

**School of Electrical Engineering, Computing and
Mathematical Sciences**

Centre for Transforming Maintenance Through Data Science

**Bayesian Hierarchical Modelling of Equipment Reliability
in Mining: A Pragmatic Approach**

Ryan K. Leadbetter

This thesis is presented for the Degree of
Doctor of Philosophy
of
Curtin University of Technology

November 2023

To the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledgement has been made. This thesis contains no material which has been accepted for the award of any other degree or diploma in any university.

Ryan K. Leadbetter

“The Quote”

— AUTHOR

Acknowledgements

Write your acknowledgements here

Abstract

From pit to port, the consistent and efficient operation of iron ore machinery is essential for maximizing profits. To this end, reliability modelling is an invaluable tool for improving the design and execution of the maintenance strategies that ensure the reliable operation of mining machinery. There are well established reliability models in the literature, but there is a discrepancy between this literature and what is actually done by practitioners in the mining industry; a theory-practice gap. This gap exists because of the imperfect reality of collecting data in the field—data sets that are small, incomplete, noisy, or all three—and the lack of methods for expanding reliability modelling to account for these imperfections. My industry-linked PhD has aimed to reduce this gap by demonstrating how Bayesian statistical modelling framework can address some of the common problems faced when fitting models to such reliability data in mining applications.

In the first part of the work, I adapt and evaluate a method for constructing an informative joint prior distribution for the parameters of weibull lifetime analysis to combat bias introduced through heavily censored lifetime data. I first illustrate the bias caused by heavy censoring and then show how encoding domain information into a joint prior for the two Weibull parameters constrains the bias. Secondly, I evaluate the proposed method through a simulation study. Finally, I provide recommendations on applying the method in practice and demonstrate on an industry data set from an overland iron ore conveyor.

In the second part of the work I focus on degradation modelling. Particularly, how the Bayesian hierarchical framework can extend the gamma stochastic

degradation process to noisy observations and then to the degradation of surfaces. In doing so, I simplify some of the literature on noisy gamma processes by demonstrating how separating the observation-degradation process into two separate conditional models removes the need for complicated inferential algorithms. Furthermore, I show the hierarchical models implementation using flexible tools that are accessible to a wider reliability audience. I also show how reparametrisation can make the gamma process more interpretable and therefore simplify prior specification and further expansions of the model. Taking this one step further, I expand the noisy gamma process to functional data analysis in order to model the degrading surface of conveyor belting.

Throughout the work I emphasise how complicated reliability processes found in practice can be broken down into manageable sub-models and how these models can be fit, evaluated, expanded, and compared using Bayesian workflow considered to be good statistical practice. In doing so hope to contribute at a larger level by providing an applied case study of the the Bayesian workflow in a reliability setting that can be used by other applied reliability practitioners to develop solutions of their own for new problems.

Contents

Acknowledgements	vii
Abstract	ix
1 Introduction	1
1.1 Maintenance decision making	3
1.2 Reliability modelling	6
1.3 Industry Examples	8
1.4 Bayesian reliability modelling	13
1.5 Structure of this thesis	23
I Part one: lifetime analysis	25
2 Heavily censored lifetime data	29
2.1 Background	31
2.1.1 Lifetime distribution	32
2.1.2 The Weibull distribution	33
2.1.3 Censoring	34
2.1.4 Left-truncation	37
2.1.5 Left-truncation and right-censoring	38

Unknown truncation time	40
2.2 Imputing truncation times	41
2.3 Informative joint prior	42
2.4 Analysis of simulated data	45
2.4.1 Simulation method	47
2.4.2 Weakly informative prior	48
2.4.3 Strongly informative prior	50
2.5 Simulation experiments	53
2.5.1 Factor levels	53
2.5.2 Accuracy measures	54
2.5.3 Results	55
2.6 Concluding remarks	55
3 Case study	59
3.1 Idler frame data	59
3.2 An informative prior	59
3.3 Fitting	59
3.4 Using the posterior	60
3.5 Discussion and conclusions	61
II Part two: Degradation modelling	63
4 A noisy gamma process for modelling degradation measurements with uncertainty	67
4.1 The gamma process	68
4.2 A noisy gamma process	70
4.3 Reparametrisation	72

4.4	Constructing the prior	73
4.5	Fitting the noisy gamma process	79
4.5.1	Data simulation	80
4.5.2	Computation	82
4.5.3	Results and diagnostics	84
4.5.4	Solutions to computational issues	90
4.6	Discussion	93
5	Noisy gamma process with unit-to-unit variability	99
5.1	Models for multiple units	102
5.1.1	The complete pooling model	104
5.1.2	Partial pooling models	106
varying μ model	107	
varying ν model	107	
varying μ and ν model	107	
5.2	Computation, posteriors, and predictive distributions	108
Complete pooling	109	
Varying μ	110	
Varying ν	113	
varying μ and ν	116	
Posterior predictive checks	117	
5.2.1	Model comparison	120
Leave-one-unit-out cross-validation	120	
Step-ahead cross-validation	121	
5.3	Failure time distributions	123
5.4	Discussions	125

6 Conveyor belt wear forecasting	129
6.1 Functional data analysis (the data model)	134
B-splines	135
Functional time series	135
6.2 Process models	137
6.2.1 Noisy gamma process	138
6.2.2 Linear model	138
6.3 Parameter model	139
6.4 Posterior sampling and inference	143
Marginal posteriors of parameters	144
Intermediate quantities	147
Posterior predictive distributions	148
6.5 Forecasting degradation curves	151
6.6 Comparison of methods	152
6.7 Failure time distributions	157
6.8 Discussions	160
7 Discussions	165
7.1 Overview and discussions	165
7.2 Strengths and limitations	165
7.3 Future directions	165
7.4 Implications for industry practitioners	165
Appendices	167
A Appendix Title	169
B Copyright Information	171

List of Figures

1.1	An annotated image of an overland iron ore conveyor (Creagh, 2020) showing the belting, idlers, and idler frames.	9
1.2	(a) frame lifetimes. (b) eCDF of lifetimes	11
1.3	Belt UT measurement data.	12
2.1	Hazard function for $\beta = 0.9$, $\beta = 1$, and $\beta = 1.1$	34
2.2	35
2.3	38
2.4	39
2.5	46
2.6	The simulated lifetimes data.	48
2.7	The prior distributions.	51
2.8	The prior distributions.	52
2.9	Comparison of the prior and posterior for F_{t_1} and F_{t_2} to show how both the elicited distributions have been updated.	53
2.10	The simulation results	56
2.11	The simulation results	57
4.1	A simulated degradation trace generated from a gamma process with parameters $\mu = 10$ and $\nu = 1.119$ and the simulated noisy observations of the trace. The black line is the underlying degradation path and the points are the noisy observations.	74

4.2 One hundred realizations from prior predictive distributions of a noise-free gamma process with mean degradation rate of 1 unit per unit time and different prior distributions for parameters: (a) shape, rate $\sim \text{Ga}(1, 0.001)$, (b) shape, rate $\sim \text{Ga}(0.001, 0.001)$, and (c) parameterization mean/coefficient of variation—see text for details.	76
4.3 The simulated degradation trace with the subset of observations selected for the small dataset highlighted in Red.	80
4.4 The Cain energy diagnostics for each chain when sampling from the posterior of the big data set (left) and small data set. Each plot compares the marginal energy distribution, π_E , with the first differenced distribution, $\pi_{\Delta E}$	83
4.5 The marginal posterior distributions of the parameters σ , μ , and ν when the BHM for the noisy gamma process is fitted to the ‘large’ and ‘small’ simulated data in Fig. 4.1. The points and intervals shown in each distribution represent, respectively, the median and 95% and 66% credible intervals. The values used to simulate the data are shown in the top row.	85
4.6 The posterior predictive distribution of the filtered degradation path compared to the true degradation path when the model is fitted using the big (left) and small (right) data set. The median of the predictive distribution is shown as a grey line and the corresponding uncertainty intervals as blue ribbons.	86
4.7 Pairs plot showing the MCMC draws from the posterior distribution of the parameters σ , μ , $\log \nu$, and the filtered value Δz_1 when the BHM of the noisy gamma process is fitted to the small dataset. The red points indicate divergences, which congregate at the end of the funnel in the pairwise plots of $\log \nu / \Delta z_1$ and $\log \nu / \mu$	88

4.8	The parallel coordinate plot of the draws from the posterior conditioned on the small dataset with the divergent traces plotted in red.	89
4.9	The divergent traces overlaid (in red) on the posterior predictive distribution of the degradation path for the small data set.	89
4.10	90
4.11	The chain energy diagnostics when the model is fit with a stronger prior on σ	92
4.12	The pairs plot of the MCMC draws when using a stronger prior on σ	92
4.13	The marginal posterior distributions of the parameters σ , μ , and ν when information is added into the analysis either through a stronger prior or supplementary data compared to the original two posteriors.	94
5.1	Crack-growth propagation data of Rodríguez-Picón, Rodríguez-Picón, Méndez-González, Rodríguez-Borbón, and Alvarado-Iniesta (2018): original data (solid lines); noisy data (dashed lines).	101
5.2	Four non-noisy prior predictive simulations from the complete pooling model. Each simulation contains the same number of units and observations as the crack growth data in Fig. 5.1.	105
5.3	Marginal posterior distributions of the underlying gamma process from a BHM where all parameters are completely pooled.	110
5.4	Marginal posterior distributions of parameters from the model where mean degradation rate μ_j varies from unit-to-unit.	112
5.5	A pairs plot of the posterior samples of μ_1 , μ_2 , μ_μ , and $\log(\sigma_\mu)$ from the varying μ model. In each bivariate plot, divergencies are plotted in red.	114

5.6 Parallel coordinate plot for the parameters and hyper parameters of the varying μ model. The divergent traces are plotted in red. . .	115
5.7 Marginal posterior distributions of the underlying gamma process from a BHM where mean degradation rate μ_j varies from unit-to-unit.	115
5.8 Parallel coordinate plot for the parameters and hyper parameters of the varying ν model. The divergent traces are plotted in red. . .	117
5.9 Posterior predictive simulations from each of the four models. Each simulated dataset, is indicated by a different colour and line type. Plot (a) shows three simulated datasets from the fitted complete pooling model, (b) from the model with varying μ , (c) from the varying ν model, and (d) from the model where both μ and ν vary from unit-to-unit.	119
5.10 Posterior distributions of the failure time distributions from the complete pooling model (left) and the varying μ model (right). . .	125
5.11 Posterior distributions of the predictive failure time distributions from the complete pooling model (left) and the varying μ model (right) for Unit 3.	126

6.1 The growing wear profile of a conveyor's belt over time. Ultrasonic thickness measurements are taken at $N = 20$ measurement locations across the width of the belt at repeated times. The vertical axis is wear in mm , calculated by subtracting the measured thickness from the original thickness of the belt. The horizontal dashed line at $25mm$ of wear indicates the maximum allowable wear before the belt needs to be replaced, i.e. the soft failure threshold. The colour gradient indicates the time the belt has been in operation. Measurements taken at the beginning of the belt's life are shown in light yellow, and the most recent set of measurements are plotted in dark red	131
6.2 A demonstration of some typical approaches used to trend belt wear. (a) and (b) demonstrate the method of (Webb, Sikorska, Khan, & Hodkiewicz, 2020), where the measurements at the fastest wearing location which are circled in (a) are trended into the future in figure (b). (c) and (d) demonstrate an alternative but similar approach used by some reliability software, where the maximum wear measurement in each profile is trended instead.	132
6.3 (a) shows a fitted wear profile for the fifth observation; (b) shows the weighted set of basis functions that make up the profile in (a); and (c) shows the unweighted set of basis functions	136
6.4 Two historic belt wear data sets from the same conveyor	140
6.5 The estimated average wear rate for each spline coefficients calculated from the datasets in Fig. 6.4. The colour of each fit corresponds to the basis functions in Fig. 6.3 (c). The horizontal scale of each plots is in tonnes and the vertical scale is the value of the spline coefficient.	141

6.6 Sixteen prior predictive simulations of the average wear profile generated from the linear path (dashed line) and gamma process (solid line) process models with the priors defined in Sec. 6.3. The values of the hyper parameters for ν in the gamma process that were used to generate each simulation are displayed on each plot.	142
6.7 Marginal posterior distributions of the gamma process model parameters.	144
6.8 Marginal posterior distributions of the linear general path model parameters.	146
6.9 The posterior draws of the filtered spline coefficients for (a) the gamma process model and (b) the linear general path model.	148
6.10 Four posterior predictive distributions, two form each model. (a) and (c) shows the predicted smooth functional observation of the wear profiles underlying the sets of UT measurements from the gamma process and general path model, respectively. The UT measurements are also plotted as points as a reference. (b) and (d) shows the posterior predictive distribution of a new observation of the wear profile along the length of the belt at each observation time for the gamma process and general path model, respectively. In (b) and (d), the observed wear profiles are also plotted for comparison.	150
6.11 Forecasts of the ninth, withheld, observation from (a) the gamma process model and (b) the linear general path model. The true, observed, measurements are plotted for reference.	151

Chapter 1

Introduction

The work presented in this thesis is part of an industry-linked PhD under the Centre for Transforming Maintenance Through Data Science (CTMTDS), a centre comprised of both academic and industry partners. One of the Centre's goals is to develop methods that support reliability engineers in managing uncertainty during the maintenance decision making process—i.e. how and when they should maintain an asset based on their understanding of the asset, the available data, and what this data tells them about the current condition of the asset. As part of the industry-linked PhD, I spent 900 hours working on industry placement projects between two different industry partners to outline research topics that are not only novel in an academic sense but also useful for applied problems in reliability that are faced by practitioners when making maintenance decisions in the mining and mineral processing industry.

