# Semiparametric Estimation of a Gaptime-Associated Hazard Function

**Timothy Teräväinen**

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

**COLUMBIA UNIVERSITY**

2014

# ABSTRACT

## Semiparametric Estimation of a Gaptime-Associated Hazard Function

## Timothy Teräväinen

This dissertation proposes a suite of novel Bayesian semiparametric estimators for a proportional hazard function associated with the gaptimes, or inter-arrival times, of a counting process in survival analysis. The Cox model is applied and extended in order to identify the subsequent effect of an event on future events in a system with renewal. The estimators may also be applied, without changes, to model the effect of a point treatment on subsequent events, as well as the effect of an event on subsequent events in neighboring subjects.

These Bayesian semiparametric estimators are used to analyze the survival and reliability of the New York City electric grid. In particular, the phenomenon of "infant mortality," whereby electrical supply units are prone to immediate recurrence of failure, is flexibly quantified as a period of increased risk. In this setting, the Cox model removes the significant confounding effect of seasonality. Without this correction, infant mortality would be misestimated due to the exogenously increased failure rate during summer months and times of high demand. The structural assumptions of the Bayesian estimators allow the use and interpretation of sparse event data without the rigid constraints of standard parametric models used in reliability studies.

# Table of Contents

# List of Figures

Notation

| | |
|---|---|
| $t_j$ | $j^{\text{th}}$ aggregated failure time |
| $i(t_j)$ | unit to fail at time $t_j$ |
| $t_{i,j}$ | $j^{\text{th}}$ failure time of unit $i$ alone |
| $u_i$ | grid points for finite approximations of calendar time |
| $\tau_{i,t}$ | most recent prior failure time for unit $i$ at time $t$ |
| $\nu_{i,t}$ | time since failure; $\nu_{i,t} = t - \tau_{i,t}$ |
| $v_i$ | grid points for finite approximations of gaptimes |
| $t_{i,j,\nu}$ | calendar time at which unit $i$ had its $j$th failure exactly $\nu$ ago |
| $\lambda_i(t)$ | hazard rate for unit $i$ at time $t$ |
| $\Lambda_i(t) = \int_0^t \lambda_i(t)\ \mathrm{dt}$ | cumulative hazard |
| $f_i(t); F_i(t); S_i(t)$ | density; distribution; and survival functions |
| $m_i(t)$ | mean residual life at time $t$ |
| $\lambda_0(t)$ | baseline hazard |
| $\lambda^*, \lambda_0^*$ | prototype hazards for prior distributions |
| $\psi(\nu_{i,t}) = e^{\phi(\nu_{i,t})}$ | infant mortality hazard |
| $\widehat{\cdots}$ | estimator of $\cdots$ |
| $\beta$ | regression coefficients |
| $Y_i(t)$ | at-risk indicator for unit $i$ at time $t$ |
| $\mathfrak{R}(t) = \sum_i Y_i(t)$ | total at-risk population at time $t$ |

I returned, and saw under the sun,
that the race is not to the swift,
nor the battle to the strong,
neither yet bread to the wise,
nor yet riches to men of understanding,
nor yet favour to men of skill;
but *time and chance* happeneth to them all.

Eccl. 9:11

# Chapter 1

# Introduction

We analyze the survival and reliability of the New York City electric grid. Despite increasing automation and development, the reliability of the electric grid has not improved. In fact, the number of severe power outages doubles every five years, currently resulting in about $50 billion of "outage costs" every year.[1].

In New York City[1], underground distribution feeders operating at 13 kilovolts (Brooklyn and Queens) or 27kV (Bronx and Manhattan) relay electricity between substations and transformers. These transformers generate three-phase[2] 440V alternating current (AC) electricity for business and res-

[1]Staten Island is omitted from consideration because above-ground distribution feeders are used only in that borough. However, due to the significantly greater effect of Hurricane Sandy on power availability (Oct. 30, 2012) in Staten Island, Consolidated Edison is considering "undergrounding" the grid there.

[2]The term three-phase refers to the configuration of the AC power. Three wires each transmit sinusoidal AC of root-mean-square (RMS) power $\frac{x}{\sqrt{3}}$, with each phase offset by $120\,^\circ$ from any other ($240 \equiv -120 \bmod 360$). The benefit is that each wire can be thinner, thus cheaper, while two wires (phases) may be short-circuited for the full power $x$. Additionally, the total root-mean-square power of the three phases combined, $x\sqrt{\sin(t)^2 + \sin(t - 2\pi/3)^2 + \sin(t - 4\pi/3)^2 \mathrm{dt}} = x\sqrt{\frac{3}{2}}$, is stationary in time. This is of ap-

idential use. The transformers then supply this power to local substations which further distribute the electricity to nodes being manholes and supply boxes. These nodes are interconnected by mains cables, supplying power to customers. The most prevalent form of failure on the secondary is a break in a mains cable, called an "open main," a condition which often goes unreported due to the size and redundancy of the secondary network.

The feeders are cable systems which connect the transformers to the secondary network. Due to the mechanical construction of the cable and the severe fluctuations of power, feeders are the most failure-prone electrical components in the power grid. The feeder system is monitored by a Supervisory Control and Data Acquisition (SCADA) system called RMS[3] with reporting rates varying between 5 seconds and 20 minutes. Although RMS is able to report failures quickly, the causes of failure are complex and only partially known. These include mechanical failures and the engaging of automatic circuit breakers which isolate the system, preventing further burnout of further components. The causes of failure are varied and often impossible to isolate, at least without an extremely expensive updating of the RMS SCADA system.

Additionally, feeder failures are associated with failures in the secondary network (personal communication). Due to the size and complexity of the secondary networks, there is no formal real-time monitoring system. Although an inspection program is in place to audit these networks, the city-wide scale, involving a total of five million nodes, is so large that any single component may remain unexamined for years.

---

plication in converting to DC power ("rectifying"). The stationarity of three-phase power is also ideal for powering electrical motors, as the flux of the three phases together produces a stationary rotating magnetic field.

[3]This is not to be confused with root-mean-square.

However, adverse events are reported by customers for maintenance in "trouble tickets." These tickets provide partial information about failures mostly in the form of "remarks" left by utility personnel. The text of these remarks have been data-mined using various machine learning techniques under the rubrics of "knowledge discovery" and "feedback elicitation."[19, p. 719][21, p. 1]

Since the failures on the secondary network are associated with the failure of the associated feeder, it is of further interest to quantify and predict the failure of feeders. In addition to preventing damage to feeders, proper scheduling of feeder maintenance will promote the stability of the secondary network.

However, there is only partial information available about the feeders themselves. Given this lack of information about the specific causes of failure, rules of the form:

> "Repair the feeder up to three times, or until a failure occurs within 24 hours of the previous failure. In either case, replace the feeder entirely."

would be useful. Such rules would need to be inferred from associations between the precondition of the rule and the frequency of failures of the feeder.

Quantifying the immediate effect of a failure is also useful; for example, if, on average, a feeder is eight times as likely to fail immediately after a repaired failure, this can be used to prioritize further maintenance on that feeder, and as a guideline to curtail customer load in the secondary network supplied by the recently-repaired feeder.

This effect is commonly called "infant mortality" in epidemiology and reliability theory. Here it is extended to an online setting, where decisions are made in real-time and subject to partial information. We also desire a framework compatible with machine learning approaches, as there is a significant

amount of explanatory information available with no clear *a priori* model to relate the information to the observed frailty of the feeders being considered.

Finally, a large amount of information is available about the composition of individual feeders and mains. For each feeder, 133 covariates[4] are available, split roughly evenly between component makeup, e.g., *percentage of paper-insulated cables* and *total number of associated transformers*; and observational data, e.g., *average historical load reported by the RMS SCADA system.*

## 1.1 Modeling of Failure Rate

The objective of interest is often the reduction of the *mean time between failures*, or MTBF, of a feeder. A related quantity is the *mean time to failure*, or MTTF. The MTTF is a more general quantity; for example, one may consider the MTTF given that one week has elapsed since the last failure. The MTBF is, specifically, the MTTF at the time that a failure has just occurred. Note that both MTBF and MTTF may be subject to further conditions; for example, one may estimate the MTTF during summer months, or among feeders of a certain composition, or both.

Treatments are suggested and evaluated based on the expected increase in the MTBF, controlling for covariates and correcting for past treatment history. Although a simple sounding objective, faithful modeling of failure rates is difficult under realistic assumptions, as for example when explanatory covariates and temporal heterogeneity are known to exist.

To illustrate these difficulties, consider a simple case, where to evaluate the

[4]Sixty features are removed due to perfect correlation with other features, while another 47 are removed due to high correlation ($\rho > 0.95$) with similar features. For example, *Proportion of Paper Joints* and *Number of Paper Joints* will be redundant since *Total Number of Joints* is also present. After removing these features, only 26 remain.

effect of a treatment, the number of failures before and after the treatment are compared.

Specifically, the number of failures in a given feeder during predetermined periods before treatment, of length $t_L$, and after, of length $t_R$, is recorded. Letting the number of recorded failures in each period be $n_L$ and $n_R$, estimates of the failure rate before and after treatment are given by

$$\hat{\lambda}_L = \frac{n_L}{t_L},$$
$$\hat{\lambda}_R = \frac{n_R}{t_R},$$

and a simple summary statistic

$$\hat{\delta}_\lambda = \hat{\lambda}_L - \hat{\lambda}_R$$

records the empirical change in failure rate.

Although this change is well-defined, it does not necessarily estimate the effect of the treatment. For example, if one of the $n_L$ failures causes unnoticed damage to some of the components of the feeder, there will be a lurking increase in the number $n_R$ of failures after treatment, which will depress (or even negate) the estimate $\hat{\delta}_\lambda$. Additionally, if the treatment is administered on the basis of the observed data $n_L$, for example if an unacceptably high number of failures due only to chance is observed, the estimate $\hat{\delta}_\lambda$ will be inflated. For these reasons, certain conditions must hold for the estimate $\hat{\delta}_\lambda$ to be informative about the effect of the treatment.

It is also known that failures occur more frequently in the summer months, due to thermal stress and increased customer demand for air conditioning. Thus, although a strong treatment effect may be estimated, it is likely due to these confounding factors, rather than the treatment itself.

Apart from fundamental issues of bias and confounding, the difficulty in deciding on failure rate models among practitioners has been noted.[7] Many of

the parametric models used in practice are unjustified and, worse, inadequate for complex, repairable components. Apart from the difficulty in modeling the effect of complex interacting predictors, the assumptions behind parametric models are often stronger than realized or wanted.

For instance, the Weibull model, a staple of reliability studies, appears particularly ill-suited to the problem at hand (as indicated subsequently), producing flawed estimates of survival time, particularly in the long-term. That is, the estimated continued survival time for a feeder which has already survived for a long time will be more and more inaccurate.

On the other hand, elaborate machine learning methods have been developed, but typically only within the paradigm of binary classification and a few of its extensions. Thus, these methods are rarely directly applicable to handling recurring failure data and thus the data must be preprocessed, sometimes severely, to allow their use.

Finally, the use of these methods often ignores causality; failures affect maintenance policy, which has an effect on future failures. These considerations are non-trivial.

## 1.2 The Smart Grid

The concerns above apply to any electric network (and with slight modifications, many other networks). One particular development in this field is, however, the so-called "Smart Grid." The Smart Grid is an evolution of the power grid to accomodate two primary novel technologies: ubiquitous real-time sensors and external information, and the distribution of energy from alternative power sources which lack the reliability of fossil fuel and nuclear power plants.

The sensors provide continuous immediate data on the conditions of the

grid; for instance, the failure of components, and the current local power demand. External information may consist of weather conditions and other quantities which predict power demand or affect the grid, and components which have failed or which are in critical condition.

Alternative energy sources such as wind or solar can supplement the grid when the weather is suitable, and replenish energy storage devices such as batteries, capacitors, and mechanical stored-energy systems such as pumped-storage hydroelectricity, flywheels, and molten salt thermal storage.

Batteries, although relatively inefficient, may be used to store and provide excess energy and to power external resources. For the latter reason, batteries are expected to be widely available and of low marginal cost for the utility company, since the utility company will be leasing a resource which they are supplying.

For instance, a fleet of self-driving electric vehicles can serve both as a system of transport and an off-line energy storage system. The vehicles would generally store power and perform deliveries during off-peak usage hours, while supplying power to the grid during peak hours.

The scale of such a system would provide flexibility to the power system, while posing many difficulties of coordination and scheduling for both the energy suppliers and the operators of the fleet. These difficulties may be ameliorated through accurate forecasting using the sensors and external information above. The proper use of these data motivates this thesis.

Larger privately-owned on-site batteries, solar panels, and mechanical or thermal storage devices can also provide lesser flexibility with greater reliability, in a role complementary to both traditional power and the fleet of batteries described above.

### 1.2.1 Criteria and Actions

A usable system for forecasting and scheduling would require:

1. Information about the state of the grid and its auxiliary components; within the grid, the capacity of generators and the reliability of components; outside the grid, information about the capacity and demand of batteries and other storage devices as above.

2. Forecasts of uncertain exogenous factors. These factors include local weather, which affects both customer demand and the supply capacity of green sources; the overall electricity market, which will influence electricity trading strategy and thus demand, and finally local customer demand.

3. A model of the decisions and actions available to various actors. The central power utility may curtail power either in an emergency, or by using economic incentives. Emergency generators and repair services may also be deployed in response to adverse grid conditions. Batteries may be charged or discharged, or prices may be set accordingly to affect such behavior by external actors as above.

4. The ability to simulate future states of the system, and to optimize certain objective functions, both to guarantee a contracted service level and to provide exploratory predictions for either human use or automatic planning.

These criteria are implemented in a feedback-control system, described below.

### 1.2.2 Prior Work

This work is a contribution to the Adaptive Stochastic Control Dynamic Treatment Controller codeveloped at the Columbia Center for Computational Learning Systems with Princeton's Castle Laboratory for decision theory and optimization. The controller uses approximate dynamic programming to predict problems and recommend optimal actions to be undertaken in response.[18]

The Controller aims to predict the expected change in a feeder's longevity based on the following information:

1. The identity of the open main in the secondary network associated with the feeder's failure, as obtained from the customer report above.

2. A "centrality score" which summarizes the topological proximity of the feeder to the open main.

3. Static attributes associated to the main and feeder, as described above.

## 1.3 Approaches

The ranking approach is popular in machine learning (ML) applications, partly due to the development of general algorithms based on various models, both generative and non-generative. In particular, methods based on support vector machines (SVMs) and decision trees are popular.

A common limitation of these methods is that they are direct adaptations of algorithms intended for classification of stationary data. In particular, although the data under consideration can naturally be modeled as a stochastic process, these algorithms require the reduction of data to sets of training and test examples (and possibly even additional sets for the selection of parameters by cross-validation).

Adaptations have been made to these methods in an effort to accomodate non-stationarity of the data-generating model. In the machine learning setting, this non-stationarity is called "concept drift."[5] The adaptation of fixed classifier models generally proceeds by either "windowing" the data or applying ensemble methods.

### 1.3.1 Windowing

Windowing is simply the training of a model on a subset of data, hopefully representing the conditions for the period of interest. For example, in the electric grid setting, predictions for summer months will be based on models which are trained only on historical summer months.

Windowing often produces an incidental benefit of reduced run-time (due to the reduction of the dataset) and can indeed improve predictive ability. However, the limitation is obvious; in this example, the data from the summer and winter months are completely isolated. It is natural to imagine adding a hyper-parameter to the model, which interpolates from the summer predictions to the winter predictions. In the absence of this hyper-parameter, it is approximated by restricting the fit of the model through the data. An example of a hyper-parameter would be a weighting function, which would make use of the winter data when it is helpful to do so. A mixture of models as a continuous function of time would be a generalization of windowing.

[5]This is due to the originating application to information retrieval and search engines, where the keywords associated with a fixed concept may nonetheless change over time. For example, the artificial intelligence field has drifted from rule-based systems, to neural networks, to genetic algorithms, to SVMs and Bayesian and/or boosted mixtures of models. Likewise, many AI subfields have been relabeled as ML.

### 1.3.2 Boosting

Boosting is a more advanced concept based principally on an adaptive training regime. Boosting operates, again, on the principle of adjusting the model by weighting the data. Specifically, boosting will iteratively reweight subsets of data in *inverse* proportion to the predictive accuracy of the current model on those subsets. These weights are then used to train a new classifer which is then combined into the ensemble of models. A variety of perspectives on boosting have been developed in the computer science literature, particularly in game theory as the steps of pessimistic data weighting followed by model optimization correspond to an approximate minimax algorithm. Recently, statistical perspectives on boosting have been developed.

Boosting was developed for the stationary setting, to adjust models as necessary by repeated iterations over the same data, applying at each iteration a greater weight to the misclassified instances. Boosting has been adapted to longitudinal data. By first discretizing time, and then applying boosting at time $t_{i-1}$ to adjust the model of time $t_i$, the boosted model adapts to changing conditions, albeit with a lag. However without further restrictions or modeling methods, this lag will be persistent and cannot exploit seasonal trends which have been noted in the electric grid.

## 1.4 Interpretability

Sophisticated machine learning models are often criticized for a lack of "interpretability." We contend that combining these machine learning models with the counting process framework for survival analysis provides a new and more flexible formulation which also allows a pragmatic interpretation.

Specifically, since the details of a classifier are not of operational interest,

they can be used to generate failure rates and failure probabilities, which are both interpretable and of direct utility in optimization and policy design. Since the binary classification is often not of primary interest, statistical methods can combine classification models to generate probabilities and thus sometimes provide an interpretation of interest.

## 1.5 New York City

The New York City feeder grid is the focus of current investigations. The power distribution network of New York City bridges the high-voltage grid, supplied from power generation stations, through step-down transformers to the lower-voltage household voltage secondary system. The network is hierarchical; power flows from stations to substations and to local transformers, which feed local grids supplying power to residences and businesses.

The feeder system comprises the cables which feed the local transformers. The feeder cables are prone to disconnection (opening) for many reasons, including activation of automatic safety relays on the substation-side, intentional disconnection due to failure of test conditions, and scheduled and unscheduled maintenance. In general, we refer to these events, excepting scheduled maintenance, collectively as "failures," with further specification when necessary.

The emergency relay disconnections, called "Open Autos" or simply "Autos," are of particular interest, as they must be predicted in order to safely schedule the other forms of maintenance. The feeder cables comprise a bottleneck and their failures, both automatic and intentional, adversely affect the downstream (consumer-side) network. Further, since these adverse effects are due to power flows damaging equipment, these effects are intensified in the event of multiple simultaneous failures. Thus, the modeling and prediction of feeder failure is a significant problem, the solution of which would provide an

effective knowledge-base for scheduling and optimization.

# Chapter 2

# Introduction to Survival Analysis

We begin with a review of standard survival analysis. The objective of survival analysis is to study the time to an event, when that time is not necessarily fully observed for each subject. This loss of information is known as *censoring*, and must be modeled correctly if meaningful inferences about the time to event are to be made.

The most prevalent form of censoring is *right-censoring*. Here, subjects enter a study at a common time, but may be lost to follow-up before the event of interest occurs. For instance, the study may end at a pre-determined time, at which not all of the subjects have experienced the event. Similarly, a subject may withdraw or be removed from the study for various reasons, in which case the mechanism causing the removal must itself be modeled, as well as the time to event.

Other forms of data reduction are frequently modeled as well. For example, *interval censoring*, wherein for each subject it is only observed that each time of event falls within an interval $(a, b]$. This interval may itself be random, as

was the case with withdrawals resulting in right-censoring. In fact for right-censoring, the corresponding interval is $(0, t]$ when the event is observed to occur at $t$, and $(t, \infty)$ when, instead, the subject withdraws at $t$, before the event occurs.

The standard survival framework assigns to each of $n$ subjects, two latent *independent* variables $(\widetilde{T}_i, C_i)$, denoting the failure and censoring times, respectively. As usual, it is assumed that subjects are independent of one another and, unless otherwise specified, that they are identically distributed. Specifically, we assume that the $(\widetilde{T}_i, C_i)$ are mutually independent, so that for any sequence of sets $(A_{i,j})_{i\in\{1,2,\ldots,n\},j\in\{1,2\}}$,

$$\Pr\left[\bigcap_{i=1}^{n}\{\widetilde{T}_i \in A_{i,1}\} \cap \{C_i \in A_{i,2}\}\right] = \prod_{i=1}^{n} \Pr[\widetilde{T}_i \in A_{i,1}] \times \Pr[C_i \in A_{i,2}].$$

A subject is considered to be censored when $C_i \leq \widetilde{T}_i$ and otherwise ($C_i > \widetilde{T}_i$) to have had an observed event. This convention is meaningful in a discrete setting, where it means that, for example, the censoring for day $i$ of observation occurs before the possibility of an event on day $i$.

The event $C_i > \widetilde{T}_i$ is indicated by the random variable $\Delta_i$ below, which will be called the *event indicator*. As usual, denote by $\delta_i$ an observation of the random variable $\Delta_i$.

An event indicator takes the value 1 if the event occurs and 0 otherwise. Thus, indicator variables are independent Bernoulli random variables, with common parameter $p_e$ being the probability of an event, $\Delta_i = 1$. Analogously, define $p_c = 1 - p_e$ as the probability of censoring, $\Delta_i = 0$. For ease of notation, define $\overline{\Delta}_i = 1 - \Delta_i$ and $\overline{\delta}_i = 1 - \delta_i$.

Due to the independence of subjects, each event indicator $\Delta_i$ is independent of the other indicators, $\Delta_{j\neq i}$, with parameter $p_e$ the probability of the event. Analogously, define $p_c = 1 - p_e$ the probability of censoring, so that each $\overline{\Delta}_i$

is a Bernoulli random variable with parameter $p_c$.

The other piece of observed information, the time of either censoring or event, is given as $T_i$. To summarize,

$$\begin{aligned} \Delta_i &= \mathbb{I}_{\widetilde{T}_i < C_i} \\ T_i &= \min(\widetilde{T}_i, C_i) = \widetilde{T}_i \wedge C_i. \end{aligned} \tag{2.1}$$

Note that $(\Delta, T)$ can be computed from $(\widetilde{T}, C)$, but not vice-versa, i.e., there is a loss of information due to censoring. It is convenient to separate the data into observed units ($\delta_i = 1$) and censored units ($\delta_i = 0$), in sets $\mathcal{E}$ (event) and $\mathcal{C}$ (censored), respectively, of sizes $n_e$ and $n_c$.

For censored subjects, with $\delta_i = 0$, only $C_i$ is observed and $T_i = C_i$. For non-censored subjects, with $\delta_i = 1$, only $\widetilde{T}_i$ is observed and $T_i = \widetilde{T}_i$. The likelihood of a censored subject corresponds to the probability of

$$\{C_i = t_i\} \cap \{\widetilde{T}_i \geq t_i\},$$

while the likelihood of a subject with observed event corresponds to the probability of

$$\{C_i > t_i\} \cap \{\widetilde{T}_i = t_i\}.$$

Define $S_\lambda$ as the survivor function of each $\widetilde{T}_i$,

$$S_\lambda(t) = \Pr[\widetilde{T}_i > t],$$

$F_\lambda$ as the distribution function

$$F_\lambda(t) = \Pr[\widetilde{T}_i \leq t],$$

and $f_\lambda$ as the density,

$$f_\lambda(t) = \frac{\partial F_\lambda(t)}{\partial t},$$

when it exists, as it will in most cases considered here.

Likewise, define $S_\theta(t)$, $F_\theta(t)$, $f_\theta(t)$ as the respective functions for each censoring variable $C_i$.

The full likelihood of the parameters $\lambda$, $\theta$ is

$$\prod_i \left(S_\lambda(t_i) f_\theta(t_i)\right)^{\overline{\delta}_i} \left(f_\lambda(t_i) S_\theta(\widetilde{t}_i)\right)^{\delta_i}, \tag{2.2}$$

due to the independence of each $(\widetilde{T}_i, C_i)$. The corresponding loglikelihood is then

$$\sum_i \overline{\delta}_i \left(\log S_\lambda(t_i) + \log f_\theta(t_i)\right) + \tag{2.3}$$

$$\delta_i \left(\log f_\lambda(t_i) + \log S_\theta(t_i)\right), \tag{2.4}$$

where, notably, the terms involving $\theta$ and $\lambda$ are linearly separated. Thus the ancillary statistics, $f_\theta$ and $S_\theta$, are *ignorable* for inference.

