Flowgraph Models: Models for Multistate Time-to-Event Data

Aparna V. Huzurbazar
Department of Mathematics and Statistics
University of New Mexico
Albuquerque, NM 87131-1141
USA
aparna@stat.unm.edu

Abstract

Semi-Markov models play an important role in the analysis of time to event data. However, in practice, data analysis for semi-Markov processes can be quite difficult and many simplifying assumptions are made. Markovian multistate models are popular for event history analysis and repeated events analysis for survival data. Semi-Markov processes provide a rich class of models applicable to this area. Flowgraph models are multistate models that provide a data analytic method for semi-Markov processes. Flowgraphs are useful for estimating Bayes predictive densities, predictive reliability functions, and predictive hazard functions for waiting times of interest in the presence of censored and incomplete data. While multistate models have been used primarily in medical research, flowgraph models have been used to model complex systems such as cellular telephone networks, construction engineering projects, and manufacturing systems, in addition to modeling disease progression for diseases such as cancer and AIDS, and for degenerative diseases such as diabetic retinopathy and kidney failure. I will present an introduction to flowgraph models, discuss Bayesian inference based on flowgraph models, and the relationship to multistate models and semi-Markov processes. Recent work with flowgraphs concerns posterior sampling based on constructed likelihoods in the presence of incomplete data.

1 Background on Flowgraph Models

Flowgraph models are useful for modeling time to event data that result from a stochastic process. A flowgraph is a graphical representation of a stochastic system that models potential outcomes, probabilities of outcomes, and waiting times for those outcomes to occur. Figure 1 is a flowgraph model for a system in a process plant. State 0 represents the functioning state. State 1 represents a noncritical failure. Noncritical

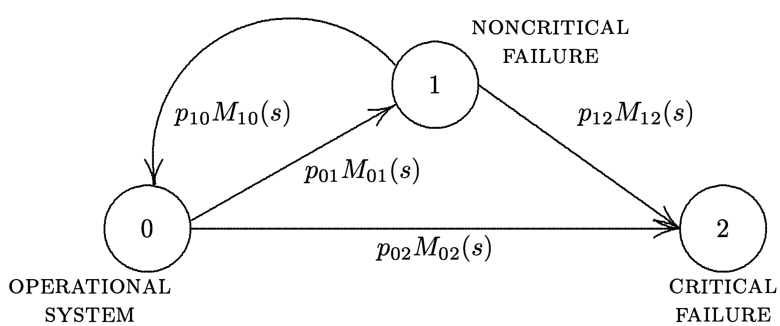

Figure 1: Flowgraph model for system failure in a process plant.

failures are failures that can be repaired without system downtime. Transition from state 1 to state 0 represents repair, returning the system to a fully operational state. State 2 represents critical failure. Critical failures are failures that cause system downtime. Such a model is useful for modelling a variety of systems at the macro level for example, cellular telephone networks or emergency response networks.

While block diagrams and signal flowgraphs are widely used to represent engineering systems, they do not incorporate probabilities, waiting times, or data analysis. The literature on flowgraph methods in engineering is vast beginning with Mason (1953). Introductions to flowgraph methods are contained in most circuit analysis or control systems textbooks such as D'Azzo and Houpis (1981), Gajic and Lelic (1996) and Whitehouse (1977). Statistical flowgraph models are based on flowgraph ideas but unlike their antecedents, flowgraph models can also be used to model and analyze data from complex stochastic systems. Early work on statistical flowgraph models focused on applications to survival analysis and includes Butler and Huzurbazar (1997) and Huzurbazar (1999). Yau and Huzurbazar (2002) present flowgraph methods for a detailed application to diabetic retinopathy. Huzurbazar (2004) presents a review of flowgraph models for use in survival analysis. Huzurbazar (2000a and b) provide applications to engineering systems. In particular, Huzurbazar (2000a) gives details on using flowgraph models for computing the total and partial system failure for cellular telephone networks. The article also deals with data analysis for general queueing systems using flowgraph models and considers M/M/q queues as a special case, although the methods presented are for general semi-Markov processes, including M/G/1 queues. Flowgraphs are also distinct from graphical models in that the states represent outcomes rather than variables. For example, while feedback loops are an integral part of flowgraph models, they are redundant in a graphical model.