From my placement time, it was apparent that there is a disconnect between the reliability modelling literature and the methods used by reliability practitioners in the mining and mineral processing industry; a theory-practice gap. There are many well-established models for reliability data, such as the Weibull distribution for lifetime data or stochastic process models for degradation data. There is also a desire by mine, processing plant, and refinery operators to use these models since once an asset is put into service, completely new information becomes avail-

able in the form of infield reliability data, which can be utilised so that operators can make better maintenance decisions based on the specific reliability characteristics of their assets, rather than estimates from the manufacturer (Jardine & Tsang, 2013). However, these in-field data are typically noisy, messy, incomplete and uncertain as a result of unique observational processes that need their own sophisticated modelling approaches before reliability practitioners in the mining and mineral processing industry can take advantage of the well-established reliability models in the literature. Two examples of issues confronting reliability engineers that I tackle in this thesis are 1) obtaining sensible estimates of lifetime distributions when lifetime data are heavily censored and truncated due to the pre-emptive replacements of assets and the way in which data is recorded, and 2) forecasting complex degradation processes with noisy and sparsely observed condition monitoring data. In this thesis, I develop novel extensions of some of the well-established reliability methods through the Bayesian model building approach (Gelman, Vehtari, et al., 2020) and demonstrate how they can be applied to observational industry data sets from overland iron ore conveyors provided by the Centre's industry partners.

The Bayesian paradigm became a strong backbone of this thesis because Bayesian methods provide a formal structure to build complicated models and incorporate multiple sources of information, such as domain expert knowledge (Meeker, Escobar, & Pascual, 2022). Furthermore, the resulting full posterior distribution obtained from Bayesian analysis allows us to easily produce estimates and uncertainty intervals of complicated functions of the model parameters (Meeker et al., 2022), which is extremely useful for propagating uncertainty through a decision-making process. While there is a well-developed subfield of Bayesian analysis in the reliability literature (Hamada, Wilson, Reese, & Martz, 2008a; Meeker et al., 2022), the Bayesian framework is underutilised in industry. This underutilisation is most likely because, for most cases, inference must be obtained through Monte Carlo simulation, and in the past, this has meant con-

structing Markov Chain Monte Carlo (MCMC) algorithms manually. However, the recent increase in popularity of Bayesian methods is due to the development of flexible and accessible probabilistic programming languages such as BUGS (Lunn, Jackson, Best, Thomas, & Spiegelhalter, 2013), JAGS (Plummer, 2003) and Stan (Stan Development Team, 2022) (which in many cases alleviate the analyst from the need to construct bespoke MCMC algorithms). The result is a newfound ability to fit and explore complex models relatively quickly and simply.

To harness these new aspects of statistical modelling more effectively, the applied Bayesian statistical community has started to develop a more rigorous workflow for building, fitting, checking, and comparing Bayesian models (Gelman, Vehtari, et al., 2020). Throughout this thesis, I clearly emphasise the components of this workflow and demonstrate them in a reliability setting. In doing so, I hope this thesis may also be used as a template for how to carry out and report the results of Bayesian analyses for reliability and maintenance problems in the field.

The remainder of this chapter provides a background to the rest of the thesis. First, in Section 1.1, I provide some context around maintenance decision-making in the mining and mineral processing industry. Then, in Section 1.2, I give a high-level overview of reliability modelling and how it informs maintenance decisions. Section 1.4 outlines Bayesian methods and the key components of the Bayesian model-building workflow, which will be a strong thematic thread throughout the remainder of the thesis. Finally, in Section 1.5, I lay out the structure of the thesis.

1.1 Maintenance decision making

The maintenance of an asset can be considered as “*all activities aimed at keeping an [asset] in, or restoring it to, the physical state considered necessary for the fulfilment of its production function*” (Geraerds, 1985). In other words, the main objective of maintenance actions is to fix/replace an asset’s components to ensure that it can perform its desired duty at an acceptable level of performance. In this

context, the only consideration when deciding when to maintain the asset is whether or not the asset is performing its duty at an acceptable level. However, in reality, the maintenance of any single asset exists in the much larger context of a company (Jardine & Tsang, 2013). There are finite resources, budget, and time that can be allocated to the maintenance of any specific asset, and some assets are more critical to production than others. This ‘big picture’ management of an asset’s maintenance is what we refer to as asset health management. It is in this bigger context that reliability engineers and planners must make their decisions about how and when to maintain an asset. Asset health management requires foresight, planning, and—most importantly—risk management.

Maintenance strategies help to allocate resources and plan maintenance schedules ahead of time. There are three general strategies: reactive, preventative, and predictive maintenance (Jardine & Tsang, 2013). I provide a more detailed overview of these strategies below, but first, note that an asset can have different strategies for its different components, and typically, the choice of strategy is dictated by how critical the component is, how expensive it is, and what type of data we can collect. But even with a maintenance strategy, once an asset is put into service, we start to gather new reliability or condition monitoring data that can be used to refine/inform the maintenance strategy. For instance, Chapter 3 uses failure time data to inform the timing of a bulk-replacement strategy.

Reactive vs Preventative vs Condition-based maintenance strategy

The simplest replacement strategy is a reactive maintenance strategy whereby components are only replaced once they fail (Heng, Zhang, Tan, & Mathew, 2009). Reactive strategies are used mostly for non-critical components. They are not typically used for mechanical components in mining because the cost due to lost production when an asset fails unexpectedly is orders of magnitude greater than the cost of planned maintenance. On the other hand, a preventative replacement strategy is when components are replaced pre-emptively after a designated

period of time or operation. This proactive approach to maintenance is suitable for cheap components whose reliability decreases with time, i.e., components that wear out (most mechanical components). A disadvantage of preventative maintenance is that it can result in overmanning assets (replacing components too frequently when they still have remaining useful life), which is a waste of money and resources. If a component is costly and critical, and it is possible to monitor its condition, then a condition-based maintenance strategy should be used. Condition-based maintenance balances using as much of the component's useful life as possible with the reduced risk of lost production by monitoring the degradation of a component and replacing it when it gets to a predetermined, unacceptable level.

There are obvious ways in which statistical modelling can inform preventative and condition-based strategies. Implementing a preventative replacement strategy requires choosing a pre-emptive time to replace the component. The better that the choice balances the cost of maintenance with the cost of unplanned failure, the better the strategy will perform. A component's specific environmental and operating conditions affect its reliability (Meeker et al., 2022). So, if it is possible to use data to 'tune' the replacement time to the component's reliability under the specific operating conditions, then the preventative policy will be more successful than one where the manufacturer's default recommendations are used. Condition-based strategies, on the other hand, are more useful if we can forecast the degradation through time to predict the failure time (useful life) of the component. More detailed and accurate forecasts will result in better maintenance plans and reduce the risk of an unexpected failure. In both cases, the more accurately we can estimate the reliability quantities, the better the strategies will perform. We can estimate these quantities and manage uncertainty around the estimates by fitting reliability models to data.

1.2 Reliability modelling

In the engineering context, reliability is the “*ability of an item to perform a required function under given conditions for a given time interval*” (ISO, 2016). This definition is very closely tied to the definition of maintenance actions in Section 1.1; maintenance actions are to ensure reliability. In the reliability modelling context, the definition of reliability is slightly different. It is the “*probability for an item to perform a required function under given conditions over a given time interval $(0, t)$* ” (ISO, 2013). In other words, it quantifies the engineering definition of reliability as a probability that a unit will not fail before t , that is, $P(T > t)$, where T is the time of failure. Here, time t can be calendar time, operating time, or some other exposure, such as loading cycles, distance travelled, or throughput (Lee & Whitmore, 2006).

The modelling definition of reliability focuses on binary outcomes (i.e., success/failure data) for a given time interval (Hamada et al., 2008a). But typically, we have more detail in data and instead want to estimate the reliability at all values of $t = [0, \infty)$. This representation is the reliability function, $R(t)$. Reliability can alternatively be expressed as the complement of the reliability function, $P(T \leq t)$, which is the cumulative failure time distribution, $F(t)$ (Meeker et al., 2022). Reliability analysis aims to estimate these functions from data. Two general approaches are taken, depending on the type of observations available: Lifetime modelling (also referred to as failure time models) and degradation models (sometimes referred to as repeat repeated-measures degradation models).

Lifetime modelling The most common form of reliability data are lifetime data. These are the recorded installation and failure times of units in operation. Lifetime modelling, therefore, aims to estimate the failure time distribution from lifetime data. These lifetime data can come from repeated failures of an asset or the lifetimes of a population of assets. The estimated failure time distribution from lifetime analysis allows the analyst to make general statements about the

reliability of a population conditional on some exposure time t and sometimes on covariates Moore (2016). However, it is common for reliability datasets to be limited in size or the number of observed failures (Meeker et al., 2022). For example, a particular asset may only have a small number of failures, or in a population of highly reliable assets, only a few may fail over the period of observation. In these cases, the data are not very informative of the parameters in a lifetime model. To combat a lack of information, the analyst can either supplement the analysis with other sources of information (which we elaborate on in Part I) or use a degradation model if they have access to measurements of the degradation process that drives failure (the focus of Part II).

Degradation modelling To use an example from Meeker et al. (2022), consider the case that in a lifetime dataset, only two out of one hundred units fail. In this case, the ninety-eight units that did not fail provide no information about how close they were to failure. If, in addition, there are repeated measurements of the level of the degradation that drives the failure, then degradation analysis allows us to look inside the other ninety-eight units and, more precisely, estimate the failure time distribution. In fact, degradation modelling can be used to derive failure time distributions if there are no failures or even for a single unit that has not yet failed (Part II). The connection between degradation models and failure time distribution is well explored (Bae, Kuo, & Kvam, 2007; Lawless & Crowder, 2004; C. Lu, Meeker, & Escobar, 1996; Meeker et al., 2022), and the connection is typically made using soft failure.

Soft failure is defined by a predetermined threshold of degradation, compared to hard failures, which are when the component can no longer operate (Hamada et al., 2008a). Hard failures are more uncertain in nature, since the level of degradation is only indicative of when the component will catastrophically fail, making them more risky, hence why soft failure is often used. However, regardless of whether a soft or hard failure definition is used, when the failure time distribution is derived from a degradation model, it is important to note that

the distribution is conditional on the particular degradation failure mode we are modelling.