Specifically, consider the score in $\lambda$, the parameter of interest,

$$\sum_i \delta_i \frac{f'_\lambda(t_i)}{f_\lambda(t_i)} - \overline{\delta}_i \frac{f_\lambda(t_i)}{S_\lambda(t_i)}, \tag{2.5}$$

which does not depend on $\theta$. The observed Fisher information is

$$\sum_i \delta_i \left(\frac{f''_\lambda(t_i)}{f_\lambda(t_i)} - \frac{f'_\lambda(t_i)^2}{f_\lambda(t_i)^2}\right) - \overline{\delta}_i \left(\frac{f'_\lambda(t_i)}{S_\lambda(t_i)} + \frac{f_\lambda(t_i)^2}{S_\lambda(t_i)^2}\right).$$

Note, however, that the true Fisher information will generally depend on $\theta$, as will be seen below.

If $\Pr[\widetilde{T} < \infty] = 1$, the standard inference setting is recovered by assuming $\Pr[C = \infty] = 1$. In this case, $T = \widetilde{T}$, so the observed $T_i = t_i$ contains all of the information available. In the presence of censoring, however, inference based only on those cases with $\delta_i = 1$ is not consistent for $\lambda$.

## 2.1 The Right-Censored Exponential Model

The exponential distribution with rate $\lambda$, denoted Expo($\lambda$), is a baseline model for the time to event, $\widetilde{T}_i$, with survivor and density functions

$$f_\lambda(t) = \lambda e^{-\lambda t}, \tag{2.6}$$

$$S_\lambda(t) = e^{-\lambda t}. \tag{2.7}$$

The importance of this distribution comes from its memorylessness property:

$$\Pr[\widetilde{T} \in [t, t+dt) | \widetilde{T} \geq t] = \lambda \mathrm{dt}, \tag{2.8}$$

that the conditional probability of failure, given survival to any time $t$, is constant.

The survival function can be derived by solving

$$S'(t) = -\lambda S(t)$$

under the initial condition $S(0) = 1$.

### 2.1.1 Inference Without Censoring

The loglikelihood in the uncensored case is

$$\log \prod_i f(\lambda; t_i) = \sum_i \log f(\lambda; t_i),$$

which in the case

$$T = \widetilde{T} \sim \mathrm{Expo}(\lambda),$$

is

$$n \log \lambda - \lambda \sum t_i,$$

with score

$$\frac{n}{\lambda} - \sum t_i,$$

giving the maximum likelihood estimator $\widehat{\lambda} = n/\sum t_i$ with Fisher information $n/\lambda^2$.

Modeling the event and censoring time distributions above as independent Expo($\lambda$) and Expo($\theta$) distributions, respectively, gives a basic model for censored data which will be elaborated below.

### 2.1.2 Physical Example of Censoring

To illustrate the use of the exponential distribution, consider the situation of an alpha-decay detector observing one atom, which decays at rate $\lambda$, where $\lambda$ is the parameter of interest. However, at the same time, the detector itself may be destroyed by the emission of a higher-energy gamma-ray from an unwanted external source, which decays at rate $\theta$. If the detector is destroyed, no observation is made.

In this case the alpha decay is the event informative for the parameter of interest, while the time of the gamma emission is the censoring variable.

Thus, we have $\widetilde{T} = t$, and $\Delta = 1$ if the alpha-decay occurs first. Otherwise, the detector is disabled prematurely and we observe only that the alpha-decay had not occurred yet, or: $\widetilde{T} \geq t$ and $\Delta = 0$.

This is one of very few situations where this model is strictly correct, though it is a useful approximation more generally, and under slightly more general conditions.

### 2.1.3 Inference

In addition to $\widetilde{T}_i$ being distributed as Expo($\lambda$), let the $C_i$ be independent and distributed as

$$C_i \sim \text{Expo}(\theta).$$

Note that in this case, $T_i$ is the minimum of two independent exponentially-distributed variables with rates $\lambda$ and $\theta$, and thus

$$T_i \sim \text{Expo}(\lambda + \theta),$$

and in fact is independent of each Bernoulli-distributed $\Delta_i$, which have distribution

$$\Delta_i \sim \text{Bern}\left(\frac{\lambda}{\lambda+\theta}\right).$$

For censored cases ($\delta_i = 0$), only the fact that $\widetilde{T} > c = t$ is observed, while for observed cases ($\delta_i = 1$), $\widetilde{T} = t$ is observed exactly.

Thus the complete likelihood is

$$L(\lambda, \theta; \mathbf{t}, \delta) = \prod_{i \in \mathcal{E}} f_\lambda(t_i) S_\theta(t_i) \prod_{i \in \mathcal{O}} f_\theta(t_i) S_\lambda(t_i), \tag{2.9}$$

whence the loglikelihood

$$\begin{aligned} l(\lambda, \theta; \mathbf{t}, \delta) = &\sum_{i \in \mathcal{E}} \log f_\lambda(t_i) + \log S_\theta(t_i) \\ &+ \sum_{i \in \mathcal{O}} \log f_\theta(t_i) + \log S_\lambda(t_i). \end{aligned} \tag{2.10}$$

Thus, inference for $\lambda$ can ignore the terms of the likelihood associated with $C$, $f_\theta$ and $S_\theta$. That is, the *partial likelihood* in $\lambda$ is sufficient for first-order inference on $\lambda$. This is an instance of the general concept of *ignorable censoring*, which will be described later.

Specifically, the score in $\lambda$ is

$$\frac{\partial l(\lambda, \theta; \mathbf{t}, \delta)}{\partial \lambda} = \sum_{i \in \mathcal{E}} \frac{f'_\lambda(t_i)}{f_\lambda(t_i)} + \sum_{i \in \mathcal{O}} \frac{S'_\lambda(t_i)}{S_\lambda(t_i)}. \tag{2.11}$$

The likelihood for $\lambda$ is

$$l(\lambda; \mathbf{t}, \delta) \propto \prod_{i \in \mathcal{E}} f_\lambda(t_i) \prod_{i \in \mathcal{C}} S_\lambda(t_i), \tag{2.12}$$

since as above the censoring is ignorable.

With i.i.d. exponential distributions of $\tilde{T}_i$, the likelihood 2.12 is

$$\prod_{i \in \mathcal{E}} \lambda e^{-\lambda t_i} \prod_{i \in \mathcal{C}} e^{-\lambda t_i},$$

giving the score

$$\frac{n_e}{\lambda} - \sum_{i=1}^{n} t_i,$$

and the maximum likelihood estimator

$$\widehat{\lambda} = \frac{n_e}{\sum t_i}.$$

The $\Delta_i$ and $T_i$ are independent. Thus the maximum likelihood estimator (hereafter mle) is the product of (an observation of) a binomial variable $N_e = \sum \Delta_i$ with parameter $(n, \frac{\lambda}{\lambda+\theta})$ and an inverse-gamma variable, $1/\sum t_i$ with parameter $(n, \lambda + \theta)$, which are mutually independent. Thus,

$$\mathbb{E}[\widehat{\lambda}] = n\frac{\lambda}{\lambda+\theta} \cdot \frac{\lambda+\theta}{n-1} = \frac{n}{n-1},$$

so the estimator is asymptotically unbiased, with variance

$$\operatorname{var}(\widehat{\lambda}) = \frac{n\lambda}{(n-1)(n-2)} \left( \theta + \frac{n\lambda}{(n-1)} \right) \tag{2.13}$$

$$\approx \frac{\lambda}{n}(\theta + \lambda). \tag{2.14}$$

To represent the case of no censoring, let $\theta = 0$. This presents no difficulty in the inference above for $\lambda$, the parameter of interest. Note that in this case the expected Fisher information for $\lambda$ is

$$\frac{n}{\lambda^2}.$$

Thus, asymptotically, the variance of $\widehat{\lambda}$ under censoring will be $\approx \frac{\lambda+\theta}{\lambda} = 1 + \frac{\theta}{\lambda}$ times greater than in the uncensored case.

### 2.1.4 Another Partial Likelihood

Consider the naive estimation scheme where only the observed events ($\delta = 1$) are included in the likelihood. By the independence of $\widetilde{T}$ and $C$, one might (incorrectly) expect this 'likelihood' to produce a consistent estimator of $\lambda$.

However, the conditioning event is $\Delta = 1$, which is a function of both $\widetilde{T}$ and $C$ and is thus not independent of $\widetilde{T}$.

In fact, the "partial" likelihood, which conditions on the event $\Delta = 1$, i.e., $\widetilde{T} < C$, is that corresponding to the conditional probability

$$\Pr[\widetilde{T} = \widetilde{t} | C > \widetilde{T}] = \frac{\lambda e^{-(\lambda+\theta)\widetilde{t}}}{\frac{\lambda}{\lambda+\theta}} = (\lambda+\theta)e^{-(\lambda+\theta)\widetilde{t}} = \text{Expo}(\lambda+\theta).$$

This conditional probability will be the same if conditioned on the event $\Delta = 0$: $C \leq \widetilde{T}$. In fact, it is the same as the marginal distribution of $\widetilde{T}$ since the minimum of two exponential variables with rates $\lambda$, $\theta$ is exponential with rate $\lambda + \theta$.

However in this case, the loglikelihood is taken only over the cases with $\delta_i = 1$:

$$l = \sum_{i:\delta_i=1} \log(\lambda+\theta) - \lambda t_i - \theta t_i. \tag{2.15}$$

This generates the score equation

$$\frac{\partial l}{\partial \lambda} = \frac{n_e}{\lambda+\theta} - \sum_{i:\delta_i=1} t_i = 0,$$

which does not give a unique estimate for the parameter $\lambda$, but does give the following estimator for the parameter $\lambda + \theta$:

$$\frac{n_e}{\sum_{i:\delta_i=1} t_i}. \tag{2.16}$$

However, the estimator 2.16 is not usually of interest, as $\theta$ is a nuisance parameter. It is also inefficient, as described below.

The loglikelihood gives the observed Fisher information

$$\frac{n_e}{(\lambda+\theta)^2},$$

with expectation

$$n\left(\frac{\lambda}{\lambda+\theta}\right)\left(\frac{1}{(\lambda+\theta)^2}\right).$$

Since the probability distributions of $T_i$ are the same whether marginal or conditioned on either $\Delta$ or $\overline{\Delta}$, estimators of the same quantity may be constructed using either only the censored cases, or all cases. The expected Fisher informations will be, respectively,

$$n\left(\frac{\theta}{\lambda+\theta}\right)\left(\frac{1}{(\lambda+\theta)^2}\right)$$

or

$$n\left(\frac{1}{(\lambda+\theta)^2}\right).$$

It is clear that the latter quantity is larger, thus

$$\frac{n}{\sum_i t_i}$$

is a superior estimator for the quantity $\lambda+\theta$, if such a thing were desired. In fact, this is the maximum likelihood estimator and thus asymptotically optimal.

### 2.1.5 Independence and Ignorability

Above, the score equation 2.11 for $\lambda$, the parameter of interest, depends only on the terms of the likelihood corresponding to the event process (i.e., $S_\lambda$ and $f_\lambda$). This is a consequence of the independence of the variables $\widetilde{T}$ and $C$, as well as the variational independence of the parameters $\lambda$, $\theta$.

#### 2.1.5.1 Violation of Independence

Consider the case with

$$\widetilde{T} \sim \text{Expo}(\lambda) \tag{2.17}$$

$$C|\widetilde{T} = \widetilde{t} \sim \text{Expo}(\theta\widetilde{t}). \tag{2.18}$$

In this case, although the true parameter is $\lambda$, the censoring time and event time are positively correlated. Intuitively, this would positively bias any estimation.

In this case, the likelihood corresponding to the censored case $\delta = 0$ is

$$\begin{aligned}\Pr[C = t, \widetilde{T} \geq t] &= \lambda\theta \int_c^\infty u e^{-(\lambda+\theta t)u} \text{du} \\ &= \lambda\theta \frac{e^{-(\lambda+\theta t)t}(\lambda t + \theta t^2 + 1)}{(\lambda + \theta t)^2},\end{aligned}$$

so that the loglikelihood in this case includes a term

$$\log(\lambda t + \theta t^2 + 1),$$

making the censoring here nonignorable.

#### 2.1.5.2 Violation of Noninformativeness

Consider also the case

$$\widetilde{T} \sim \text{Expo}(\lambda)$$

$$C \sim \text{Expo}(\lambda)$$

with $\widetilde{T}$ and $C$ independent.

In this case the loglikelihood associated to the observed data is

$$\sum_i \log\lambda - 2\lambda t_i,$$

producing the estimator

$$\widehat{\lambda} = \frac{n}{2\sum t_i}.$$

Note that this is natural since, in this case, the data corresponds to observations of $\widetilde{T}_i \wedge C_i$, which are distributed i.i.d. as Expo($2\lambda$), and the survival estimator corresponds with the m.l.e.

Unlike the original example, this likelihood includes the survivor term $S_{\theta=\lambda}(t_i)$ for observed cases ($\delta_i = 1$), as well as the density term $f_{\theta=\lambda}(t_i)$ for censored cases.

## 2.2 Distributions

It is necessary and useful to introduce another representation of a continuous random variable, the hazard, or failure rate. The hazard at time $t$, $\lambda(t)$ is "the probability of failure in the interval $[t, t + \mathrm{dt})$, given survival up to time $t$,"

$$\lambda(t) = \frac{f(t)}{S(t)} = \frac{-\mathrm{d}\,\ln(S(t))}{\mathrm{dt}}.$$

The cumulative hazard, simply

$$\Lambda(t) = \int_0^t \lambda(u)\mathrm{du},$$

is also of importance below. Note that given a survival variable with a density, the hazard may be expressed as

$$\frac{f(t)}{S(t)} = \lambda(t) \tag{2.19}$$

$$\frac{-\partial \log(S(t))}{\partial t} = \lambda(t), \tag{2.20}$$

and thus by integrating and applying the initial condition $S(0) = 1$,

$$S(t) = \exp\left(-\int_0^t \lambda(u)\mathrm{du}\right) = \exp(-\Lambda(t)), \tag{2.21}$$

from which differentiation gives

$$f(t) = \frac{-\partial S(t)}{\partial t} = \lambda(t)\exp(-\Lambda(t)). \tag{2.22}$$

### 2.2.1 Exponential

The family of exponential distributions, with parameter $\lambda$, is

$$F(t) = 1 - \exp(-\lambda t); \tag{2.23}$$

$$f(t) = \lambda \exp(-\lambda t); \tag{2.24}$$

$$\lambda(t) = \lambda. \tag{2.25}$$

The mean lifetime is

$$\frac{1}{\lambda},$$

and the quantiles are

$$\frac{-\ln(1-q)}{\lambda}.$$

### 2.2.2 Weibull

The Weibull family is a straight-forward generalization of the exponential, parametrized in $\lambda$ and $\alpha$, with

$$F(t) = 1 - \exp(-\lambda^\alpha t^\alpha);$$

$$f(t) = \alpha\lambda^\alpha t^{\alpha-1}\exp(-\lambda^\alpha t^\alpha);$$

$$\lambda(t) = \alpha\lambda^\alpha t^{\alpha-1}\ (\alpha, \lambda > 0).$$

The mean lifetime is

$$\frac{1}{\lambda}\Gamma\left(1 + \frac{1}{\alpha}\right),$$

and the quantiles are

$$\frac{(-\ln(1-q))^{1/\alpha}}{\lambda}.$$

In the special case $\alpha = 1$, the Weibull is the exponential distribution with rate $\lambda$, whereas for $0 < \alpha < 1$, more mass is distributed to the beginning of the curve. For this reason, the Weibull family is used to model "infant mortality," such as when manufactured items may, due to variation in quality, initially fail quickly. This is particularly clear in the hazard function, which monotonically decreases to 0 if $0 < \alpha < 1$.

Conversely, with $\alpha > 1$, mass is distributed toward the tail, modeling an increased failure rate later in the life of a component, as may be due to aging or material fatigue. As before, this is clear from the hazard function, which monotonically diverges if $\alpha > 1$.

We continue with an overview of inference for these parameters of interest.

## 2.3 Inference

Assuming independent ignorable right-censoring, the partial loglikelihood of $n$ i.i.d. Weibull variables, $T_i$, observed at $(t_i)_{i=1}^n$, is given by

$$\log L(\lambda, \alpha) = \sum_i \log \left( f(t; \lambda, \alpha)^{\delta_i} S(t; \lambda, \alpha)^{1-\delta_i} \right)$$

$$= \sum_i \delta_i (\log \alpha + \alpha \log \lambda + (\alpha - 1) \log t_i) - \sum_i \lambda^\alpha t_i^\alpha.$$

The estimating equations are then

$$\sum \delta_i = \lambda^\alpha \sum t_i^\alpha;$$

$$\sum \delta_i \left( \frac{1}{\alpha} + \log \lambda + \log t_i \right) = \lambda^\alpha \sum t_i^\alpha \log(\lambda t_i),$$

which lack a solution in closed form.

The observed Fisher information in $\lambda$ is

$$\alpha \left( \frac{n_e}{\lambda^2} + \lambda^{\alpha-2} (\alpha - 1) \sum t_i^\alpha \right),$$

and in $\alpha$ is

$$\frac{n_e}{\alpha^2} + \sum_i \log^2(\lambda t_i)(\lambda t_i)^\alpha,$$

while the true Fisher information will depend on the random variable modeling the censoring.

## 2.4 Regression

To move forward, it is necessary to establish a regression framework for these problems. The natural approach is to use the log-link; letting $T$ be a lifetime random variable as above, let $Y = \log T$ and

$$Y = \log T = \mu + \sigma W,$$

with $W$ a noise term (independent of all other variables if present). This procedure requires a conventional association of $W$ with a particular member of the family. Specifically, $W = \log T_0$, with $T_0 \sim F_0$, the choice of $T_0, F_0$ being specific to each family.

Clearly, the form of $W$ is determined by the distribution of $T$. In our case, $T$ has support $(0, \infty)$, so that $W$ will have support $(-\infty, \infty)$.

The regression context is obtained by moving from the distribution of a lifetime, to the conditional distribution of a lifetime, given explanatory variables $X = x$.

By analogy with linear regression, where the mean $\mu$ is replaced by a linear form $\beta^\top \mathbf{x} = \sum_i \beta_i x_i$ in the explanatory variables, the mean of the linked variable above can be rewritten

$$\mu(x) = \mathbb{E}[Y|x] = \beta^\top \mathbf{x}.$$

In the following section, we will examine regression in the case of a Weibull

variable, illustrating the usefulness of the accelerated failure time model, where the parameter $\mu$ corresponds to a scaling of time in the survival function.

In fact, this scaling is linear, giving a parametric family of distributions given by $S(\mu t)$ or, in the regression case, a corresponding family in $\beta$ given by

$$S\left(e^{\beta^\top \mathbf{x}} t\right).$$

This interpretation can be useful in reliability analysis and public health, as it allows the intuitive comparison of two populations in terms of "effective lifetime."

Note that an "accelerated lifetime" property may be added to any distribution by introducing a $\mu$ parameter to the survival function. However, the true accelerated lifetime property concerns a linear scaling of the survival function by the mean parameter of the log-linked form above.

The analogue of this property in the hazard function, called "relative hazard," has the form

$$\lambda(t) = \lambda_0(\mu t),$$

which will be examined in a later section.

As it turns out, the Weibull is unique in having both properties.

## 2.5 Accelerated Failure Time

Let $T_0$ denote a Weibull-distributed variable with parameters $\alpha = \lambda = 1$, with corresponding distribution $F_0$. Consider the variable $W = \log T_0$, so $e^W \sim F_0$. The distribution of $W$ is then specified by the cumulative distribution function

$$F_w(w) = 1 - e^{-e^w},$$

which is called the type-I extreme-value distribution.

Expressing a Weibull variable $T$ in regression form as $\log(T) = \mu + \sigma W$, the distribution $F_T(t)$ is then

$$1 - e^{-e^{\frac{\log t - \mu}{\sigma}}}$$

$$= 1 - e^{-t^{1/\sigma} e^{-\mu/\sigma}},$$

which is a reparametrization of the original Weibull with $\sigma = 1/\alpha$ and $\mu = -\log\lambda$.

Note that

$$F_T(t) = F_0\left(\left(te^{-\mu}\right)^{1/\sigma}\right),$$

so that a change in $\mu$ results in a scaling of the effective age of the subject.

For example, taking the exponential case ($\sigma = 1$) for ease, and letting $\mu = -\log\kappa$, we see that $F_T(t) = F_0(\kappa t)$. So, in the actuarial (distribution) sense, a cohort with $\mu = -\log 2$ can be considered to "age twice as fast" as the baseline cohort given by $X_0$. A distribution family with this property of always allowing a subfamily of distributions which scales the time argument of $F$ is called an *accelerated failure time* model.

Now to obtain the standard linear regression of the mean, we simply replace the mean $\mu$ with a value conditional on $Z$ and depending on the conditioning value $Z = z$.

$$\log T | Z = z = \sigma W,$$

in which case,

$$\mathbb{E}[\log T | Z = z] = \beta^\top \mathbf{z},$$

while the linked Weibull variable will have mean

$$\mathbb{E}[T | Z = z] = e^{e^{\mathbf{z}^\top \beta}} \Gamma(1 + \sigma).$$

One final characteristic of the Weibull emerges from the hazard, which in regression terms is

$$\lambda(t) = \frac{1}{\sigma} e^{-e^{\mathbf{z}^\top \beta}/\sigma} t^{1-\sigma} \sigma$$

Taking $\sigma = 1$ for ease again, analogously to the accelerated failure time property above, this induces a subfamily which scales $\lambda$ rather than $F$:

$$\lambda(t) = \lambda_0 \left(te^{-\mu}\right) = \lambda_0 \left(te^{-e^{\mathbf{z}^\top \beta}}\right).$$

### 2.5.1 Residual Life

For many questions it is natural to estimate quantities concerning phenomena observed at an intermediate time $s$, conditioned on the continued observation of the subject to this time.

In general, these quantities concern what is called the "residual life" of the subject: the random remaining time to event conditioned on continued observation (neither censoring nor event) through time $s$, or

$$\tilde{T}_s = \tilde{T}|T > s \tag{2.26}$$

$$= \tilde{T}|\tilde{T} > s, C > s, \tag{2.27}$$

where under independence of the event and censoring times,

$$\tilde{T}_s = \tilde{T}|\tilde{T} > s, C > s \tag{2.28}$$

$$= \tilde{T}|\tilde{T} > s. \tag{2.29}$$

For instance, the prognostic estimand of the *mean time between failures*, i.e. the mean lifetime, may be adapted to observation through $s$. This can be called the mean time until failure, and $\mathbb{E}[\tilde{T}_s - s|T > s]$ is the quantity to be estimated.

Although $\tilde{T}$ is the variable of interest, the conditioning event also involves the censoring variable. Analogously say

$$C_s = C|\tilde{T} > s, C > s, \tag{2.30}$$

noting of course that as above,

$$C_s = C|C > s, \tag{2.31}$$

under independence.

Consider $\tilde{T} \sim F_\lambda$ with ignorable censoring The density of the conditional variable pair $(\tilde{T}_s = t, C_s = u)$ is

$$\frac{f_\lambda(t) f_\theta(u)}{S_\lambda(s) S_\theta(s)} \mathbb{I}_{t>s, u>s}.$$

Introducing the notation

$$f(t; s) = \frac{f(t)}{S(s)} \mathbb{I}_{t>s},$$

the conditional variables are clearly independent by factorization. Note that the conditional survival $S(t; s)$ may be written as

$$S(t; s) = \Pr[\tilde{T} > t|\tilde{T} > s] = \frac{\exp(-\Lambda(t))}{\exp(-\Lambda(s))} = \exp(-(\Lambda(t) - \Lambda(s))), \tag{2.32}$$

identifying the residual cumulative hazard

$$\Lambda(t; s) = \Lambda(t) - \Lambda(s). \tag{2.33}$$

Of course $F(t; s) = 1 - S(t; s)$ and

$$\lambda(t; s) = \lambda(t),$$

covering all of the functions of interest.

