Outline (English)

This cumulative dissertation, containing three independent articles, contributes new methodologies for the estimation of combined effects across transition-types in the field of multi-state modelling. Two new algorithms are introduced: *Boosting Multi-State Models*, and *Structured Fusion Lasso Penalized Multi-State Models*. A flexible framework for simulation of event histories as the realization of a multi-state model is additionally provided.

Multi-state models are a statistical model class that allows to work with observations of processes where categorical scaled quantities, such as individual health states, change their states (called *transition/event*) over the course of a continuous time scale, for instance time since study origin. The goal of multi-state modelling is the quantification of timely dynamics that lead to the observed event histories. The concept in the analysis of event times that corresponds to the timely dynamics is the *hazard rate function*. In the multiplicative parametrization, the transition-type specific hazard rate function is modelled as a product of baseline hazard functions and a model term that proportionally accelerates or slows down this rate, or in other words: the effects of covariates are described by factors of proportionality. From here, several modelling paths are ready to be pursued. In the most prominent model class (*Cox-type semi-parametric partial-likelihood models*), the event specific baseline hazard rate functions are kept unspecified. This preserves misspecification with regard to the functional form.

This latter model class provides the basis to the 1st article, which is the establishment of the component-wise functional gradient descent boosting algorithm (short *boosting*) for multi-state models. Without making further assumptions on the relationship between covariates and transition-specific hazard rate functions, the number of parameters to be estimated here results from the product of the number of transitions and the number of covariates – if all covariates are binary or continuously scaled. With multi-level categorically scaled covariates, and/or potentially non-linear effects, the complexity of the parameter space is further increased, making numerical instability a reasonable problem. Furthermore, model choice, performed in classical supervised fashion, is very time as well as labour consuming. As a solution, the developed boosting algorithm for multi-state models overcomes these difficulties by the collective performance of decentralized weak estimators, where the minimization of the model's loss function (as the generalization concept to residuals sum of squares in Gaussian regression models) is performed by stepwise descending its gradient into the direction with respective current steepest descent. This is achieved by combining single component-wise and transition-type specific (or transition-type combined) regression models on a sequentially updated response. The use of the established R add-on package `gamboostMSM` is demonstrated, and results on simulated and real application data are given.

A second modelling approach (2nd article) is to use penalization methods by the application of suitable penalty terms. The aim here is to the estimate covariate effect coefficients equal to zero, or on an equal value for two coefficients that belong to the same covariate, but to two different transition-types. The first issue can be approached by the penalisation of the absolute values of the covariate coefficients during the estimation. The penalisation of absolute differences between coefficients of effects of the same covariate on different transition-types leads to sparse competing risk relations within a multi-state model, and concerns the second issue of equality of covariate effect coefficients. Using piece-wise exponential models, this concept is expandable to baseline hazard rate functions. Throughout this article, a new estimation approach providing sparse multi-state modelling by the above principles is established, based on the estimation of multi-state models and simultaneously penalising the L_1-norm of covariate coefficients and their differences in a structured way. The use of the established R add-on package `penMSM` is demonstrated, and results on real application data are given.

During the process of establishing a new method for multi-state modelling, or while analysing event history data as the observed outcome of a multi-state model, it is essential to verify the results by using simulated event history data of which the underlying ground truth is known. Therefore, one needs to draw random event histories as artificial observations. While it is clear, from a theoretical perspective, how to achieve the simulation of event histories as observations of a multi-state model, there exists a shortage in the availability of user-friendly and general software. The R add-on package `simMSM`, which is introduced throughout the 3rd article, contributes to the task of filling this gap. `simMSM` provides means to simulate event histories from a general multi-state model set-up, where the underlying multi-state model is parametrized by the products of the transition-type specific baseline hazard rate functions and the exponential of transition-type specific linear predictor terms. The latter factors potentially consist of time-varying effect functions and time-change covariates. The baseline hazard rate and (time-varying) covariate effect functions are potentially taken into account as non-linear functions. It is furthermore possible to incorporate dependencies on the past of the simulated event sequences.

Überblick (Deutsch)

Diese kumulative Dissertation setzt sich aus drei unabhängigen Artikeln zusammen und hat das Ziel neue Methoden für die Schätzung von Kovariableneffekten auf übergangsartspezifische Hazardratenfunktionen im Bereich der Mehrstadienmodellierung beizutragen. Hierfür werden insbesondere zwei neue Schätzalgorithmen eingeführt: *Boosting Multi-State-Models* und *Structured Fusion Lasso Penalized Multi-State Models*. Zusätzlich wird ein flexibler Rahmen für die Simulation von Ereignisgeschichten als Realisierungen eines Mehrstadienmodells gegeben.

Mehrstadienmodelle erlauben die statistische Analyse von Beobachtungen bei denen eine kategorialskalierte Größe, wie etwa individuelle Gesundheitszustände, ihren Zustand im Verlauf einer kontinuierlichen Zeitskala, zum Beispiel Zeit seit Studienbeginn, ändert (genannt Übergang/Ereignis). Das Ziel der Mehrstadienmodellierung ist die Quantifizierung der zeitlichen Dynamik die zu den beobachteten Ereignisgeschichten führt. Die Hazardratenfunktion ist dabei das Konzept der Ereigniszeitanalyse dass dieser zeitlichen Dynamik entspricht. In einer multiplikativen Parametrisierung wird sie modelliert durch das Produkt der Baselinehazardratenfunktionen mit einem Term der diese proportional beschleunigt oder verlangsamt. Von hier aus sind mehrere Modellierungsmöglichkeiten gegeben: In der meistangewandten Modellklasse (*Cox-type semi-parametric partial-likelihood models*) werden die ereignisartspezifischen Baselinehazardratenfunktionen nicht spezifiziert, was vor funktionaler Fehlspezifikation der Baselinehazardratenfunktionen bewahrt.