Statistical degradation models can be divided into general path models or stochastic process models (Pandey & Yuan, 2006; Si, Wang, Hu, & Zhou, 2011). General path models assign a functional or empirical form to the degradation path of a unit, typically a theoretically motivated function such as in (Robinson & Crowder, 2000). This method assumes that the functional form can sufficiently represent the underlying degradation and that measurement error can completely account for any deviations in measurements from this fixed path. Heterogeneity between units can be added through regression/random effects (Robinson & Crowder, 2000). On the other hand, stochastic processes do not assume a fixed path (Pandey & Yuan, 2006). In stochastic processes, the jumps in degradation are modelled as random variables, meaning that they account for random variation in the degradation process over time. The stochastic process model we focus on is the Gamma process, as in (Lawless & Crowder, 2004). A rough comparison of the two methods is made in (Ye & Xie, 2015); however, no comprehensive comparisons have been explored.

1.3 Industry Examples

In this thesis, I show two examples of industry problems. Both problems relate to the components of an overland iron ore conveyor. One example is a preventative maintenance problem, and the other is a condition-based one. In Part I of the thesis, I look at the preventative replacement of idlers, whereas in Part II I focus on forecasting the degradation of the conveyor belting to inform condition-based decisions. The two components are shown in Figure 1.1.

Idlers Idlers (sometimes called rollers) support the weight of the belt and ore. They are relatively cheap components, and there can be hundreds or thousands of them on a single conveyor. Idlers are organised in frames, usually consisting



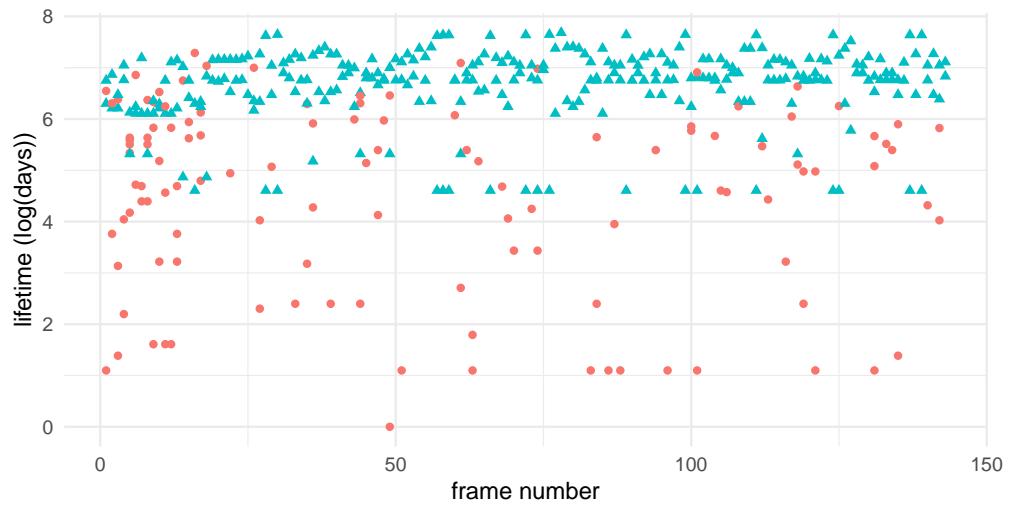
Figure 1.1: An annotated image of an overland iron ore conveyor (Creagh, 2020) showing the belting, idlers, and idler frames.

of three idlers: one central idler directly under the belt and two wing idlers supporting the sides of the belt to create the cupped shape. The idlers and idler frames are shown in Figure 1.1. Idlers are mechanical components and, therefore, wear out with operation. When an idler fails, it does not cause a direct impact on production; however, failed idlers can damage the belt, and damage to the belt results in major downtime. Reliability engineers need to manage the replacement of the idlers to minimise the risk of them failing and damaging the belt while simultaneously minimising the maintenance cost. It is not yet financially viable to monitor the condition of all idlers on a single conveyor, let alone all conveyors on a mine site. Therefore, a preventative maintenance strategy is used.

The survival data of idler frames I use in this thesis, derived from their installation and replacement times, can be used to inform the preventative maintenance strategy. Unfortunately, the failure data is only reliably recorded down to the frame number level, not the position of the idler in the frame. However, when one of the idlers in a frame fails, usually all the idlers in that frame are replaced, meaning that we can model the reliability of the idler frames to inform the preventative maintenance strategy. The data set is shown in Figure 1.2 (a). The figure shows the frame lifetimes for a single overland conveyor over six years. Because idlers are long-lasting components and also because they are preventatively replaced, there are many cases where we do not observe the entire lifetime of an idler, either because it had not failed by the time we stopped observing it, it was pre-emptively replaced, or it was in operation before we started reliably capturing failure and installation data. These partially observed lifetimes are censored lifetimes and are shown in blue in Figure 1.2 (a). We discuss censoring in more detail in Figure I. Figure 1.2 (b) shows the non-censored and censored lifetimes plotted cumulatively. In the ordered plot, we can see that many of the shorter lifetimes are fully observed, while most long lifetimes are obscured by censoring.

Belt The belt of an overland conveyor is much more costly to replace, both in terms of time and money. Furthermore, if the belt fails, then major downtime

(a)



(b)

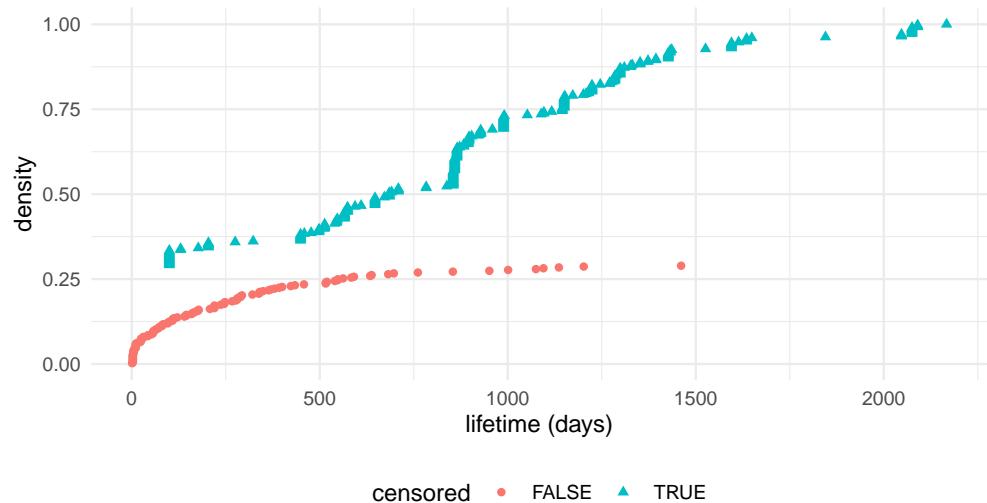


Figure 1.2: (a) frame lifetimes. (b) eCDF of lifetimes

is inevitable. Over time, the constant loading of ore onto the belt wears away a protective topcoat of rubber, exposing the structural components of the belt to the risk of being damaged by the ore. To assess the structural integrity of the belt, engineers stop the belt occasionally to take ultrasonic-thickness (UT) measurements of the topcoat and make sure that it is thick enough to provide adequate protection. An example of the UT data is shown in Figure 1.3.

Engineers use these UT measurements to estimate the soft failure time of the belt and plan its replacement, i.e. forecast when the top coat will no longer be thick enough to protect the belt. However, forecasting the wear of the belt to inform decisions requires principled statistical modelling to properly quantify the many different sources of uncertainty, which can be done in the Bayesian framework.

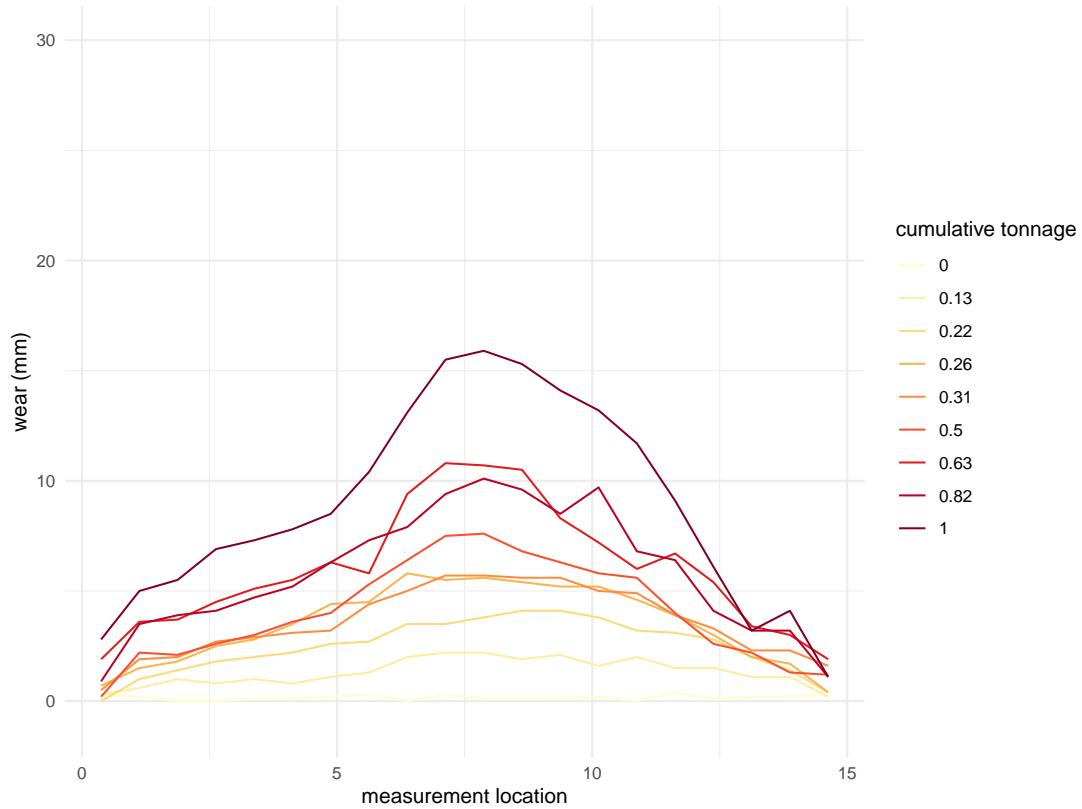


Figure 1.3: Belt UT measurement data.

1.4 Bayesian reliability modelling

In the Bayesian statistical framework, probabilities are subjective statements about our uncertainty. In a Bayesian analysis, we aim to construct probability statements about parameters of interest, θ , conditioned on the data y and (implicitly) on the values of covariates (Gelman, Carlin, et al., 2020). That is, we wish to obtain $P(\theta|y)$, known as the posterior distribution. To do this, Bayesian inference relies on Bayes rule,

$$P(\theta|y) = \frac{P(y|\theta)P(\theta)}{P(y)}, \quad (1.1)$$

where $P(y|\theta)$ is the likelihood of the data conditional on θ , $P(y)$ is the marginal distribution of the data, and $P(\theta)$ is the prior distribution of the parameters. The prior distribution encodes our belief about the parameters before observing the data and, therefore, encodes any additional information that we may have about the phenomenon we are modelling, either from historical data or domain expertise. Alternatively, we can write the un-normalised posterior as

$$P(\theta|y) \propto p(y|\theta)P(\theta). \quad (1.2)$$

In practice, however, the posterior distribution is rarely available in closed form, and we need to simulate draws from the posterior distribution using Markov chain Monte-Carlo methods to perform inference. There are several powerful and flexible probabilistic programming languages, such as STAN, which allow us to easily implement MCMC algorithms for complex model structures.

Bayesian workflow Surrounding Bayesian inference is a larger workflow of good statistical practice. Just as there is good practice for statistical modelling, the applied Bayesian community has developed its own workflow tailored to the specifics of Bayesian analysis (Gelman, Vehtari, et al., 2020), the main components of which focus on model construction, drawing from the posterior using

MCMC methods and diagnosing issues with computation, sense checking the model with simulated data, evaluating and using the posterior distribution, and comparing and expanding models. Here, I give a very high-level overview of the components in the workflow most relevant to this thesis. See (Gelman, Vehtari, et al., 2020) for an in-depth description of a Bayesian workflow. The components of this workflow that I use within this thesis are conditional modelling, prior predictive checking, sampling and diagnostics, posterior predictive checking and posterior inference, and model comparison.