Using the residual cumulative hazard above, we can restate the residual life density as

$$f(t; s) = \lambda(t) \exp(-(\Lambda(t) - \Lambda(s))).$$

For example, the mean residual life (i.e. the expected survival time following $s$) is convenient to compute as

$$\mathbb{E}[T - s|T \geq s] = \int_s^\infty S(t; s) \mathrm{dt} \tag{2.34}$$

$$= \frac{\int_s^\infty S(t) \mathrm{dt}}{S(s)}, \tag{2.35}$$

by exchanging the order of integration in

$$\begin{aligned}&\int_s^\infty S(t)\mathrm{dt}\\ =&\int_s^\infty \mathbb{E}[\mathbb{I}_{t<T_s}]\mathrm{dt},\end{aligned}$$

where $T_s$ is a random variable constructed with the distribution $F(t;s)$ above, giving

$$\begin{aligned}&\mathbb{E}\left[\int_s^\infty \mathbb{I}_{t<T_s}\mathrm{dt}\right]\\ =&\mathbb{E}[T_s - s].\end{aligned}$$

The median residual life is simply the solution in $m$ of

$$\exp(-(\Lambda(m+s)-\Lambda(s))) = \frac{1}{2},$$

or the infimum of the solution set.

# Chapter 3

# Exploratory Analyses of the New York City Electric Grid Data

## 3.1 Data

The data at hand cover the time period March 30, 2005 to September 31, 2007, with 939 total feeder units spanning the boroughs of Manhattan, Brooklyn, Queens and the Bronx.

| Feeder Distribution by Borough | |
|---|---|
| Manhattan | 564 |
| Brooklyn | 164 |
| Queens | 115 |
| Bronx | 94 |

The borough of Staten Island is omitted because the power distribution system there is mostly above-ground and unrepresentative of the greater NYC

area, which uses underground distribution.[1]

The dataset includes 3,022 failures (defined as unscheduled outages not due to direct operator action), for a rate of ≈1.36 per feeder per year. Despite this seemingly low number, there is significant concern about cascading effects of failed feeders on neighboring units, as well as the need for systematic scheduling of maintenance and prophylaxis. Further, the failures are far from uniformly distributed both in time and across feeders. Some exploratory statistics are reported below.

Particular focus is given to the Long Island City area in Queens in connection with a development project there. This area is labelled as subnetwork 01Q and comprises 22 feeders. In the data under consideration, the Long Island City area includes 186 failures, for a higher rate of ≈3.57 failures per feeder per year.

## 3.2 Historical Outages

Although the failure probabilities are small, the effects of power failure can be dramatic, and the cumulative effect of a maintenance strategy can translate into large sums of money in the long run.

## 3.3 Failure Rates Across Boroughs

The distribution of the number of failures associated to each feeder in four boroughs is reported below.

[1]The disproportionate effect of Hurricane Sandy, in late October 2012, on Staten Island has prompted Consolidated Edison to consider replacing the electric grid there with an underground system as used elsewhere in New York City.

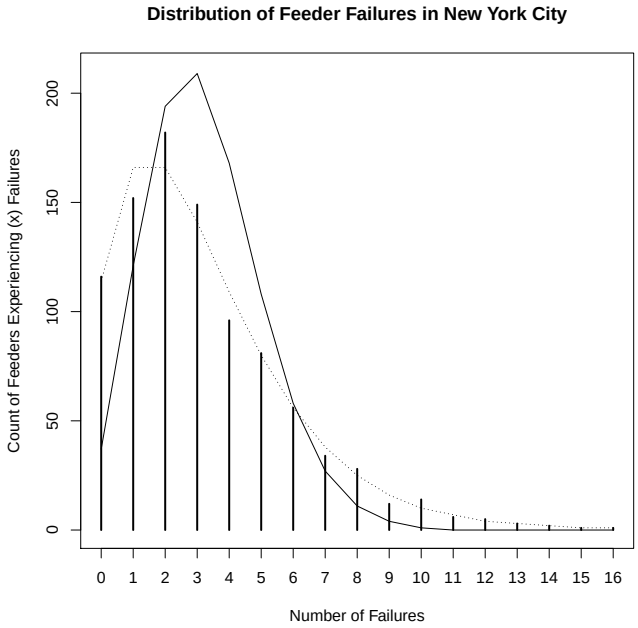



Figure 3.1: Distribution of Observed Feeder Failures.
Solid: Poisson model fit; Dotted: Negative-binomial model fit.

All of the feeders are under observation for the entire recorded period. Thus, if the feeders were identical in terms of failure rate, the distribution of the number of failures should be approximately Poisson, as given by the solid line in the figures.

To the contrary, the distribution is much better captured by the negative-binomial distribution, which can be interpreted as a mixture of Poisson distributions generated by heterogenous failure rates.

This phenomenon is called "overdispersion" of the failure counts, as it results in a wider distribution than the Poisson. Note that the failure rates may vary feeder-by-feeder or by calendar time; since the components used in feeders vary widely and failures are more common in the summer months, both are likely. A simple distribution of counts cannot distinguish between the two.

Further, heterogeneity is not the only cause of overdispersion. It may also be caused by "contagion" between failures, as for example if one failure causes an increase in failure rate, either in the same feeder or neighboring feeders. This might be expected, as a failure causes voltage spikes which can damage equipment and which are propagated to neighboring units through the grid. Although the negative-binomial model is not derived from an assumption of contagion, it will nonetheless capture the overdispersion.[2]

Since overdispersion can result from a mixture of Poisson distributions with different parameters, it is natural to stratify the data in order to isolate possible sources of this variation. Differences in the equipment and service policies between boroughs of New York City make stratification by boroughs a natural candidate. However, overdispersion persists within each borough, as seen in the following figures.

[2]Specifically, the negative-binomial distribution results when the failure rates are Gamma-distributed and the feeders are independent. However, the model can capture overdispersion when this condition is only approximately fulfilled. Thus this statistic cannot itself distinguish between heterogeneity in failure rate, and contagion between failure events.

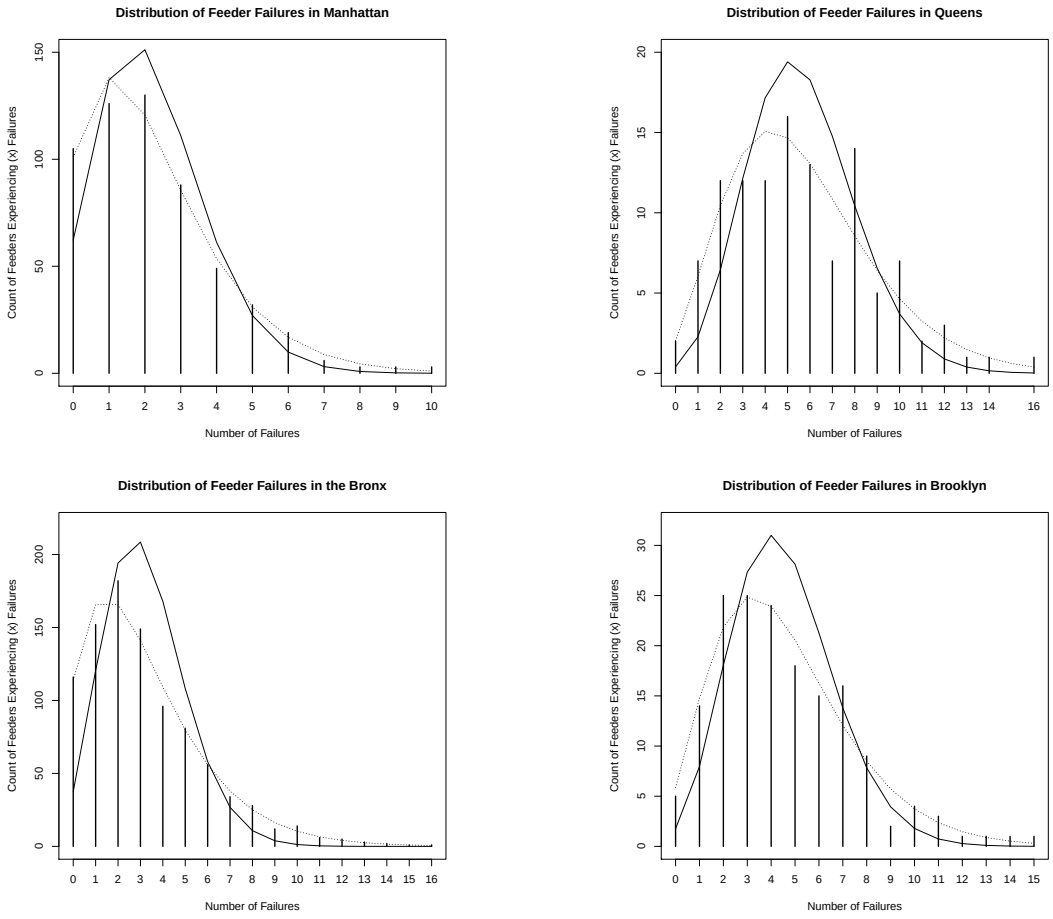



Figure 3.2: Distribution of Observed Feeder Failures by Borough.
Solid: Poisson model fit; Dotted: Negative-binomial model fit.

In each case, except perhaps Manhattan, there is significant overdispersion. better-captured by the negative-binomial model.

## 3.4 Failure Rates with Respect to Time

The methods presented herein can, under certain conditions, identify the effects of true contagion as opposed to seasonality. The methods are compatible with standard regression methods, which can be used to further explain and

isolate the variation in failure rate produced by variation in feeder composition as well as in usage and load patterns.

An examination of the failure times can proceed by calendar time or gap time, the lengths of the periods between subsequent failures. The effect of calendar time through seasonality is evident below, with periods of increased failure rates occurring every year during the summer. Additionally, an anomalous event is present at $t \approx 470$ days.

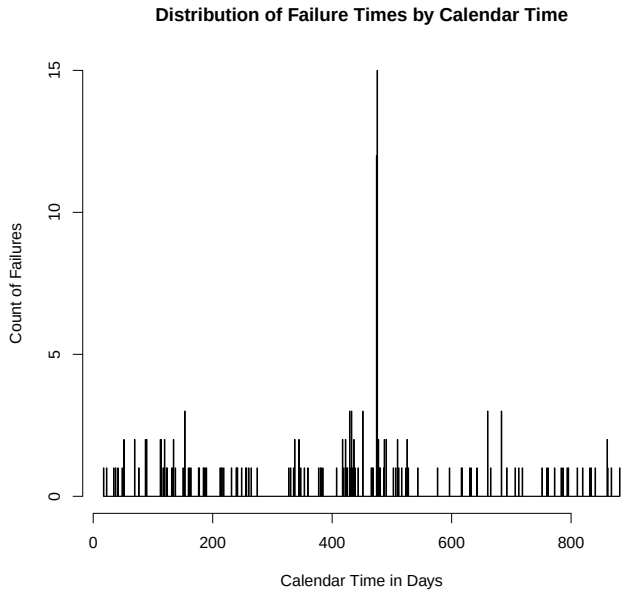



Figure 3.3: Empirical Distribution of Observed Feeder Failures in Calendar Time.

Solid: Poisson model fit; Dotted: Negative-binomial model fit.

This information about the seasonal effect is useful for certain applications, such as scheduling repairs and emergency power backup systems for summer months, and even making power purchase agreements with external suppliers of electricity.

However, for other planning purposes this effect is a source of confounding.

One may be interested in separating this effect from the "intrinsic" lifetime of a feeder: the time between repairs in a controlled environment. In this case, the seasonal effect is a confounding factor and, assuming a multiplicative model, may be controlled for by the standard Cox partial likelihood (see ch. xx).

Similarly, the estimation of effect of certain natural covariates, such as customer power demand, may be negatively affected by seasonal correlation. The effect of power demand, which is significantly greater in the summer, may be overestimated unless the seasonal effect evident above is controlled for.

In some cases, it may be important to model and control for the effect of both calendar and gaptime. For instance, the adverse effect of power demand on residual lifetime may be greater both in the summer and in the later part of a feeder's duty cycle. While the Cox partial likelihood to remove confounding by a calendar-time effect is easily constructed and solved due to the $\lambda_0(t)$ term common to all units, the equivalent partial likelihood to remove the gap time effect is not obvious since the gap time varies between units.

Consider the diagram below, where each $\star$ indicates an event with immediate restoration. At calendar time $t = 5$, under the Cox model each unit is subject to an unknown calendar time hazard multiplier $\lambda_0(5)$. However, with the equivalent gap time model, at time $t = 5$, the units are subject to the gap time hazard multipliers $\psi(2)$, $\psi(1)$, $\psi(3)$.

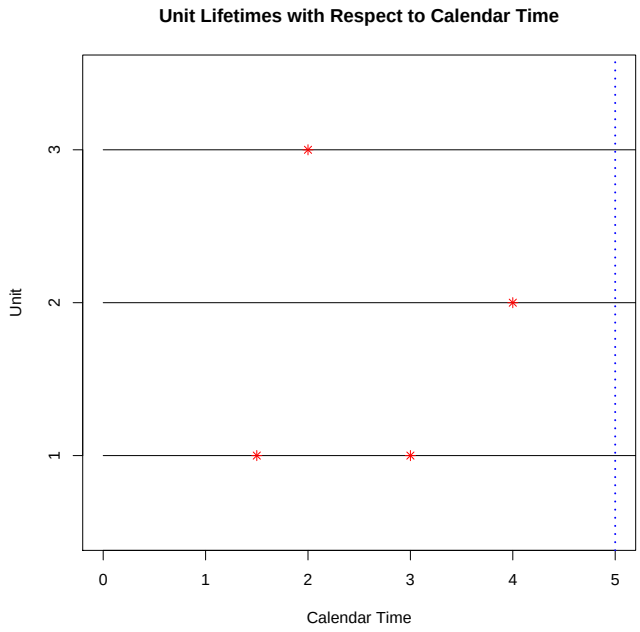



Figure 3.4: Hypothetical Example of Gap Times in Calendar Time

The crude empirical distribution of gap times, without controlling for calendar time effect, is below.

Two visualizations of feeder lifetimes are presented below, with gap time in linear and $\log_{10}$ scales, respectively. Each lifetime is represented as a line segment which begins at calendar time $t_1$ and gap time 0, being immediately after the previous failure. This line segment terminates at calendar time $t_2$ and gap time $t_2 - t_1$. The lifetimes in this diagram are aggregated across all units under observation; one feeder may contribute several lifetimes to the graph.

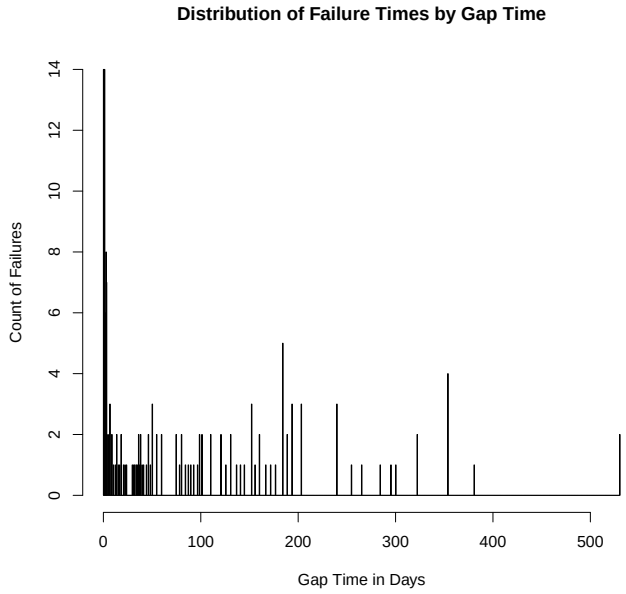



Figure 3.5: Empirical Distribution of Observed Feeder Failures in Gaptime.

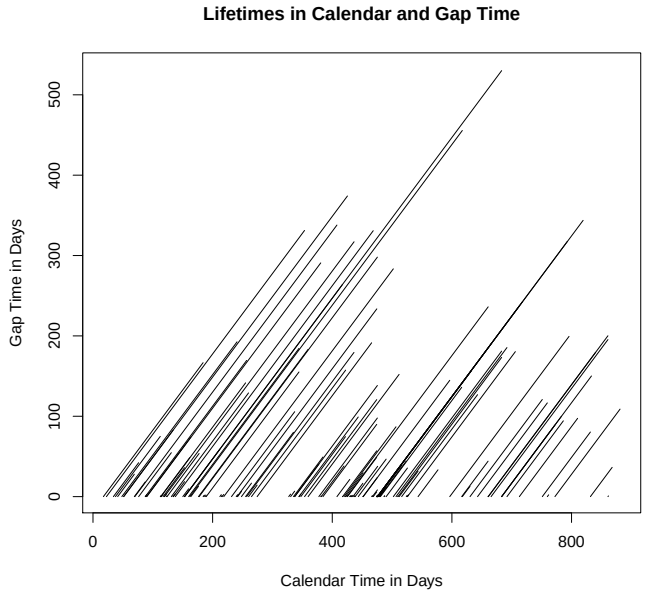



Figure 3.6: Empirical Distribution of Lifetimes against Calendar Time.

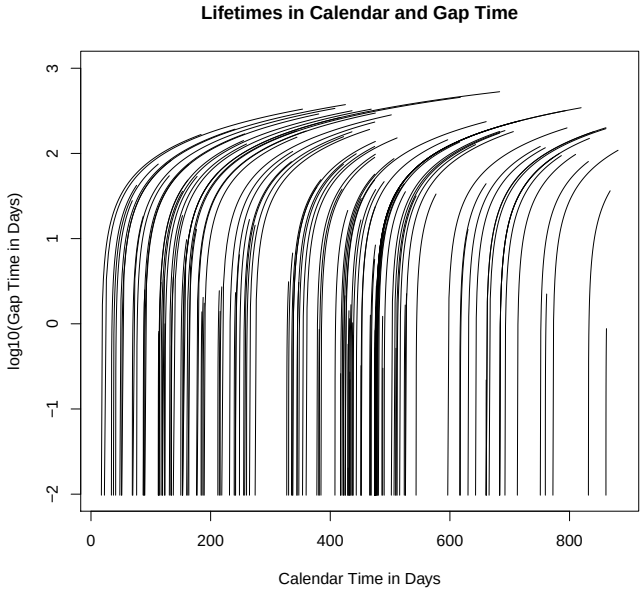



Figure 3.7: Empirical Distribution of $\log_{10}$(Lifetimes) against Calendar Time.

# Chapter 4

# Longitudinal and Nonparametric Methods

The preceding methods can be interpreted in a more general framework, where $N$ units are observed from $t = 0$ to $t = T$.[1]

Consider a partition of the interval $[0, T)$ into $[u_0, u_1), [u_1, u_2), \ldots, [u_{J-1}, u_J)$ with $u_0 = 0$ and $u_J = T < \infty$. Assume that $T > \max_i t_i$. Instead of considering the data $(t_i, \delta_i)$ in and of itself, one may consider the entire lifetime of each corresponding subject.

This approach has significant advantages, for example a subject may be temporarily censored and return to the study after some time has passed. This flexibility is important for clinical trials as well as industrial applications, where a repair necessarily removes a unit from observation.

In some cases, one may cobble together a path around this approach when it is called for; however, this is usually unnatural, and often removes information. For example, consider an industrial unit $i$ activated at time $t = 0$ days, brought offline (censored) at $t = 10$, brought back online at $t = 15$, and which fails at

[1] Generally, it is possible that $T = \infty$, however $T$ will be assumed finite here.

$t = 30$. One may dummy-code a sequence (in $k$) of standard right-censored trials for this unit as in the following.

| $k$ | $t_{i,k}$ | $\delta_{i,k}$ |
|---|---|---|
| 1 | 10 | 0 |
| 2 | 15 | 1 |

This coding removes information; particularly, the duration of censoring as well as the calendar time of the subsequent failure. A longitudinal approach may easily incorporate this information, as will be seen later.

First, we will consider the models from earlier chapters in terms of this approach.

## 4.1 Traditional Likelihood

To examine the earlier models, it is necessary to introduce notation expressing the periods of observation in terms of the intervals $[u_{j-1}, u_j)$. First, note that the set

$$\{j > 1 \mid u_j < t_i\}.$$

contains the indices of all intervals through which the subject $i$ survives under observation. Note that each element $j$ is strictly less than $J$.

Denote by $j(t)$ the index of the right-hand endpoint of the interval containing the time $t$:

$$j(t) = \min\{k \mid u_k > t\},$$

so the interval $\left[u_{j(t_i)-1}, u_{j(t_i)}\right)$ always contains $t_i$.

Informally, the likelihood is approximated by taking the following proba-

bilistic model

$$\prod_{j>1,u_j<t_i} \Pr\big[\text{no event or censoring in the interval } [u_{j-1},u_j) \;\big|\; \text{no event or censoring in intervals } [u_{k-1},u_k), k<j\big] \times \Pr\big[\text{event in the interval } [u_{j(t_i)},u_{j(t_i)+1})\big]^{\delta_i} \times \Pr\big[\text{censoring in the interval } [u_{j(t_i)},u_{j(t_i)+1})\big]^{\overline{\delta_i}} \tag{4.1}$$

This represents the survival data as an approximation of a continuous observation, which is recovered as the grid becomes infinitely dense. Some of the continuous results are given below; however, immediately following are some examples of the discrete approximation which re-express earlier representations in this more general form.

Following the previous chapter, let the observed data be $(t_i, \delta_i)$, with general likelihood

$$\prod_i S(t_i)^{\overline{\delta_i}} f(t_i)^{\delta_i} = \prod_i S(t_i)\lambda(t_i)^{\delta_i}. \tag{4.2}$$

Also continue to take $u_0 = 0$.

Consider a censored case, $T_i = t_i, \delta_i = 0$, with observation restricted to a grid as above with fixed $J$. That is, the unit is fully observed without event or censoring through the intervals

$$[0,u_1), [u_1,u_2), \ldots, [u_{j(t_i)-2}, u_{j(t_i)-1}),$$

and is removed from observation at some time in $[u_{j(t_i)-1}, u_{j(t_i)})$.

### 4.1.1 Exponential

Assuming independence between $\tilde{T}_i$ and $C_i$, that $\tilde{T}_i \sim \text{Expo}(\lambda)$, $C_i \sim \text{Expo}(\theta)$, and that $\delta_i = 1$, the likelihood determined by observed data $T_i = t_i$, $\Delta_i = \delta_i$

has likelihood

$$\begin{aligned} & e^{-\lambda u_1} \prod_{j=2}^{j(t_i)-1} \frac{e^{-\lambda u_j}}{e^{-\lambda u_{j-1}}} && \times \\ & e^{-\theta u_1} \prod_{j=2}^{j(t_i)-1} \frac{e^{-\theta u_j}}{e^{-\theta u_{j-1}}} && \times \\ & \left(1 - e^{-\lambda(u_{j(t_i)} - u_{j(t_i)-1})}\right). \end{aligned}$$

When the grid is small, $u_{j(t_i)-1} \to t_i$ and $u_{j(t_i)} - u_{j(t_i)-1} \approx \mathrm{dt}$, using the approximate identity $e^{-\epsilon} \approx 1 - \epsilon$ gives the limiting continuous likelihood

$$e^{-\theta t_i} \lambda e^{-\lambda t_i} \mathrm{dt},$$

which is the same as before. The case for $\delta_i = 0$ follows similarly with only the last line of (above) changed, and gives the desired likelihood

$$e^{-\lambda t_i} \theta e^{-\theta t_i} \mathrm{dt}.$$

### 4.1.2 General

The likelihood for the general case of continuous hazard functions is similar to the exponential case. Keeping independence but allowing arbitrary densities $f_\lambda$ and $f_\theta$, the likelihood is

$$\begin{aligned} & S_\lambda(u_1) \prod_{j=2}^{j(t_i)-1} \frac{S_\lambda(u_j)}{S_\lambda(u_{j-1})} && \times \\ & S_\theta(u_1) \prod_{j=2}^{j(t_i)-1} \frac{S_\theta(u_j)}{S_\theta(u_{j-1})} && \times \\ & \frac{F_\lambda(u_j) - F_\lambda(u_j)}{S_\lambda(u_{j-1})} && = \\ & S_\theta(u_{j(t_i)-1}) \left(F_\lambda(u_j) - F_\lambda(u_{j(t_i)-1})\right) \end{aligned}$$

giving in the limit, as before,

$$S_\theta(t_i) f_\lambda(t_i) \mathrm{dt}.$$

Again, it is similar for $\delta_i = 0$.

## 4.2 Indicator Functions

As mentioned above, this methodology allows a generalization of the usual statistics $(\delta_i, t_i)$. It is useful to introduce an indicator variable $Y_i(t)$, as a function of time, indicating whether the unit $i$ is under observation at time $t$. Again, we will use the notation $y_i(t)$ for an observation of this variable. In the case of the traditional statistics $\delta_i, t_i$,

$$y_i(t) = \begin{cases} 1 & \text{if } t \leq t_i \\ 0 & \text{if } t > t_i, \end{cases}$$

independent of the value of $\delta_i$. The interpretation is that a unit is susceptible to an (observed) event until $t_i$, at which time it is removed either from observation $(\delta_i = 0)$ or entirely $(\delta_i = 1)$.

