases. Alternatively, the point predictions can be smoothed into a continuous function with off-shelf software, for example with polynomial local regression smoothing (Figure 12.3 (c)) or generalised linear smoothing (Figure 12.3 (d)). Whichever method is chosen, the survival function is now non-increasing right-continuous and the (R7) reduction is complete.

12.5.6.7. R8) Deterministic Survival → Probabilistic Classification

Predicting a deterministic survival time from the multi-label classification predictions is relatively straightforward and can be viewed as a discrete analogue to (C3) (Section 12.4.3). For the discrete hazard representation, one can simply take the predicted time-point for an individual to be time



(a) Point Predictions

(b) Survival Step Function

(c) Local polynomial regression smoothing

(d) Generalised linear smoothing

Figure 12.3.: Survival function as a: point prediction (a), step function assuming constant risk (b), local polynomial regression smoothing (c), and generalised linear smoothing (d). (c) and (d) computed with `ggplot2` (Wickham 2016).

at which the predicted hazard probability is highest however this could easily be problematic as there may be multiple time-points at which the predicted hazard equals 1. Instead it is cleaner to first cast the hazard to a pseudo-survival probability (Section 12.5.6.6) and then treat both representations the same.

Let $\tilde{S}_i$ be the predicted multi-label survival probabilities for an observation i s.t. $\tilde{S}_i(\tau)$ corresponds with $\hat{P}(Y_{i;\tau} = 1)$ for label $\tau \in \mathcal{K}$ where $Y_{i;\tau}$ is defined in Section 12.5.6.2 and $\mathcal{K} = \{1, ..., K\}$ is the set of labels for which to make predictions. Then the survival time transformation is defined by

$$T_2(\tilde{S}_i) = \inf\{\tau \in \mathcal{K} : \tilde{S}_i(\tau) \leq \beta\}$$

for some $\beta \in [0, 1]$.

This is interpreted as defining the predicted survival time as the first time-point in which the predicted probability of being alive drops below a certain threshold β. Usually $\beta = 0.5$, though this can be treated as a hyper-parameter for tuning. This composition can be utilised even if predictions are not non-increasing, as only the first time the predicted survival probability drops below the threshold is considered. With this composition the (R8) reduction is now complete.

12.6. Choices and Defaults

Before concluding the chapter, this brief section describes a common problem that occurs when programming pipelines and how this thesis (and implementation in `mlr3proba`) addresses this.

Many Choices

Implementation of any of these pipelines leads to an important trade-off between user-choice and sensible decisions. When programming any software, the more choice that is given to the user, the higher the potential to make less sensible decisions; in the extreme as the number of user possibilities tends to infinity, the probability of a user selecting a sensible decision will tend to zero. On the other hand, if decisions are fully-restricted to sensible decisions then the user's choice is also fully-restricted by the subjective concept of 'sensible'.

To illustrate the problem, below are three possible choices that could be made with the compositors in Section 12.4:

- A linear predictor predicted by a CPH could be composed with a PH-ANN-predicted baseline and AFT model form to a full distribution.
- A survival time predicted by a regression SSVM could be composed with a Gompertz baseline and PO model form to a full distribution.
- A survival time could be composed by taking the 42nd quantile from a survival distribution predicted by a random survival forest.

Each choice lacks a meaningful interpretation however there is no apriori reason why they should yield 'bad' predictions and all could be considered in a benchmark experiment. Dismissing these examples as 'not sensible' may lead to dismissing the optimal model with respect to predictive performance.

Sensible Defaults

It has been demonstrated that the choice of defaults vastly influences human decision making (E. J. Johnson and Goldstein 2003), which is known as the '(endogenous) default effect'. This effect extends to computer science and parameter defaults. Setting sensible defaults for parameters encourages users towards using these defaults in their code and this 'sensible defaults' design principle is routinely used in programming software.[8]

This thesis advocates for a slight adaptation to the 'sensible defaults' design principle: non-proprietary open-source software should apply the sensible defaults principle whilst allowing users to make any choice that is possible (even if not sensible); whereas proprietary software should only allow sensible choices. This distinction is important from an ethical standpoint: in the latter case users may not be domain-experts and therefore the developer could be considered liable for negative consequences of building models from non-sensible choices.

12.7. Conclusions

This chapter introduced composition and reduction to survival analysis and formalised specific strategies. Formalising these concepts allows for better quality of research and most importantly improved transparency. Clear interface points for hyper-parameters and compositions allow for reproducibility that was previously obfuscated by unclear workflows and imprecise documentation for pipelines.

Additionally, composition and reduction improves accessibility. Reduction workflows vastly increase the number of machine learning models that can be utilised in survival analysis, thus opening the field to those whose experience is limited to regression or classification. Formalisation of workflows allows for precise implementation of model-agnostic pipelines as computational objects, as opposed to functions that are built directly into an algorithm without external interface points.

Finally, predictive performance is also increased by these methods, which is most prominently the case for the survival model averaging compositor (C4) (as demonstrated by RSFs).

All compositions in this chapter, as well as (R1)-(R6), have been implemented in `mlr3proba` with the `mlr3pipelines` (M. Binder et al. 2019) interface. The reductions to classification will be implemented in a near-future update. Additionally the `discSurv` package (Welchowski and Schmid 2019) will be interfaced as a `mlr3proba` pipeline to incorporate further discrete-time strategies.

The compositions (C1) and (C3) are included in the benchmark experiment in R. E. B. Sonabend (2021) so that every tested model can make probabilistic survival distribution predictions as well as deterministic survival time predictions. Future research will benchmark all the pipelines in this chapter and will cover algorithm and model selection, tuning, and comparison of performance. Strategies from other papers will also be explored.

[8] No specific reference for the 'sensible defaults' principle could be found, though it is often seen as a direct consequence of the 'convention over configuration' principle.

13. Alternative Methods

TODO (150-200 WORDS)

This survey has focused on reviewing machine learning models according to the three key themes of this thesis (Section 1.1.2) and within the thesis scope (Section 3.2). Therefore this survey has not exhaustively covered all machine learning models and entire model classes have been omitted; this short section briefly discusses these classes.

Bayesian Models

As stated in the thesis scope, only frequentist frameworks are considered in this thesis. In terms of accessibility, many more off-shelf survival model implementations exist in the frequentist framework. Despite this, there is good evidence that Bayesian survival models, such as Bayesian neural networks (Bakker et al. 2004; Faraggi et al. 1997), can perform well (Bishop 2006) and a survey of these models may be explored in future work.

Gaussian Processes

Gaussian Processes (GPs) are a class of model that naturally fit the survival paradigm as they model the joint distribution of random variables over some continuous domain, often time. The simplest extension from a standard Cox model to GP is given by the non-linear hazard

$$h(\tau|X_i) = h_0(\tau)\phi(g(\tau|X_i)); \quad g(\cdot) \sim \mathcal{GP}(0, k)$$

where ϕ is a non-negative link function, $\mathcal{GP}$ is a Gaussian process (Rasmussen and Williams 2004), and k is a kernel function with parameters to be estimated (Kim and Pavlovic 2018). Hyper-parameters are learnt by evaluating the likelihood function (Bishop 2006) and in the context of survival analysis this is commonly performed by assuming an inhomogeneous Poisson process (Fernández, Rivera, and Teh 2016; Saul 2016; Vehtari and Joensuu 2013). For a comprehensive survey of GPs for survival, see Saul (2016) (Saul 2016). There is evidence of GPs outperforming Cox and ML models (Fernández, Rivera, and Teh 2016). GPs are excluded from this survey due to lack of implementation (thus accessibility) and poorer transparency. Future research could look at increasing off-shelf accessibility of these models.

Non-Supervised Learning

As well as pure supervised learning, there are also survival models that use active learning (Nezhad et al. 2019), transfer learning, or treat survival analysis as a Markov process. As with GPs, none of these are currently available off-shelf and all require expert knowledge to be useful. These are not discussed in detail here but a very brief introduction to the Markov Process (MP) set-up is provided to motivate further consideration for the area.

(8) visualises the survival set-up as a Markov chain. In each discrete time-point $t_1, ..., t_{K-1}$, an individual can either move to the next time-point (and therefore be alive at that time-point), or move to one of the absorbing states ('Dead' and 'Censored'). The final time-point, t_K, is never visited as an individual must be dead or censored at the end of a study, and hence are last seen alive at t_{K-1}. In this set-up, data is assumed sequential and the time of death or censoring is determined by the last state at which the individual was seen to be alive, plus one, i.e. if an individual transitions from t_k to 'Death', then they died at t_{k+1}. This setting assumes the Markov property, so that the probability of moving to the 'next' state only depends on the current one. This method lends itself naturally to competing risks, which would extend the 'Dead' state to multiple absorbing states for each risk. Additionally, left-censoring can be naturally incorporated without further assumptions (Abner, Charnigo, and Kryscio 2013).