In a flowgraph model, the states, representing outcomes, are connected by directed line segments called *branches*. These branches are labeled with *transmittances*. These transmittances are labelled with the "transition probability $\times$ moment generating function (MGF) the of waiting time distribution in the previous state" which is a quantity called the *branch transmittance*. In the figure, probabilities and moment generating functions of the waiting time distributions are shown as branch transmittances. The waiting times on the branches can be any parametric distributions that admit moment generating functions. Hence, the model is quite general in that exponential assumptions or assumptions that the waiting for the various branches be from the same family of distributions (for tractability) are not made. For example, in Figure 1, the branch waiting time for state 0 to state 1 could be inverse Gaussian, the branch waiting time for state 0 to state 2 could be gamma, and the branch waiting for state 1 to state 2 could be Weibull.

We use the branch transmittances of a flowgraph model along with *flowgraph algebra* to solve for the MGF of the distribution of the waiting time of interest. This MGF is converted to a density, reliability, or hazard function using a saddlepoint approximation (cf. Daniels (1954)). The end result from a flowgraph model is a Bayesian predictive distribution of the quantity of interest. Quantities of interest include predicting the distribution of the total time to critical failure, $0 \to 2$; predicting the waiting time to repair, say, $1 \to 0$; predicting the time to noncritical failure $0 \to 1$; or predicting the total number of times the system reaches noncritical failure and is repaired, i.e. the total number of times the transition $1 \to 0$ is made.

Flowgraphs models can be viewed as a type of semi-Markov multistate model. Multistate models are used in survival analysis to describe longitudinal, time to event data. Hougaard (1999) provides a review. They model stochastic processes that progress through various stages. In terms of data analysis, multistate models have been restricted to the realm of Markov models. In a Markov multistate model, given the current state of the process, the transition time to a future state does not depend on the past history of the process. At the initial state of a Markov process, the transition time is a minimum of the waiting time distributions corresponding to all possible transitions from the initial state. Hence, in practice, for tractability and analytical convenience, exponential distributions are assumed. Occasionally, with appropriate parametric restrictions, Weibull distributions are used (cf. Wilson and Soloman (1994)), exploiting the fact that the minimum of independent and identically distributed Weibulls is again a Weibull distribution.

A semi-Markov multistate model allows the transition time to a future state to depend on the duration of time spent in the current state. In practice, it is quite difficult to analyze data for semi-Markov multistate models. One method of analysis for multistate models consists of combining independent submodels for each *transition intensity*, a method that restricts the analysis to models with unidirectional or progressive flow (cf. Andersen and Keiding (2002)). In fact, Hougaard (1999) states that for non-unidirectional or non-progressive multistate models, it is impossible to obtain general formulas for transition probabilities for models where the hazard is allowed to depend on the history in any way. Flowgraphs work in the MGF domain and circumvent this difficulty.

Another method is the use of the popular proportional hazards model. The proportional hazards model as used in medical statistics was developed by Cox (1972); however, the idea of assuming proportional hazards to fit more parsimonious models dates back to the operations research literature (cf. Allen (1963)). Cox's model is semi-parametric and assumes that the intensity of the counting process is a product of a parametric function of the covariates and an arbitrary function of time, hence its designation as "proportional hazards". This method is also restricted to a unidirectional or progressive multistate model. The obvious restriction of the proportional hazards model is that hazards are not always proportional. In both methods, the key approach for analyzing such multistate models is based on modeling the hazard function, a quantity that is not directly observable. The end result is a hazard function model based on a set of covariates which can be converted to a survival function if required. There are many extensions of this model and further details can be found in survival analysis texts such as Klein and Moeschberger (1997) and Therneau and Grambsch (2000).

The flowgraph methodology allows for a variety of distributions to be used within the stages of the multistate model and also easily handles reversibility. In a medical context, this means that in a progressive disease, a patient is allowed to improve at times. In the engineering context, this means that a failed component can be repaired. Flowgraphs model the observable waiting times rather than the hazards and as such, they do not directly make any assumptions about the shape of the hazard. The end result from a flowgraph analysis are Bayes predictive densities, CDFs, survivor or reliability functions, and hazard functions of the waiting times of interest. If one prefers, maximum likelihood estimation is also available. Flowgraphs also handle censored and incomplete data.