The example of an industrial unit given above may be naturally represented with an indicator function. Recall that the unit is activated at $t = 0$ days, is brought offline from $t = 10$ through $t = 15$, and fails at $t = 30$. This history is easily represented by the indicator function

$$y_i(t) = \begin{cases} 1 & \text{if } t \leq 10 \\ 0 & \text{if } 10 < t \leq 15 \\ 1 & \text{if } 15 < t \leq 30 \\ 0 & \text{if } t > 30, \end{cases}$$

which captures the history more accurately than the dummy-coding approach used above. In this way, the indicator $Y_i$ is obviously more expressive than the standard statistics. It is, however, incomplete, and requires a record of event times as well to form a complete likelihood.

The principal restriction on $Y_i(t)$ is that it depend only on information prior to time $t$, or in other words that it is a *predictable* process.

Although we will use standard results later, for purposes of exposition consider a finite grid as above. Predictability in this case implies that the value $Y_i(s)$ on $[u_{j-1}, u_j)$ can depend only on events which occur in the intervals $[u_{k-1}, u_k)$, $k \leq j-1$.

Particularly, using the definition

$$Y_i(t) = \begin{cases} 1 & \text{if } t \leq u_{j_{i,J}} \\ 0 & \text{if } t > u_{j_{i,J}}, \end{cases}$$

it is possible to restate the traditional likelihood above as

$$\prod_{j=1}^{J} \left( e^{-\lambda_j \delta_j \mathbb{I}_{j<j_{i,J}}} \lambda_j^{\mathbb{I}_{j=j_{i,J}}} \right)^{Y_i(u_j)}.$$

### 4.2.1 Continuous Theory

The full likelihood of this process, assuming a Poisson process with modeled intensity, is

$$\prod_i e^{\int_0^\infty \lambda_0(u) Y_i(u) e^{\beta^\dagger x_{i,u}} \psi(u-\tau_{i,u}) \mathrm{du}} \prod_j \lambda_0(t_{i,j}) e^{\beta^\dagger x_{i,t_{i,j}}} \psi(t_{i,j} - \tau_{i,t_{i,j}}).$$

The aggregate likelihood is obtained by summing the modeled intensity across all units and observing events without knowledge of which unit produced them. It is given by:

$$e^{\int_0^\infty \lambda_0(u) \sum_i Y_i(u) e^{\beta^\dagger x_{i,u}} \psi(u-\tau_{i,u}) \mathrm{du}} \prod_t \lambda_0(t) \sum_i e^{\beta^\dagger x_{i,t}} \psi(t - \tau_{i,t}).$$

## 4.3 Likelihood

The full likelihood of this process, assuming a Poisson process with modeled intensity, is

$$\prod_i e^{\int_0^\infty \lambda_0(u) Y_i(u) e^{\beta^\dagger x_{i,u}} \psi(u-\tau_{i,u}) \mathrm{du}} \prod_j \lambda_0(t_{i,j}) e^{\beta^\dagger x_{i,t_{i,j}}} \psi(t_{i,j} - \tau_{i,t_{i,j}}).$$

### 4.3.1 Partial Likelihood

The standard Cox partial likelihood is derived by conditioning the full likelihood on this aggregated "information," and yields

$$\frac{\prod_t e^{\beta^\dagger x_{i(t),t}} \psi(t_{i(t),t} - \tau_{i(t),t})}{\prod_t \sum_i e^{\beta^\dagger x_{i,t}} \psi(t - \tau_{i,t}) Y_i(t)}.$$

This partial likelihood is also a partial justification for applying a prior to $\psi$ and thus obtaining a "partial posterior" for $\psi$, derived using only the *identity* of failed units.

However, the usual argument for the partial likelihood includes, on some level, the non-assumption that $\lambda_0()$ is "fully flexible" and thus that the aggregate likelihood cannot provide information about $\beta$. On a heuristic level, the absence of failures at any time $t$ can be "explained" by allowing $\lambda_0(t) = 0$. With no constraints (smoothness) or prior on $\lambda_0$, this is in fact a maximum likelihood estimate.

On the other hand, it is not clear that the aggregate likelihood is not informative about $\psi$, and in fact there is reason to think that it is.

Among non-parametric estimators, the general assumption here is that *units are independent* and *time periods are exchangeable.*

## 4.4 Product-Limit Estimator

Given a finite set of intervals

$$\{U_i = [u_{i-1}, u_{i+1})\}_{i=1}^{I},$$

with $u_0 = 0, u_{I+1} = \infty$, the product-limit estimator for the survivor function is defined for $i \in \{0, 1, \ldots, I\}$ by

$$\hat{S}(u_i) = \prod_{j=1}^{i} \frac{|\mathfrak{R}(u_j)| - d_j}{|\mathfrak{R}(u_j)|},$$

with

$$d_i = |\{j : u_{i-1} \leq \tilde{t}_j < u_i\}|$$

the number of units which die during interval $U_i$.

Some issues emerge with censoring in the final interval $U_I$; if any censoring occurs, the formal estimate $\hat{S}(t) > 0$ for any large $t$. Typically, $u_I$ is chosen to represent some *a priori* "natural end of life," with $\hat{S}(t) = 0$ for $t \geq u_I$. These issues of course have no general answers.

## 4.5 Nelson-Aalen Estimator

Another reasonable estimand is the cumulative hazard, $\int_0^t \lambda(u)\mathrm{du} = \Lambda(u)$, which is estimated by

$$\hat{\Lambda}(u) = \sum_{i=1}^{I} \frac{1}{|\mathfrak{R}(u_i)|}.$$

A slightly more general application of this formula occurs when, under the model

$$\lambda_i(u) = \lambda_0(u)\mu(\mathbf{x}_{i,u})Y_i(u),$$

with $Y_i(u)$ the at-risk indicator and $\mu$ a regression function, is to have

$$\hat{\Lambda}_0(t) = \sum_{i=1}^{I} \frac{1}{\sum_j \mu(\mathbf{x}_{j,t_i})Y_j(t_i)} = \int_0^T \frac{\sum_j \mathrm{dN}_j(t)}{\sum_j \hat{\mu}(x_j)Y_j(t)},$$

a two-step estimator where the estimates $\hat{\mu}$ have been obtained by some other method (e.g. the Cox partial likelihood). Given a finite set of intervals $[t_i, t_{i+1})$, the product-limit estimator for the survivor function is defined by

$$\hat{S}_i = \prod_{j=1}^{i} \frac{|R(t_i)| - d_i}{|R(t_i)|},$$

One approach to fitting a survival model is to use the Cox partial likelhood. Assuming that there are no ties, and letting $t_\cdot \in \mathfrak{T}$ index the failure times and denoting by $i(t)$ the unit to fail at time $t$, and denoting by $\psi(j,t)$ the hazard of unit $j$ at time $t$, this partial likelihood is

$$\prod_{t \in \mathfrak{T}} \frac{\psi(i(t),t)}{\sum_{j \in R(t)} \psi(j,t)}.$$

Notably, using this partial likelihood identifies any component of the hazard which depends only on the calendar time:

$$\prod_{t \in \mathfrak{T}} \frac{\phi(i(t),t)\lambda_0(t)}{\sum_{j \in R(t)} \phi(j,t)\lambda_0(t)}$$

$$= \prod_{t \in \mathfrak{T}} \frac{\phi(i(t),t)}{\sum_{j \in R(t)} \phi(j,t)}.$$

Given a loglinear model, this generates the score

$$\sum_t x_{i(t)} - \frac{\sum_j x_j(t) e^{\beta^\dagger x_j(t)}}{\sum e^{\beta^\dagger x_j(t)}}$$

and information

$$-\sum_t \frac{\left(\sum_j e^{\beta^\dagger x_j(t)}\right)\left(\sum_j x_j x_j^\dagger e^{\beta^\dagger x_j(t)}\right) - \left(\sum_j x_j e^{\beta^\dagger x_j(t)}\right)\left(\sum_j x_j^\dagger e^{\beta^\dagger x_j(t)}\right)}{\left(\sum_j e^{\beta^\dagger x_j(t)}\right)^2},$$

which is positive by a trivial extension of the Cauchy inequality. It's worth noticing that the failure never formally enters this quantity. Specifically, this appears to just be an accident of the loglinear model.

# Chapter 5

# Counting Process

A general formulation of survival problems is found in the counting process framework as reviewed in Andersen, et al.[2] We will follow their formulation, introducing the relevant concepts and applying them to re-examine the traditional facts reviewed in chapters before. Finally we will use the technique to extend into the non-parametric domain and introduce novel process estimators for the infant mortality by applying these methods to the gap-time.

## 5.1 Process

A stochastic process is a random variable indexed by both time and Borel sets of the relevant probability space. The essence of stochastic processes is the interplay of these two dimensions; just as the flow of time generally reveals more information to an observer, this is represented by an increasing sequence of probability spaces, $\mathcal{F}_t$, indexed by time $t$.

Increasing simply means that any set measurable in $\mathcal{F}_s$ is also measurable in $\mathcal{F}_t$ for $t > s$. Each $\mathcal{F}_t$ uses the same universe $\Omega$. Often, the probability spaces involved will represent the minimal information needed to measure

the variables of the process at index times, perhaps coupled with some side information added to $\mathcal{F}_0$ (and thus to all $\mathcal{F}_t$). In this case, the sequence is said to be "generated" by the process $X_t$. This terminology is justified by the Kolmogorov extension theorem.

However, in all interesting continuous cases, this sequence of increasing probability spaces cannot be explicitly constructed. Thus, applications are generally approached through some standard ideas (and theorems) which impose a general and well-understood structure. These ideas often correspond roughly to our intuitive ideas of "predictability," "signal vs. noise," and so on.

### 5.1.1 Substantiation

Note that all processes considered here are continuous from the right and have limits from the left.

Consider the simple Poisson process in time, $N(t)$, with fixed rate $\lambda$. The Poisson process obeys two fundamental properties. First, that for each $t \geq s$,

$$N(t) - N(s) \sim \Lambda(t) - \Lambda(s),$$

which in this case is

$$N(t) - N(s) \sim \lambda(t - s).$$

Second, that for each $s \leq t < u \leq v$, $N(t) - N(s)$ is independent of $N(v) - N(u)$. Finally, it is assumed that $N(0) = 0$.

Thus, for the simple process with $\lambda(t) = \lambda$, since the marginal distributions obey

$$N(t) \sim \text{Pois}(\lambda t),$$

it follows that $\mathbb{E}[N(t)] = \lambda t$ and, further, by independent increments,

$$\mathbb{E}[N(t)|N(s)] = N(s) + \lambda(t - s).$$

It follows that

$$\lim_{\epsilon \to 0} \mathbb{E}[N(t)|N(t-\epsilon)] = N(t_-) + \lambda \mathrm{dt}.$$

The expected increment in the standard Poisson process is then $\lambda$dt per time increment dt. Note that, upon defining

$$\lim_{\epsilon \to 0} \mathbb{E}[N(0)|N(-\epsilon)] = 0,$$

the expectation of $N(t)$ above can be expressed as

$$\int_0^t \lim_{\epsilon \to 0} \mathbb{E}[N(s)|N(s-\epsilon)] = \int_0^t \lambda \mathrm{dt} = \lambda t,$$

giving a more general way to compute $\mathbb{E}[N(t)]$.

The generality of this method can be seen in a simple example, with rate bounded by $M$ for convenience:

$$N(t)|N(t_-) \sim \mathrm{Pois}(M \wedge \lambda \mathrm{dt}); \;\; N(0_-) = n_0 \text{ fixed}$$

In other words, it is a "generalized" Poisson process with stochastically varying rate

$$\lambda(t; M)|N(t_-) = \lambda[M \wedge N(t_-)],$$

dependent only on the immediate history of $N$. In this simple model of population growth, the rate of growth at $t$ is proportional to the number of cells alive immediately before, at $t - \epsilon$, with a carrying capacity of $M$.

The expected population increment per time is then given by

$$\lim_{\epsilon \to 0} \mathbb{E}[N(s) - N(s-\epsilon)|N(s-\epsilon)] = N(s_-) + \lambda N(s_) - N(s_-) = \lambda N(s_-)\mathrm{ds}$$

Similarly to before, we can integrate using the initial condition $N(0) = n_0$ to obtain

$$\mathbb{E}[N(t)] = n_0 e^{\lambda t}.$$

In general, by linearity, the process $M(t) = N(t) - \mathbb{E}[N(t)]$ has mean zero.

### 5.1.2 Variation

Having a process analogue of mean above, the incremental variation can be defined similarly as the *predictable variation process*:

$$\langle N(t)\rangle = \int_0^t \mathbb{E}[(N(s) - \mathbb{E}[N(s)]^2)|N(s_-)]\mathrm{ds} = \int_0^t M(s)^2\mathrm{ds}$$

## 5.2 Traditional Setting

The connection with the survival analysis framework before is made clear through the variable $Y_i(t)$, which in general is an indicator of whether the unit $i$ is at-risk of failure at time $t_-$. Being defined as a left-hand limit, $Y_i(t)$ is predictable in the sense of being well-defined as a random variable with respect to the information at $t$. Additionally, set $Y(t) = \sum_i Y_i(t)$, giving the size of the population at risk of event at time $t$.

In the traditional survival setting, $Y_i(t) = 1$ if and only if the unit has not been censored or failed at time $t$.

$$Y_i(t) = \mathbb{I}_{N_i(t_-)=0} = \mathbb{I}_{T_i>t_-}.$$

Note that $Y_i$ is predictable and depends only on the observed data, $T_i = t_i$. Further, note that $\mathbb{E}[Y_i(t)] = S(t_-)$ which is equal to $S(t)$ if the distribution function is continuous.

### 5.2.1 Exponential

Consider the variables $N(t)$ and $Y(t)$ as above, for the standard survival case of one unit with constant failure rate $\lambda$ and no censoring. As before, we compute the expected increment of $Y(t)$; since (by convention) $Y(t)$ is adapted to $\mathcal{F}_{t_-}$, we work instead with the increment

$$Y(t_+) - Y(t) = \lim_{\epsilon\to 0} Y(t+\epsilon) - Y(t).$$

Note,

$$\mathbb{E}[Y(t_+)|Y(t)=0]=0; \mathbb{E}[Y(t_+)|Y(t_-)=1]=Y(t)(1-\lambda \mathrm{dt}),$$

thus,

$$\mathbb{E}[Y(t_+)-Y(t)|Y(t)]=-Y(t)\lambda \mathrm{dt}.$$

This, with initial condition $Y(0)=1$ gives

$$\mathbb{E}[Y(t)]=e^{-\lambda t}.$$

To check, note that

$$\mathbb{E}[T]=\int_0^\infty S(t)\mathrm{dt}=\int_0^\infty e^{-\lambda t}\mathrm{dt}=1/\lambda.$$

The expression

$$\int_0^t \mathrm{dN}(s)Y(s)$$

is thus a random variable indicating whether the unit has failed by time $t$. Thus to compute the probability,

$$F^*(t)=\int_0^t \mathbb{E}[\mathrm{dN}(s)Y(s)|\mathcal{F}_{s_-}]=\int_0^t \lambda\frac{e^{-\lambda s}}{\lambda}=1-e^{-\lambda t}.$$

Particularly, $F^*(\infty)=1$ since there is no censoring.

#### 5.2.1.1 Censoring

Consider the case above, but with independent right-censoring at rate $\theta$. Since $N(t)$ only counts the event, $\mathbb{E}[\mathrm{dN}(t)|N(t_-)]=\lambda \mathrm{dt}$ as before, but as removal from risk occurs at rate $\lambda+\theta$ due to independent censoring, the predicted $Y$-term has $\lambda$ replaced with $\lambda+\theta$. Thus the probability of observing an event by $t$ is

$$F^*(t)=\int_0^t \mathbb{E}[\mathrm{dN}(s)Y(s)|\mathcal{F}_{s_-}]=\int_0^t \lambda\frac{e^{-(\lambda+\theta)s}}{\lambda}=\frac{\lambda}{\lambda+\theta}\left(1-e^{-(\lambda+\theta)t}\right).$$

In particular, the probability of an event occurring before censoring is $F^*(\infty) = \frac{\lambda}{\lambda+\theta}$, as derived earlier.

Note, however, that this mechanism can handle any predictable censoring mechanism, i.e. $Y(t)$ must only be measurable in $\mathcal{F}_{t_-}$.

## 5.3 Cox Model

One approach to fitting a survival model is to use the Cox partial likelihood. Assuming that there are no ties, and letting $t_\cdot \in \mathfrak{T}$ index the failure times and denoting by $i(t)$ the unit to fail at time $t$, and denoting by $\psi(j,t)$ the hazard of unit $j$ at time $t$, this partial likelihood is

$$\prod_{t\in\mathfrak{T}} \frac{\psi(i(t),t)}{\sum_{j\in\mathfrak{R}(t)}\psi(j,t)}.$$

Notably, using this partial likelihood identifies any component of the hazard which depends only on the calendar time:

$$\prod_{t\in\mathfrak{T}} \frac{\phi(i(t),t)\lambda_0(t)}{\sum_{j\in\mathfrak{R}(t)}\phi(j,t)\lambda_0(t)}$$

$$= \prod_{t\in\mathfrak{T}} \frac{\phi(i(t),t)}{\sum_{j\in\mathfrak{R}(t)}\phi(j,t)}.$$

Given a loglinear model, this generates the score

$$\sum_t x_{i(t)} - \frac{\sum_{j\in\mathfrak{R}(t)} x_j(t)e^{\beta^\dagger x_j(t)}}{\sum_{j\in\mathfrak{R}(t)} e^{\beta^\dagger x_j(t)}}$$

and information

$$-\sum_t \frac{\left(\sum_{j\in\mathfrak{R}(t)} e^{\beta^\dagger x_j(t)}\right)\left(\sum_{j\in\mathfrak{R}(t)} x_j x_j^\dagger e^{\beta^\dagger x_j(t)}\right) - \left(\sum_{j\in\mathfrak{R}(t)} x_j e^{\beta^\dagger x_j(t)}\right)\left(\sum_{j\in\mathfrak{R}(t)} x_j^\dagger e^{\beta^\dagger x_j(t)}\right)}{\left(\sum_{j\in\mathfrak{R}(t)} e^{\beta^\dagger x_j(t)}\right)^2},$$

which is positive by a trivial extension of the Cauchy inequality. It's worth noticing that the failure does not appear in this quantity. This appears to just be an accident of the loglinear model.

Finally, note that by fitting a model with the Cox partial likelihood, we may call this in itself a proportional hazards model.

# Chapter 6

# Gap Times

We introduce models for estimating infant mortality through the Cox partial likelihood by introducing a nonparametric term indexed by "time since failure". The model of the hazard rate function is

$$\begin{aligned}\lambda_i(t) &= \lambda_0(t)\psi(t-\tau_{i,t})Y_i(t)\\ &= \lambda_0(t)\psi(\nu_{i,t})Y_i(t),\end{aligned} \tag{6.1}$$

generating the likelihood

$$L(\lambda_0,\psi) = \prod_i \prod_{j:0<t_{i,j}<\infty} \exp\left(-\int_{t_{i,j-1}}^{t_{i,j}} \lambda_0(s)\psi(\nu_{i,s})y_i(s)\mathrm{ds}\right)\lambda_0(t_{i,j})\psi(\nu_{i,t_{i,j}}), \tag{6.2}$$

with the convention that $t_{i,0}=0$ and $t_{i,j}$ is the $j^{\text{th}}$ failure time of unit $i$.

The corresponding Cox partial likelihood is

$$\prod_{j:0<t_j<\infty} \frac{\psi(\nu_{i(t_j),t_j})}{\sum_i \psi(\nu_{i,t_j})y_{i,t_j}}, \tag{6.3}$$

where, again, $t_0=0$ and $t_j$ is the $j^{\text{th}}$ failure time of all units aggregated. As the model produces no ties with probability zero, let $i(t_j)$ be the unit

corresponding to the $j^{\text{th}}$ aggregated failure. Also note that the $y_{i(t_j),t_j} = 1$ by definition of an event occurring at $t_j$ in unit $i_{t_j}$.

The $\psi$ term models the "natural lifetime" of the unit through the hazard rate function, while the $\lambda_0$ term models exogenous effects related to calendar time. Examples of such exogenous effects include seasonality and unfavorable conditions which affect units more-or-less uniformly, such as a heat wave. The partial likelihood aims to remove this effect.

## 6.1 Observed Gaptimes

We will use the term *observed gaptime* to refer to gaptimes for which the likelihood function gives information, the details of which will vary according to the likelihood function used.

Take, as an example, two units with simultaneous initial failures at time 0, and the recorded events:

$$(0, 3, 5, 10+),$$
$$(0, 2, 9, 10+),$$

with observation halted at time 10. Suppose that the Cox partial likelihood is used.

The first unit has observed events with gaptimes 3 and 2, while the second unit has events at gaptimes 2 and 7. Thus the set of observed gaptimes is a superset of $\{2, 3, 7\}$.

The gaptime 4 is also observed according to the definition above, since the term in the Cox partial likelihood corresponding to the second unit's failure at time 9 is informative for the gaptime 4 through the first unit.

Likewise, when the first unit fails at calendar time 3, the second unit has observed gaptime $3 - 2 = 1$.

The full set of observed gaptimes is $\{1, 2, 3, 4, 7\}$

## 6.2 Examples

### 6.2.1 Example One

To see what "information" this model gives, examine the minimal case of two independent on-line units which each fail once, at distinct observed times, $t_1$, $t_2$ respectively, with $t_2 > t_1$. We will suppose that the units renew immediately upon failure, so that $Y_i(t) = 1$ at all times.

The likelihood is then

$$\exp\left[-\int_0^{t_1} 2\lambda_0(s)\psi(s)\mathrm{ds} - \int_{t_1}^{t_2} \lambda_0(s)(\psi(s) + \psi(s - t_1))\right] \lambda_0(t_1)\psi(t_1)\lambda_0(t_2)\psi(t_2 - t_1)$$

and the partial likelihood is

$$\frac{\psi(t_1)}{\psi(t_1) + \psi(t_1)} \times \frac{\psi(t_2)}{\psi(t_2) + \psi(t_2 - t_1)} \propto \frac{\psi(t_2)}{\psi(t_2) + \psi(t_2 - t_1)}.$$

giving two "parameters":

$$\beta_1 = \log(\psi(t_2 - t_1)) = \varphi(t_2 - t_1),$$
$$\beta_2 = \log(\psi(t_2)) = \varphi(t_2).$$

The gradient of the partial likelihood in $\beta$ is then

$$\frac{\partial \ell}{\partial \beta_1} = \frac{-e^{\beta_1}}{e^{\beta_1} + e^{\beta_2}},$$
$$\frac{\partial \ell}{\partial \beta_2} = 1 - \frac{e^{\beta_2}}{e^{\beta_1} + e^{\beta_2}} = \frac{-\partial \ell}{\partial \beta_1}.$$

Solved directly (but heuristically) as a score equation, this gives $\beta_2 = -\infty$ with $\beta_1$ undetermined. However, the maximum partial likelihood estimator is "achieved" with $\beta_1 = -\infty$ and $\beta_2 = \infty$, corresponding to

$$\widehat{\psi(t_2)} = \infty;\ \widehat{\psi(t_2 - t_1)} = 0.$$

### 6.2.2 Example Two

Consider two units, which both come online at time $t = 0$ and which are observed daily through time $t = 4$. Units one and two fail at the following calendar times

$$\text{Unit one: } (1, 3),$$
$$\text{Unit two: } (2, 4).$$

The full likelihood is then

$$e^{-[2\psi(1)\lambda_0(1)+(\psi(1)+\psi(2))(\lambda_0(2)+\lambda_0(3)+\lambda_0(4))]} \times$$
$$\psi(1)\psi(2)^3\lambda_0(1)\lambda_0(2)\lambda_0(3)\lambda_0(4).$$

This can only be solved in $\psi(\cdot)$ insofar as $\widehat{\psi(1)} = 0$; $\psi(2)$ is left undetermined. From the equation $\lambda_0(1) = \frac{1}{2\psi(1)}$, it would then be required that $\lambda_0(1) = \infty$. The interpretation of this estimate for any practical application is unclear, and certainly cannot be taken literally.