This set-up has been considered in survival both for Markov models and in the context of reinforcement learning (Data Study Group Team 2020), though the latter case is underdeveloped and future research could pursue this further.

14. Survival Software

TODO (150-200 WORDS)

15. Conclusions

References

Aalen, Odd. 1978. "Nonparametric Inference for a Family of Counting Processes." *The Annals of Statistics* 6 (4): 701–26.

Abadi, Martín, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, et al. 2015. "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems." https://www.tensorflow.org/.

Abner, Erin L, Richard J Charnigo, and Richard J Kryscio. 2013. "Markov chains and semi-Markov models in time-to-event analysis." *Journal of Biometrics & Biostatistics* Suppl 1 (e001): 19522. https://doi.org/10.4172/2155-6180.S1-e001.

Akaike, Hirotugu. 1974. "A New Look at the Statistical Model Identification." *IEEE Transactions on Automatic Control* 19 (6): 716–23. https://doi.org/10.1093/ietfec/e90-a.12.2762.

Akritas, Michael G. 1994. "Nearest Neighbor Estimation of a Bivariate Distribution Under Random Censoring." *Ann. Statist.* 22 (3): 1299–1327. https://doi.org/10.1214/aos/1176325630.

Andres, Axel, Aldo Montano-Loza, Russell Greiner, Max Uhlich, Ping Jin, Bret Hoehn, David Bigam, James Andrew Mark Shapiro, and Norman Mark Kneteman. 2018. "A novel learning algorithm to predict individual survival after liver transplantation for primary sclerosing cholangitis." *PLOS ONE* 13 (3): e0193523. https://doi.org/10.1371/journal.pone.0193523.

Bakker, Bart, Tom Heskes, Jan Neijt, and Bert Kappen. 2004. "Improving Cox survival analysis with a neural-Bayesian approach." *Statistics in Medicine* 23 (19): 2989–3012. https://doi.org/10.1002/sim.1904.

Bello, Ghalib A, Timothy J W Dawes, Jinming Duan, Carlo Biffi, Antonio de Marvao, Luke S G E Howard, J Simon R Gibbs, et al. 2019. "Deep-learning cardiac motion analysis for human survival prediction." *Nature Machine Intelligence* 1 (2): 95–104. https://doi.org/10.1038/s42256-019-0019-2.

Bennett, Steve. 1983. "Analysis of survival data by the proportional odds model." *Statistics in Medicine* 2 (2): 273–77. https://doi.org/https://doi.org/10.1002/sim.4780020223.

Biganzoli, E M, F Ambrogi, and P Boracchi. 2009. "Partial logistic artificial neural networks (PLANN) for flexible modeling of censored survival data." In *2009 International Joint Conference on Neural Networks*, 340–46. https://doi.org/10.1109/IJCNN.2009.5178824.

Biganzoli, Elia, Patrizia Boracchi, Luigi Mariani, and Ettore Marubini. 1998. "Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach." *Statistics in Medicine* 17 (10): 1169–86. https://doi.org/10.1002/(SICI)1097-0258(19980530)17:10%3C1169::AID-SIM796%3E3.0.CO;2-D.

Binder, Harald, and Martin Schumacher. 2008. "Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models." *BMC Bioinformatics* 9 (1): 14. https://doi.org/10.1186/1471-2105-9-14.

Binder, Harold. 2013. "CoxBoost: Cox models by likelihood based boosting for a single survival endpoint or competing risks." CRAN.

Binder, Martin, Florian Pfisterer, Bernd Bischl, Michel Lang, and Susanne Dandl. 2019. "mlr3pipelines: Preprocessing Operators and Pipelines for 'mlr3'." CRAN. https://cran.r-project.org/package=mlr3pipelines.

Bishop, Christopher M. 2006. *Pattern recognition and machine learning*. springer.

Blanche, Paul, Jean-François Dartigues, and Hélène Jacqmin-Gadda. 2013. "Review and comparison of ROC curve estimators for a time-dependent outcome with marker-dependent censoring." *Biometrical Journal* 55 (5): 687–704. https://doi.org/10.1002/bimj.201200045.

Blanche, Paul, Aurélien Latouche, and Vivian Viallon. 2012. "Time-dependent AUC with right-censored data: a survey study," October. https://doi.org/10.1007/978-1-4614-8981-8_11.

Bland, J Martin, and Douglas G. Altman. 2004. "The logrank test." *BMJ (Clinical Research Ed.)* 328 (7447): 1073. https://doi.org/10.1136/bmj.328.7447.1073.

Bou-Hamad, Imad, Denis Larocque, and Hatem Ben-Ameur. 2011. "A review of survival trees." *Statist. Surv.* 5: 44–71. https://doi.org/10.1214/09-SS047.

Bower, Hannah, Michael J Crowther, Mark J Rutherford, Therese M.-L. Andersson, Mark Clements, Xing-Rong Liu, Paul W Dickman, and Paul C Lambert. 2019. "Capturing simple and complex time-dependent effects using flexible parametric survival models: A simulation study." *Communications in Statistics - Simulation and Computation*, July, 1–17. https://doi.org/10.1080/03610918.2019.1634201.

Breiman, Leo. 1996. "Bagging Predictors." *Machine Learning* 24 (2): 123–40. https://doi.org/10.1023/A:1018054314350.

Breiman, Leo, and Philip Spector. 1992. "Submodel Selection and Evaluation in Regression. The X-Random Case." *International Statistical Review / Revue Internationale de Statistique* 60 (3): 291–319. https://doi.org/10.2307/1403680.

Brier, Glenn. 1950. "Verification of forecasts expressed in terms of probability." *Monthly Weather Review* 78 (1): 1–3.

Brilleman, Sam. 2019. "simsurv: Simulate Survival Data." CRAN. https://cran.r-project.org/package=simsurv.

Buckley, Jonathan, and Ian James. 1979. "Linear Regression with Censored Data." *Biometrika* 66 (3): 429–36. https://doi.org/10.2307/2335161.

Buhlmann, Peter. 2006. "Boosting for high-dimensional linear models." *Ann. Statist.* 34 (2): 559–83. https://doi.org/10.1214/009053606000000092.

Buhlmann, Peter, and Torsten Hothorn. 2007. "Boosting Algorithms: Regularization, Prediction and Model Fitting." *Statist. Sci.* 22 (4): 477–505. https://doi.org/10.1214/07-STS242.

Bühlmann, Peter, and Bin Yu. 2003. "Boosting With the L2 Loss." *Journal of the American Statistical Association* 98 (462): 324–39. https://doi.org/10.1198/016214503000125.

Chambless, Lloyd E, and Guoqing Diao. 2006. "Estimation of time-dependent area under the ROC curve for long-term risk prediction." *Statistics in Medicine* 25 (20): 3474–86. https://doi.org/10.1002/sim.2299.

Chen, Hung Chia, Ralph L. Kodell, Kuang Fu Cheng, and James J. Chen. 2012. "Assessment of performance of survival prediction models for cancer prognosis." *BMC Medical Research Methodology* 12. https://doi.org/10.1186/1471-2288-12-102.

Chen, Tianqi, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, et al. 2020. "xgboost: Extreme Gradient Boosting." CRAN. https://cran.r-project.org/package=xgboost.

Chen, Yen-Chen, Wan-Chi Ke, and Hung-Wen Chiu. 2014. "Risk classification of cancer survival using ANN with gene expression data from multiple laboratories." *Computers in Biology and Medicine* 48: 1–7. https://doi.org/https://doi.org/10.1016/j.compbiomed.2014.02.006.

Ching, Travers. 2015. "Cox-Nnet." https://github.com/lanagarmire/cox-nnet.

Ching, Travers, Xun Zhu, and Lana X Garmire. 2018. "Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data." *PLOS Computational Biology* 14 (4): e1006076. https://doi.org/10.1371/journal.pcbi.1006076.

Choodari-Oskooei, Babak, Patrick Royston, and Mahesh K. B. Parmar. 2012a. "A simulation study of predictive ability measures in a survival model I: Explained variation measures." *Statistics in Medicine* 31 (23): 2627–43. https://doi.org/10.1002/sim.4242.