Model specification The first step of any Bayesian analysis is to postulate a joint probability distribution for the model. For complicated processes, this first step can be simplified by using a Bayesian Hierarchical (multi-level) Modelling approach, which uses the fundamental notion of the *the law of total probability*, $P(A, B, C) = P(A|B, C)P(B|C)P(C)$, to decompose a complicated joint probability into a string of simpler conditional probabilities (Wikle, Zammit-Mangion, & Cressie, 2019, p. 13). Berliner (1996) proposed Bayesian Hierarchical Modelling (BHM) as a way of studying an underlying latent process by breaking the joint probability of the data, process, and parameters down into three sub-models;

$$\begin{aligned} p(\text{data, process, parameter}) &= p(\text{data}|\text{process, parameter}) && \text{data model} \\ p(\text{process}|\text{parameter}) & && \text{process model} \\ p(\text{parameter}) & && \text{parameter model} \end{aligned}$$

The first level is the data model, $p(\text{data}|\text{process, parameter})$, which describes the observation process. The second level in the hierarchy is the process model, $p(\text{process}|\text{parameter})$. It describes the underlying process that is of scientific interest. The third level in the hierarchy, $p(\text{parameter})$, is the parameter model, and, in a Bayesian setting, refers to the prior distribution. Each of these different levels in the hierarchy can also comprise smaller constituent conditional models. Cressie and Wikle (2011) advocate using the BHM approach for study-

ing underlying latent spatial and spatio-temporal processes, but the same general approach is used to break down models for nested data structures under the term multi-level modelling (Gelman, Carlin, et al., 2020).

In the last level of this hierarchical structure, the prior distribution summarizes any a priori beliefs the analyst has about the process they are trying to study before having observed the data. There are two different ways in which this information is encoded into the parameter model: the choice of distribution, and the values of the hyperparameters. Before the advent of contemporary sampling algorithms, Bayesian analysis relied on conjugate prior distributions, or convenient prior distributions that facilitated the use of Gibbs samplers or conventional Metropolis-Hastings algorithms (Gilks, Richardson, & Spiegelhalter, 1996). However, with the development of more efficient sampling algorithms such as Hamiltonian Monte Carlo (M. Betancourt, 2017), we are no longer limited by such requirements and can select priors that reflect our state of knowledge, facilitate efficient computation, and that can be justified and evaluated in a principled way. A useful tool for choosing the parameter model and understanding how it interacts with the process and data models is to simulate data from the full Bayesian model.

Simulation for model checking Bayesian analysis generally uses a fully generative model, so long as the prior is proper. When using a generative model, the model can not only be run "backwards" to perform inference but also "forward" to simulate fictitious data. For example, likelihood methods require a distribution for the data given the parameters, $P(y|\theta)$, but since there is no distribution for the parameters, there is no way of simulating data from this model unless we supply some reasonable values of the parameters. Bayesian analysis, on the other hand, specifies a distribution for both the data, y , and the parameters, θ ; $P(y, \theta) = P(y|\theta)P(\theta)$. Using this generative characteristic of Bayesian models to simulate data is useful for understanding unfamiliar or complicated models.

Prior predictive simulation is when we simulate data from the model before

conditioning on the observed data and is one of the key steps in the ‘Bayesian workflow’ (Gelman, Vehtari, et al., 2020, Figure 1). Prior predictive simulation can be used to understand the plausibility of a parameter model in the context of the likelihood (Gelman, Simpson, & Betancourt, 2017). Prior predictive simulations can also be a useful tool to elicit domain expert knowledge on the measurable outcome in order to develop an informative prior, rather than specifying domain knowledge directly on the parameters of the model. In the terminology of Gabry, Simpson, Vehtari, Betancourt, and Gelman (2019b), priors that when combined with the likelihood lead to simulated data that could be plausibly observed are known as *weakly informative priors*. According to Gabry et al. (2019b), such weakly informative priors should, for the most part, lead to plausible simulated data but may have some mass around extreme, but not completely implausible, realizations. Nevertheless, when using prior predictive checks to evaluate priors and to find sensible ones, the idea is *not* to try different values of the hyperparameters until the realizations are concentrated around the data that we are analyzing; instead, as Gabry, Simpson, Vehtari, Betancourt, and Gelman (2019a) write, the analyst “should have enough familiarity with the subject matter to look at prior predictive simulations . . . without needing to make direct comparisons with the data that will be used for model fitting.” They go on to say that “a *reasonable* [my emphasis] prior is a prior that yields a reasonable prior data-generating process, not that the researcher should tailor the prior to suit the particular observations in hand.”

Furthermore, simulating data from the model and then re-fitting the model to the simulated data is another useful way in which prior predictive simulation can help us better understand our Bayesian model. By fitting the model to simulated data for which we know the true parameter values, we gauge an understanding of what our model is capable of learning from the data. For example, in Chapter 4 we explore the limitation of a noisy gamma process model for noisy degradation data when only a few degradation measurements are observed.

HMC and diagnostics Throughout this thesis, I use the No-U-Turn sampler (Hoffman & Gelman, 2014) implemented in the probabilistic programming language Stan (Stan Development Team, 2022) to draw samples from the posterior distributions of Bayesian models. The No-U-Turn sampler is an adaptive variant of the successful Hamiltonian Monte Carlo (HMC) algorithm (Neal, 2011). HMC borrows the idea of Hamiltonian dynamics from physics to improve the random walk behaviour of traditional MCMC methods in order to move much more rapidly through the target distribution (Gelman, Carlin, et al., 2020). The No-U-Turn sampler improves the HMC algorithm by alleviating the user from the difficult task of choosing the step size and number of steps used to approximate the Hamiltonian trajectories (Hoffman & Gelman, 2014). The theoretical foundations of HMC are formulated in differential geometry, an advanced field of mathematics, so I do not discuss the details of HMC in this thesis. M. Betancourt (2017) provides a very nice conceptual introduction to HMC, and a more rigorous overview is given in Gelman, Carlin, et al. (2020, p. 300); any reader interested in the specifics should look to M. J. Betancourt (2015).

One added advantage of using a variant of HMC is the useful within chain diagnostics. For most general MCMC methods, we can check that chains have mixed using numerical summaries such as the potential scale reduction factor, \hat{R} (Gelman & Rubin, 1992), and follow up with trace plots of the individual chains, and we can check for inefficient exploration of the posterior using auto-correlation functions of each chain. However, it is difficult to diagnose why sampling is poorly behaved. Alternatively, when using HMC or one of its variants, one of the requirements for the algorithm to work efficiently is that the geometry of the set that contains the bulk of the target distribution is fairly smooth (Gabry et al., 2019b). While it is most often not possible to check for this condition mathematically, it can be checked numerically during sampling. When this set is not smooth, the leapfrog algorithm used to approximate the Hamiltonian trajectories diverges from the energy-conserving trajectory in the areas of high curvature (non-smooth

areas) and races off to infinity. Using a threshold energy, above which trajectories are considered divergent trajectories, we can diagnose problematic areas in the posterior (Gabry et al., 2019b), referred to by some as degeneracies (M. Betancourt, 2020). Sometimes, these degeneracies—and the poor sampling that results—can be resolved by re-parametrising the model (M. Betancourt & Girolami, 2015) while in other cases it cannot. In the latter, the degenerate behaviour may indicate an issue with the model.

If divergent transitions are present, then visually plotting the divergent trajectories alongside the non-divergent trajectories highlights the troublesome areas of high curvature in the posterior that obstructs exploration, since the true divergent transitions will be clustered around the problematic areas of the parameter space (Gabry et al., 2019b). Two useful visual diagnostics are the bivariate scatter plot, also called a pairs plot, and the parallel coordinate plot, see (Gabry et al., 2019b) for examples and descriptions. Since divergent transitions are flagged using a threshold, some reported divergences may be false positives. If this is the case, their distributions should match that of the non-divergent samples in either type of plot. However, if there are, in fact, areas of high curvature in the posterior, then the divergent transitions should be spatially correlated with these areas. Throughout this thesis, I use both pairs plots and parallel coordinate plots for checking sampling but only present them if they add to the discussion of the thesis (Chapters 4 and 5).

Evaluating and using the posterior Once we have postulated the model, generated samples from the posterior, and are confident that the samples sufficiently represent the posterior, we can use the posterior samples to perform inference and inform decisions. The result of fitting the model with MCMC methods is that we obtain S simulations of the parameters θ from their posterior distribution,

$$\theta_s \sim_{\theta|y} \pi. \quad (1.3)$$

Using the posterior draws of the parameters, we can find not only the estimated expected values of the parameters and credible intervals but also posterior predictive distributions for new data and uncertainty estimates for new functions of the parameters, such as failure time distributions.

Posterior predictive checking (Gelman, Carlin, et al., 2020) is a visual method for evaluating if the model has fit the data. After obtaining the posterior distribution, we can simulate replications of the data conditioned on the observed data. The predictive distribution for which these replicated datasets arise is the posterior predictive distribution. If the model fits the data, then the replications under the fitted model should look similar to the observed data (Gelman, Carlin, et al., 2020, p. 143). Put another way, the observed data should be plausible under the posterior predictive distribution. The posterior predictive distribution can be expressed as

$$p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)d\theta \quad (1.4)$$

where θ is the set of parameters and latent variables in the model. Computationally, we obtain draws from $p(\tilde{y}|y)$ by simulating from the data model conditional on the posterior draws of the parameters,

$$\tilde{y}^s|y, \theta^s \sim_{\tilde{y}|\theta^s} \pi \quad (1.5)$$

where the superscript s signifies a particular draw from the posterior obtained through MCMC simulation. For example, if we had assumed that data are independent and normally distributed conditional on the underlying parameters, then we could simulate draws of \tilde{y} by simulating from $N(\mu^s, \sigma^s)$. In the same way, we can generate predictive distributions for any latent variables in a hierarchical model. If the data look plausible under these predictive distributions then this is a sign that the model has reasonably fit the data.

In a similar fashion, estimates of functions of the parameters and uncertainty intervals for these ‘utility’ functions can be obtained. Two examples of useful

utility functions are failure time distributions and cost functions. To calculate the distribution of the function, we calculate the value of the function conditional on the MCMC draws of the parameters

$$F_X \sim F(x|\theta^s). \quad (1.6)$$

Comparing models Once we have fit a series of suitable models for a data set, we next want to evaluate how well they describe the true data-generating process and to compare them. To do so, we evaluate their ability to predict new observations. In the absence of an independent, external test set, it is conventional to use *information criteria* to compare models. These criteria, such as AIC, DIC, and others, are used to seek a compromise between goodness-of-fit and model complexity and to assess out-of-sample prediction accuracy. AIC and DIC are easy to calculate, but they are not fully Bayesian; hence, criteria such as WAIC (Watanabe-Akaike Information Criterion) and leave-one-out cross-validation (LOO-CV) are to be preferred (Vehtari, Gelman, & Gabry, 2017).

To compare models, we use LOO-CV, where the measure of distributional predictive accuracy is the *log score*. The log score is the log-likelihood of a new observation \tilde{y}_i given the posterior distribution of the parameters. It is also the probability of the new observation under the posterior predictive density, and it can be written as,

$$\text{lpd} = \log \int p(\tilde{y}_i|\theta)p(\theta|y)d\theta = \log p(\tilde{y}_i|y), \quad (1.7)$$

where θ is the set of parameters and y is the observed data. The parameters can also include unobserved latent variables. The measure in eq. (1.7) is called the *log posterior density (lpd)*. If we observe multiple new data points $\tilde{y} = (\tilde{y}_1, \dots, \tilde{y}_I)$, this can be dealt with in a point-wise fashion using the *log point-wise posterior*

density (*lppd*),

$$\text{lppd} = \sum_{i=1}^I \log p(\tilde{y}_i|y). \quad (1.8)$$

In cases where each of the new observations are independent of one another given the parameters and latent variables, which is the case for the majority of cases in this thesis, the point-wise predictive density is equal to the joint predictive density of the set of new observations; $\log p(\tilde{y}|y) = \sum_{i=1}^I \log p(\tilde{y}_i|y)$. 1.7 and 1.8 are defined for a given set of new observations, but the new unobserved data points \tilde{y}_i arise from the true data generating process and so are a random variable with distribution

$$\tilde{Y}_i = f(\tilde{y}_i). \quad (1.9)$$

Hence, a better measure of predicted accuracy is the expectation of the lppd, the elppd, which is obtained by integrating over \tilde{Y}_i

$$\text{elppd} = \sum_{i=1}^I \int \log p(\tilde{y}_i|y) f(\tilde{y}_i) d\tilde{y}_i. \quad (1.10)$$

In the context of Bayesian models fit with MCMC, the computed elppd can be calculated by averaging over the $s = 1, \dots, S$ MCMC draws from the posterior,

$$\text{computed elppd} = \sum_{i=1}^I \int \log \frac{1}{S} \sum_{s=1}^S p(\tilde{y}_i|\theta^s) f(\tilde{y}_i) d\tilde{y}_i. \quad (1.11)$$

Although this would be the best measure of predictive accuracy for our Bayesian models, we obviously do not know the true data-generating process, and so cannot define $f(\tilde{y}_i)$. We can, however, approximate the expectation above by using cross-validation whereby we iteratively withhold a portion of the observed data, sample from the posterior conditioned on the rest of the data, and then calculate the log-likelihood of the withheld portion of the data given the samples from the posterior. The simplest form of cross-validation is leave-one-out (LOO-CV),

where we withhold each observation,

$$\text{elppd}_{\text{LOO-cv}} = \sum_{i=1}^I \log \frac{1}{S} \sum_{s=1}^S p(y_i | [\theta]_{-[i]}^s). \quad (1.12)$$

$[\theta]_{-[i]}^s$ is the posterior draws for the set of parameters and latent variables conditioned on all the observed data except the withheld observation y_i .