From the equations

$$\lambda_0(j) = \frac{1}{\psi(1) + \psi(2)} = \frac{1}{\psi(2)},\ j \in \{2, 3, 4\},$$

it can be seen that the model cannot distinguish the effects of calendar time and gaptime. Each calendar time $j \in \{2, 3, 4\}$ is associated with an event of gaptime 2; the multiplicative model having only the term $\lambda_0(2)\psi(2)$ can only identify that the calendar-time effect (if any) is homogenous in the sample, and reciprocal to the gaptime effect.

#### 6.2.2.1 Cox Partial Likelihood in $(\lambda_0, \psi)$-Model

We turn to the Cox partial likelihood method to remove the dependency on calendar time,

$$L = \left(\frac{\psi(1)}{\psi(1) + \psi(1)}\right)\left(\frac{\psi(2)}{\psi(1) + \psi(2)}\right)^3.$$

Here, the first failure is uninformative, as opposed to the full-likelihood. The remaining events always occur 2 time units after the failure of the affected unit, and 1 time unit after the failure of the other unit.

The partial loglikelihood is

$$\ell = -\log(2) + 3\varphi(2) - 3\log(\psi(1) + \psi(2)),$$

giving the score equations

$$\frac{d\ell}{d\varphi(1)} = \frac{-3\psi(1)}{\psi(1) + \psi(2)}$$
$$\frac{d\ell}{d\varphi(2)} = 3 - \frac{3\psi(2)}{\psi(1) + \psi(2)} = -\frac{d\ell}{d\varphi(1)}$$

These equations have no finite solution. However, as the derivatives are strictly negative and positive, respectively, and equal in magnitude, it can be said that $\widehat{\psi}(1) = 0$ and $\widehat{\psi}(2) = \infty$.

#### 6.2.2.2 Full Likelihood in $\psi$-Model

Finally, we remove $\lambda_j$ completely from the model, or, equivalently, set $\lambda_j = 1$, and consider the full likelihood

$$L = e^{-5\psi(1)-3\psi(2)}\psi(1)\psi(2)^3.$$

The loglikelihood,

$$\ell = -5\psi(1) - 3\psi(2) + \varphi(1) + 3\varphi(2),$$

gives the score equations

$$\frac{d\ell}{d\varphi(1)} = 1 - 5\psi(1)$$
$$\frac{d\ell}{d\varphi(2)} = 3(1 - \psi(2))$$

and the estimates $\widehat{\psi}(1) = 0.2$, $\widehat{\psi}(2) = 1$. The Fisher information is $5\psi(1), 0, 3\psi(2), 0$.

### 6.2.3 Example Three

Suppose that $n_f$ failures occur among $n$ units, according to a Poisson process with rate $0 < \psi(\nu)\lambda_0(t) < \infty$. Let $0 = t_{(0)} < t_{(1)} < t_{(2)} < \cdots < t_{(n_f)}$ be the ordered failure times.

The full likelihood may easily be given, but skipping to the partial likelihood for reasons of identifiability, note that it will have form

$$\frac{1}{n}\frac{\psi_2}{\sum_j \psi_{2,j}}\frac{\psi_3}{\sum_j \psi_{3,j}}\cdots,$$

with $\psi_i = \psi_{i,j_i}$ for some $\psi_{i,j_i}$ in its corresponding denominator.

Note first that each fraction term is either equal to $1/n$, or contains at least two different $\psi_{i,j}$ terms in the denominator sum. Note also that the $i^{\text{th}}$ failure occurs at time $t_{(i)}$ and has associated failure time $t_{(i)} - t_{(j)}$ for some $0 \leq j < i$.

Once a term $\psi(t_{(i)} - t_{(j)})$ occurs in the numerator of a term, it will never (w.p. 1) occur again anywhere in the partial likelihood. Otherwise, it would be necessary that

$$\begin{aligned} t_{(i+1)} - t_{(j')} &= t_{(i)} - t_{(j)} \\ t_{(i+1)} - t_{(i)} &= t_{(j')} - t_{(j)},\ j < j' < i+1. \end{aligned}$$

However, since the increments are drawn from a continuous distribution, the result follows.

As a result, any gradient is either identically zero or has form

$$\frac{\partial \ell}{\partial \varphi_i} = \mathbb{I}_i - \frac{n_i \psi_i}{n_i \psi_i + c_i},$$

where $0 < n_i < n$; $\mathbb{I}_i$ takes the value 1 if $\psi_i$ occurs in a numerator and is 0 otherwise; and $c_i > 0$ is a sum of terms, not all of which are equal to $\psi_i$.

It follows that $\widehat{\varphi}_i \in \{-\infty, \infty\}$ for each $i$ for which the gradient is not identically zero. Thus, $\widehat{\psi(\nu)}$ for any observed gaptime $\nu$ is either $\widehat{\psi(\nu)} = \infty$ if a failure occurs at gaptime $\nu$, or is $\widehat{\psi(\nu)} = 0$ otherwise.

Thus, the nonparametric partial likelihood in $\varphi$ with gaptimes in $\mathbb{R}$ is of no direct use in standard estimation, at least if it is assumed that the failure times arise from a Poisson process. The same argument applies under the standard assumption of no ties in the data.

### 6.2.4 Conceptual

To illustrate that a gaptime is partly a matter of interpretation, consider

$$\lambda_0(t) = \lambda_1 \text{ if } t \in [0, 1),$$
$$\lambda_0(t) = \lambda_2 \text{ if } t \in [1, 2),$$

with time in years. This model allows yearly variation in the hazard. Suppose that $\psi(\nu)$ is unknown and that $n$ units enter a trial at time 0, while another $n$ units from the same population enter an identical trial at time 1.

For simplicity, suppose that in each trial there is a fixed censoring time, say 0.75 and 1.75 respectively. Note that this implies that trial $i$ will be finished within a year and thus involve only the year-specific $\lambda_i$. Alternatively, one may consider a general censoring time $C_{i,j}$ as long as it is independent and uninformative for $i$ and $\lambda_i$ or $\psi$ are large enough that

$$\Pr[T_{i,j} > i] \approx 0.$$

Let the observed times be denoted as $t_{i,j}$, with $i \in \{1, 2\}$ denoting the year and $1 \leq j \leq n$ indexing the units under study, with corresponding observation indicators $\delta_{i,j}$ and $n_i^F = \sum_j \delta_{i,j}$.

Thus the loglikelihood function is

$$\sum_{i,j} \delta_{i,j} \log \lambda_i + \sum_{i,j} \delta_{i,j} \log \psi(t_{i,j}) - \sum_{i,j} \lambda_i \int_0^{t_{i,j}} \psi(s)\mathrm{ds}$$
$$= \sum_i n_i^F \log \lambda_i + \sum_{i,j} \delta_{i,j} \log \psi(t_{i,j}) - \sum_{i,j} \lambda_i \int_0^{t_{i,j}} \psi(s)\mathrm{ds},$$

so the score "estimator" for $\lambda_i$ (which of course depends on $\psi$) is

$$\widehat{\lambda_i} = \frac{n_i}{\sum_j \int_0^{t_{i,j}} \psi(s)\mathrm{ds}}.$$

However, note that this example is formally identical to a standard Cox regression problem, with the gaptime taking the role of calendar time and the gaptime hazard $\psi$ taking the place of the baseline hazard $\lambda_0$, if we introduce the parameter $\beta_i = \log \lambda_i$ to represent the "effect" of the year of the trial.

Although a contrived example, it does illustrate that the gaptime is partly just a matter of interpretation. In the general case, the gaptime effect is a nuisance parameter[1] for estimation of the baseline hazard, while the baseline hazard is a nuisance parameter for the gaptime effect.

The basic idea of this example will be useful in later, more involved, problems. In general: the gaptime may be viewed as resetting to 0 at each failure; however, it is equivalent to imagine that each epoch is a separate trial (linked by the covariates and, maybe, prior information), and that the calendar time is simply shifted by the time under observation (which is, after all, just a sample from a distribution).

### 6.2.5 Infeasibility

We note that in the case of gaptimes in discrete units $(1, 2, \ldots)$ certain likelihood functions will never be generated. For example,

$$\frac{1}{2}\left(\frac{\psi(2)}{\psi(1)+\psi(2)}\right)^2\left(\frac{\psi(1)}{\psi(1)+\psi(2)}\right)^2$$

will never occur.

[1]Note, however, that the gaptime $\nu_t$ at $t$ is itself a random variable adapted to $\mathcal{F}_{t^-}$ and predictable on $\mathcal{F}_t$. This is important to keep in mind.

To see this, take two units failing at times $(1, 3, 4)$ and $(2)$. This generates the likelihood

$$\frac{1}{2}\left(\frac{\psi(2)}{\psi(1)+\psi(2)}\right)^2\left(\frac{\psi(1)}{\psi(1)+\psi(2)}\right).$$

There is now no way to extend the sequence of failure times to add another $\frac{\psi(1)}{\psi(1)+\psi(2)}$ term. The only way to get the numerator is to extend the first unit's failures to $(1, 3, 4, 5)$. However, this can only introduce the term $\frac{\psi(1)}{\psi(1)+\psi(3)}$.

Inspection will assure that there is no other ordering of failure times to generate the likelihoods mentioned above, apart from simply exchanging units one and two.

This property is interesting but is beyond the scope of this thesis.

### 6.2.6 Traditional

Note that, in the absence of staggered entry times, this method is inapplicable when $Y_i(t) = \mathbb{I}_{\tilde{T}>t}$. Specifically, with failure times ordered as

$$t_{(1)} = \nu_{i(1),t_{(1)}} < t_{(2)} = \nu_{i(2),t_{(2)}} < \cdots < t_{(N)} = \nu_{i(N),t_{(N)}},$$

the partial likelihood is

$$\prod_i \frac{\psi(t_{(i)})}{(N-i+1)\psi(t_{(i)})} = \frac{1}{N!},$$

so that no information is available.

Thus, with this $Y_i(t)$ indicator, consider the case where one unit enters the study, in health, immediately at $t = 0$ and fails at $t_1$, while another enters, in health, after a delay $\delta$ and fails at $\delta + t_2$. We can immediately exclude the case $\delta > t_1$, as this case is equivalent to two separate studies, each yielding no information. Now, excluding a tie, there are two cases, $t_1 < \delta + t_2$ and $t_1 > \delta + t_2$.

The same manipulations as above give for the first case

$$\widehat{\psi(t_1)} = \infty;\ \widehat{\psi(t_1 - \delta)} = 0,$$

and for the second case

$$\widehat{\psi(t_2)} = \infty;\ \widehat{\psi(t_2 + \delta)} = 0.$$

In this scenario, with $N > 2$ units available for study (and under a reasonable failure model), it is clear that a staggered-entry schedule will identify at most $N - 1$ points, $\nu$, where $\psi(\nu) = \infty$ and $\frac{(N-1)N}{2}$ points where $\psi(\nu) = 0$. In fact, given $N_D$ *distinct* staggered entry points, $N_D \geq 2$, there will (w.p. 1 under a reasonable failure model) be $N - 1$ points with $\psi(\nu) = \infty$ and $\frac{(N_D-1)N_D}{2}$ points with $\psi(\nu) = 0$.

## 6.3 Summary

Although the model is informative under reasonable conditions, particularly when the units are subject to renewal, it is clear that the solution requires regularization.

First, the model only identifies $\psi$ at $\mathcal{O}(|T|N)$ points, which cannot be designed in advance. Second, $\widehat{\psi(\cdot)}$ will be either 0 or $\infty$, where it is defined.

We also note that the use of the at-risk indicator, $Y_i(t)$, handles some conceptually distinct cases in a general way. First, using $Y_i(t) = \mathbb{I}_{\tilde{T}>t}$ gives the standard survival analysis context, although this is restricted in utility as described above.

However, on-line systems, such as electrical feeders, can be modeled substantively. Given detailed maintenance information, the $Y_i(t)$ can be used to indicate the times at which the unit is operational, and thus susceptible to failure. The latter case is, of course, preferable since the method is data-intensive.

Specifically, the set of observed renewal times, $\{(t-\tau_{i,t})\}$, grows as $N|T|$, while this set is restricted to $\mathcal{O}(N^2)$ elements in the standard right-censoring setting.

In the following sections, we examine two complementary Bayesian approaches to the identifiability problems above.

We conclude with a heuristic explanation of the derivation of the Cox likelihood, i.e. marginalizing hazard estimation onto the set of times with failure.

## 6.4 Marginalizing Times without Failure

We consider the contribution to the likelihood from the observation of no failures between times $t_{i-1}, t_i$, assuming a Poisson process without censoring and that $\varphi(\cdot) < \infty$:

$$L = e^{-\int_{t_{i-1}}^{t_i} \lambda_0(u) \sum_{j \in \Re(u)} e^{\varphi(i-\tau_{u,j})} \mathrm{d}u}.$$

Taking the derivative with respect to $\lambda_0$ at time $s \in (t_{i-1}, t_i)$:

$$\frac{\partial L}{\partial \lambda_0(s)} = \left(e^{-\int_{t_{i-1}}^{t_i} \lambda_0(u) \sum_{j \in \Re(u)} e^{\varphi(i-\tau_{u,j})} \mathrm{d}u}\right) \times \left(-\lambda_0(s) \sum_j e^{\varphi(s-\tau_{s,j})}\right), \tag{6.4}$$

which is negative for all positive values of $\lambda_0(s)$. Since $\lambda_0 \geq 0$ by definition, the maximum likelihood estimate of baseline hazard is $\widehat{\lambda_0(s)} = 0$, which gives the mle of failure rate

$$\hat{\lambda_0}(s) \sum_j e^{\varphi(s-\tau_{s,j})} = 0.$$

Substituting this into the likelihood, we see that it does not depend on $\varphi$ when there are no failures, reducing the estimation problem to event times. This result, derived more formally, is also valid under random censoring, as shown by Cox and given in Andersen.[2].

Thus, since intervals without failures give no information about $\varphi$, we can reduce the problem of estimating $\varphi$ by conditioning on failures having occurred at the observed times. The probabilities under consideration are then the probabilities of each *observed unit* failing at time $t$, given that *some unit* (under risk) failed at time $t$, which is:

$$\prod_t \frac{\text{unit } i \text{ fails at } t}{\text{some unit fails at } t} = \prod_t \frac{\lambda_0(t)e^{\varphi(t-\tau_{t,i})}}{\lambda_0(t)\sum_j e^{\varphi(t-\tau_{t,j})}} = \prod_t \frac{e^{\varphi(t-\tau_{t,i})}}{\sum_j e^{\varphi(t-\tau_{t,j})}},$$

which gives the Cox likelihood for $\varphi$ at those values $t-\tau_{t,j}$, which are observed.

After the estimate of $\varphi$ is obtained, we can derive an estimate of $\Lambda_0 = \int_0^t \lambda_0$ through the weighted non-parametric Nelson-Aalen estimator.[14]. This $\Lambda_0$ is smoothed and used directly in computing the test-penalty or, if desired, $\lambda_0$ may be approximately estimated by differentiating the smoothed version.

# Chapter 7

# Score Derivation

We derive the score function for the gaptime. Assume a hazard rate of form

$$\lambda(t) = \lambda_0(t)\psi(\nu_t),$$

which for a fixed observation period $[0, T]$, without (for now) censoring or other difficulties, gives the likelihood

$$L = \prod_i e^{-\int_0^T \lambda_0(s)\psi(\nu_{i,s})\mathrm{ds}} \prod_j \lambda_0(t_j)\psi(\nu_{i,t_j}),$$

and the log-likelihood

$$l = \sum_i \left( \sum_j \log \lambda_0(t_j) + \log \psi(\nu_{i,t_j}) \right) - \int_0^T \lambda_0(s)\psi(\nu_{i,s})\mathrm{ds}.$$

Differentiating this gives the score function for $\psi(\nu_\cdot)$ at a fixed $\nu_\cdot$:

$$S = \sum_i \left[ \frac{|\{j : \nu_{i,t_j} = \nu_\cdot\}|}{\psi(\nu_\cdot)} - \sum_j \mathbb{I}_{\nu_{i,t_j} \geq \nu_\cdot} \lambda_0(t_{i,\nu_\cdot}) \right],$$

which depends on the nuisance parameter $\lambda_0$.

A greater problem than mere dependence lies in the fact that (see below) the estimator $\lambda_0(\cdot) = 0$ at many of the values $t_{i,\nu_\cdot}$.

This estimator is applied to the Consolidated Edison dataset in the later chapters.

### 7.0.1 Estimation of Calendar Time an Gap Time Effects

One apparent solution is to estimate $\lambda_0$. Unfortunately, this estimator will depend on $\psi$.

Nonetheless the "score" is

$$S = \sum_i \frac{|\{j : t_j = t_.\}|}{\lambda(t_.)} - \psi(\nu_{i,t_.}),$$

which when solved for 0 gives the "estimator"

$$\widehat{\lambda_0(t_.)} = \frac{\sum_i |\{j : t_j = t_.\}|}{\sum_i \psi(\nu_{i,t_.})}.$$

Notably, when this partial estimator is integrated, under the Poisson assumption of no simultaneous failures, a weighted version of the Nelson-Aalen estimator is obtained:

$$\widehat{\Lambda_0(t)} = \int_0^t \frac{\mathrm{dN}_.(s)}{\sum_i \psi(\nu_{i,t_.})}.$$

Of course, the problem of estimation is the presence of two nonparametric estimands in the score equations.

### 7.0.2 Simple Solution

One straight-forward solution is to iterate between the , using the $i^{\text{th}}$ estimate of $\lambda_0$ as a plug-in estimator for the $(i+1)^{\text{th}}$ estimate of $\psi$. However, the estimators as stated are unsuitable for this purpose, as they take non-zero values only at event times. Thus it is necessary to plug a smoothed version of $\widehat{\lambda_{0;i}}$ into the estimator $\widehat{\psi_{i+1}}$ for $\psi$, and vice versa.

The simplest way to achieve this smoothing is to apply a kernel to the integrated version of $\widehat{\lambda_0}$ (resp. $\widehat{\psi}$), and then to differentiate the kernel to obtain the smoothed $\widehat{\lambda_{0;s}}$.

### 7.0.3 Complementary Solution

One might hope to obtain an estimator of $\lambda_0$ which corrects for $\psi$ automatically.

If only an estimator of $\lambda_0$ is desired, this can solve the problem immediately. If a full solution is desired, the corrected estimator of $\lambda_0$ can be smoothed and used to in the estimator of $\psi$ without the need to reiterate.

In other situations, when the gaptime is of principal interest, the roles of $\lambda_0$ and $\psi$ would be reversed in the first step above.

## 7.1 Nuisance Tangent Space Method

In this section we follow closely the method used in Tsiatis[24] to derive the Cox proportional hazards model, to derive corrected estimators for $\psi$ and $\lambda$.

### 7.1.1 Intuition

Before heading this way, we note briefly some intuition, analogous to that associated with the Cox model, which may be used to anticipate the general form of these estimators.

The Cox proportional hazards model can be understood as estimating regression parameters, $\beta$, by conditioning on sets of times at which the ancillary "baseline hazard" $\lambda_0$ is constant among units. Specifically, by conditioning on the failure times $\{t_k\}$, the partial (conditional) likelihood

$$\prod_k \frac{e^{\beta^\dagger \mathbf{x}_{i(k)}(t_k)}\lambda_0(t_k)}{\sum_j Y_j(t_k)e^{\beta^\dagger \mathbf{x}_j(t_k)}\lambda_0(t_k)} = \prod_k \frac{e^{\beta^\dagger \mathbf{x}_{i(k)}(t_k)}}{\sum_j Y_j(t_k)e^{\beta^\dagger \mathbf{x}_j(t_k)}} \tag{7.1}$$

is generated. Note that, at failure time $t_k$, the units in the riskset apart from $i(k)$ have not failed. In fact, they are remarkable only in that they match the failed unit in calendar time $t_k$.

Analogously, if we consider an estimator (of potentially, say, $\beta$ or $\lambda_0$) which attempts to compensate for *gaptimes*, we might expect the estimator to match among units which match the failed unit in terms of its most recent gaptime, $\nu_{i(k),k}$, rather than the calendar time $t_k$.

Further note that this matching may include one unit several times, and may even match the failed unit to itself, say by including any unit at any calendar times $t$ such that

$$\nu_{i,t} = \nu_{i,t_k}.$$

Particularly, note that the number of matching units will be roughly proportional to the number of gaptimes larger than that of the failed unit.

### 7.1.2 Notes

It is worth noting that, intuitively, any model space generated by functions of the calendar time, $\{\alpha(t)\}$, will necessarily contain any model generated by functions of the gaptimes. After all, one can, it would seem, always reduce the calendar time $t$ to the gaptime by taking $t - \tau_t$.

However, this is not quite correct. Note that $\tau_t$ necessarily involves information in the measure space $\mathcal{F}_t$; particularly, it requires not only the information in $\mathcal{F}_{\tau_t}$ that a failure occurred at this time, but also that there have been, in the intervening $(\tau_t, t)$, no further failures. It is the dependence of the intensity of the process being studied here on the past of the process which leads to it being called sometimes a *generalized* Poisson process, rather than merely an "inhomogeneous" one.

Note, further, that the baseline hazard function itself is a free function of calendar time, but one which does not depend on any information of the process itself. Thus, we can say that $\lambda_0(t)$ is, for all $t$, $\mathcal{F}_0$ measurable. In fact in most cases, $\lambda_0(t)$ will be effectively a parameter; however, sometimes, for

example when units have unobserved characteristics pertinent to their survival (e.g. the type of cabling in feeders, or genetic factors in humans), it may be useful to induce a distribution over $\lambda_0$ in the spirit of latent stratification.

### 7.1.3 Derivation

With the above notes in mind, we can adapt the methods of [Tsiatis] to the problem at hand. Note, now, that neither of the model spaces of

$$\alpha_1(t) \in \mathcal{F}_0 \text{ or}$$

$$\alpha_2(\nu_t) \in \mathcal{F}_t$$

properly contain the other. This would not be the case if $\alpha_1(t)$ were replaced by

$$\alpha_1'(t) \in \mathcal{F}_t;$$

in which case, the two model spaces would be equal to each other.

It is easy to see this, since $\nu_t = t - \tau_t$, and $\tau_t$ is predictable within $\mathcal{F}_t$, since $\tau_t$ depends only on the time of *past* failures. !!!(make more precise)

Since the model space of $\alpha_1(t)$ (evaluated at the truth $\alpha_1 = \lambda_0$) does not include information about $\nu_t$ for any of the units, the adaptation of the Tsiatis approach involves an estimate of $\nu_t$ and thus $\psi(\nu_t)$ from the data.

We continue with a review of some semiparametric theory, after which the derivation will be clear.

#### 7.1.3.1 Discrete Version

Take the discrete model that

$$\psi(\nu) = \sum_l \psi_l \mathbb{I}_{\nu \in [l_l, l_r)}$$

with $L \in \{1, \cdots, L\}$; $1_l = 0$; and $L_r = \infty$, with the intervals fixed. We will use the methods above to find the estimator for the parameters $\psi$.

From the methods above, the nuisance tangent space is spanned by functions of the calendar time, $\alpha(t)$, which do not depend on any information associated with the process. That is, deterministic functions of calendar time, which is notable here if only because $\psi(\nu(t))$ is, if interpreted as a function of $t$ alone, a stochastic function. Specifically, the span is over functions $\alpha \in \mathcal{L}^2$ below:

$$\int_0^t \alpha(s)\mathrm{dN}(s) - \alpha(s)\lambda_0(s)\psi_0(\nu_s)\mathrm{ds} = \int_0^t \alpha(s)\mathrm{dM}_0(s),$$

where $\mathrm{dM}_0$ is to emphasize that the martingale is defined on the true values $\lambda_0$, $\psi_0$.

# Chapter 8

# Radial Basis Function Gaussian Prior

## 8.1 Definition

Returning to the partial likelihood in $\psi$,

$$\prod_t \frac{\psi(\nu_{i(t),t})Y_{i(t),t}}{\sum_j \psi(\nu_{j,t})Y_{j,t}},$$

we consider adding a prior to the function $\psi(\nu)$ to provide smoothing between observed gap times.

Analogously to the loglink commonly used in inference of a Poisson rate in a generalized linear model, we introduce a log-Gaussian process generating the gaptime effect $\psi$. Specifically, introduce the marginal transformation

$$\varphi(\nu) = \log \psi(\nu)$$

and define $\varphi$ as a Gaussian process.