———. 2012b. "A simulation study of predictive ability measures in a survival model II: explained randomness and predictive accuracy." *Statistics in Medicine* 31 (23): 2644–59. https://doi.org/10.1002/sim.4242.

Ciampi, Antonio, Sheilah A Hogg, Steve McKinney, and Johanne Thiffault. 1988. "RECPAM: a computer program for recursive partition and amalgamation for censored survival data and other situations frequently occurring in biostatistics. I. Methods and program features." *Computer Methods and Programs in Biomedicine* 26 (3): 239–56. https://doi.org/https://doi.org/10.1016/0169-2607(88)90004-1.

Ciampi, Antonio, Johanne Thiffault, Jean Pierre Nakache, and Bernard Asselain. 1986. "Stratification by stepwise regression, correspondence analysis and recursive partition: a comparison of three methods of analysis for survival data with covariates." *Computational Statistics and Data Analysis* 4 (3): 185–204. https://doi.org/10.1016/0167-9473(86)90033-2.

Clevert, Djork-Arné, Thomas Unterthiner, and Sepp Hochreiter. 2015. "Fast and accurate deep network learning by exponential linear units (elus)." *arXiv Preprint arXiv:1511.07289*.

Collett, David. 2014. *Modelling Survival Data in Medical Research*. 3rd ed. CRC.

Collins, Gary S., Joris A. De Groot, Susan Dutton, Omar Omar, Milensu Shanyinde, Abdelouahid Tajar, Merryn Voysey, et al. 2014. "External validation of multivariable prediction models: A systematic review of methodological conduct and reporting." *BMC Medical Research Methodology* 14 (1): 1–11. https://doi.org/10.1186/1471-2288-14-40.

Colosimo, Enrico, Fla´vio Ferreira, Maristela Oliveira, and Cleide Sousa. 2002. "Empirical comparisons between Kaplan-Meier and Nelson-Aalen survival function estimators." *Journal of Statistical Computation and Simulation* 72 (4): 299–308. https://doi.org/10.1080/00949650212847.

Cortes, Corinna, and Vladimir Vapnik. 1995. "Support-Vector Networks." *Machine Learning* 20: 273–97. https://doi.org/10.1007/BF00994018.

Cox, D. R. 1972. "Regression Models and Life-Tables." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 34 (2): 187–220.

———. 1975. "Partial Likelihood." *Biometrika* 62 (2): 269–76. https://doi.org/10.1080/03610910701884021.

Cui, Lei, Hansheng Li, Wenli Hui, Sitong Chen, Lin Yang, Yuxin Kang, Qirong Bo, and Jun Feng. 2020. "A deep learning-based framework for lung cancer survival analysis with biomarker interpretation." *BMC Bioinformatics* 21 (1): 112. https://doi.org/10.1186/s12859-020-3431-z.

Data Study Group Team. 2020. "Data Study Group Final Report: Great Ormond Street Hospital." https://doi.org/10.5281/zenodo.3670726.

Dawid, A P. 1984. "Present Position and Potential Developments: Some Personal Views: Statistical Theory: The Prequential Approach." *Journal of the Royal Statistical Society. Series A (General)* 147 (2): 278–92. https://doi.org/10.2307/2981683.

Dawid, A Philip. 1986. "Probability Forecasting." *Encyclopedia of Statistical Sciences* 7: 210——218.

Dawid, A Philip, and Monica Musio. 2014. "Theory and Applications of Proper Scoring Rules." *Metron* 72 (2): 169–83. https://arxiv.org/abs/arXiv:1401.0398v1.

Demler, Olga V, Nina P Paynter, and Nancy R Cook. 2015. "Tests of calibration and goodness-of-fit in the survival setting." *Statistics in Medicine* 34 (10): 1659–80. https://doi.org/10.1002/sim.6428.

Demšar, Janez. 2006. "Statistical comparisons of classifiers over multiple data sets." *Journal of*

Machine Learning Research 7 (Jan): 1–30.

Dietterich, Thomas G. 1998. "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms." *Neural Computation* 10 (7): 1895–1923. https://doi.org/10.1162/089976698300017197.

Du, Xian, and Sumeet Dua. 2011. "Cancer prognosis using support vector regression in imaging modality." *World Journal of Clinical Oncology* 2 (1): 44–49. https://doi.org/10.5306/wjco.v2.i1.44.

Efron, Bradley. 1988. "Logistic Regression, Survival Analysis, and the Kaplan-Meier Curve." *Journal of the American Statistical Association* 83 (402): 414–25. https://doi.org/10.2307/2288857.

Evers, Ludger, and Claudia-Martina Messow. 2008. "Sparse kernel methods for high-dimensional survival data." *Bioinformatics* 24 (14): 1632–38.

Faraggi, David, and Richard Simon. 1995. "A neural network model for survival data." *Statistics in Medicine* 14 (1): 73–82. https://doi.org/10.1002/sim.4780140108.

Faraggi, David, R Simon, E Yaskil, and A Kramar. 1997. "Bayesian Neural Network Models for Censored Data." *Biometrical Journal* 39 (5): 519–32. https://doi.org/10.1002/bimj.4710390502.

Fernández, Tamara, Nicolas Nicolás Rivera, and Yee Whye Teh. 2016. "Gaussian Processes for Survival Analysis." In *Advances in Neural Information Processing Systems*. Vol. 29. Nips. Curran Associates, Inc. http://arxiv.org/abs/1611.00817 https://proceedings.neurips.cc/paper/2016/file/ef1e491a766ce3127556063d49bc2f98-Paper.pdf.

Fleming, Thomas R, Judith R O'Fallon, Peter C O'Brien, and David P Harrington. 1980. "Modified Kolmogorov-Smirnov Test Procedures with Application to Arbitrarily Right-Censored Data." *Biometrics* 36 (4): 607–25. https://doi.org/10.2307/2556114.

Fotso, Stephane. 2018. "Deep Neural Networks for Survival Analysis Based on a Multi-Task Framework." *arXiv Preprint arXiv:1801.05512*, January. http://arxiv.org/abs/1801.05512.

Fouodo, Cesaire J K, I Konig, C Weihs, A Ziegler, and M Wright. 2018. "Support vector machines for survival analysis with R." *The R Journal* 10 (July): 412–23.

Freund, Yoav, and Robert E Schapire. 1996. "Experiments with a new boosting algorithm." In. Citeseer.

Friedman, Jerome. 1999. "Stochastic Gradient Boosting." *Computational Statistics & Data Analysis* 38 (March): 367–78. https://doi.org/10.1016/S0167-9473(01)00065-2.

Friedman, Jerome H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *The Annals of Statistics* 29 (5): 1189–1232. http://www.jstor.org/stable/2699986.

Friedman, Michael. 1982. "Piecewise exponential models for survival data with covariates." *The Annals of Statistics* 10 (1): 101–13.

Fritsch, Stefan, Frauke Guenther, and Marvin N. Wright. 2019. "neuralnet: Training of Neural Networks." CRAN. https://cran.r-project.org/package=neuralnet.

Gelfand, Alan E, Sujit K Ghosh, Cindy Christiansen, Stephen B Soumerai, and Thomas J McLaughlin. 2000. "Proportional hazards models: a latent competing risk approach." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 49 (3): 385–97. https://doi.org/https://doi.org/10.1111/1467-9876.00199.

Gensheimer, Michael F., and Balasubramanian Narasimhan. 2018. "A Simple Discrete-Time Survival Model for Neural Networks," 1–17. https://doi.org/arXiv:1805.00917v3.

Gensheimer, Michael F, and Balasubramanian Narasimhan. 2019. "A scalable discrete-time survival model for neural networks." *PeerJ* 7: e6257.

Georgousopoulou, Ekavi N, Christos Pitsavos, Christos Mary Yannakoulia, and Demosthenes B Panagiotakos. 2015. "Comparisons between Survival Models in Predicting Cardiovascular Dis-

ease Events : Application in the ATTICA Study (2002-2012)." *Journal of Statistics Applications & Probability* 4 (2): 203–10.

Gerds, Thomas A, and Martin Schumacher. 2006. "Consistent Estimation of the Expected Brier Score in General Survival Models with Right-Censored Event Times." *Biometrical Journal* 48 (6): 1029–40. https://doi.org/10.1002/bimj.200610301.

Giunchiglia, Eleonora, Anton Nemchenko, and Mihaela van der Schaar. 2018. "Rnn-surv: A deep recurrent model for survival analysis." In *International Conference on Artificial Neural Networks*, 23–32. Springer.