In hierarchical models, the definition of a new observation and the likelihood of those observations depends on what aspect of the model's predictive performance we are trying to assess. For example, in a degradation dataset with multiple units and multiple observations per unit, we could obtain new observations for the same units at the same observation times, $\tilde{y}_{n,i}|z_{n,i}, \sigma$ (although this case is a bit unrealistic), new observations for an observed unit at some time in the future, $\tilde{y}_{n,I+1}|\tilde{z}_{n,I+1}, \sigma$, or we could observe an entirely new unit, $\tilde{y}_{n+1}|\tilde{z}_{n+1}, \sigma$, where $\tilde{y}_{n+1} = [\tilde{y}_{n+1,1}, \dots, \tilde{y}_{n+1,I}]$. In all three cases, the likelihoods of the observations conditional on the draws from the posterior predictive distribution are much the same since, for the most part, we assume in our data models that observations are independent given the underlying degradation parameters. However, the method used for constructing the predictive distributions of the parameters and intermediate quantities will differ depending on the model's hierarchical structure.

Using cross-validation to estimate elppd involves repeatedly re-fitting the model to different subsets of the data, which is computationally inefficient. A much more efficient method is to approximate the $\text{elppd}_{\text{LOO-cv}}$ (Vehtari et al., 2017); however, in hierarchical modelling cases where the 'left out' portions of the data are nested, and as the size of these nested portions increases, the approximations are less likely to work well (Vehtari et al., 2017). All of the cases in this thesis where $\text{elppd}_{\text{LOO-cv}}$ is used to compare models is in the context of such nested data structures. Furthermore, most of the datasets are reasonably small. Therefore, I use the full cross-validation scheme and incur the slight computational overheads.

1.5 Structure of this thesis

This chapter has introduced the industry-derived motivation for the work in this thesis and provided a high-level introduction to the threads that flow through the body of work: reliability analysis and Bayesian model building. The body of the thesis is divided into two parts and unified at the end by a general discussion/concluding chapter. Part I addresses lifetime analysis, whereas Part II addresses degradation modelling. At the beginning of each part, I have included a preamble that provides a background on the industry placement project/s that motivated the works and points out which chapters have been published or submitted for publication. I hope these short sections of metadiscourse provide a glimpse into the extra work that has gone into defining novel research problems whose solutions are useful to reliability practitioners in the industry.

Part I of the thesis focuses on lifetime analysis. Specifically, how we can obtain reliable inference from Weibull analysis of data that...

Part II of the thesis is more loosely structured. The three chapters–four, five, and six–show an iterative model-building process centred around a Gamma stochastic process for degradation. Chapter 4 demonstrates how the Gamma process can be extended through the Bayesian Hierarchical modelling (BHM) framework to account for noisy observations. Chapter 5 then shows the expansion of the noisy gamma process model through the same BHM structure to simultaneously model the degradation traces of multiple units and incorporate unit-to-unit variability. Chapter 6 looks at the applied problem of modelling the wearing surface of an overland conveyor’s belt. In the chapter, I show how functional data analysis can be used alongside degradation modelling to model the degradation of a wearing surface. I use the model structures proposed in Chap. 4 and 5 for the gamma process to show their usefulness in an applied scenario in the mining industry. I also compare the stochastic process method with a simpler general path model.

The concluding chapter, Chapter 7, ties the two parts of work in the thesis

back to the overarching topics of reliability and maintenance in the iron ore and mineral processing industry. The thesis concludes with a discussion of the strengths and limitations of the work, areas of future work, and the implications of this work for industry practitioners.

Part I

Part one: lifetime analysis

A preamble about which chapters have been published and which chapters came from industry placements.

“Since the initial approach was published in Leadbetter et al. (2021), I have taken an alternative approach that does...”

Chapter 2

Heavily censored lifetime data

Computerised maintenance management systems (CMMS) such as SAP (SAP SE, 2023) are now embedded in companies maintenance procedures, meaning that these companies now posses large scale datasets of component installation and replacement times. A natural use of these personalised failure time data sets is for tailoring replacement strategies for the companies specific operating environments (Meeker et al., 2022, p. 13), rather than solely relying on the manufacturers recommendations. One problem however, is that these large observational datasets collected through CMMS are much messier than the experimental ones used by manufacturers in traditional reliability/warranty analysis. This messiness comes about because of reporting issues, incomplete historic records, and the fact that most components are pre-emptively replaced before they fail because of the risk to production and employee safety. The result is that many of the valuable data sets stored in CMMS systems are incomplete in the form of censoring and left-truncation. Censoring is when the true lifetime of a failed component is not known but either an upper bound, lower bound, or both are known. Whereas left-truncation arises when only units that have lasted more than some truncation time are observed.

The censored and left-truncated nature of such data make what would otherwise be a very strait forward analysis far more complicated. Worse yet, the

incompleteness of the data is not always obvious and mistreatment during analysis can lead to biased or miss informed results. In this part of the thesis I focus on the lifetime analysis of the idler-frame dataset shown in 1.3. In this case, right censoring arises due to the set of idlers in a frame either being preventatively replace or still being in operation when the data were analysed. A portion of the lifetimes are also left-truncated, since any idlers that were installed and failed before t_{start} , when failures started being recorded in the CMMS, are not captured in the dataset. Treatment of right censored and left-truncated data was addressed by Hong, Meeker, and McCalley (2009). A further complicating factor of the idler frame dataset is that the installation time of idler-frames that were already in operation when data started being captured in the CMMS are unknown, meaning that the left-truncated lifetimes are also censored and have unknown truncation times. This issue is sometime referred to as unknown initial conditions () or unknown exposure history (). In this chapter, I propose a method for handling such cases in a Bayesian framework by imputing censored lifetimes and along with them, the truncation times.

Incompleteness of a dataset due to censoring and truncation increase uncertainty in lifetime analysis, particularly with respect to the upper tail of the lifetime distribution, since the data only contains partial information about the longer lifetimes. To improve analysis, domain knowledge can be used to fill the gap left by censoring, however only if the prior is constructed properly to do so. Kaminskiy and Krivtsov (2005) propose a method for constructing a joint prior for the two parameters of the Weibull distribution by eliciting domain expert knowledge on the CDF…

In this chapter of the thesis, I show how right censored and left-truncated lifetimes with unknown exposure history can be handled using fully a Bayesian treatment, and how the methods of Kaminskiy and Krivtsov (2005) for developing a joint prior for the two parameters of the Weibull distribution can be incorporated into the model to supplement the analysis with additional sources

of information and hence ‘fill in the gaps’ left by the incomplete lifetime data. In this chapter I demonstrate the proposed methods using fictitious lifetime data simulated from known parameter values in a way that emulates the idler frame lifetime data to show that the method is capable of reclaiming the true parameter values. In the next chapter, I apply the method to the real idler frame dataset, and show how the fully Bayesian treatment has advantages for... (Follow Meekers paper on the transformers.)

In the remainder of this chapter is structured . . .

- Background on Weibull lifetime, censoring, left-truncation, left-truncation with unknown exposure history, constructing a joint prior.
- Demonstrate through simulation that the method accounts for the incompleteness of the data without introducing much bias.
- Extend the prior proposed in Kaminskiy and Krivtsov (2005) and demonstrate it’s usefulness in this context.

2.1 Background

Lifetime analysis, also called survival analysis, is the analysis of failure time data from a population of particular components/assets to derive the risk of failure of a component dependent on it’s level of exposure (usually some form of time) and sometimes other covariates (Moore, 2016). From here on I will use the general term unit/s to refer to individual/groups of the same asset or component. Lifetime analysis of a population of units typically takes place by first specifying a sampling distribution for the lifetimes by choosing some parametric lifetime distribution for the units and incorporating any observational characteristics of the data—for example censoring—then, estimating the parts of the distribution from failure time data using an appropriate inferential mechanism, and finally using the fitted model to derive useful reliability measures about the population

which can be used to inform asset management plans. When done in a Bayesian context, the first step of this process also includes specifying a prior distribution. From the resulting inference, we can devise optimal replacement strategies that minimise the risk of unplanned failures, and hence the risk of lost production, and also the cost of the maintenance strategy.

2.1.1 Lifetime distribution

The lifetimes of the units are modelled as a random variable defined in terms of t , the exposure time, on $[0, \infty)$. t is some continuous or discrete exposure time from a clearly defined origin, the installation of the component, to a well defined event, the failure of the component. In reliability analysis, the exposure is typically absolute time, the operating time of the unit, or cycles of operation. In this analysis I use absolute time since operating time was not available. Next, a specific parametric lifetime distribution is chosen for the random variable t , $p(t|\theta)$, expressed as the probability density function (PDF) and the parameters of the lifetime distribution are estimated from the data. Once the estimates are obtained, different specifications of the lifetime distribution can be used to draw useful insights in order to inform decisions:

- **Cumulative distribution function** (CDF), $F(t)$, is the probability that a unit will have failed by time T , i.e. $P(t \leq T)$. It is also sometimes called the cumulative risk function.
- **Survival function**, $S(t)$, is the complement of the CDF and defines the probability of a unit surviving up to an exposure time t .
- **Hazard function**, $h(t)$, which is the instantaneous failure rate, i.e. the probability that given a unit has survived up to time T it will fail in the next small interval.

For example, the CDF quantifies the risk of unplanned failures given a chosen preventative maintenance interval, and the Hazard function identifies if a units

risk of failure increases as it ages and therefore if a preventative maintenance strategy is even suitable at all.

2.1.2 The Weibull distribution

In the analysis that follows, I use the Weibull distribution to model the idler frame lifetimes, that is

$$y|\beta, \eta \sim \text{Weibull}(\beta, \eta), \quad (2.1)$$

where β is the shape parameter and η is the scale. The Weibull distribution is a commonly used lifetime distribution because of its ability to capture an increasing, constant, or decreasing risk of failure. In addition, the weibull distribution is the limiting distribution for the minimum value in a sample when the sample space is lower bounded; such as lifetimes, which must be greater than zero. This characteristic of the weibull distribution gives it a convenient interpretation in component reliability; the lifetime of a unit is the time of the first occurring catastrophic failure mode of the unit. In my analysis that follows, I use the coupled parameterization of the two-parameter Weibull distribution, which has PDF

$$f(t) = \frac{\beta}{\eta} \left(\frac{t}{\eta} \right)^{\beta-1} \exp^{-\left(\frac{t}{\eta} \right)^\beta}, \quad (2.2)$$

CDF

$$F(t) = 1 - \exp^{-\left(\frac{t}{\eta} \right)^\beta}, \quad (2.3)$$

and hazard function

$$h(t) = \frac{\beta}{\eta} \left(\frac{t}{\eta} \right)^{\beta-1}. \quad (2.4)$$

The shape parameter β dictates whether the hazard increases, $\beta > 1$, decreases, $\beta < 1$, or stays constant, $\beta = 1$. The effect of the shape parameter on the hazard function is demonstrated in *Figure 2.1*. Practically speaking, if the hazard function increases with exposure, then this corresponds to a wear out failure mechanism, where as if it decreases then this corresponds to infant mortality.