Complete data on a flowgraph model consists of having every intermediate transition for each observation. The associated waiting time for the transitions may be censored, however, the transition information is complete. In Figure 1, complete data consists of observations such as $0 \to 1 \to 0 \to 1 \to 2$ or $0 \to 2$ where we know that the system trasited directly to a critical failure.

Incomplete data consists of data that have complete information on observed waiting times but incomplete information on the associated transitions. For example, in Figure 1, if we observe a waiting time such as $0 \to 2$ but know that the system experienced a noncritical failure before a critical failure, we expect that the transition is $0 \to 1 \to 2$ but that the transition 1 is incomplete. This leads to another type of data called *unrecognizably incomplete data*. Unrecognizably incomplete data are data that appear to have complete information on the transitions and waiting times but in reality are incomplete with respect to transition information. For example, suppose that we observe $0 \to 1 \to 2$, we might assume that the data are complete. But perhaps the true observation transitions were $0 \to 1 \to 0 \to 1 \to 2$ with the first 1 and second 0 missing.

Current research on flowgraph models is concerned with various strategies for posterior simulation basedon constructed likelihoods. This talk will present an introduction to multistate models and flowgraph models in the context of semi-Markov processes. The focus will be on data analysis for such processes in the presence of censored and incomplete data. Time permitting, methods for uncrecognizably incomplete data will also be presented.

References

Allen, W. R. (1963). A note on conditional probability of failure when hazards are proportional. *Operations Research*, 11, 658-659.

Andersen. P. K., and Keiding, N. (2002). Multi-state models for event history analysis. *Statistical Methods for Medical Research*, 11, 91-115.

Butler, R.W. and Huzurbazar, A.V. (1997). Stochastic network models for survival analysis. *Journal of the American Statistical Association*, *92*, 246-257.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society Series B*, 34, 187-220.

Daniels, H. (1954). Saddlepoint approximations in statistics. *Annals of Mathematical Statistics*, *25*, 631-650.

D'Azzo, J. and Houpis, C. (1981). *Linear Control System Analysis and Design : Conventional and Modern*, New York : McGraw-Hill.

Dorf, R. and Bishop, R. (1995). *Modern control systems.* Reading, Mass. : Addison-Wesley.

Gajic, Z., and Lelic, M. (1996). *Modern control systems engineering.* New York : Prentice Hall.

Hougaard, P. (1999). Multi-state models: a review. *Lifetime Data Analysis*, 5, 239-264.

Huzurbazar, A. V. (1999). Flowgraph models for generalized phase type distributions with non-exponential waiting times. *Scandinavian Journal of Statistics*, *26*, 145-157.

Huzurbazar, A. V. (2000a). Modeling and analysis of engineering systems data using flowgraph models. *Technometrics,* 42, 300-306.

Huzurbazar, A. V. (2000b). Modeling Time-to-Event Data using Flowgraph Models. In *Advances on Methodological and Applied Aspects of Probability and Statistics*, edited by N. Balakrishnan, 561-571. New York: Taylor and Francis.

Huzurbazar, A. V. (2004a). Modeling survival data using flowgraph models. In *Advances in Survival Analysis (Handbook of Statistics, Volume 23)*, 729-746. Edited by N. Balakrishnan and C. R. Rao. Amsterdam: Elsevier.

Klein, J., and Moeschberger, M. (1997). *Survival Analysis: Techniques for Censored and Truncated Data.* New York: Springer-Verlag.

Mason, S. J. (1953). Feedback theory—Some properties of signal flow graphs. *Proceedings of the Institute of Radio Engineers*, 41, 1144-1156.

Therneau, T., and Grambsch, P. (2000). *Modeling Survival Data: Extending the Cox Model.* New York: Springer.

Whitehouse, G. (1977). *Systems analysis and design using network techniques.* New Jersey: Prentice-Hall.

Wilson S., and Solomon P. (1994). Estimates for different stages of HIV/AIDS disease. *Computations in Applied Biosciences*, 10(6), 681-683.

Yau, C. L., and Huzurbazar, A. V. (2002). Analysis of censored and incomplete survival data using flowgraph models. *Statistics in Medicine*, 21 , 3727-3743.