Die letztgenannte Modellklasse bildet die Grundlage des ersten Artikels, der die Entwicklung eines komponentenweisen, funktionellen Gradientenabstiegsalgorithmus (genannt *Boosting*) für Mehrstadienmodelle zum Ziel hat. Ohne weitere Annahmen über den Zusammenhang zwischen Kovariablen und übergangsartspezifischen Hazardratenfunktionen ergibt sich die Anzahl der zu schätzenden Parameter des Mehrstadienmodells hier als das Produkt der Anzahl der Übergangsarten und der Anzahl der Kovariablen – falls alle Kovariablen binär oder kontinuierlich skaliert sind. Mit mehrkategorialskalierten Kovariablen und/oder potentiell nichtlinearen Effekte wird die Komplexität des Parameterraums weiterhin vergrößert, so dass numerische Instabilität ein ernstzunehmendes Problem wird. Außerdem ist hier klassische Modellwahl sehr zeit- und arbeitsaufwendig. Als Lösung überwindet der entwickelte Boostingalgorithmus für Mehrstadienmodelle diese Schwierigkeiten durch die kollektive Leistung von dezentralen (kovariablen- und übergangsartspezifisch oder kovariablenspezifisch und übergangsartübergreifend) und für sich sehr einfachen Lernalgorithmen, wobei die Minimierung der Verlustfunktion des Modells (als Verallgemeinerungskonzept der Residuenquadratsumme des Gaußschen Regressionsmodells) durch einen schrittweisen Abstieg des Gradienten in die Richtung mit dem jeweils aktuell steilsten Abstieg erfolgt. Die Nutzung des implementierten R Zusatzpaketes `gamboostMSM` wird demonstriert und es werden Ergebnisse auf simulierten und echten Anwendungsdaten gegeben.

Eine weitere Möglichkeit (zweiter Artikel) um zu einem sparsamen Mehrstadienmodell zu gelangen ist die Bestrafung der (partiellen) Likelihood mit einem geeigneten Strafterm der die Schätzung von Kovariableneffekten auf den Wert Null, oder auf den gleichen Wert für zwei Koeffizienten die zur selben Kovariablen und zu zwei verschiedenen Übergangsarten gehören, ermöglicht. Die erste Eigenschaft kann durch die Bestrafung der Absolutwerte der Effektkoeffizienten erreicht werden. Die Bestrafung der absoluten Differenzen zwischen den Effektkoeffizienten derselben Kovariablen für verschiedene Übergangsarten führt zu sparsameren Übergangsartbeziehungen innerhalb eines Mehrstadienmodells, und bezieht sich auf die zweite Eigenschaft der Modellregulariserung. Durch die Verwendung eines stückweise exponentiellen Modellansatzes ist dieses Konzept erweiterbar auf Baselinehazardratenfunktionen. In diesem Artikel wird ein neuer Schätzansatz zur sparsamen Mehrstadienmodellierung nach den oben genannten Grundsätzen definiert, basierend auf der Bestrafung der absoluten Kovariablekoeffizienten und deren Differenzen auf strukturierte Weise. Die Nutzung des implementierten R Zusatzpaketes `penMSM` wird demonstriert. Weiterhin wird der neue Modellierungsansatz an einer realen Anwendungssituation illustriert.

Während der Methodenentwicklung zur Mehrstadienmodellierung, oder auch bei der Analyse von beobachteten Ereignisgeschichten, ist es hilfreich die Ergebnisse anhand simulierter Beobachtungen zu verifizieren, da hier der datengenerierende Prozess bekannt ist. Es müssen dafür Ereignisgeschichten als künstliche Beobachtungen gezogen werden, wofür aus der Praxisperspektive ein Mangel an benutzerfreundlicher und allgemeiner Software vorherrscht. Das R Zusatzpaket `simMSM`, welches im dritten Artikel beschrieben wird, hat zur Aufgabe diese Lücke zu schließen: `simMSM` stellt Funktionen zur Simulation von Ereignisgeschichten eines allgemeinen Mehrstadienmodellaufbaus zur Verfügung, wobei das zugrunde liegende Mehrstadienmodell durch die Produkte der übergangsartspezifischen Baselinehazardratenfunktionen und der exponentiellen übergangsartspezifischen linearen Prädiktoren parametrisiert wird. Die letzteren Faktoren können dabei potenziell aus zeitlich variierenden Effektfunktionen und sich zeitlich ändernden Kovariablen bestehen, die Baselinehazardratenfunktionen und (zeitvariablen) Kovariableneffektfunktionen können als nichtlineare Funktionen in Betracht gezogen werden. Es ist weiterhin möglich, Abhängigkeiten von der Vergangenheit der simulierten Ereignissequenzen zu integrieren. Die Nutzung des implementierten R Zusatzpaketes `penMSM` und dessen kompletter Funktionalität wird anhand mehrerer verschiedener Simulationsszenarien demonstriert.