Gneiting, Tilmann, and Adrian E Raftery. 2007. "Strictly Proper Scoring Rules, Prediction, and Estimation." *Journal of the American Statistical Association* 102 (477): 359–78. https://doi.org/10.1198/016214506000001437.

Goli, Shahrbanoo, Hossein Mahjub, Javad Faradmal, Hoda Mashayekhi, and Ali-Reza Soltanian. 2016. "Survival Prediction and Feature Selection in Patients with Breast Cancer Using Support Vector Regression." Edited by Francesco Pappalardo. *Computational and Mathematical Methods in Medicine* 2016: 2157984. https://doi.org/10.1155/2016/2157984.

Goli, Shahrbanoo, Hossein Mahjub, Javad Faradmal, and Ali-Reza Soltanian. 2016. "Performance Evaluation of Support Vector Regression Models for Survival Analysis: A Simulation Study." *International Journal of Advanced Computer Science and Applications* 7 (June). https://doi.org/10.14569/IJACSA.2016.070650.

Gompertz, Benjamin. 1825. "On the Nature of the Function Expressive of the Law of Human Mortality, and on a New Mode of Determining the Value of Life Contingencies." *Philosophical Transactions of the Royal Society of London* 115: 513–83.

Good, I J. 1952. "Rational Decisions." *Journal of the Royal Statistical Society. Series B (Methodological)* 14 (1): 107–14. http://www.jstor.org/stable/2984087.

Gordon, Louis, and Richard A Olshen. 1985. "Tree-structured survival analysis." *Cancer Treatment Reports* 69 (10): 1065–69.

Graf, Erika, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. 1999. "Assessment and comparison of prognostic classification schemes for survival data." *Statistics in Medicine* 18 (17-18): 2529–45. https://doi.org/10.1002/(SICI)1097-0258(19990915/30)18:17/18%3C2529::AID-SIM274%3E3.0.CO;2-5.

Graf, Erika, and Martin Schumacher. 1995. "An Investigation on Measures of Explained Variation in Survival Analysis." *Journal of the Royal Statistical Society. Series D (The Statistician)* 44 (4): 497–507. https://doi.org/10.2307/2348898.

Greenwell, Brandon, Bradley Boehmke, Jay Cunningham, and. GBM Developers. 2019. "gbm: Generalized Boosted Regression Models." CRAN. https://cran.r-project.org/package=gbm.

Gressmann, Frithjof, Franz J. Király, Bilal Mateen, and Harald Oberhauser. 2018. "Probabilistic supervised learning." https://doi.org/10.1002/iub.552.

Habibi, Danial, Mohammad Rafiei, Ali Chehrei, Zahra Shayan, and Soheil Tafaqodi. 2018. "Comparison of Survival Models for Analyzing Prognostic Factors in Gastric Cancer Patients." *Asian Pacific Journal of Cancer Prevention : APJCP* 19 (3): 749–53. https://doi.org/10.22034/APJCP.2018.19.3.749.

Haider, Humza, Bret Hoehn, Sarah Davis, and Russell Greiner. 2020. "Effective ways to build and evaluate individual survival distributions." *Journal of Machine Learning Research* 21 (85): 1–63.

Han, Ilkyu, June Hyuk Kim, Heeseol Park, Han-Soo Kim, and Sung Wook Seo. 2018. "Deep learning approach for survival prediction for patients with synovial sarcoma." *Tumor Biology* 40 (9): 1010428318799264. https://doi.org/10.1177/1010428318799264.

Harrell, F E Jr, K L Lee, R M Califf, D B Pryor, and R A Rosati. 1984. "Regression modelling

strategies for improved prognostic prediction." *Statistics in Medicine* 3 (2): 143–52. https://doi.org/10.1002/sim.4780030207.

Harrell, Frank E., Robert M. Califf, and David B. Pryor. 1982. "Evaluating the yield of medical tests." *JAMA* 247 (18): 2543–46. http://dx.doi.org/10.1001/jama.1982.03320430047030.

Harrell, Frank E., Kerry L. Lee, and Daniel B. Mark. 1996. "Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors." *Statistics in Medicine* 15: 361–87. https://doi.org/10.1002/0470023678.ch2b(i).

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2001. *The Elements of Statistical Learning*. Springer New York Inc.

Heagerty, Patrick J., Thomas Lumley, and Margaret S. Pepe. 2000. "Time-Dependent ROC Curves for Censored Survival Data and a Diagnostic Marker." *Biometrics* 56 (2): 337–44. https://doi.org/10.1111/j.0006-341X.2000.00337.x.

Heagerty, Patrick J, and Yingye Zheng. 2005. "Survival Model Predictive Accuracy and ROC Curves." *Biometrics* 61 (1): 92–105. https://doi.org/10.1111/j.0006-341X.2005.030814.x.

Henderson, and Velleman. 1981. "Building multiple regression models interactively." *Biometrics* 37: 391——411.

Herrmann, Moritz, Philipp Probst, Roman Hornung, Vindi Jurinovic, and Anne-Laure Boulesteix. 2021. "Large-scale benchmark study of survival prediction methods using multi-omics data." *Briefings in Bioinformatics* 22 (3). https://doi.org/10.1093/bib/bbaa167.

Hielscher, Thomas, Manuela Zucknick, Wiebke Werft, and Axel Benner. 2010. "On the Prognostic Value of Gene Expression Signatures for Censored Data BT - Advances in Data Analysis, Data Handling and Business Intelligence." In, edited by Andreas Fink, Berthold Lausen, Wilfried Seidel, and Alfred Ultsch, 663–73. Berlin, Heidelberg: Springer Berlin Heidelberg.

Hosmer, David W, and Stanley Lemeshow. 1980. "Goodness of fit tests for the multiple logistic regression model." *Communications in Statistics-Theory and Methods* 9 (10): 1043–69.

Hosmer Jr, David W, Stanley Lemeshow, and Susanne May. 2011. *Applied survival analysis: regression modeling of time-to-event data*. Vol. 618. John Wiley & Sons.

Hothorn, Torsten, Peter Buehlmann, Thomas Kneib, Matthias Schmid, and Benjamin Hofner. 2020. "mboost: Model-Based Boosting." CRAN. https://cran.r-project.org/package=mboost.

Hothorn, Torsten, Peter Bühlmann, Sandrine Dudoit, Annette Molinaro, and Mark J Van Der Laan. 2005. "Survival ensembles." *Biostatistics* 7 (3): 355–73. https://doi.org/10.1093/biostatistics/kxj011.

Hothorn, Torsten, Kurt Hornik, and Achim Zeileis. 2006. "Unbiased Recursive Partitioning: A Conditional Inference Framework." *Journal of Computational and Graphical Statistics* 15 (3): 651——674.

Hothorn, Torsten, and Berthold Lausen. 2003. "On the exact distribution of maximally selected rank statistics." *Computational Statistics & Data Analysis* 43 (2): 121–37. https://doi.org/10.1016/S0167-9473(02)00225-6.

Hothorn, Torsten, Berthold Lausen, Axel Benner, and Martin Radespiel-Tröger. 2004. "Bagging survival trees." *Statistics in Medicine* 23 (1): 77–91. https://doi.org/10.1002/sim.1593.

Hothorn, Torsten, and Achim Zeileis. 2015. "partykit: A Modular Toolkit for Recursive Partytioning in R." *Journal of Machine Learning Research* 16: 3905–9. http://jmlr.org/papers/v16/hothorn15a.html.

Huang, Shigao, Jie Yang, Simon Fong, and Qi Zhao. 2020a. "Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges." *Cancer Letters* 471: 61–71. https://doi.org/https://doi.org/10.1016/j.canlet.2019.12.007.

———. 2020b. "Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges." *Cancer Letters* 471: 61–71. https://doi.org/https://doi.org/10.1016/j.canlet.2019.12.

007.

Hung, Hung, and Chin-Tsang Chiang. 2010. "Estimation methods for time-dependent AUC models with survival data." *The Canadian Journal of Statistics / La Revue Canadienne de Statistique* 38 (1): 8–26. http://www.jstor.org/stable/27805213.

HURVICH, CLIFFORD M, and CHIH-LING TSAI. 1989. "Regression and time series model selection in small samples." *Biometrika* 76 (2): 297–307. https://doi.org/10.1093/biomet/76.2.297.

Ishwaran, By Hemant, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. 2008. "Random survival forests." *The Annals of Statistics* 2 (3): 841–60. https://doi.org/10.1214/08-AOAS169.