This is important from a maintenance perspective because if the component does not wear out, then preventative replacement policy is not suitable (Jardine & Tsang, 2013). In other words, we want to be sure that $\beta > 1$ before implementing a preventative policy.

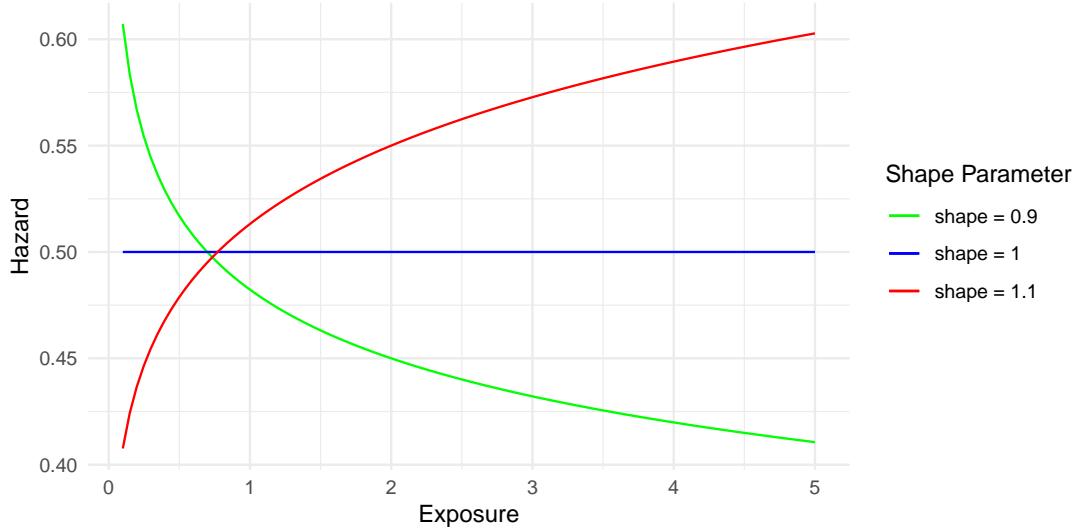


Figure 2.1: Hazard function for $\beta = 0.9$, $\beta = 1$, and $\beta = 1.1$.

2.1.3 Censoring

It is very common for lifetime data to be censored. Censoring occurs when we only partly observe the lifetime of a unit, or in other words we only observe upper and lower bounds for the lifetime. There are three types of censoring: left, interval, and right censoring, but all are treated in much the same way. *Figure 2.2* demonstrates these three types of censoring. In the figure, three units are installed at time t_0 and then two follow up inspections are performed at times t_1 and t_2 . The true, partially observed, failure times of each unit are shown as crosses in the figure. Left censoring occurs when we only observe an upper bound of the lifetime, an example is if we know the install time of a unit and observe it as failed at t_1 and so we only know that the true value of the lifetime must be less than t_1 . Interval censoring occurs if we only know an upper and lower

bound for the lifetime, i.e. if a unit fails between inspection times then we know that the lifetime must be greater than t_1 but less than t_2 . Right censoring occurs when we only know a lower bound for the lifetime. For example, if a unit is still in operation when we perform our analysis, e.g. it has lasted longer than t_2 , then we only know that the true value of the lifetime must be greater than t_2 . Right and left censoring are special cases of interval censoring where the upper or lower bound of the lifetime are infinity or zero respectively. Left censoring is fairly uncommon and so In the discussions that follow, I focus on right and interval censoring but all of the methods can be easily extended to accommodate left censored data as well.

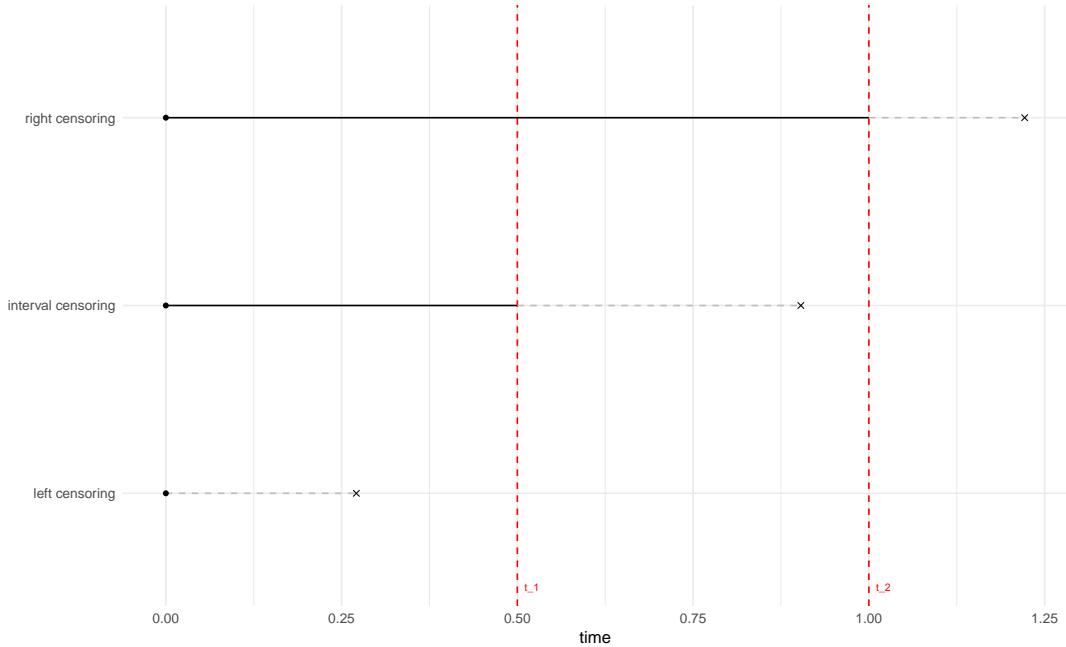


Figure 2.2

One way of handling censored data is to treat the censored lifetimes as missing data, which in a Bayesian framework is to treat them as a random variable in the model (Reich & Ghosh, 2019, p.211), and constrain their values to fall within the upper and lower censoring bounds (Stan Development Team, 2024). This is easily done in probabilistic programming languages like Stan. Treating the missing data as random variables in the model requires us to specify a distribution for the

censored lifetimes, which we assume is the same distribution as the rest of the population,

$$y_i^C | \beta, \eta \sim \text{Weibull}_{t_i^{\text{Lower}}}^{t_i^{\text{Upper}}}(\beta, \eta). \quad (2.5)$$

Here y_i^C is an unobserved censored lifetimes, and the superscript t^{Upper} and subscript t^{Lower} indicate that the distribution is constrained by the upper and lower censoring times. Once the missing lifetimes have been imputed, the likelihood of the observed and imputed lifetimes can be calculated in the same way as a typical lifetime dataset with no censoring. This approach of imputing the censored lifetimes is not unique to Bayesian methods. The same can be done using an Expectation Maximisation algorithm and maximum likelihood (Mitra, 2013). However, using the Bayesian approach, along with MCMC methods, it is very simple to derive uncertainty intervals for the parameters, imputed values, and useful quantities.

An alternative approach is to simply integrate out the censored observations. The probability that a censored observation falls between the upper and lower censoring times is

$$\Pr [t^{\text{Lower}} < y_i^C \leq t^{\text{Upper}}] = \int_{t^{\text{Lower}}}^{t^{\text{Upper}}} f(y_i^C) dy_i^C = F(t^{\text{Upper}}) - F(t^{\text{Lower}}). \quad (2.6)$$

By integrating out the censored observations, the likelihood can be written as

$$L(\theta | y^O, t^{\text{Upper}}, t^{\text{Lower}}) = \prod_{i=1}^{N^O} f(y_i^O) \prod_{j=1}^{N^C} [F(t_j^{\text{Upper}}) - F(t_j^{\text{Lower}})], \quad (2.7)$$

where θ are the parameters of the lifetime distribution, N^O and N^C are the number of fully observed and censored observations respectively, and y_i^O are the fully observed lifetime. This second approach is much more commonly used, particularly in the reliability literature (). However, as I show later, for the particular problem when data are also left-truncated with unknown installation times, it is convenient to frame the model in the form of the first imputation

approach.

2.1.4 Left-truncation

Truncation arises when a sample comes from an incomplete population, or in other words, there is some criteria that part of the population must appease in order to be observable (Guo, 1993). Left-truncation, for example, arises when some units must survive up to a certain time in order to be included in the dataset. It is also possible for data to be right- or doubly-truncated, but left-truncation is the most common in lifetime data. The definition of left-truncation and left-censoring may seem very similar, however, they are distinctly different (Mitra, 2013). Censoring is a characteristic of the sample, i.e. we know the number of left-censored observations but not the exact values of their lifetimes, whereas truncation is a characteristic of the population because we do not know how many units were not included in the dataset because they did not survive past the truncation time(the time from the installation of the unit to the start of observation) and hence our sample is not representative of the true population. An example of a left-truncated dataset is shown in *Figure 2.3*.

In *Figure 2.3*, units that were installed prior to t_{start} come from a left-truncated distribution since any other unit that was installed at the same time but did not last until t_{start} are not included in the sample. The left-truncated cases caused by the start of the observation period (units three, four, five, and seven in *Figure 2.3*) tend to over-represent low-risk cases since any high risk case installed at the same time is absent from the sample (Guo, 1993). Observations that arise from a left-truncated distribution can be included in the analysis by re-normalising their lifetime distribution by dividing by the probability of surviving past the truncation time;

$$L(\theta|y_i^T) = \frac{f(y_i^T)}{1 - F(\tau_i^{\text{Left}})}, \quad (2.8)$$

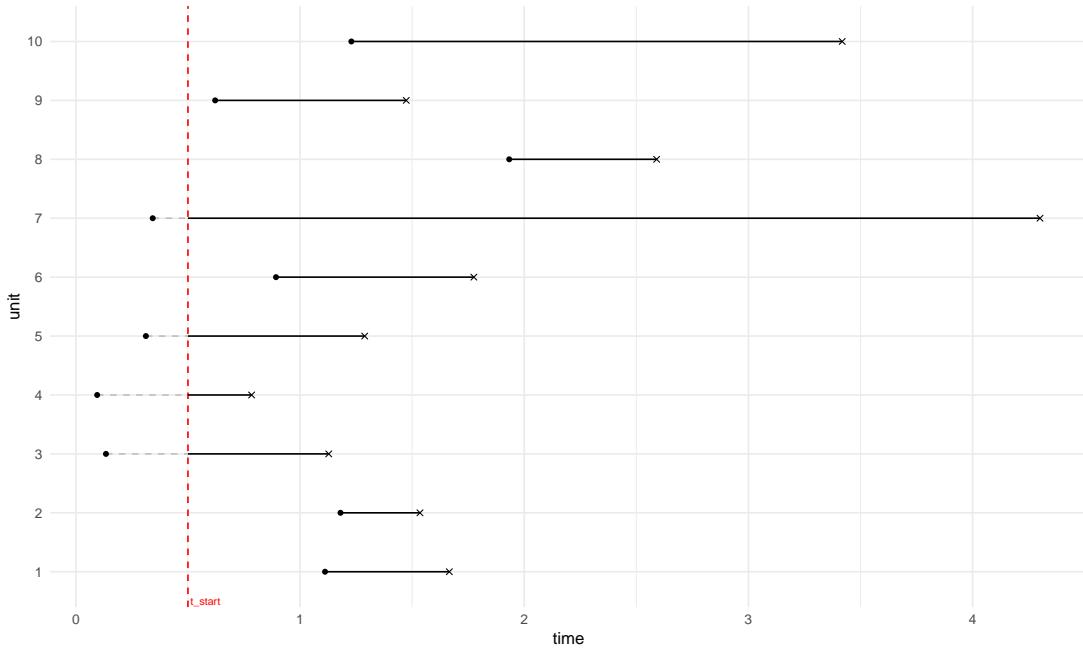


Figure 2.3

where y_i^T is the left-truncated lifetime, and τ_i^{Left} is the truncation time.