A univariate Gaussian process is characterized as having Gaussian marginal distributions,

$$\varphi(\nu) \sim \mathcal{N}(\mu(\nu), \sigma^2(\nu)),$$

and a specified positive-semidefinite covariance function

$$\mathrm{cov}(\nu_i, \nu_j) = \Sigma(\nu_1, \nu_2).$$

The general properties of a covariance function are symmetry and positive-semidefiniteness, the latter being the property of having

$$\int \Sigma(\nu_1, \nu_2) f(\nu_1) f(\nu_2) \mathrm{d}\nu_1 \mathrm{d}\nu_2 \geq 0$$

for any $f(\nu)$ satisfying

$$\int |f(\nu)|^2 \mathrm{d}\nu < \infty.$$

The parametrization of the covariance function will be indicated in a subscript, e.g. $\Sigma_\rho(s,t)$ parametrized in $\rho$.

The stationary Gaussian process is simpler, requiring identical distribution under shifts:

$$\varphi(\nu) =_{\mathcal{D}} \varphi(\nu + \delta)$$

for all $\delta$, whenever both sides are defined. The stationary process is often adequate or even desirable for use as a prior in statistical modeling as often there is no reason to anticipate that the process varies temporally in a known way. Of course, the posterior process need not be stationary and may need to be sampled or otherwise estimated if its full distribution is desired.

The stationary Gaussian process is specified with $\mu(\nu) = \mu$ constant and a covariance function,

$$\Sigma(\nu_1, \nu_2) = \Sigma(\nu_2 - \nu_1),$$

depending only on the distance $\nu_2 - \nu_1$.

## 8.2 Radial Basis Function

We use the standard two-parameter *radial basis* function for covariance, which is specified by

$$\Sigma_{a,b}(s,t) = a^2 \exp\left(\frac{-(s-t)^2}{b^2}\right).$$

The parameters $a^2$ and $b$ are often called, respectively, the marginal variance and *characteristic time scale*.

Standard results allow the marginalization of the Gaussian process onto the observed gap times, $(\nu_i)$, and first-order methods are used to fit these marginals, passed through the smooth transformation above.

## 8.3 Left-Censoring

Since this dataset was obtained by observation beginning at a fixed arbitrary time, the value $\nu_{i,t}$ is unknown for all $t < t_{i,(1)}$. This left-censoring should be uninformative, so we set $Y_i(t) = 0$ for all such $t$ and begin the analysis at the first time where $\sum_i Y_i(t) > 1$.

## 8.4 Fitting

The "partial posterior probability" in $\psi$ is the product of the marginalized Gaussian process prior, $\pi$, and the Cox partial likelihood $L$.

For a first pass at estimation, we consider "log-partial posterior probability" in $\varphi$,

$$\begin{aligned}\log L = l + \log \pi = \; & \textstyle\sum_t \varphi(t - \tau_{i,t}) - \\ & \log \textstyle\sum_{j \in \Re(t)} \psi(t - \tau_{j,t}) - \\ & \tfrac{1}{2}\varphi^\dagger K^{-1} \varphi.\end{aligned}$$

We apply the Newton-Raphson method to find the maximum partial a-posteriori estimate. The gradient with respect to $\varphi$ is

$$\nabla(l + \log \pi) = \sum_t \left[ e_{\tau_{i(t),t}} - \frac{\sum_j \psi(t - \tau_{j,t}) e_{\tau_{j,t}}}{\sum_j \psi(t - \tau_{j,t})} \right] - K^{-1}\varphi.$$

Since these quantities involve mostly the total hazard at time $t$, the correct notation simplifies it greatly. Introducing the notation

$$\begin{aligned} s_t &= \textstyle\sum_\cdot \psi(t - \tau_{\cdot,t}) \\ s_{t;i} &= \textstyle\sum_\cdot \psi(t - \tau_{\cdot,t}) e_{\tau_i} \\ s_{t;i,j} &= \textstyle\sum_\cdot \psi(t - \tau_{\cdot,t}) e_{\tau_i} e_{\tau_j}, \end{aligned}$$

this may be rewritten as

$$\nabla(l + \log \pi) = \sum_t \left[ e_{\tau_{i(t),t}} - \frac{s_{t;(j,t)}}{s_t} \right] - K^{-1}\varphi.$$

with Hessian

$$\nabla\nabla(l + \log \pi) = -\left[ \sum_t \frac{s_t s_{t;(i,t),(j,t)} - s_{t;(i,t)} s_{t;(j,t)}}{s_t^2} \right] - K^{-1}.$$

The Hessian is negative-definite (see below), so Newton-Raphson optimization is guaranteed to converge to the maximum a posteriori estimate. The step-size is dynamically adjusted and is stopped on a relative improvement of the partial posterior probability by less than $10^{-8}$.

## 8.5 Model Selection

A fully Bayesian analysis would introduce a hyperprior and sample the joint posterior in $a^2, b$, and $\varphi$. Instead, we use model selection or empirical Bayes, where the selection of the hyperparameters is performed using the fitted model above to evaluate a criterion in terms of the hyperparameters. With fixed $a^2$ and $b$, call the resulting MAP estimate above $\hat{\varphi}(a^2, b)$.

Conceptually, a criterion in $\varphi$, $C(\varphi)$, is used to define a criterion in the hyperparameters, $C_2(a^2, b)$, as

$$C_2(a^2, b) = C(\hat{\varphi}(a^2, b)),$$

which is maximized in terms of the hyperparameters. Note, however, that $\varphi$ is an estimate and thus depends on the sample, $D$, as well. This will be important, so denote this explicitly as $\varphi_D(a^2, b)$. In addition, the criterion $C$ is estimated, typically on a subset of data not used in the estimation of $\phi$, so likewise denote the criterion as

$$C_2(a^2, b, D, D') = C_{D'}(\hat{\varphi}_D(a^2, b)).$$

That is, the data $D'$ is used to *validate* the estimate $\hat{\varphi}_D(a^2, b)$.

The procedure is fixed by selecting a function $C$ and an algorithm to use the data to search over values of the hyperparameters. The utilization of the data in evaluating the function $C$ will be important, so to this explicit, calling the subset of data used in evaluation $D$, denote the criterion function as $C(\varphi, D)$.

In this case, the function $C(\varphi, D)$ the partial likelihood of $\varphi$ in the selected data $D$. The parameters will be selected by defining a grid of values $(a_i^2, b_j)$ and, for each point in the grid, estimating the value $C_2(a_i^2, b_j)$, at which point the $a^2, b$ corresponding to the maximum $\widehat{C_2}$ is selected.

### 8.5.1 Cross-Validation

An accepted method for dividing the data is to iteratively divide the data into a training set, $D$, and disjoint validation set $D'$.

Typically, the $D$ and $D'$ partition the full dataset, in which case the method is called cross-validation, with each sampled partition called a "fold." The number of folds is typically limited by computational considerations; the spe-

cial case of $N$-fold cross-validation will exhaust the data in the sense of validating the model on every datapoint by iteratively assigning to $D'$ every singleton set. This special case is termed "Leave-One-Out Cross-Validation," or LOOCV, and often considered ideal.

The application of this method to longitudinal data requires a specific definition of the division of the data. The units of partitioning will be taken as the individual units under observation, meaning that the units will not be further sub-divided by time. This is at least somewhat reasonable since the estimand is a function of the gaptime, which cannot easily be localized in time.

Recall, however, that the partial likelihood requires at least $N = 2$ units to be informative. In light of this, the "ideal" is to conduct "Leave-Two-Out Cross-Validation," L2OCV, by selecting for each fold each of the $\binom{N}{2}$ possible pairs of feeders for the validation set $D'$.

The sampling of $C_2$ is implemented by taking the mean of the partial likelihood sampled over all $\frac{N(N-1)}{2}$ folds. The $(a_i^2, b_j)$ achieving the greatest mean is selected as the model.

Using all folds is tractable on our current dataset, however if necessary the hold-out pairs may be randomly sampled.

## 8.6 Baseline Estimation

Under this model, the cumulative hazard rate is identified as

$$\Lambda_i(t) = \int_0^t \lambda_0(u)\psi(\nu_i(u))Y_i(u)\mathrm{du}.$$

Using the first order equation

$$\mathrm{dN}_i(u) \approx \lambda_i(u) = \lambda_0(u)\psi(\nu_i(u))Y_i(u)\mathrm{du}$$

obtain

$$\lambda_0(u) \approx \frac{\mathbb{I}_{\sum Y_i(u)>0}}{\sum Y_i(u)} \sum_i \frac{\mathrm{dN}_i(u)}{\psi(\nu_i(u))},$$

giving an immediate non-parametric point estimate for $\Lambda_0$ as

$$\widehat{\Lambda_0(t)} = \int_0^t \frac{\mathbb{I}_{\sum Y_i(u)>0}}{\sum Y_i(u)} \sum_i \frac{\mathrm{dN}_i(u)}{\psi(\nu_i(u))}.$$

Kernel methods can be applied to estimate $\lambda_0(t)$ from this estimate.

Note that this estimator adjusts for infant mortality in units and thus gives an estimate of $\hat{\Lambda}_0(t)$ as a pure exogenous effect. The effect of this is uncl

Specifically, in the presence of significant infant mortality, we would expect this $\hat{\lambda}_0(t)$ to be attenuated, compared to the direct Nelson-Aalen estimate.

## 8.7 Application

We apply this method to the Long Island, Queens subnetwork of the New York City electric grid system.

The failure gap times are reduced to percentiles, and the Gaussian process applied across the mean time of each bin. The figure following, which compares the smoothed and unsmoothed versions, illustrates the value of smoothing in this problem.

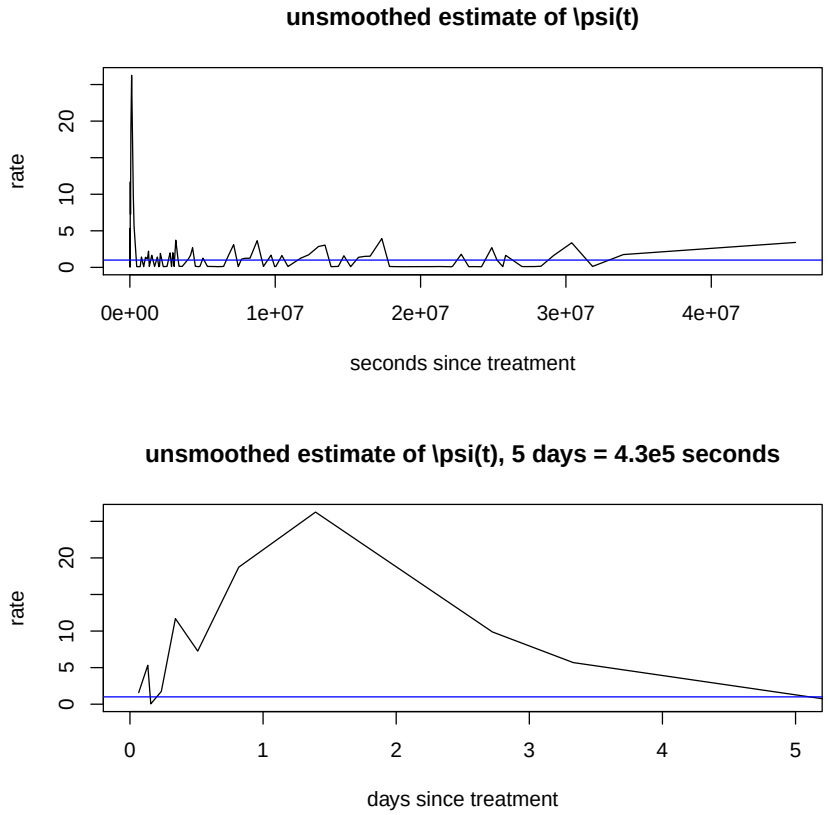
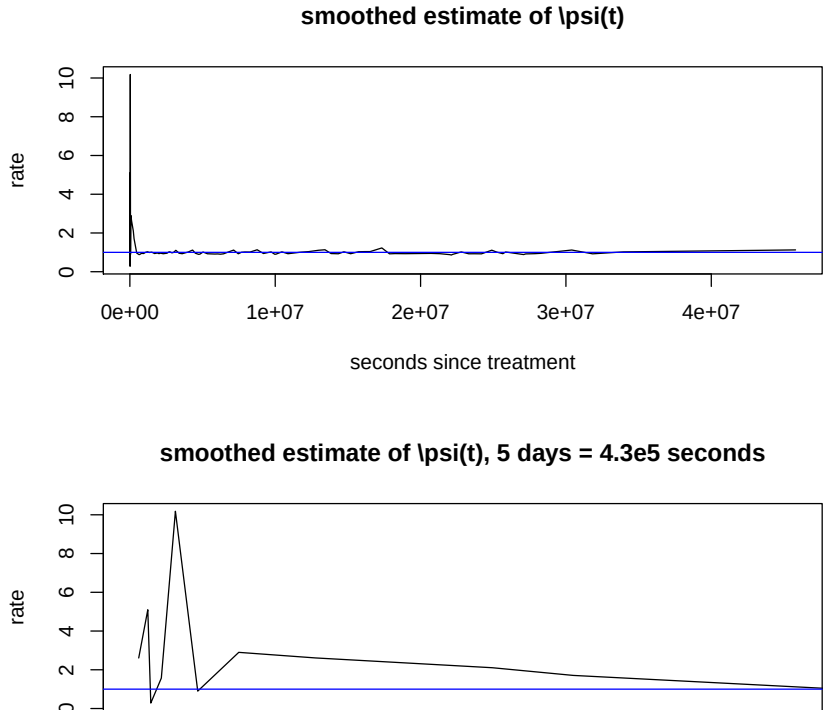



Figure 8.1: Unsmoothed Cox Estimator of Subsequent Effect of Failure



# Chapter 9

# Ornstein-Uhlenbeck Prior

## 9.1 Definition

The Ornstein-Uhlenbeck process is well-known and can be interpreted physically as the displacement (in time) of a spring with constant $\theta$; and (possibly random) initial displacement $\varphi_0$, subject to shocks modeled as a Brownian motion with intensity $\sigma^2$.

The process may be specified by

$$\mathrm{d}\varphi_\nu = -\theta\varphi_\nu \mathrm{d}\nu + \sigma \mathrm{dW}_\nu,$$

with $W_t$ a standard Brownian motion. For now we will take $\varphi_0$ as a fixed parameter, to be inferred. Since $\mathrm{d}\varphi_\nu$ is predictable, the process is Gaussian and its mean and covariance ($\nu_s < \nu_t$) are given by Ito's formula as

$$\begin{aligned} \mu_\nu &= \mathbb{E}[\varphi_\nu] = \varphi_0 \exp(-\theta t), \qquad (9.1) \\ \mathrm{cov}(\varphi_{\nu_s}, \varphi_{\nu_t}) &= \frac{\sigma^2}{2\theta}\left[\exp(\theta_{\nu_s} - \theta_{\nu_t}) - \exp(-(\theta_{\nu_s} + \theta_{\nu_t}))\right]. \end{aligned}$$

As this is just a Gaussian process with specified covariance and mean, we can marginalize out values of $\nu$ which are unobserved, giving the mean vector

and covariance matrix $\mu(\varphi_0, \theta)$, $\Sigma(\theta, \sigma^2)$ and the "partial posterior probability" (variable dependence suppressed):

$$= \frac{1}{\sqrt{\det(\Sigma)}} \exp\left[\frac{-(\varphi-\mu)^\dagger \Sigma^{-1} (\varphi-\mu)}{2}\right] \times \prod_t \frac{\exp(\varphi_{i(t),\nu_{i,t}})}{\sum_{j\in\Re_t} \exp(\varphi_{j,\nu_{j,t}})}.$$

## 9.2 Fitting

We begin by fitting a "deterministic" Ornstein-Uhlenbeck model, which is then relaxed and updated sequentially. The loglikelihood in $\varphi_0$ and $\theta$ is plotted below.

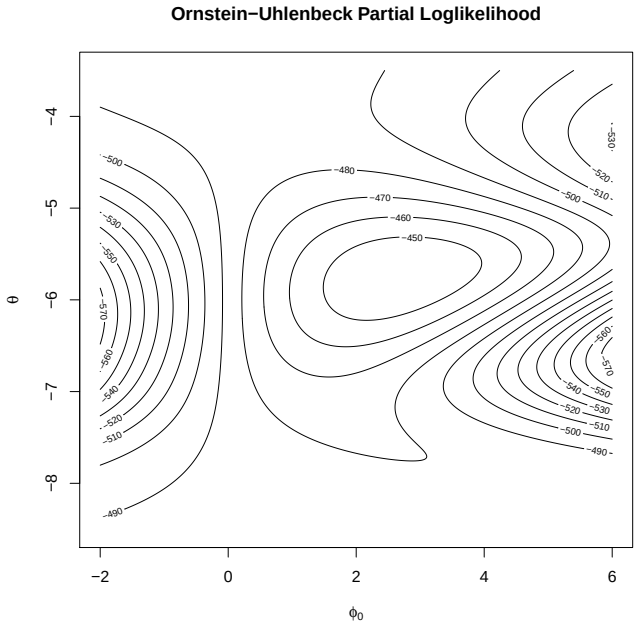



Figure 9.1: Ornstein-Uhlenbeck Partial Loglikelihood in Hyperparameters

The estimated parameter values are

| Hyperparameter estimates. | |
|---|---|
| $\widehat{\varphi_0}$ | 2.69 |
| $\widehat{\psi_0}$ | 14.73 |
| $\widehat{\theta}$ | $10^{-5.74} \approx 1.82 \times 10^{-6}$ |

which generate the infant mortality estimate below.

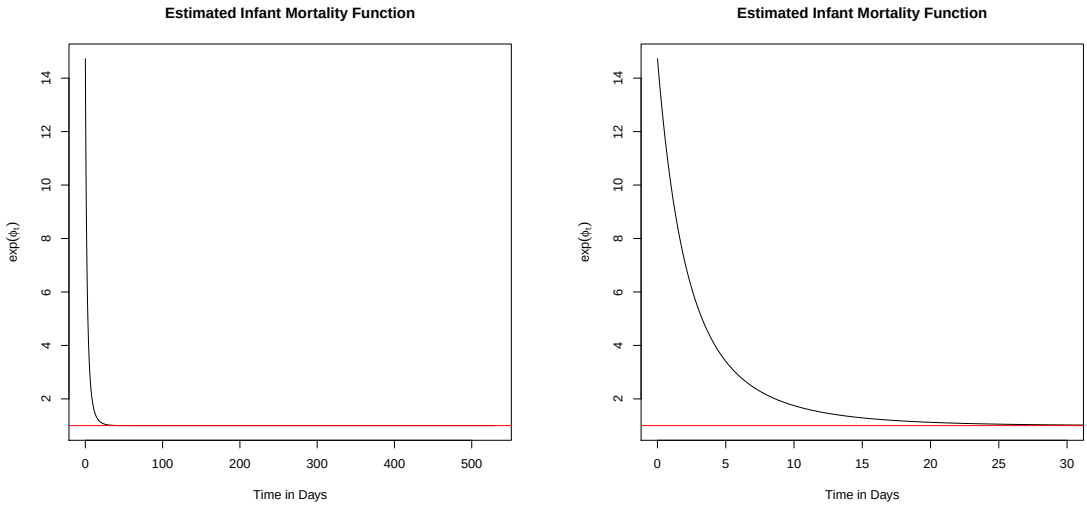



Figure 9.2: Mean Subsequent Effect of Failure according to Selected Ornstein-Uhlenbeck Model

## 9.3 Posterior Probability

Moving toward a complete Bayesian model, we have a distribution with the schematic form:

$$p(\text{data}|\varphi)p(\varphi|\text{hyperparameters}),$$

the hyperparameters being $\{\theta, \sigma^2, \varphi_0\}$.

Without a hyperprior on the hyperparameters, this gives a full distribution proportional to

$$\prod_{t\in\mathbb{T}} \frac{(2\theta)^{|T|}}{\sigma^{2|T|}|\det(\Sigma_{\theta;0})|} \exp\left[-\frac{\theta}{\sigma^2}\left((\varphi-\varphi_0\exp(-\theta\mathbf{t}))^\dagger \Sigma_{\theta;0}^{-1}(\varphi-\varphi_0\exp(-\theta\mathbf{t}))\right)\right]\times \frac{\exp\left(\varphi_{i(t),t}\right)}{\sum_{j\in\Re(t)}\exp\left(\varphi_{j,t}\right)}, \tag{9.2}$$

where the matrix $\Sigma_{\theta;0}$ is specified by

$$\sigma_{\theta;0_{i,j}} = \exp(\theta(t_i - t_j)) - \exp(-\theta(t_i + t_j)) \ (t_i < t_j).$$

The $\Sigma_{\theta;0}$ matrix will be called "skeletal," and is related to the actual marginal Ornstein-Uhlenbeck matrix $\Sigma_{\theta,\sigma^2}$ by

$$\begin{aligned} \sigma_{\theta,\sigma^2{}_{i,j}} &= \sigma_{\theta;0_{i,j}}\frac{\sigma^2}{2\theta} && \text{or} \\ \Sigma_{\theta,\sigma^2} &= \text{Diag}\left(\frac{\sigma^2}{2\theta}\right)\Sigma_{\theta;0} && . \end{aligned} \tag{9.3}$$

## 9.4 Form and Conjugacy

We will consider the form of the model for both inference of $\varphi$ and model selection.

We momentarily interpret the model above as a fully Bayesian model with independent diffuse priors on the hyperparameters, i.e.

$$p(\text{data}|\varphi)p(\varphi|\text{hyperparameters})p(\text{hyperparameters}) \tag{9.4}$$

$$\begin{aligned} &= p(\text{data}|\varphi)p(\varphi|\varphi_0,\theta,\sigma^2)p(\varphi_0)p(\theta)p(\sigma^2) \\ &\propto \prod_{t\in\mathbb{T}} \frac{\exp\left(\varphi_{i(t),t}\right)}{\sum_{j\in\Re(t)}\exp\left(\varphi_{j,t}\right)} \times \frac{(2\theta)^{|T|}}{\sigma^{2|T|}|\det(\Sigma_{\theta;0})|}\times \\ &\quad \exp\left[-\frac{\theta}{\sigma^2}\left((\varphi-\varphi_0\exp(-\theta\mathbf{t}))^\dagger \Sigma_{\theta;0}^{-1}(\varphi-\varphi_0\exp(-\theta\mathbf{t}))\right)\right], \end{aligned} \tag{9.5}$$

although we keep the expression $p(\text{data}|\varphi)$ for convenience.

The standard manipulations give

$$p(\varphi_0, \theta, \sigma^2|\text{data}, \varphi)p(\text{data}|\varphi)p(\varphi) = p(\text{data}|\varphi)p(\varphi|\varphi_0, \theta, \sigma^2)p(\varphi_0)p(\theta)p(\sigma^2) \tag{9.6}$$

$$p(\varphi_0, \theta, \sigma^2|\text{data}, \varphi) = p(\varphi|\varphi_0, \theta, \sigma^2)p(\varphi_0)p(\theta)p(\sigma^2)/p(\varphi) \tag{9.7}$$

$$p(\varphi_0, \theta, \sigma^2|\text{data}, \varphi) = p(\varphi|\varphi_0, \theta, \sigma^2)/p(\varphi) \tag{9.8}$$

$$p(\varphi_0, \theta, \sigma^2|\text{data}, \varphi) \propto p(\varphi|\varphi_0, \theta, \sigma^2). \tag{9.9}$$

We now consider the posterior distributions of the hyperparameters in detail, noting that they depend formally only on $\varphi$ and not the data.

The full expression is

$$\frac{1}{\frac{\sigma^{|T|}}{(2\theta)^{|T|/2}}\sqrt{|\Sigma_{\theta,0}|}} \times \exp\left[\frac{-1}{2\left(\frac{\sigma^2}{2\theta}\right)}\left((\varphi - \varphi_0 e^{-\theta\mathbf{t}})^\dagger \Sigma_{\theta,0}^{-1}(\varphi - \varphi_0 e^{-\theta\mathbf{t}})\right)\right].$$

It is worth noting that, with the other variables held fixed, the variance term $\sigma^2$ is inverse-gamma distributed,

$$\sigma^2|\varphi \sim \text{Inv}-\text{Gamma}\left(\frac{|T|-2}{2}, \theta\left((\varphi - \varphi_0 e^{-\theta\mathbf{t}})^\dagger \Sigma_{\theta,0}^{-1}(\varphi - \varphi_0 e^{-\theta\mathbf{t}})\right)\right).$$

Likewise, again holding the other variables fixed, by completing the square and ignoring proportional terms, $\varphi_0$ is shown to be normally distributed,

$$\varphi_0|\varphi \sim \mathcal{N}\left(\frac{e^{-\theta\mathbf{t}^\dagger}\Sigma_{\theta,0}^{-1}\varphi}{e^{-\theta\mathbf{t}^\dagger}\Sigma_{\theta,0}^{-1}e^{-\theta\mathbf{t}}}, \frac{\sigma^2}{2\theta e^{-\theta\mathbf{t}^\dagger}\Sigma_{\theta,0}^{-1}e^{-\theta\mathbf{t}}}\right).$$

### 9.4.1 Derivations

We seek to marginalize out whatever variables we can. Since inference for $\theta$ is clearly the most computationally-challenging aspect above, a great hope would be to derive a workable form for the marginal distribution of $\theta|\text{data}$. This would allow the application of model selection to remove $\theta$ from further consideration and derive an at least-approximate inference for the remaining parameters.