Ishwaran, Hemant, Eugene H Blackstone, Claire E Pothier, and Michael S Lauer. 2004. "Relative Risk Forests for Exercise Heart Rate Recovery as a Predictor of Mortality." *Journal of the American Statistical Association* 99 (467): 591–600. https://doi.org/10.1198/016214504000000638.

Ishwaran, Hemant, and Udaya B Kogalur. 2018. "randomForestSRC." https://cran.r-project.org/package=randomForestSRC.

Jackson, Christopher. 2016. "flexsurv: A Platform for Parametric Survival Modeling in R." *Journal of Statistical Software* 70 (8): 1–33.

Jackson, Dan, Ian R. White, Shaun Seaman, Hannah Evans, Kathy Baisley, and James Carpenter. 2014. "Relaxing the independent censoring assumption in the Cox proportional hazards model using multiple imputation." *Statistics in Medicine* 33 (27): 4681–94. https://doi.org/10.1002/sim.6274.

Jager, Kitty J, Paul C van Dijk, Carmine Zoccali, and Friedo W Dekker. 2008. "The analysis of survival data: the Kaplan–Meier method." *Kidney International* 74 (5): 560–65. https://doi.org/https://doi.org/10.1038/ki.2008.217.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An introduction to statistical learning*. Vol. 112. New York: Springer.

Jiao, Yasen, and Pufeng Du. 2016. "Performance measures in evaluating machine learning based bioinformatics predictors for classifications." *Quantitative Biology* 4 (4): 320–30.

Jing, Bingzhong, Tao Zhang, Zixian Wang, Ying Jin, Kuiyuan Liu, Wenze Qiu, Liangru Ke, et al. 2018. "RankDeepSurv." https://github.com/sysucc-ailab/RankDeepSurv.

———, et al. 2019. "A deep survival analysis method based on ranking." *Artificial Intelligence in Medicine* 98: 1–9. https://doi.org/https://doi.org/10.1016/j.artmed.2019.06.001.

Johnson, Brent A, and Qi Long. 2011. "Survival ensembles by the sum of pairwise differences with application to lung cancer microarray studies." *Ann. Appl. Stat.* 5 (2A): 1081–101. https://doi.org/10.1214/10-AOAS426.

Johnson, Eric J, and Daniel Goldstein. 2003. "Do Defaults Save Lives?" *Science* 302 (5649): 1338 LP–1339. https://doi.org/10.1126/science.1091721.

Kalbfleisch, John D, and Ross L Prentice. 2011. *The statistical analysis of failure time data*. Vol. 360. John Wiley & Sons.

Kamarudin, Adina Najwa, Trevor Cox, and Ruwanthi Kolamunnage-Dona. 2017. "Time-dependent ROC curve analysis in medical research: Current methods and applications." *BMC Medical Research Methodology* 17 (1): 1–19. https://doi.org/10.1186/s12874-017-0332-6.

KATTAN, MICHAEL W. 2003. "Comparison of Cox Regression With Other Methods for Determining Prediction Models and Nomograms." *Journal of Urology* 170 (6S): S6–10. https://doi.org/10.1097/01.ju.0000094764.56269.2d.

Katzman, Jared L, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. 2018. "DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network." *BMC Medical Research Methodology* 18 (1): 24.

https://doi.org/10.1186/s12874-018-0482-1.

Katzman, Jared, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. 2016. "Deep Survival: A Deep Cox Proportional Hazards Network," June.

Kent, John T., and John O'Quigley. 1988. "Measures of dependence for censored survival data." *Biometrika* 75 (3): 525–34. https://doi.org/10.1093/biomet/75.3.525.

Khan, Faisal M., and Valentina Bayer Zubek. 2008. "Support vector regression for censored data (SVRc): A novel tool for survival analysis." *Proceedings - IEEE International Conference on Data Mining, ICDM*, 863–68. https://doi.org/10.1109/ICDM.2008.50.

Kim, Minyoung, and Vladimir Pavlovic. 2018. "Variational Inference for Gaussian Process Models for Survival Analysis." *UAI*, 435–45.

Király, Franz J., Markus Löning, Anthony Blaom, Ahmed Guecioueur, and Raphael Sonabend. 2021. "Designing Machine Learning Toolboxes: Concepts, Principles and Patterns." *arXiv*, January. http://arxiv.org/abs/2101.04938.

Király, Franz J, Bilal Mateen, and Raphael Sonabend. 2018. "NIPS - Not Even Wrong? A Systematic Review of Empirically Complete Demonstrations of Algorithmic Effectiveness in the Machine Learning and Artificial Intelligence Literature." *arXiv*, December. http://arxiv.org/abs/1812.07519.

Kirmani, S N U A, and Ramesh C Gupta. 2001. "On the Proportional Odds Model in Survival Analysis." *Annals of the Institute of Statistical Mathematics* 53 (2): 203–16. https://doi.org/10.1023/A:1012458303498.

Klein, John P, and Melvin L Moeschberger. 2003. *Survival analysis: techniques for censored and truncated data*. 2nd ed. Springer Science & Business Media.

Kohavi, Ron. 1995. "A study of cross-validation and bootstrap for accuracy estimation and model selection." *Ijcai* 14 (2): 1137–45.

Korn, Edward L., and Richard Simon. 1990. "Measures of explained variation for survival data." *Statistics in Medicine* 9 (5): 487–503. https://doi.org/10.1002/sim.4780090503.

Korn, Edward L, and Richard Simon. 1991. "Explained Residual Variation, Explained Risk, and Goodness of Fit." *The American Statistician* 45 (3): 201–6. https://doi.org/10.2307/2684290.

Koziol, James A., and Zhenyu Jia. 2009. "The concordance index C and the Mann-Whitney parameter Pr(X>Y) with randomly censored data." *Biometrical Journal* 51 (3): 467–74. https://doi.org/10.1002/bimj.200800228.

Kvamme, Håvard. 2018. "Pycox." https://pypi.org/project/pycox/.

Kvamme, Håvard, Ørnulf Borgan, and Ida Scheel. 2019. "Time-to-event prediction with neural networks and Cox regression." *Journal of Machine Learning Research* 20 (129): 1–30.

Laan, Mark K van der, Eric C Polley, and Alan E Hubbard. 2007. "Super Learner." *Statistical Applications in Genetics and Molecular Biology* 6 (1). https://doi.org/10.2202/1544-6115.1309.

Land, Walker H, Xingye Qiao, Dan Margolis, and Ron Gottlieb. 2011. "A new tool for survival analysis: evolutionary programming/evolutionary strategies (EP/ES) support vector regression hybrid using both censored / non-censored (event) data." *Procedia Computer Science* 6: 267–72. https://doi.org/https://doi.org/10.1016/j.procs.2011.08.050.

Langford, John, Paul Mineiro, Alina Beygelzimer, and Hal Daume. 2016. "Learning Reductions that Really Work." *Proceedings of the IEEE* 104 (1).

Lao, Jiangwei, Yinsheng Chen, Zhi-Cheng Li, Qihua Li, Ji Zhang, Jing Liu, and Guangtao Zhai. 2017. "A Deep Learning-Based Radiomics Model for Prediction of Survival in Glioblastoma Multiforme." *Scientific Reports* 7 (1): 10353. https://doi.org/10.1038/s41598-017-10649-8.

Lawless, Jerald F, and Yan Yuan. 2010. "Estimation of prediction error for survival models." *Statistics in Medicine* 29 (2): 262–74. https://doi.org/10.1002/sim.3758.

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. "The Parable of

Google Flu: Traps in Big Data Analysis." *Science* 343 (6176): 1203 LP–1205. https://doi.org/10.1126/science.1248506.

LeBlanc, Michael, and John Crowley. 1992. "Relative Risk Trees for Censored Survival Data." *Biometrics* 48 (2): 411–25. https://doi.org/10.2307/2532300.

———. 1993. "Survival Trees by Goodness of Split." *Journal of the American Statistical Association* 88 (422): 457–67. https://doi.org/10.2307/2290325.

Lee, Changhee, William Zame, Jinsung Yoon, and Mihaela Van der Schaar. 2018. "DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks." *Proceedings of the AAAI Conference on Artificial Intelligence* 32 (1). https://doi.org/10.1609/aaai.v32i1.11842.

Lee, Donald K K, Ningyuan Chen, and Hemant Ishwaran. 2019. "Boosted nonparametric hazards with time-dependent covariates." https://arxiv.org/abs/arXiv:1701.07926v6.