2.1.5 Left-truncation and right-censoring

A common scenario in reliability datasets is a combination of both left-truncation and right-censoring. This case naturally arises in historic observational datasets, such as those found in CMMS, where units are repeatedly replaced once they fail and any units that were installed and failed before the start of the observation process (which might be the date a new CMMS was adopted) are absent in the dataset. *Figure 2.4* shows a toy example of this case, where three units are repeatedly replaced when they fail and we start to observe their failures at t_{start} and stop at t_{end} . Any lifetimes that fail before t_{start} are unobserved and so are greyed out, resulting in the first observed lifetime of each unit being a left-truncated sample. Lifetimes that surpass t_{end} are only partially observed (right censored), hence the portion of these lifetimes that sits to the right of t_{end} is also greyed out. Hong et al. (2009); Kundu and Mitra (2016); Mitra (2013) analyse a dataset of electrical transformer failures that follows this general structure and

Mittman (2018) looks at a similar case for computer hard drives. The idler frame failure data is also an example of this type of dataset.

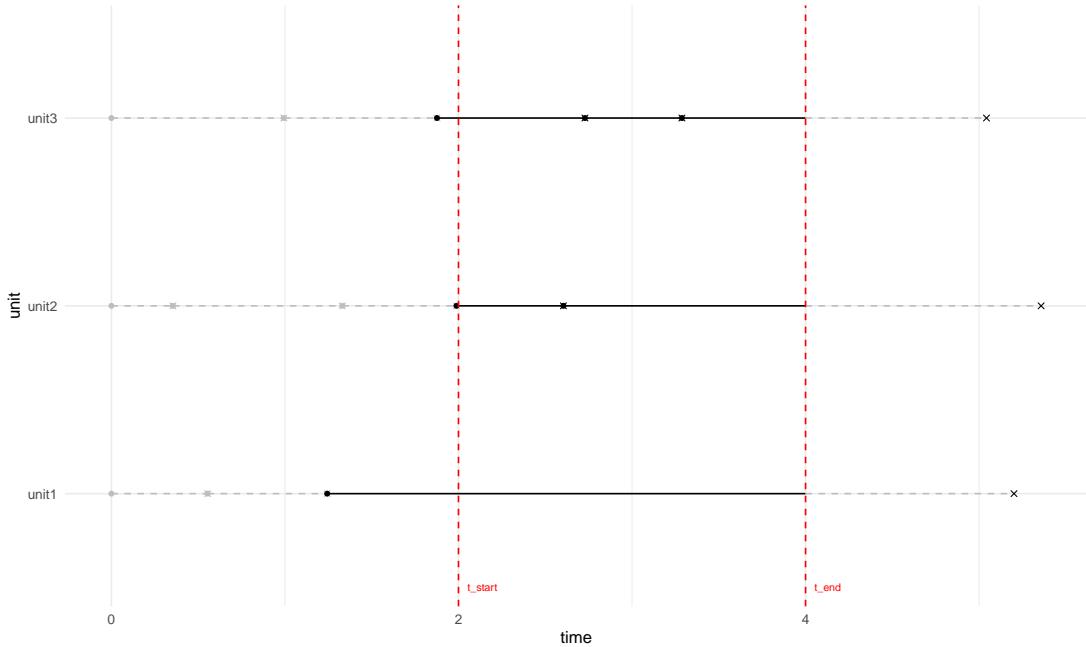


Figure 2.4

Hong et al. (2009) shows how a likelihood for data that is left-truncated and right-censored can be constructed using the general approach of integrating out the censored observations,

$$L(\theta|y^O, t_{\text{Lower}}^L, \tau_i^L) = \prod_{i=1}^{N^O} \left[\frac{f(y_i^O)}{1 - F(\tau_i^L)} \right] \prod_{j=1}^{N^C} \left[\frac{1 - F(t_j^{\text{Lower}})}{1 - F(\tau_j^L)} \right], \quad (2.9)$$

where τ^L are the truncation times and if the lifetime is not truncated $\tau^L = 0$. Kundu and Mitra (2016) then implements the same approach in a Bayesian framework using a gibbs sampling algorithm to draw samples from the posterior. Mitra (2013) takes the alternative approach of imputing the censored lifetimes using an expectation maximisation algorithm and the complete data likelihood

$$L(\theta|y^O, \hat{y}^C, \tau_i^L) = \prod_{i=1}^{N^O} \left[\frac{f(y_i^O)}{1 - F(\tau_i^L)} \right] \prod_{j=1}^{N^C} \left[\frac{f(\hat{y}_j^C)}{1 - F(\tau_j^L)} \right], \quad (2.10)$$

where the \hat{y}_j^C are the imputed values of the censored observations. I express the same approach as 2.10 in a Bayesian framework as

$$\begin{aligned} y_i^O | \beta, \eta &\sim \text{Weibull}(\beta, \eta) \quad T[\tau_i^L,] \\ y_j^C | \beta, \eta &\sim \text{Weibull}_{t_j^{\text{Lower}}}(\beta, \eta) \quad T[\tau_j^L,] \\ \beta, \eta &\sim \pi(\theta_{\beta, \eta}), \end{aligned}$$

where $\text{Weibull}_{t_j^{\text{Lower}}}$ indicates that the random variable y_j^C has a Weibull distribution and is constrained to be greater than the censoring time, and $T[\tau^L,]$ indicates that the distributions are re-normalised by the probability $P(y > \tau^L)$ (Note: I stole this notation from the stan code). For the moment, I express a joint prior for the parameters in its most general form.

Unknown truncation time A problem arises when the installation time of the left-truncated lifetimes is unknown, since to normalise the truncated lifetime distribution of the left-truncated observation τ^{Left} must be known. For example, if for the dataset shown in Figure 2.4 there was no information at all prior to t_{start} , then we could not use the likelihood in 2.9 or 2.10. This is the case for the idler frame data in. This problem is referred to either as unknown exposure history () or initial conditions (). In these case, there are two approaches that can be taken (Guo, 1993). The first, discard all of the left-truncated samples (), in which case the parameter estimates are still unbiased (). However, in doing so we throw away a large amount of information. In most cases of left-truncation and right censoring, the right censoring masks any information about longer lifetimes and so the left-truncated samples are the only source of information about the upper tail of the lifetime distribution. The second approach is to assume a constant hazard, i.e. $\beta = 1$, since in this case the Weibull distribution reduces to the exponential and, no matter the age of a unit, the probability of it surviving a given period is constant (this is the memoryless trait of the exponential distribution). However, assuming a constant hazard is very restrictive and often one of the

aims of performing lifetime analysis in the first place is to determine if $\beta > 1$. Furthermore, assuming an exponential distribution when the data do not have a constant hazard may lead to severe bias in the parameter estimates (?). In *Section ??* I show how treating the unknown installation times as a case of censoring and using the method of imputing the censored data, the missing truncation times can also be imputed and reasonable parameter estimates can be obtained.

2.2 Imputing truncation times

Using the toy example in *Figure 2.4*, say we do not observe any of the installation or failure times to the left of t_{start} . In this case, we know that the first, partially, observed lifetime from each unit started some time between $t = 0$ and $t = t_{\text{start}}$. This is a case of interval censoring, where the lower censoring bound is the time from the beginning of observation to the failure time and the upper bound is from $t = 0$ to the failure time. If we did not know the origin time $t = 0$ with respect to $t = t_{\text{start}}$ then it would be a case of right censoring, but the following logic would still apply. Let t_i^{failure} be the failure time of the i^{th} observation. Treating the observations as interval censored, the left-truncated lifetime can be imputed as in *Section ??* by sampling from

$$\hat{y}_i^L | \beta, \eta \sim \text{Weibull}_{t_i^{\text{failure}} - t_{\text{start}}}^{t_i^{\text{failure}}} (\beta, \eta). \quad (2.11)$$

Using the imputed values of the lifetime it is then possible to calculate the truncation time by

$$\tau_i^L = \hat{y}_i^L - (t_i^{\text{failure}} - t_{\text{start}}). \quad (2.12)$$

A complication arises when the lifetime is both interval censored by the start of the observation period and right censored by the end, such as unit one in *Figure 2.4*. In this case the truncation times value is unknown but is between

$\tau^L = 0$ and $\tau^L = \min(t_{\text{start}}, \hat{y}_i^L - (t_{\text{end}} - t_{\text{start}}))$. Since we have no reason to expect that the truncation time is not uniform, we can impute the truncation time using the uniform distribution

$$\tau_i^L \sim \text{Uniform}(0, \min(t_{\text{start}}, \hat{y}_i^L - (t_{\text{end}} - t_{\text{start}}))). \quad (2.13)$$

Sampling the value of τ_i^L in this way should incorporate the extra uncertainty in our Bayesian estimates.

2.3 Informative joint prior

In cases where data only provide partial information, such as lifetime data where some lifetimes are masked by censoring, informative priors can help to ‘fill in the gaps’ and inform areas of the model that the data cannot. For example, right censoring masks the upper tail of the lifetime distribution, and if there are no observed failures beyond some censoring time then the data do not contain any information that informs this upper tail of the lifetime. This is clear in Section ?? when I fit a Weibull distribution to censored lifetimes after discarding any truncated lifetimes. An added motivation for using an informative prior in these cases is that non-informative priors can place mass in unreasonable parts of the parameter space and when combined with a weak likelihood, this leads to spurious parameter estimates (Tian, Lewis-Beck, Niemi, & Meeker, 2024).

Kaminskiy and Krivtsov (2005) proposes a method for encoding an informative joint prior for the two parameters of the Weibull distribution by eliciting information about cross sections of the weibull CDF. In the proposed method, the analyst provides their expected value of the CDF at two exposure times, t_1 and t_2 , and their level of uncertainty around these estimates. The pair of estimate and uncertainty level at each exposure time $(\mu_{\hat{F}_{t_i}}, \sigma_{\hat{F}_{t_i}})$ are encoded as the mean and standard deviation of a beta distribution. By then sampling realisations of the CDF at each exposure time from the two distributions, ensuring that $\hat{F}_{t_1} < \hat{F}_{t_2}$,

the parameters of the Weibull CDF that passes through the two realisations can be calculated to obtain a draw from the informative joint prior. Kaminskiy and Krivtsov (2005) then show how to obtain a Bayesian point estimate of the CDF at some new exposure time where binomial failure data are available by first calculating the parameters of a new beta distribution that describes the CDF at the new exposure time using the joint prior draws and then updating the parameter estimates with the data, taking advantage of the beta-binomial conjugacy.

The method that Kaminskiy and Krivtsov (2005) propose to obtain joint draws from an informative prior for the two weibull parameters falls under the recommendations of Gelman, Vehtari, et al. (2020), who suggest to elicit information on the outcome space—which is more familiar to partitioners—and then translate this information into an informative prior also that indirectly describes how the parameters should be allowed to covary with one another. An alternative method proposed by (Meeker et al., 2022) to elicit an informative prior for the Weibull parameters is to reparameterise the distribution in terms of the shape β and some quantile, p_r , of the distributions and then specify independent priors for β and p_r . Both methods have proven to be useful in practice and been implemented in reliability software due to their usability (?). However, it can be difficult to elicit information about β for practitioners not very familiar with Weibull analysis. Therefore, I use a variation of Kaminskiy and Krivtsovs method for obtaining draws from a joint informative prior for the Weibull parameters and build upon their method by showing...

- Need to discuss Krivtsov’s development of Kaminskiy and Krivtsov (2005) somewhere...

In their proposed procedure, the prior is updated with the data using the conjugacy of the binomial likelihood and beta prior for the CDF at the new exposure time t_3 . Doing so only provides point estimates of the updated parameter values of the beta distribution at t_3 . To obtain corresponding values of the Weibull parameters, the analyst must once again sample paris of realisations along the

CDF—now at either t_3 and t_1 or t_3 and t_2 —and calculate the parameters of the Weibull CDF that passes through the two points. In doing so they ‘reuse’ the prior distribution at either t_1 or t_2 to generate the posterior and so the ‘prior belief’ about the CDF at which ever time is used is not really updated at all. Furthermore, the resulting joint draws will be sensitive to whether t_1 or t_2 is used to regenerate the joint distribution of the Weibull parameters. By instead implementing the method for obtaining draws from the joint prior within the HMC sampling algorithm, I show how to properly filter the prior through the likelihood to obtain the full posterior. Doing so also means that the prior can be updated with detailed lifetime data(i.e. failure times rather than binomial trial data), censoring and truncation information can be included in the likelihood, and the resulting full posterior contains proper uncertainty quantification.