#### 9.4.1.1 $\theta, \sigma^2|\varphi$

For simplicity and since a flat prior on $\varphi_0$ is reasonable, we begin with integrating out $\varphi_0$ to derive $\theta, \sigma^2|\varphi$, giving

$$\sqrt{\frac{|\Sigma_{\theta,0}^{-1}}{e^{-\theta\mathbf{t}^\dagger}\Sigma_{\theta,0}^{-1}e^{-\theta\mathbf{t}}\left(\frac{\sigma^2}{2\theta}\right)^{|T|-1}}} \times \exp\left(\frac{-1}{2\frac{\sigma^2}{2\theta}}\right)\left[\varphi^\dagger\Sigma_{\theta,0}^{-1}\varphi - \frac{\left(\varphi^\dagger\Sigma_{\theta,0}^{-1}e^{-\theta\mathbf{t}}\right)^2}{e^{-\theta\mathbf{t}^\dagger}\Sigma_{\theta,0}^{-1}e^{-\theta\mathbf{t}}}\right].$$

#### 9.4.1.2 $\theta|\varphi$

From this we can recognize (holding $\theta$-dependent functions fixed), that the ($\varphi_0$-removed) distribution of $\sigma$ is formally Inv$-$Gamma$(\alpha, \beta)$ with $\alpha = |T| - 2$, which assists calculation and the future introduction of a hyperprior on $\sigma$.

Performing this calculation, we obtain the marginal distribution of $\theta|\varphi$:

$$\propto \theta \times \left(\frac{\left(\frac{|\Sigma_{\theta,0}|^{-1}2^{|T|-1}}{e^{-\theta\mathbf{t}^\dagger}\Sigma_{\theta,0}^{-1}e^{-\theta\mathbf{t}}}\right)^{1/2}}{\left(\varphi^\dagger\Sigma_{\theta,0}^{-1}\varphi - \frac{\left(\varphi^\dagger\Sigma_{\theta,0}^{-1}e^{-\theta\mathbf{t}}\right)^2}{e^{-\theta\mathbf{t}^\dagger}\Sigma_{\theta,0}^{-1}e^{-\theta\mathbf{t}}}\right)^{\frac{|T|-3}{2}}}\right) \times \Gamma\left(\frac{|T|-3}{2}\right).$$

This is clearly unpleasant, but things are not so bad. A fairly accurate linear expansion of $\Sigma_{\theta,0}^{-1}$ exists, and a good (empirical) approximation of $|\Sigma_{\theta,0}|$ exists as well.

It is worth mentioning that, since we are in an abstract setting already, there is no reason to use the same parameter $\theta$ for both the deterministic and correlation components. The model can be straight-forwardly expanded by replacing $\theta$ by $(\theta_1, \theta_2)$ with mean and covariance given by

$$\begin{aligned} \mathbb{E}[\varphi_t] = \mu_{t;\theta_1} &= \varphi_0 e^{-\theta_1} && , \\ \operatorname{cov}(\varphi_s, \varphi_t) &= \frac{\sigma^2}{2\theta_2} e^{-\theta_2|s-t|}e^{-\theta_2(s+t)} && . \end{aligned} \tag{9.10}$$

#### 9.4.1.3 Brief Note

Viewing this function in terms of $\varphi$, i.e. as the forward distribution $\varphi|\theta$, an obvious interpretation emerges. Removing multiplicative constants independent of $\varphi$, we have

$$\begin{aligned}\varphi|\theta &\propto \frac{1}{\left(\varphi^\dagger \Sigma_{\theta,0}^{-1}\varphi - \frac{\left(\varphi^\dagger \Sigma_{\theta,0}^{-1} e^{-\theta \mathbf{t}}\right)^2}{e^{-\theta \mathbf{t}^\dagger} \Sigma_{\theta,0}^{-1} e^{-\theta \mathbf{t}}}\right)^{\frac{|T|-3}{2}}}, \\ &= \frac{1}{\|\varphi - \frac{\langle \varphi, e^{-\theta \mathbf{t}}\rangle}{\|e^{-\theta \mathbf{t}}\|} e^{-\theta \mathbf{t}}\|_2^{|T|-3}},\end{aligned}$$

with implicit respect to the inner product given by $\Sigma_{\theta,0}$.

To summarize, for $|T| > 4$, this is a power distribution with parameter $|T| - 3$ in the norm of the projection of $\varphi$ onto the subspace orthogonal to the deterministic component of the Ornstein-Uhlenbeck process, $e^{-\theta \mathbf{t}}$, weighted along the directions indicated by $\Sigma_\theta$.

With even a modest number $|T|$ of observed gap times $\nu$, this density is practically singular on the class $\varphi \propto e^{-\theta \mathbf{t}}$, in the sense that $f(\varphi|\theta)$ takes a value smaller than machine-zero outside a neighborhood of radius smaller than machine-zero. For practical purposes, the choice of $\theta$ determines the "shape" of the log-infant mortaity curve.

Thus with diffuse independent priors on $\theta, \sigma^2, \phi_0$, it is now justified to use the deterministic Ornstein-Uhlenbeck approximation above. Specifically, if there is some kind of matrix norm bound on $\Sigma_\theta$ as $|T|$ increases, i.e. roughly as $N \cdot I$ increases, then the covariance structure is asymptotically ignored.

### 9.4.2 Inference for $\theta$

We now proceed to practical methods of inference. Unfortunately, the $\theta$ term is not as simple to update. To perform full inference, a grid of skeletal $\Sigma_{\theta,0}$ matrices must be precomputed and inverted. Although one may be tempted

to use the linear expansion

$$\Sigma_{\theta+\mathrm{d}\theta,0}^{-1} \approx \Sigma_{\theta,0}^{-1}\left(\mathbb{I} - \frac{\partial\Sigma_{\theta,0}}{\partial\theta}\Sigma_{\theta,0}^{-1}\,\mathrm{d}\theta\right)$$

for interpolation, this is unfortunately only valid in a neighborhood of order $10^{-8}$, making it mostly useless since a grid of this precision would be prohibitive to compute, and because this is many orders smaller than the error.

One option is to ignore the dependence of $\Sigma_{\theta,0}$ on $\theta$, and perform posterior inference solely on the deterministic part of the Ornstein-Uhlenbeck process. As a compromise, in sampling it may be feasible to update the full matrix $\Sigma_{\theta,0}$ (and its inverse) only every $N$ updates, using the current value of $\theta$ as obtained by the approximation.

In this case, there is still no closed form, but the posterior

$$\exp\left[\frac{-1}{2\sigma^2/2\theta}\left(\left(\varphi - \varphi_0 e^{-\theta\mathbf{t}}\right)^\dagger \Sigma^{-1}\left(\varphi - \varphi_0 e^{-\theta\mathbf{t}}\right)\right)\right]$$

is relatively simple. There is not even a closed form for a first-order estimating equation since it depends on both $\theta$ and the norm of $e^{-\theta\mathbf{t}}$ and its product with $\varphi$. Nonetheless, a grid approximation to the distribution can be used together with, say, slice sampling, for an approximate posterior.

Additionally, the inverting of the $\Sigma_{\theta,0}$ matrix, which is still required a significant number of times, is very time-consuming and numerically unstable.

## 9.5 Model Selection, or, Empirical Bayes for $\theta$

Interestingly, we can partially bypass this difficulty by applying empirical Bayes methods, or model selection. Since the application is computationally difficult, we will use the term "model selection."

The efficiency of this procedure depends on requiring, computationally, only the square root of a matrix and enough time to run several Monte Carlo samples.

This might not be viable, since integrating out $\varphi_0$ and $\sigma$ leaves an unpleasant expression involving the determinant $|\Sigma_{\theta,0}|$.

# Chapter 10

# Gamma Prior

It may be of interest to generalize a given model. Particularly, there is a tendency in reliability to focus on the Weibull even when it is not necessarily appropriate, or even completely inappropriate.

In this chapter, we consider a Bayesian modeling approach for generalizing a specified family of hazard functions.[1] This model is used to improve the fit from a Weibull model applied to survival analysis, and to summarize the deviation from the Weibull model in terms of a precision parameter, $\tau$.

Let $\Lambda^*(t;\gamma)$ be a fixed family of smooth cumulative hazards, parametrized by finite-dimensional vector $\gamma$. This family will be the *prototype* for a random cumulative hazard process given below.

Denote the distribution of a gamma cumulative hazard process, $\Lambda(t)$, with given prototype $\Lambda^*$ and precision parameter $\tau$ as

$$\Lambda(t) \sim \mathcal{GCH}(\Lambda^*(t);\tau),$$

[1]Note that the phrase *families of distributions* would be unnecessarily limiting. For example, it is natural for us to evaluate a Weibull assumption on mortality with respect to gaptimes, while leaving the calendar-time baseline hazard completely free. See section *Weibull* below.

with independent increments having distributions[2]

$$\Lambda(t) - \Lambda(s) \sim \Gamma(\tau(\Lambda^*(t) - \Lambda^*(s)), \tau),$$

that is, letting

$$\begin{aligned} \Delta\Lambda_t &= \Lambda(t) - \Lambda(s), \\ \Delta\Lambda_t^* &= \Lambda^*(t) - \Lambda^*(s), \end{aligned} \tag{10.1}$$

each increment $\Delta\Lambda_t = \Lambda(t) - \Lambda(s)$ is independently distributed with density

$$\frac{\tau^{\tau\Delta\Lambda_t}}{\Gamma(\tau\Delta\Lambda_t^*)}\Delta\Lambda^{\tau\Delta\Lambda^*-1}e^{-\tau\Delta\Lambda}.$$

It's worth noting that another approach exists with some similarities, which has some currency among the machine learning community, the so-called *copula process.*[26]

## 10.1 Weibull

For now, fix the $\Lambda^*$-family as Weibull with shape and scale parameters $\gamma_1, \gamma_2$, that is

$$\Lambda^*(t) = \gamma_1^{\gamma_2} t^{\gamma_2}.$$

Denote an independent increment drawn from the GCH process as, for $s < t$,

$$\Lambda(t) - \Lambda(s).$$

Specifically, we split the time interval of observation with respect to $Q$ quantiles, into

$$[0, u_1), [u_1, u_2), \cdots, [u_{Q-1}, u_Q),$$

[2]The notation $\Gamma$ is used for both the gamma function and the gamma distribution. The meaning is clear from context, or since the function takes one argument while the distribution takes two.

with $u_Q = \max t_{i,j} + \epsilon$, and denote the corresponding independent random increments as

$$\delta\Lambda_i = \Lambda(u_i) - \Lambda(u_{i-1}),$$

and analogously define

$$\delta u_i = u_i - u_{i-1}.$$

Assuming, as an approximation, that the hazard is constant in these intervals, associate to each interval the constant random hazard rate

$$\lambda_i = \frac{\delta\Lambda_i}{\delta u_i}.$$

## 10.2 Principal Modeling

As a prelude to the modeling of gaptimes, consider the use of this model to model survival in the standard case of right-censoring.

Associate to each subject $i$ the interval containing its event

$$[u_{u(i)}, u_{u(i)+1}),$$

with $u(i) = Q$ if no event is observed. and denote the observed probability of this event as the product over intervals $[u_k, u_{k+1})$,

$$\prod_i \prod_k \left( \frac{\Lambda_{k+1} - \Lambda_k}{u_{k+1} - u_k} \right)^{\mathbb{I}_{u(i)=k} y_i} e^{(\Lambda_{k+1} - \Lambda_k)\mathbb{I}_{u(i) \geq k}}.$$

Denote by $y_k$ the number of units at risk at time $u_k$ and by $n_k$ the number of units with an observed event during $[u_k, u_{k+1})$. Then, exchanging order and collapsing over $i$, the above may be written

$$\begin{aligned} &\prod_k \left( \frac{\Lambda_{k+1} - \Lambda_k}{u_{k+1} - u_k} \right)^{n_k} e^{(\Lambda_{k+1} - \Lambda_k) y_k} \qquad (10.2) \\ &\propto \prod_k (\Lambda_{k+1} - \Lambda_k)^{n_k} e^{(\Lambda_{k+1} - \Lambda_k) y_k}. \end{aligned}$$

This gives a simple posterior probability

$$\propto \prod_k \frac{\tau^{\tau\delta\Lambda_i^*}}{\Gamma\left(\tau\delta\Lambda_i^*\right)}(\Lambda_{k+1}) - \Lambda_k)^{\tau\delta\Lambda_i^* + n_k - 1} e^{-(\tau + y_k)(\Lambda_{k+1} - \Lambda_k)}.$$

#### 10.2.0.1 Model Selection

Integrate out $\Lambda_{k+1} - \Lambda_k$ terms, giving the marginal posterior probability in $\Lambda_{k+1}^* - \Lambda_k^*$ (or their corresponding hyperparameters). Denoting $\delta\lambda_k^* = \Lambda_{k+1}^* - \Lambda_k^*$, this probability is given by

$$\pi = \prod_k \pi_k = \prod_k \frac{\Gamma(\tau\delta\lambda_k^* + n_k)}{\Gamma(\tau\delta\lambda_k^*)} \left(\frac{\tau}{\tau + y_k}\right)^{\tau\delta\lambda_k^*} \frac{1}{(\tau + y_k)^{n_k}}. \tag{10.3}$$

The estimating equations for the MAP (assuming parametrization of $\lambda_k^*$ by $\theta$) are

$$\sum_k \frac{\partial \pi_k}{\partial \lambda_k^*} \frac{\partial \lambda_k^*}{\partial \theta} = 0 \tag{10.4}$$

$$\sum_k \left( \frac{\Gamma'}{\Gamma}(\tau\lambda_k^* + n_k) - \frac{\Gamma'}{\Gamma}(\tau\lambda_k^*) + \log\left(\frac{\tau}{\tau + y_k}\right)\right) \frac{\partial \lambda_k^*}{\partial \theta} = 0.$$

Unfortunately, due to the moderate tail of the gamma distribution, MAP estimation is quite inadequate. For instance, it may easily be shown that, with a saturated parametrization of $\lambda^*$ (i.e. with one free parameter per interval),

$$\lim_{\tau \to \infty} \widehat{\lambda_{k MAP}^*} \approx \frac{1}{\tau}\frac{n_k}{y_k} \approx \frac{1}{\tau}.$$

Application of the chain rule then shows that the MAP estimation of $\theta$ is subject to the same phenomenon, as the parameter $\tau$ does not appear in $\lambda_k^*(\theta)$.

Thus, fixing the model by using the MAP estimate for large $\tau$ will give inadequate estimators for $\lambda_k$ as well. Particularly, as $\tau \uparrow^{\infty}$,

$$\mathbb{E}[\lambda_k | D, \lambda_k^*] = \frac{\tau\lambda_k^* + n}{\tau + y} \approx \frac{n_k + \frac{n_k}{y_k}}{y + \tau} \approx 0,$$

$$\widehat{\lambda_{kMAP}} = (\mathbb{E}[\lambda_k|D, \lambda_k^*] - 1) \vee 0 \approx 0.$$

This phenomenon is illustrated below, with the posterior probability given for the scale and shape parameters, $\gamma_1$ and $\gamma_2$ respectively. The data is a simulated Weibull dataset, details to follow.

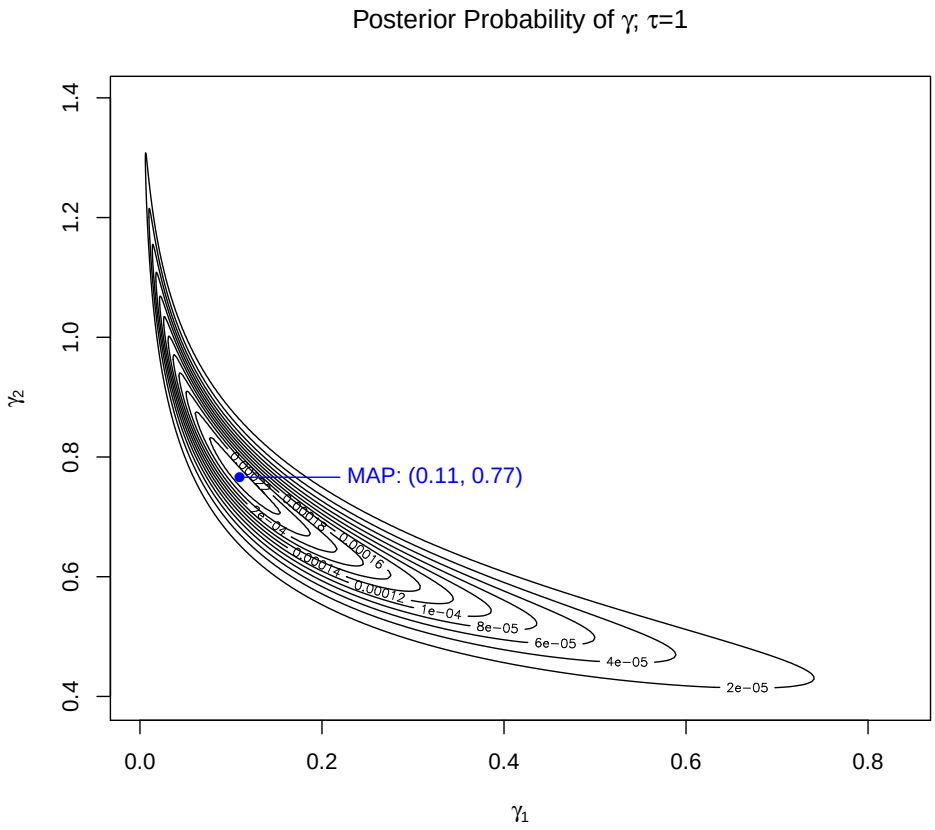



Figure 10.1: Posterior Probability in Gamma-Weibull Hyperparameters for Simulated Data

#### 10.2.0.2 Weibull Estimation

It is necessary to use a slightly more sophisticated estimation method. We begin with the posterior mean as an estimator of the hazard increments $\lambda_k$.

Two fortunate facts expedite this pursuit:

1. The mean of gamma distribution has closed form.

2. The Weibull distribution, as most standard parametric survival distributions, is two-dimensional which allows grid methods.

Letting $D$ denote the observed data, and emphasizing the dependence of $\lambda^*$ on $(\gamma_1, \gamma_2)$, note that

$$\begin{aligned} \mathbb{E}[\lambda|D] &= \int \lambda \int \mathrm{p}(\lambda|\lambda^*_\gamma, D)\mathrm{p}(\lambda^*|D)\mathrm{d}\lambda^*_\gamma \mathrm{d}\lambda \\ &= \int \mathrm{d}\lambda^*_\gamma \mathrm{p}(\lambda^*_\gamma|D) \int \mathrm{d}\lambda\, \lambda\, \mathrm{p}(\lambda|\lambda^*_\gamma, D) \\ &= \int \mathrm{d}\lambda^*_\gamma \mathrm{p}(\lambda^*_\gamma|D)\mathbb{E}[\lambda|\lambda^*_\gamma, D]. \end{aligned} \tag{10.5}$$

In the final expression, we use grid methods to approximate the distribution $\mathrm{p}(\lambda^*|D)$ by integrating out the $\lambda$ terms as done in the model selection above. The expectation has closed form

$$\mathbb{E}[\lambda_k|\lambda^*_k, D] = \frac{\tau\lambda^*_k + n_k}{\tau + y_k}.$$

an isotropic grid of $\gamma$ values and normalizing by their sum $S$, the posterior mean is given by

$$\mathbb{E}[\lambda_k|\lambda^*_k, D] = \sum_{\gamma_1,\gamma_2} \mathrm{p}(\gamma_1, \gamma_2|D) \times \frac{\tau\gamma_1(u_{k+1}^{\gamma_2} - u_k^{\gamma_1}) + n_k}{\tau + y_k}.$$

A simple adaptive algorithm is used to compute the necessary grid size by iteratively finding the maximum and expanding the region until the values of $\mathrm{p}(\gamma_1, \gamma_2|D)$ on the edge of the region are negligible, followed by refining the grid and repeating the process within the previous region.

#### 10.2.0.3 Limit of large $\tau$

Elementary limit identities give the posterior distribution $\lambda^*|D$ as

$$\lambda^* \sim \Gamma(n+1, y)$$

when $\tau \to \infty$.

The same identities apply in another parametrization; for the Weibull case this is

$$\gamma|D \sim \gamma_1^n \left(t^{\gamma_2} - s^{\gamma_2}\right)^n e^{-\gamma_1 (t^{\gamma_2} - s^{\gamma_2}) y}.$$

Particularly,

$$\gamma_1|D, \gamma_2 \sim \Gamma\left(n+1, \left(t^{\gamma_2} - s^{\gamma_2}\right) y\right).$$

Further manipulation gives the marginal conditional density of $\gamma_2|D$:

$$\gamma_2|D \propto \left(t^{\gamma_2} - s^{\gamma_2}\right)^{-1},$$

as well as the normalizing constant for $\gamma|D$,

$$\frac{y^{n+1}}{\Gamma(n+1)} \left( \int_0^\infty \mathrm{d}\gamma_2 \left(t^{\gamma_2} - s^{\gamma_2}\right)^{-1} \right)^{-1},$$

which exists for all $t > s$.

#### 10.2.0.4 Computational Considerations

The efficiency of this method is significantly higher than for the Gaussian methods, primarily due to two factors. First, the conjugacy of the gamma distribution to the generalized (discretized) Poisson likelihood. Second, the limited number of hyperparameters: $\tau$ and $\gamma$, where $\gamma$ typically has at most two parameters.

The limited number of hyperparameters and relatively tight distribution allow simple grid methods to be executed quickly. Other sampling methods would also be straight-forward; for example, Gibbs sampling which alternates as follows:

1. Sample from $\tau|\gamma, D$.

2. Sample from $\gamma|\tau, D$ using Metropolis-Hastings.

3. Sample from $\lambda|\gamma, \tau, D$, or use the closed-form equation for $\mathbb{E}[\lambda|\gamma, \tau, D]$ as above if only a point estimate is desired.

For very high precision computations, this *may* be faster than a grid method, and may be accessible to computational packages which use sampling as an atomic operator (e.g. BUGS, JAGS).

However, the non-elliptic posterior distribution of $\gamma$ as shown above is of some concern for simple MCMC. For this reason, it may in fact be more efficient to sample from the full space $(\gamma, \tau)$, as the freedom in the $\tau$ direction can help the sampler out of the tips of the distribution in $\gamma$.

An upper bound for error within an isotropic grid is

$$\epsilon \times \max \|\nabla f\|_2 = \epsilon C,$$

where $\epsilon$ is the uniform grid spacing. Note that this assumes that the mass outside the domain covered by the grid is negligible.

The computational complexity in constructing the grid is $\mathcal{O}\left(V/\epsilon^d\right)$ evaluations of the posterior probability function, where $V$ is the volume of the domain and $d$ the dimension. Each evaluation takes $\mathcal{O}(Q)$ time, giving a total complexity of

$$\mathcal{O}\left(QVC/\epsilon^d\right),$$

or

$$\mathcal{O}\left(QVC/\epsilon^2\right)$$

in the Weibull case as computed above. This is repeated for each value of $\tau$, for which a similar analysis applies.

#### 10.2.0.5 Modeling

We consider a simulated dataset parametrized by the Weibull family to illustrate the method above.

100 total event times are simulated as follows:

$$\tilde{T}_i \sim_{iid} \text{Weibull}(0.8, 0.01),$$

$$C_i \sim_{iid} \text{Expo}(0.01), C \perp \tilde{T}.$$

The total number of observed events is $\sum_i \delta_i = 55$, distributed as follows along with the survivor distribution.