Li, Liang, Tom Greene, and Bo Hu. 2018. "A simple method to estimate the time-dependent receiver operating characteristic curve and the area under the curve with right censored data." *Statistical Methods in Medical Research* 27 (8): 2264–78. https://doi.org/10.1177/0962280216680239.

Liang, Hua, and Guohua Zou. 2008. "Improved AIC Selection Strategy for Survival Analysis." *Computational Statistics & Data Analysis* 52 (5): 2538–48. https://doi.org/10.1016/j.csda.2007.09.003.

Liestol, Knut, Per Kragh Andersen, and Ulrich Andersen. 1994. "Survival analysis and neural nets." *Statistics in Medicine* 13 (12): 1189–1200. https://doi.org/10.1002/sim.4780131202.

Lundberg, Scott M, and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." *Advances in Neural Information Processing Systems* 30.

Lundin, M, J Lundin, H B Burke, S Toikkanen, L Pylkkänen, and H Joensuu. 1999. "Artificial Neural Networks Applied to Survival Prediction in Breast Cancer." *Oncology* 57 (4): 281–86. https://doi.org/10.1159/000012061.

Luxhoj, James T., and Huan Jyh Shyur. 1997. "Comparison of proportional hazards models and neural networks for reliability estimation." *Journal of Intelligent Manufacturing* 8 (3): 227–34. https://doi.org/10.1023/A:1018525308809.

Ma, Shuangge, and Jian Huang. 2006. "Regularized ROC method for disease classification and biomarker selection with microarray data." *Bioinformatics (Oxford, England)* 21 (January): 4356–62. https://doi.org/10.1093/bioinformatics/bti724.

Mani, D R, James Drew, Andrew Betz, and Piew Datta. 1999. "Statistics and data mining techniques for lifetime value modeling." In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 94–103.

Mariani, L, D Coradini, E Biganzoli, P Boracchi, E Marubini, S Pilotti, B Salvadori, et al. 1997. "Prognostic factors for metachronous contralateral breast cancer: A comparison of the linear Cox regression model and its artificial neural network extension." *Breast Cancer Research and Treatment* 44 (2): 167–78. https://doi.org/10.1023/A:1005765403093.

Mayr, Andreas, Benjamin Hofner, and Matthias Schmid. 2016. "Boosting the discriminatory power of sparse survival models via optimization of the concordance index and stability selection." *BMC Bioinformatics* 17 (1): 288. https://doi.org/10.1186/s12859-016-1149-8.

Mayr, Andreas, and Matthias Schmid. 2014. "Boosting the concordance index for survival data–a unified framework to derive and evaluate biomarker combinations." *PloS One* 9 (1): e84483–83. https://doi.org/10.1371/journal.pone.0084483.

McKinney, Scott Mayer, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, et al. 2020. "International evaluation of an AI system for breast cancer screening." *Nature* 577 (7788): 89–94. https://doi.org/10.1038/s41586-019-1799-6.

Meinshausen, Nicolai, and Peter Bühlmann. 2010. "Stability selection." *Journal of the Royal*

Statistical Society: Series B (Statistical Methodology) 72 (4): 417–73. https://doi.org/10.1111/j.1467-9868.2010.00740.x.

Moghimi-dehkordi, Bijan, Azadeh Safaee, Mohamad Amin Pourhoseingholi, Reza Fatemi, Ziaoddin Tabeie, and Mohammad Reza Zali. 2008. "Statistical Comparison of Survival Models for Analysis of Cancer Data." *Asian Pacific Journal of Cancer Prevention* 9: 417–20.

Molnar, Christoph. 2019. *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/.

Murphy, Allan H. 1973. "A New Vector Partition of the Probability Score." *Journal of Applied Meteorology and Climatology* 12 (4): 595–600. https://doi.org/10.1175/1520-0450(1973)012%3C0595:ANVPOT%3E2.0.CO;2.

N. Venables, W, and B D. Ripley. 2002. *Modern Applied Statistics with S*. Springer. http://www.stats.ox.ac.uk/pub/MASS4.

Nadeau, Claude, and Yoshua Bengio. 2003. "Inference for the Generalization Error." *Machine Learning* 52 (3): 239–81. https://doi.org/10.1023/A:1024068626366.

Nair, Vinod, and Geoffrey E Hinton. 2010. "Rectified linear units improve restricted boltzmann machines." In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 807–14.

Nasejje, Justine B, Henry Mwambi, Keertan Dheda, and Maia Lesosky. 2017. "A comparison of the conditional inference survival forest model to random survival forests based on a simulation study as well as on two applications with time-to-event data." *BMC Medical Research Methodology* 17 (1): 115. https://doi.org/10.1186/s12874-017-0383-8.

Nelson, Wayne. 1972. "Theory and Applications of Hazard Plotting for Censored Failure Data." *Technometrics* 14 (4): 945–66.

Newson, Roger B. 1983. "Comparing the predictive power of survival models using Harrell's c or Somers' D." *The Stata Journal*, no. ii: 1–19.

Nezhad, Milad Zafar, Najibesadat Sadati, Kai Yang, and Dongxiao Zhu. 2019. "A Deep Active Survival Analysis approach for precision treatment recommendations: Application of prostate cancer." *Expert Systems with Applications* 115: 16–26. https://doi.org/https://doi.org/10.1016/j.eswa.2018.07.070.

Ng, Ryan, Kathy Kornas, Rinku Sutradhar, Walter P. Wodchis, and Laura C. Rosella. 2018. "The current application of the Royston-Parmar model for prognostic modeling in health research: a scoping review." *Diagnostic and Prognostic Research* 2 (1): 4. https://doi.org/10.1186/s41512-018-0026-5.

Oh, Sung Eun, Sung Wook Seo, Min-Gew Choi, Tae Sung Sohn, Jae Moon Bae, and Sung Kim. 2018. "Prediction of Overall Survival and Novel Classification of Patients with Gastric Cancer Using the Survival Recurrent Network." *Annals of Surgical Oncology* 25 (5): 1153–59. https://doi.org/10.1245/s10434-018-6343-7.

Ohno-Machado, Lucila. 1996. "Medical applications of artificial neural networks: connectionist models of survival." Stanford University Stanford, Calif.

———. 1997. "A COMPARISON OF COX PROPORTIONAL HAZARDS AND ARTIFICIAL NEURAL NETWORK MODELS FOR MEDICAL PROGNOSIS The theoretical advantages and disadvantages of using different methods for predicting survival have seldom been tested in real data sets [1 , 2]. Althou." *Comput. Biol. Med* 27 (1): 55–65.

Patel, Katie, Richard Kay, and Lucy Rowell. 2006. "Comparing proportional hazards and accelerated failure time models: An application in influenza." *Pharmaceutical Statistics* 5 (3): 213–24. https://doi.org/10.1002/pst.213.

Pencina, Michael J., Ralph B. D'Agostino, and Linye Song. 2012. "Quantifying discrimination of Framingham risk functions with different survival C statistics." *Statistics in Medicine* 31 (15):

1543–53. https://doi.org/10.1002/sim.4508.

Peters, Andrea, and Torsten Hothorn. 2019. "ipred: Improved Predictors." CRAN. https://cran.r-project.org/package=ipred.

Polley, Eric C, and Mark J Van Der Laan. 2010. "Super learner in prediction."

Pölsterl, Sebastian. 2020. "scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn." *Journal of Machine Learning Research* 21 (212): 1——6. http://jmlr.org/papers/v21/20-729.html.

Potapov, Sergej, Werner Adler, and Matthias Schmid. 2012. "survAUC: Estimators of prediction accuracy for time-to-event data."

Puddu, Paolo Emilio, and Alessandro Menotti. 2012. "Artificial neural networks versus proportional hazards Cox models to predict 45-year all-cause mortality in the Italian Rural Areas of the Seven Countries Study." *BMC Medical Research Methodology* 12 (1): 100. https://doi.org/10.1186/1471-2288-12-100.

Qi, Jiezhi. 2009. "Comparison of Proportional Hazards and Accelerated Failure Time Models." PhD thesis.

R., Cox, and Snell J. 1968. "A General Definition of Residuals." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 30 (2): 248–75.

Rahman, M. Shafiqur, Gareth Ambler, Babak Choodari-Oskooei, and Rumana Z. Omar. 2017. "Review and evaluation of performance measures for survival prediction models in external validation settings." *BMC Medical Research Methodology* 17 (1): 1–15. https://doi.org/10.1186/s12874-017-0336-2.

Rasmussen, C. E., and C. K. I. Williams. 2004. *Gaussian processes for machine learning*. Vol. 14. 2. https://doi.org/10.1142/S0129065704001899.

Reid, Nancy. 1994. "A Conversation with Sir David Cox." *Statistical Science* 9 (3): 439–55. https://doi.org/10.1214/aos/1176348654.

Ridgeway, Greg. 1999. "The state of boosting." *Computing Science and Statistics* 31: 172——181.

Rietschel, Carl, Jinsung Yoon, and Mihaela van der Schaar. 2018. "Feature Selection for Survival Analysis with Competing Risks using Deep Learning." *arXiv Preprint arXiv:1811.09317*.

Ripley, Brian D, and Ruth M Ripley. 2001. "Neural networks as statistical methods in survival analysis." In *Clinical Applications of Artificial Neural Networks*, edited by Richard Dybowski and Vanya Gant, 237–55. Cambridge: Cambridge University Press. https://doi.org/DOI: 10.1017/CBO9780511543494.011.

Ripley, R M, A L Harris, and L Tarassenko. 1998. "Neural network models for breast cancer prognosis." *Neural Computing & Applications* 7 (4): 367–75. https://doi.org/10.1007/BF01428127.

Royston, P. 2001. "The Lognormal Distribution as a Model for Survival Time in Cancer, With an Emphasis on Prognostic Factors." *Statistica Neerlandica* 55 (1): 89–104. https://doi.org/10.1111/1467-9574.00158.

Royston, Patrick, and Douglas G. Altman. 2013. "External validation of a Cox prognostic model: Principles and methods." *BMC Medical Research Methodology* 13 (1). https://doi.org/10.1186/1471-2288-13-33.

Royston, Patrick, Mahesh K B Parmar, and Douglas G Altman. 2008. "Visualizing Length of Survival in Time-to-Event Studies: A Complement to Kaplan–Meier Plots." *JNCI: Journal of the National Cancer Institute* 100 (2): 92–97. https://doi.org/10.1093/jnci/djm265.

Royston, Patrick, and Mahesh K. B. Parmar. 2002. "Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects." *Statistics in Medicine* 21 (15): 2175–97. https://doi.org/10.1002/sim.1203.

Royston, Patrick, and Willi Sauerbrei. 2004. "A new measure of prognostic separation in survival

data." *Statistics in Medicine* 23 (5): 723–48. https://doi.org/10.1002/sim.1621.

Sashegyi, Andreas, and David Ferry. 2017. "On the Interpretation of the Hazard Ratio and Communication of Survival Benefit." *The Oncologist* 22 (4): 484–86. https://doi.org/10.1634/theoncologist.2016-0198.

Saul, Alan D. 2016. "Gaussian Process Based Approaches for Survival Analysis." University of Sheffield.

Schemper, Michael, and Robin Henderson. 2000. "Predictive Accuracy and Explained Variation in Cox Regression." *Biometrics* 56: 249–55. https://doi.org/10.1002/sim.1486.

Schmid, Matthias, Thomas Hielscher, Thomas Augustin, and Olaf Gefeller. 2011. "A Robust Alternative to the Schemper-Henderson Estimator of Prediction Error." *Biometrics* 67 (2): 524–35. https://doi.org/10.1111/j.1541-0420.2010.01459.x.

Schmid, Matthias, and Torsten Hothorn. 2008a. "Boosting additive models using component-wise P-splines." *Computational Statistics & Data Analysis* 53 (2): 298–311.

———. 2008b. "Flexible boosting of accelerated failure time models." *BMC Bioinformatics* 9 (February): 269. https://doi.org/10.1186/1471-2105-9-269.

Schmid, Matthias, and Sergej Potapov. 2012. "A comparison of estimators to evaluate the discriminatory power of time-to-event models." *Statistics in Medicine* 31 (23): 2588–2609. https://doi.org/10.1002/sim.5464.

Schwarz, Gideon. 1978. "Estimating the Dimension of a Model." *The Annals of Statistics* 6 (2): 461–64. https://doi.org/10.1214/aos/1176344136.

Schwarzer, Guido, Werner Vach, and Martin Schumacher. 2000. "On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology." *Statistics in Medicine* 19 (4): 541–61. https://doi.org/10.1002/(SICI)1097-0258(20000229)19:4%3C541::AID-SIM355%3E3.0.CO;2-V.

Segal, Mark Robert. 1988. "Regression Trees for Censored Data." *Biometrics* 44 (1): 35——47.

Seker, H, M O Odetayo, D Petrovic, R N G Naguib, C Bartoli, L Alasio, M S Lakshmi, G V Sherbet, and O R Hinton. 2002. "An artificial neural network based feature evaluation index for the assessment of clinical factors in breast cancer survival analysis." In *IEEE CCECE2002. Canadian Conference on Electrical and Computer Engineering. Conference Proceedings (Cat. No.02CH37373)*, 2:1211–1215 vol.2. https://doi.org/10.1109/CCECE.2002.1013121.

Seker, Huseyin, Michael O Odetayo, Dobrila Petrovic, Raouf N G Naguib, C Bartoli, L Alasio, M S Lakshmi, and G V Sherbet. 2002. "Assessment of nodal involvement and survival analysis in breast cancer patients using image cytometric data: statistical, neural network and fuzzy approaches." *Anticancer Research* 22 (1A): 433–38. http://europepmc.org/abstract/MED/12017328.

Shiao, Han-Tai, and Vladimir Cherkassky. 2013. "SVM-based approaches for predictive modeling of survival data." In *Proceedings of the International Conference on Data Mining (DMIN)*, 1. The Steering Committee of The World Congress in Computer Science, Computer ….

Shivaswamy, Pannagadatta K., Wei Chu, and Martin Jansche. 2007. "A support vector approach to censored targets." In *Proceedings - IEEE International Conference on Data Mining, ICDM*, 655–60. https://doi.org/10.1109/ICDM.2007.93.

Sonabend, Raphael. 2020. "survivalmodels: Models for Survival Analysis." CRAN. https://raphaels1.r-universe.dev/ui#package:survivalmodels.

Sonabend, Raphael Edward Benjamin. 2021. "A Theoretical and Methodological Framework for Machine Learning in Survival Analysis: Enabling Transparent and Accessible Predictive Modelling on Right-Censored Time-to-Event Data." PhD, University College London (UCL). https://discovery.ucl.ac.uk/id/eprint/10129352/.

Sonabend, Raphael, Franz J Király, Andreas Bender, Bernd Bischl, and Michel Lang. 2021.

"mlr3proba: an R package for machine learning in survival analysis." Edited by Jonathan Wren. *Bioinformatics* 37 (17): 2789–91. https://doi.org/10.1093/bioinformatics/btab039.

Song, Xiao, and Xiao-Hua Zhou. 2008. "A semiparametric approach for the covariate specific ROC curve with survival outcome." *Statistica Sinica* 18 (July): 947–65.

Spooner, Annette, Emily Chen, Arcot Sowmya, Perminder Sachdev, Nicole A Kochan, Julian Trollor, and Henry Brodaty. 2020. "A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction." *Scientific Reports* 10 (1): 20410. https://doi.org/10.1038/s41598-020-77220-w.

Spruance, Spotswood L, Julia E Reid, Michael Grace, and Matthew Samore. 2004. "Hazard ratio in clinical trials." *Antimicrobial Agents and Chemotherapy* 48 (8): 2787–92. https://doi.org/10.1128/AAC.48.8.2787-2792.2004.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. "Dropout: a simple way to prevent neural networks from overfitting." *The Journal of Machine Learning Research* 15 (1): 1929–58.

Stasinopoulos, Mikis, Bob Rigby, Vlasios Voudouris, and Daniil Kiose. 2020. "gamlss.add: Extra Additive Terms for Generalized Additive Models for Location Scale and Shape." CRAN. https://cran.r-project.org/package=gamlss.add.

Street, W Nick. 1998. "A Neural Network Model for Prognostic Prediction." In *Proceedings of the Fifteenth International Conference on Machine Learning*. San Francisco.

Therneau, Terry M. 2015. "A Package for Survival Analysis in S." https://cran.r-project.org/package=survival.

Therneau, Terry M., and Beth Atkinson. 2019. "rpart: Recursive Partitioning and Regression Trees." CRAN.

Therneau, Terry M., and Elizabeth Atkinson. 2020. "Concordance." https://cran.r-project.org/web/packages/survival/vignettes/concordance.pdf.

Therneau, Terry M., Patricia M. Grambsch, and Thomas R. Fleming. 1990. "Martingale-based residuals for survival models." *Biometrika* 77 (1): 147–60. https://doi.org/10.1093/biomet/77.1.147.