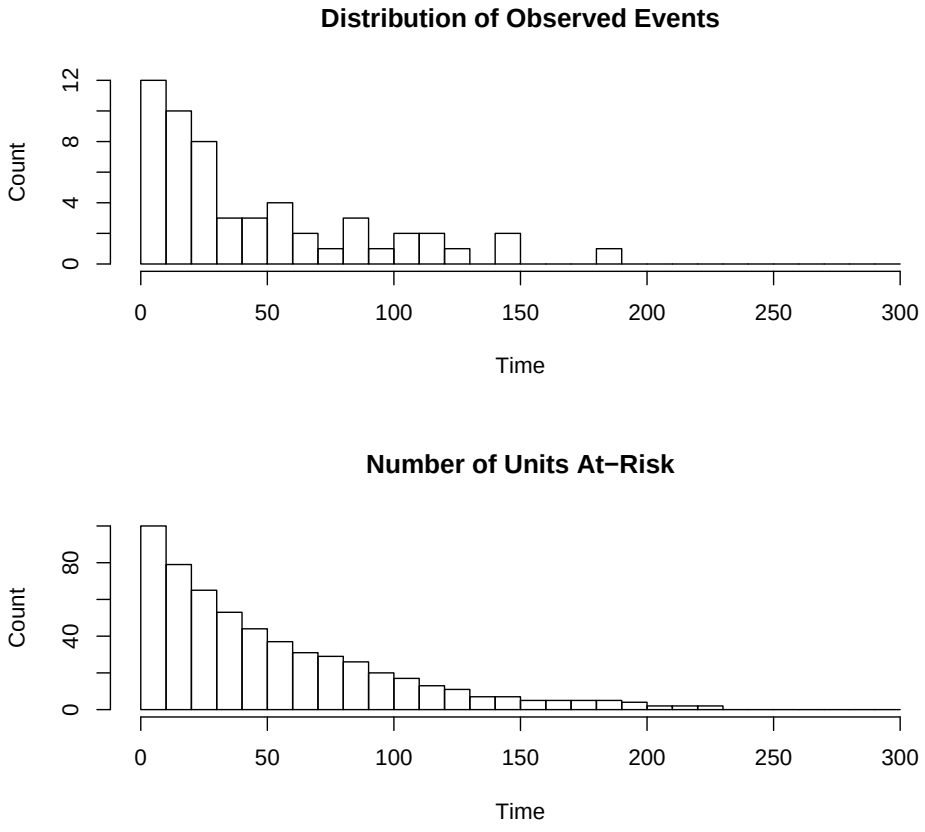



Figure 10.2: Empirical Distribution of Simulated Weibull Data

The method above is used to estimate the hazard function, illustrated below, and compared with the empirical and parametric Weibull maximum-likelihood estimators.

The hazard rate corresponding to the truth

$$\lambda(t) = \gamma_1 \gamma_2 t^{\gamma_2 - 1}$$

is plotted in black. The empirical estimates $(n_k/y_k)$ for each bin are provided as points. Note that the posterior mean estimator (plotted repeatedly as a function of the precision parameter $\tau$),

$$\widehat{\lambda_k} = \mathbb{E}[\lambda_k | D]$$

interpolates from a near-variant of the empirical estimator when $\tau = 10^{-5}$, to an apparent limiting case when $\tau = 10^6$.

Notably, the posterior mean estimator converges, as $\tau$ increases, to an estimator distinct from the standard MLE. In this case, the posterior mean estimator provides a better fit of the true hazard than does the standard MLE.

The delta method will provide a variance estimate of the hazard corresponding to the MLE. Sampling or the LaPlace approximation will provide an analogous interval estimate for the Bayesian case.

Replication studies will be conducted to test this case summarily, and will be continued in the study of gaptime by this method, below.

### 10.2.1 Gaptime Modeling

Note, at this point, having defined a piecewise constant hazard rate, recalling the earlier gaptime model

$$\lambda(t) = \lambda_0(t)\psi(t),$$

it is natural for our application to change the interpretation of the gamma process hazard to the gaptime effect, rather than the estimation of an overall

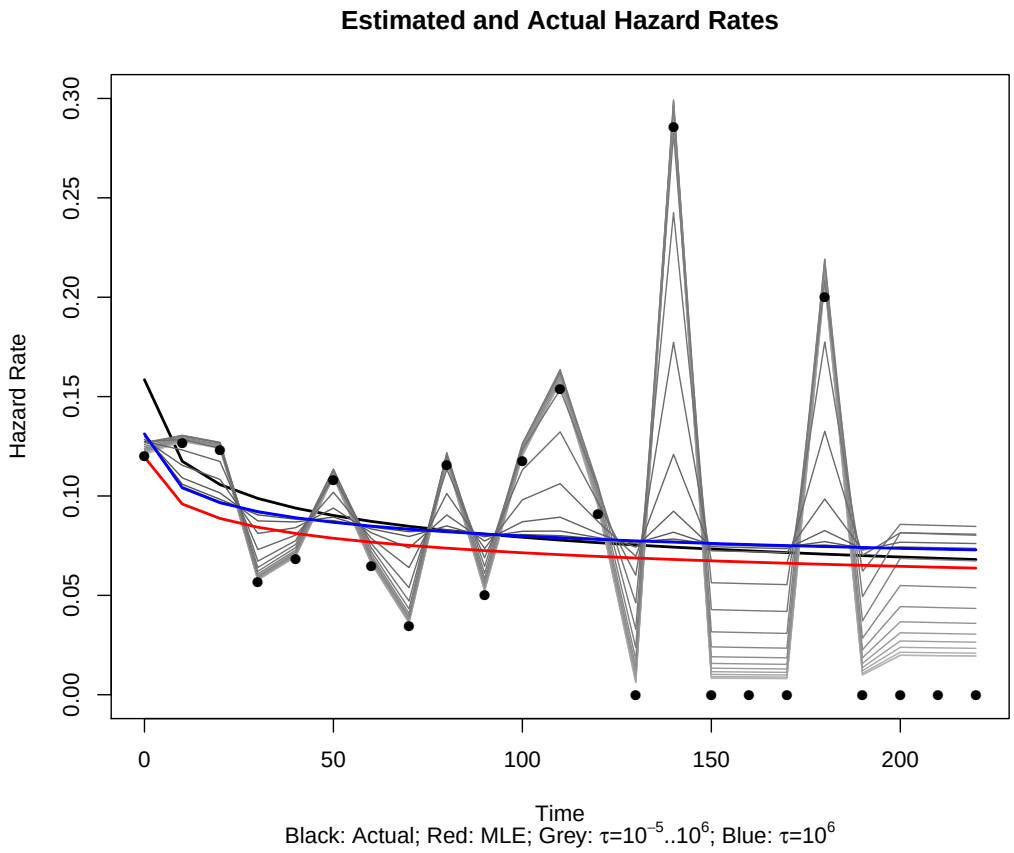



Figure 10.3: Posterior Estimators Produced by Varying Hyperparameter $\tau$.

hazard. Of course this prior may equally well be applied to $\lambda_0$ in an analogous manner. However, the objective is to allow a hazard which is "close" to Weibull, but allowed to vary in order to capture deviations present in the data.

Thus, we switch notation $\Lambda_i \mapsto \Psi_i$, $\lambda_i \mapsto \psi_i$ to reflect the modeling of a partial hazard. Likewise, change to a grid based on $Q$ quantiles of gaptimes, $(u_i)$. Everything subsequent in this section can apply to either case, as will be detailed below.

The $\delta\Psi_i$ are independent with distribution

$$\delta\Psi_i \sim \Gamma\left(\tau\gamma_1^{\gamma_2}\left(u_{i+1}^{\gamma_2} - u_i^{\gamma_2}\right), \tau\right).$$

Associate to each interval the mean partial hazard

$$\psi_i = \frac{\delta\Psi_i}{\delta u_i}.$$

The prior in terms of $\psi_i$ is then

$$\prod_q \frac{\tau^{\tau\gamma_1^{\gamma_2}(t_q^{\gamma_2} - t_{q-1}^{\gamma_2})}}{\Gamma\left(\tau\gamma_1^{\gamma_2}\left(t_q^{\gamma_2} - t_{q-1}^{\gamma_2}\right)\right)} (\Delta\Psi_q)^{\tau\gamma_1^{\gamma_2}(t_q^{\gamma_2} - t_{q-1}^{\gamma_2})-1} \exp\left(-\tau\Delta\Psi_q\right).$$

Recall, the full likelihood of the data with piecewise constant (fixed) baseline hazard,

$$\lambda_0(t) = \lambda_{0;k(t)},$$

and piecewise constant random gaptime hazard,

$$\psi(t) = \left(\frac{\Delta\Psi}{\Delta v}\right)_{q(t)},$$

is

$$\exp\left(-\sum_j \int_0^T \lambda_{0;k(t)} \left(\frac{\Delta\Psi}{\Delta v}\right)_{q(v_j(t))} Y_j(t)\mathrm{dt}\right) \times \tag{10.6}$$
$$\prod_j \prod_{k=1}^{n_j} \lambda_{0;k(t_{j,k})} \left(\frac{\Delta\Psi}{\Delta v}\right)_{q(\nu_{j,k})}.$$

This expression may be rearranged by reindexing along $q$, the index of the grid of gaptimes $(u_q)$ introduced above, producing

$$\exp\left(-\sum_q \left(\frac{\Delta\Psi_q}{\Delta u_q}\right) \sum_j \int_0^T \lambda_{0;k(t)} \mathbb{I}_{q(t)=q} Y_j(t)\mathrm{dt}\right) \times \tag{10.7}$$
$$\left(\frac{\Delta\Psi_q}{\Delta u_q}\right)^{\sum_{j,k} \mathbb{I}_{q(\nu_{j,k})=q}} \prod_{j,k} \lambda_{0;k(t_{j,k})}.$$

Note that the likelihood is the product of an unnormalized gamma density in the $\Delta\Psi/\Delta v$ terms, times a $\prod \lambda_0$ term and a $1/\Delta v$ term. The $\Delta\Psi$ terms are isolated in the unnormalized gamma density which is formally conjugate to the discrete gamma partial cumulative hazard process prior, in terms of the natural parameters of the gamma density.

Denoting the unnormalized gamma density by $\Gamma_.$, this may be written as

$$\prod_q \Gamma_. \left( \Delta\Psi_q; n_q + 1, \frac{\sum_j \int_0^T \lambda_{0;k(t)} \mathbb{I}_{q(t)=q} Y_j(t) \mathrm{dt}}{\Delta u_q} \right) \times \tag{10.8}$$
$$(\Delta u_q)^{-\sum_{j,k} \mathbb{I}_{q(\nu_{j,k})=q}} \prod_{j,k} \lambda_{0;k(t_{j,k})}.$$

Each $\Delta\Psi_q$ term in the likelihood is formally conjugate to the corresponding term in the prior, so the posterior distribution of $\Delta\Psi_q$ is proportional to an malnormalized gamma distribution.

Including the normalization term of the prior separately, the distribution of $\Delta\Psi_q$ is proportional to

$$\prod_q \Gamma_. \left( \Delta\Psi_q; n_q + 1 + \tau\gamma_1^{\gamma_2} \left( u_{i+1}^{\gamma_2} - u_i^{\gamma_2} \right), \tau + \frac{\sum_j \int_0^T \lambda_{0;k(t)} \mathbb{I}_{q(t)=q} Y_j(t) \mathrm{dt}}{\Delta u_q} \right) \times \tag{10.9}$$
$$\frac{\tau^{\tau\gamma_1^{\gamma_2}(t_q^{\gamma_2} - t_{q-1}^{\gamma_2})}}{\Gamma(\tau\gamma_1^{\gamma_2}(t_q^{\gamma_2} - t_{q-1}^{\gamma_2}))}.$$

The $\Delta\Psi_q$ terms may be integrated out, allowing straight-forward and efficient model selection. Gibbs or iterative-maximization methods may be applied, alternating between the hyperparameters $\tau, \gamma_1, \gamma_2$ and the baseline hazard $\lambda$.

This marginalized posterior distribution in $\tau, \gamma_1, \gamma_2$ and $\lambda$ is, including all

previously suppressed terms,

$$\prod_q \frac{\tau^{\tau\gamma_1^{\gamma_2}(t_q^{\gamma_2}-t_{q-1}^{\gamma_2})}}{\left(\tau + \frac{\sum_j \int_0^T \lambda_{0;k(t)}\mathbb{I}_{q(t)=q}Y_j(t)\mathrm{dt}}{\Delta u_q}\right)^{n_q+\tau\gamma_1^{\gamma_2}\left(u_{i+1}^{\gamma_2}-u_i^{\gamma_2}\right)}} \times \tag{10.10}$$

$$\frac{\Gamma\left(n_q + \tau\gamma_1^{\gamma_2}\left(u_{i+1}^{\gamma_2} - u_i^{\gamma_2}\right)\right)}{\Gamma\left(\tau\gamma_1^{\gamma_2}(t_q^{\gamma_2} - t_{q-1}^{\gamma_2})\right)} \times$$

$$(\Delta u_q)^{-\sum_{j,k}\mathbb{I}_{q(\nu_{j,k})=q}} \prod_{j,k} \lambda_{0;k(t_{j,k})}.$$

To integrate out the $\lambda_{\cdot}$ terms would require repeated integration of an elliptic integral.

Nonetheless, this expression is easily maximized in $\tau, \gamma_1, \gamma_2$, and $\lambda$ with the Broyden-Fletcher-Goldfarb-Shannon (BFGS) method,[3] by alternating maximization of the log-posterior with respect to the hyperparameters and $\lambda$.

Given the resulting estimates the individual $\psi_q$ each have the closed-form MAP estimate

$$\widehat{\psi_{q_{\mathrm{MAP}}}} = \frac{\widehat{\tau}\widehat{\gamma_1}^{\widehat{\gamma_2}}(u_{q+1}^{\widehat{\gamma_2}} - u_q^{\widehat{\gamma_2}}) + n_q - 1}{\widehat{\tau}(u_{q+1} - u_q) + \sum_j \int_0^T \widehat{\lambda_{0;k(t)}}\mathbb{I}_{q(t)=q}Y_j(t)\mathrm{dt}}\,.$$

The results of this optimization with $Q = 20$ are

| Hyperparameter estimates. | |
|---|---|
| $\widehat{\gamma_{1_{\mathrm{MAP}}}}$ | 0.0015 |
| $\widehat{\gamma_{2_{\mathrm{MAP}}}}$ | 0.224 |
| $\widehat{\tau_{\mathrm{MAP}}}$ | 0.151 |

### 10.2.2 Hyperprior

The parameters $\gamma_1, \gamma_2$ being the location parameters of the discrete gamma hazard process distribution, they are given a diffuse ($\propto 1$) hyperprior on $[0, \infty)$.

[3] available in standard numerical software packages including R, SciPy, and the GSL.

However, the precision parameter $\tau$ is strongly analogous to the precision parameter $1/\sigma^2$ of the normal distribution. The prior variance is

$$\text{var}(\Gamma(\tau\Delta\Lambda^*, \tau)) = \frac{\Delta\Lambda^*}{\tau},$$

while the weight given to the prior increment $\Delta\Lambda^*$ in the posterior distribution is $\tau$.

Thus by analogy with the limit of the noninformative inverse-$\chi^2$ prior for normal distributions with unknown variance[8], it may be reasonable to assign an improper hyperprior to $\tau$ which is inversely proportional to the variance:

$$\pi(\gamma_1, \gamma_2, \tau) \propto 1/\tau^{-1} = \tau.$$

#### 10.2.2.1 Interpretation

Clearly, the gamma process model allows significantly flexibility and gives at least exploratory evidence for the invalidity of the Weibull assumption.

To see what happened, recall that the Weibull model allows two possibilities: infant mortality with later hazard decaying to zero; and infant durability with hazard later diverging.

Since the true gaptime effect in the feeder system features infant mortality with a hazard that stabilizes in the long run around some non-zero limiting value, the full Weibull-gamma process model was able to fit the "head" of the distribution.[4] Contrariwise, the less flexible standard parametric Weibull model would under-estimate the infant mortality in order to accomodate a tail which does not decay to zero.

[4]The infant mortality corresponding to the Weibull fit of $\widehat{\gamma_2}$ is overestimated. Of course this is due to the flexibility of the gamma process model for $\psi$, but a better explanation would be preferred.

The figure below compares the Weibull fit induced by the gamma process parameter estimate $\widehat{\gamma_2}$ to a standard parametric Weibull fit without a prior model.

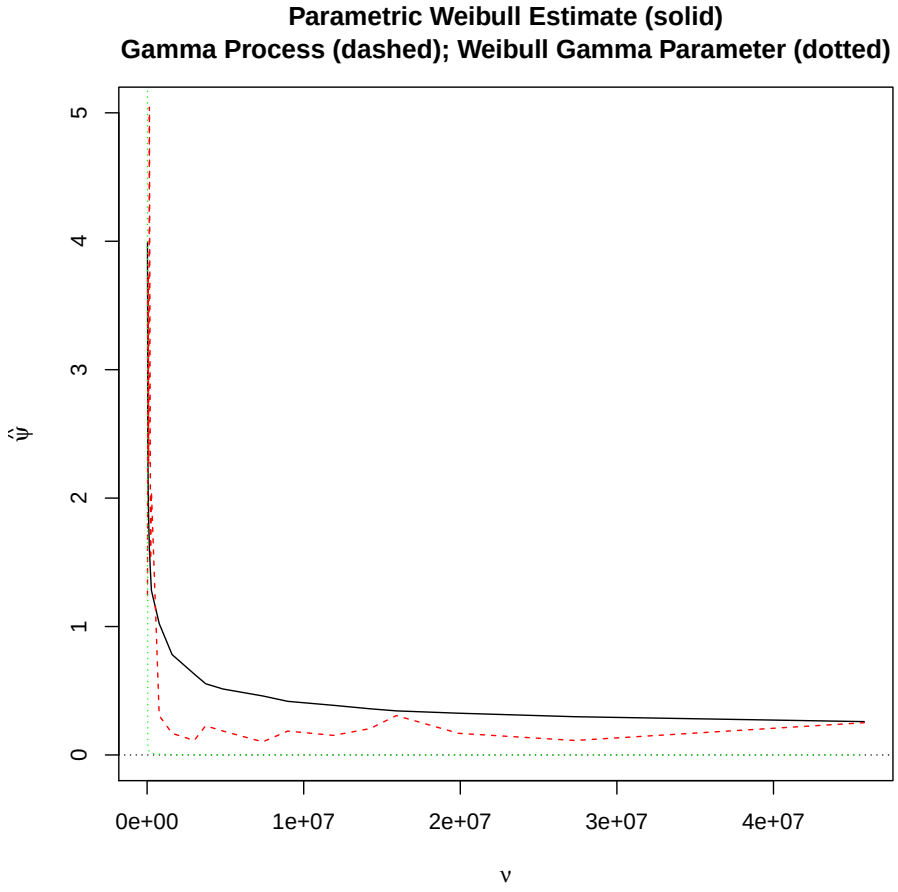



Figure 10.4: Mean Subsequent Effect of Failure according to Selected Gamma-Weibull Model.

The full gamma process captures both the severe infant mortality in the first few days while also capturing the non-zero tail behavior.

# Bibliography


[1] S.M. Amin. U.S. electrical grid gets less reliable. *Spectrum, IEEE*, 48(1):80–80, 2011.

[2] P.K. Andersen, R.D. Gill, and N. Keiding. *Statistical Models Based on Counting Processes*. Springer Series in Statistics. Springer New York, 1997.

[3] Roger N. Anderson, Albert Boulanger, Warren B. Powell, and Warren Scott. Adaptive stochastic control for the smart grid. *Proceedings of the IEEE*, 99(6):1098–1115, 2011.

[4] Xi Chen, Qihang Lin, Seyoung Kim, J. Carbonell, and E. Xing. An efficient proximal gradient method for general structured sparse learning. *Arxiv preprint arXiv*, 1005, 2010.

[5] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 187–220, 1972.

[6] Haimonti Dutta, David Waltz, Alessandro Moschitti, Daniele Pighin, Philip Gross, Claire Monteleoni, Ansaf Salleb-Aouissi, Albert Boulanger, Manoj Pooleery, and Roger Anderson. Estimating the time between failures of electrical feeders in the New York power grid. *Next Generation Data Mining Summit, NGDM*, 2009.

[7] J. Filliben et al. *NIST/SEMATECH Engineering Statistics Handbook*. National Institute of Standards and Technology, 2002.

[8] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2003.

[9] Philip Gross, Albert Boulanger, Marta Arias, David L. Waltz, Philip M. Long, Charles Lawson, Roger Anderson, Matthew Koenig, Mark Mastrocinque, William Fairechio, et al. Predicting electricity distribution feeder failures using machine learning susceptibility analysis. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 1705. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

[10] Philip Gross, Ansaf Salleb-Aouissi, Haimonti Dutta, and Albert Boulanger. Ranking electrical feeders of the New York power grid. In *Machine Learning and Applications, 2009. ICMLA'09. International Conference on*, pages 359–365. IEEE, 2009.

[11] Philip N. Gross, Ansaf Salleb-Aouissi, Haimonti Dutta, and Albert G. Boulanger. Susceptibility ranking of electrical feeders: A case study. 2008.

[12] Joseph G. Ibrahim, Ming-Hui Chen, and Debajyoti Sinha. *Bayesian Survival Analysis*. Wiley Online Library, 2005.

[13] Jeff Jones and Joe Hayes. Estimation of system reliability using a non-constant failure rate model. *Reliability, IEEE Transactions on*, 50(3):286–288, 2001.

[14] J.D. Kalbfleisch and R.L. Prentice. *The Statistical Analysis of Failure Time Data*. Wiley series in probability and mathematical statistics. Probability and mathematical statistics. J. Wiley, 2002.

[15] J.P. Klein and M.L. Moeschberger. *Survival Analysis: Statistical Methods for Censored and Truncated data*. Springer-Verlag, New York, NY, 2003.

[16] Bev Littlewood and John L Verrall. A bayesian reliability model with a stochastically monotone failure rate. *Reliability, IEEE Transactions on*, 23(2):108–114, 1974.

[17] Torben Martinussen and Thomas H. Scheike. *Dynamic Regression Models for Survival Data*. Springer, 2006.

[18] Warren B. Powell. *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. Wiley Series in Probability and Statistics. Wiley, 2011.

[19] Axinia Radeva, Cynthia Rudin, Rebecca Passonneau, and Delfina Isaac. Report cards for manholes: Eliciting expert feedback for a learning task. In *Machine Learning and Applications, 2009. ICMLA'09. International Conference on*, pages 719–724. IEEE, 2009.

[20] C.E. Rasmussen and C.K.I. Williams. *Gaussian Processes for Machine Learning*. Adaptative computation and machine learning series. University Press Group Limited, 2006.

[21] Cynthia Rudin, Rebecca J. Passonneau, Axinia Radeva, Haimonti Dutta, Steve Ierome, and Delfina Isaac. A process for predicting manhole events in Manhattan. *Machine Learning*, 80(1):1–31, 2010.

[22] Cynthia Rudin, David Waltz, Roger N. Anderson, Albert Boulanger, Ansaf Salleb-Aouissi, Maggie Chow, Haimonti Dutta, Philip N. Gross, Bert Huang, Steve Ierome, et al. Machine learning for the New York City power grid. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(2):328–345, 2012.

[23] Sam C. Saunders. *Reliability, Life Testing and the Prediction of Service Lives: For Engineers and Scientists*. Springer, 2007.

[24] Anastasios A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer, 2006.

[25] M.J. van der Laan and J.M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer Series in Statistics. Springer, 2003.

[26] Andrew Gordon Wilson and Zoubin Ghahramani. Copula processes. *arXiv preprint arXiv:1006.1350*, 2010.

[27] Leon Wu, Gail Kaiser, Cynthia Rudin, and Roger Anderson. Data quality assurance and performance measurement of data mining for preventive maintenance of power grid. In *Proceedings of the First International Workshop on Data Mining for Service and Maintenance*, pages 28–32. ACM, 2011.

[28] Leon Wu, Gail E. Kaiser, Cynthia Rudin, David L. Waltz, Roger N. Anderson, Albert G. Boulanger, Ansaf Salleb-Aouissi, Haimonti Dutta,

and Manoj Pooleery. Evaluating machine learning for improving power grid reliability. 2011.

[29] Leon Wu, Timothy Teravainen, Gail Kaiser, Roger Anderson, Albert Boulanger, and Cynthia Rudin. Estimation of system reliability using a semiparametric model. In *Energytech, 2011 IEEE*, pages 1–6. IEEE, 2011.