Tsoumakas, Grigorios, and Ioannis Katakis. 2007. "Multi-Label Classification: An Overview." *International Journal of Data Warehousing and Mining* 3 (3): 1–13. https://doi.org/10.4018/jdwm.2007070101.

Tutz, Gerhard, and Harald Binder. 2007. "Boosting Ridge Regression." *Computational Statistics & Data Analysis* 51 (February): 6044–59. https://doi.org/10.1016/j.csda.2006.11.041.

Tutz, Gerhard, and Matthias Schmid. 2016. *Modeling Discrete Time-to-Event Data*. Springer Series in Statistics. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-28158-2.

Uno, Hajime, Tianxi Cai, Michael J. Pencina, Ralph B. D'Agostino, and L J Wei. 2011. "On the C-statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival Data." *Statistics in Medicine* 30 (10): 1105–17. https://doi.org/10.1002/sim.4154.

Uno, Hajime, Tianxi Cai, Lu Tian, and L J Wei. 2007. "Evaluating Prediction Rules for t-Year Survivors with Censored Regression Models." *Journal of the American Statistical Association* 102 (478): 527–37. http://www.jstor.org/stable/27639883.

Ushey, Kevin, J J Allaire, and Yuan Tang. 2020. "reticulate: Interface to 'Python'." CRAN. https://cran.r-project.org/package=reticulate.

Van Belle, Vanya, Kristiaan Pelckmans, Johan A K Suykens, and Sabine Van Huffel. 2008. "Survival SVM: a practical scalable algorithm." In *Proceedings of the 16th European Symposium on Artificial Neural Networks (ESANN)*, 89–94.

Van Belle, Vanya, Kristiaan Pelckmans, Johan A. K. Suykens, and Sabine Van Huffel. 2007. "Sup-

port Vector Machines for Survival Analysis." In *In Proceedings of the Third International Conference on Computational Intelligence in Medicine and Healthcare*. 1.

Van Belle, Vanya, Kristiaan Pelckmans, Sabine Van Huffel, and Johan A. K. Suykens. 2011. "Support vector methods for survival analysis: A comparison between ranking and regression approaches." *Artificial Intelligence in Medicine* 53 (2): 107–18. https://doi.org/10.1016/j.artmed.2011.06.006.

Van Belle, Vanya, K Pelckmans, Johan A. K. Suykens, and Sabine Van Huffel. 2011. "Learning Transformation Models for Ranking and Survival Analysis." *Journal of Machine Learning Research* 12: 819–62.

Van Belle, V, K Pelckmans, J A K Suykens, and S Van Huffel. 2010. "Additive survival least-squares support vector machines." *Statistics in Medicine* 29 (2): 296–308. https://doi.org/10.1002/sim.3743.

Van Houwelingen, Hans C. 2000. "Validation, calibration, revision and combination of prognostic survival models." *Statistics in Medicine* 19 (24): 3401–15. https://doi.org/10.1002/1097-0258(20001230)19:24%3C3401::AID-SIM554%3E3.0.CO;2-2.

———. 2007. "Dynamic prediction by landmarking in event history analysis." *Scandinavian Journal of Statistics* 34 (1): 70–85. https://doi.org/10.1111/j.1467-9469.2006.00529.x.

Vehtari, Aki, and Heikki Joensuu. 2013. "A Gaussian processes model for survival analysis with time dependent covariates and interval censoring." https://users.aalto.fi/$/sim$ave/VehtariJoensuu_GIST_CT_timing_poster_2013.pdf.

Vinzamuri, Bhanukiran, Yan Li, and Chandan K. Reddy. 2017. "Pre-processing censored survival data using inverse covariance matrix based calibration." *IEEE Transactions on Knowledge and Data Engineering* 29 (10): 2111–24. https://doi.org/10.1109/TKDE.2017.2719028.

Vock, David M, Julian Wolfson, Sunayan Bandyopadhyay, Gediminas Adomavicius, Paul E Johnson, Gabriela Vazquez-Benitez, and Patrick J O'Connor. 2016. "Adapting machine learning techniques to censored time-to-event health record data: A general-purpose approach using inverse probability of censoring weighting." *Journal of Biomedical Informatics* 61: 119–31. https://doi.org/https://doi.org/10.1016/j.jbi.2016.03.009.

Volinsky, Chris T, and Adrian E Raftery. 2000. "Bayesian Information Criterion for Censored Survival Models." *International Biometric Society* 56 (1): 256–62.

Wang, Hong, and Gang Li. 2017. "A Selective Review on Random Survival Forests for High Dimensional Data." *Quantitative Bio-Science* 36 (2): 85–96. https://doi.org/10.22283/qbs.2017.36.2.85.

Wang, Ping, Yan Li, and Chandan K. Reddy. 2019. "Machine Learning for Survival Analysis." *ACM Computing Surveys* 51 (6): 1–36. https://doi.org/10.1145/3214306.

Wang, Zhu. 2019. "bujar: Buckley-James Regression for Survival Data with High-Dimensional Covariates." CRAN. https://cran.r-project.org/package=bujar.

Wang, Zhu, and C Y Wang. 2010. "Buckley-James Boosting for Survival Analysis with High-Dimensional Biomarker Data." *Statistical Applications in Genetics and Molecular Biology* 9 (1). https://doi.org/https://doi.org/10.2202/1544-6115.1550.

Wei, L J. 1992. "The Accelerated Failure Time Model: A Useful Alternative to the Cox Regression Model in Survival Analysis." *Statistics in Medicine* 11: 1871–79.

Welchowski, Thomas, and Matthias Schmid. 2019. "discSurv: Discrete Time Survival Analysis." CRAN. https://cran.r-project.org/package=discSurv.

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wolpert, David H. 1992. "Stacked generalization." *Neural Networks* 5 (2): 241–59. https://doi.org/https://doi.org/10.1016/S0893-6080(05)80023-1.

Wright, Marvin N., and Andreas Ziegler. 2017. "ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R." *Journal of Statistical Software* 77 (1): 1—–17.

Xiang, Anny, Pablo Lapuerta, Alex Ryutov, Jonathan Buckley, and Stanley Azen. 2000. "Comparison of the performance of neural network methods and Cox regression for censored survival data." *Computational Statistics & Data Analysis* 34 (2): 243–57. https://doi.org/https://doi.org/10.1016/S0167-9473(99)00098-5.

Yang, Yanying. 2010. "Neural Network Survival Analysis." PhD thesis, Universiteit Gent.

Yasodhara, Angeline, Mamatha Bhat, and Anna Goldenberg. 2018. *Prediction of New Onset Diabetes after Liver Transplant*.

Zare, Ali, Mostafa Hosseini, Mahmood Mahmoodi, Kazem Mohammad, Hojjat Zeraati, and Kourosh Holakouie Naieni. 2015. "A Comparison between Accelerated Failure-time and Cox Proportional Hazard Models in Analyzing the Survival of Gastric Cancer Patients." *Iranian Journal of Public Health* 44 (8): 1095–1102. https://doi.org/10.1007/s00606-006-0435-8.

Zhang, Yucheng, Edrise M Lobo-Mueller, Paul Karanicolas, Steven Gallinger, Masoom A Haider, and Farzad Khalvati. 2020. "CNN-based survival model for pancreatic ductal adenocarcinoma in medical imaging." *BMC Medical Imaging* 20 (1): 11. https://doi.org/10.1186/s12880-020-0418-1.

Zhao, Lili, and Dai Feng. 2020. "Deep Neural Networks for Survival Analysis Using Pseudo Values." *IEEE Journal of Biomedical and Health Informatics* 24 (11): 3308–14. https://doi.org/10.1109/JBHI.2020.2980204.

Zhou, Zheng, Elham Rahme, Michal Abrahamowicz, and Louise Pilote. 2005. "Survival Bias Associated with Time-to-Treatment Initiation in Drug Effectiveness Evaluation: A Comparison of Methods." *American Journal of Epidemiology* 162 (10): 1016–23. https://doi.org/10.1093/aje/kwi307.

Zhu, Wan, Longxiang Xie, Jianye Han, and Xiangqian Guo. 2020. "The Application of Deep Learning in Cancer Prognosis Prediction." *Cancers* 12 (3): 603. https://doi.org/10.3390/cancers12030603.

Zhu, X, J Yao, and J Huang. 2016. "Deep convolutional neural network for survival analysis with pathological images." In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 544–47. https://doi.org/10.1109/BIBM.2016.7822579.

A. The first appendix