Rather than beta distributions, I use normal distribution truncated between zero and one to encode prior belief about the CDF at t_1 and t_2 . This is because when specifying prior information close to the boundary, for example if our estimate of the cdf is 0.95 or 0.05, with reasonable uncertainty surrounding the estimate, a beta distribution places lot of mass very close to zero or one which causes numerical issues during sampling. Truncated normal distributions are much better behaved. The prior can be expressed as

$$\begin{aligned}\hat{F}_{t_1} &\sim N_0^1(\mu_{\hat{F}_{t_1}}, \sigma_{\hat{F}_{t_1}}) \\ \hat{F}_{t_2} &\sim N_{\hat{F}_{t_1}}^1(\mu_{\hat{F}_{t_1}}, \sigma_{\hat{F}_{t_1}}) \\ \beta &= \frac{g(\hat{F}_{t_2}) - g(\hat{F}_{t_1})}{\log(t_1/t_2)} \\ \eta &= \exp \left[\log(t_1) - \frac{\hat{F}_{t_1}}{\beta} \right]\end{aligned}$$

where $g(\hat{F}) = \log(-\log(1-\hat{F}))$ and the superscript and subscripts on the normal distribution indicate the upper and lower constraints of the support.

Depending on the choice of t_1 and t_2 the analyst encodes different information

into the joint prior, which is reflected in how the two parameters covary with one another. For demonstration, plots (a), (b), and (c) in Figure 2.5 show the draws from three different joint priors constructed using this elicitation method. To construct the prior, I specify some ‘true’ values of the parameters ($\beta = 1.1$; $\eta = 1$) and then specify the priors to reflect the true value of the CDF at different quantiles $t_i = p_r$ so that all of the priors contain the ‘true’ parameter values. (a) shows a prior where information is encoded at early exposure times ($t_1 = p_{0.05}$ and $t_2 = p_{0.20}$), (b) shows one where information is encoded at times around the median age ($t_1 = p_{0.40}$ and $t_2 = p_{0.60}$), and (c) shows one where the elicitation times are in the upper tail of the lifetime distribution ($t_1 = p_{0.90}$ and $t_2 = p_{0.99}$). Plots (d), (e), and (f) in Figure 2.5 show the corresponding uncertainty around the Weibull CDF that results from the different priors in (a), (b), and (c) respectively. We see that the joint prior allows us to express our prior belief in one area of the PDF while allowing the prior to be vague in others. This characteristic has useful application in the context of heavy censoring since censoring masks longer lifetimes and hence increases uncertainty in the upper tail of the distribution, but still contains useful information in the lower tail.

In this chapter, I compare the informative joint prior shown in Figure 2.5 (c) and (f), which encodes information in the upper tail, with the vague independent priors $\beta \sim N^+(1.1, 1)$ and $\eta \sim N^+(1, 1)$, where the superscript (+) indicates that the prior is truncated to be positive.

2.4 Analysis of simulated data

- Here I demonstrate the method with simulation.
- I simulate lifetime data in a way that emulates the observational process of the idler frame data.
- However, I save the installation times of observations that are left truncated by the start of the observation time.

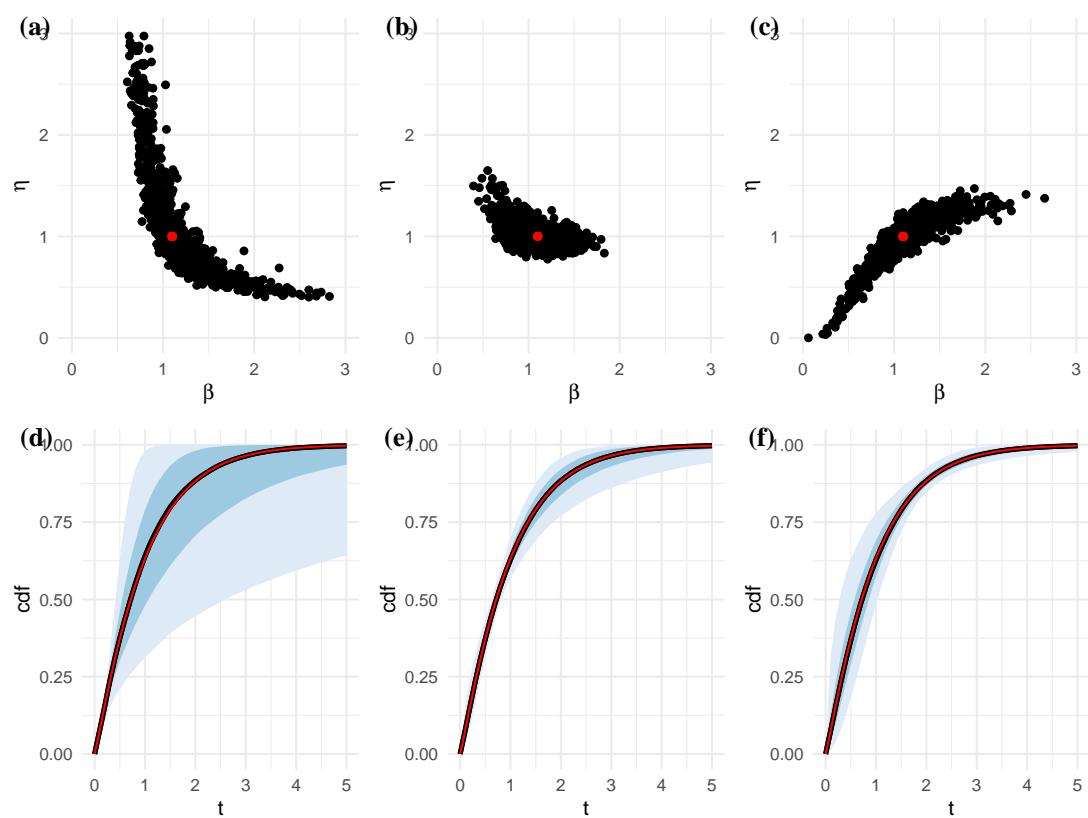


Figure 2.5

- I then compare the inference of the imputation approach with the alternative where we discard all of the left censored lifetimes and also with the case where we do in fact know exposure history of the left truncated samples.
- For a small sample, the imputation approach gives almost the same inference as the case where we know the true installation times. When we throw out all of the left truncation times, there is a massive increase in uncertainty, particularly around the upper tail of the lifetime distribution, as shown in the posterior of the CDF.
- I then refit the same models to the simulated data but with an informative joint prior encoded in a away that informs the upper tail of the distribution. All of the models now give very similar results, however, the case where we discard the left truncated samples still has larger uncertainty.

2.4.1 Simulation method

To simulate data in a way that emulates the idler frame lifetime data set, I sample $N \times M$ draws from a Weibull distribution with known shape parameter $\beta = 1.1$, and scale parameter $\eta = 1$. I then assign these lifetimes to $M = 10$ units. To calculate failure times rather than lifetimes, I take the cumulative sum of the N lifetimes assigned to each unit. The installation times are calculated by taking the lag of the failure times. I then define a start, $t_{start} = 5$, and end, $t_{end} = 6$, time for the observation window. Any lifetimes where both the install and failure times sit either before t_{start} or after t_{end} are discarded. Of the remaining lifetimes, if the install time is less than t_{start} then t_{start} is set as the observed start of the lifetime and the observation is marked as left truncated, while if the failure time is greater than t_{end} then t_{end} is substituted as the observed failure time and the observation is marked as right censored. Table 2.1 presents a simulated dataset using this method and parameter values. The simulated observations are also plotted in Figure 2.6, where the start and end of observation are show as red

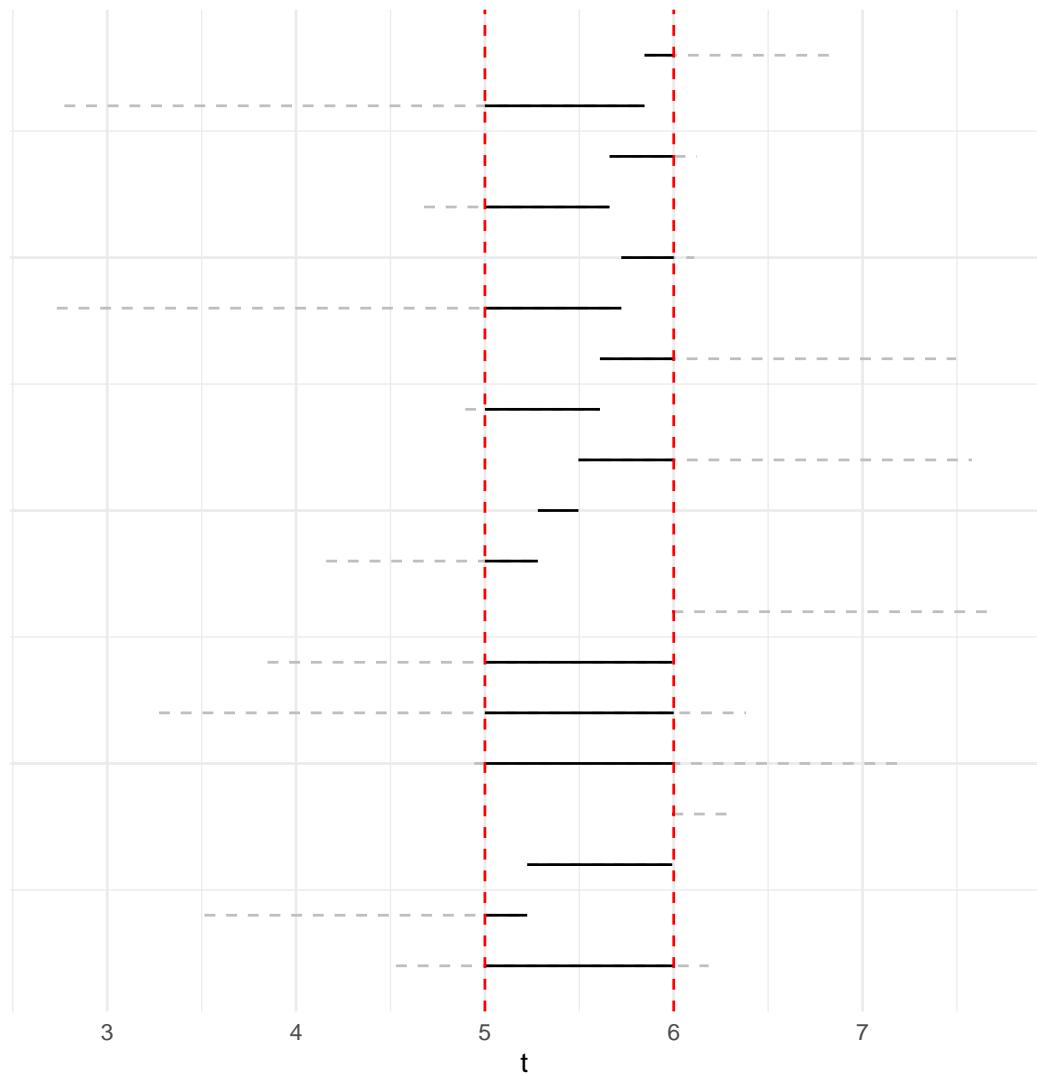


Figure 2.6: The simulated lifetimes data.

vertical lines and the observed portions of the simulated data are shown in black and the unobserved ‘incomplete’ part is shown in gray. Notice that if we were to throw away the left truncated samples, then we would be discarding more than 50% of the sample.

2.4.2 Weakly informative prior

I now fit the imputation model to the simulated data in Table 2.1 and compare the resulting posterior with the case where the installation times of the left trun-

Table 2.1: The simulated CMMS data for ten units.

Unit	Lifetime	True failure	True install	Observed install	Observed failure
1	1.66	6.18	4.53	5.00	6.00
2	1.71	5.22	3.51	5.00	5.22
2	0.77	5.99	5.22	5.22	5.99
2	0.31	6.31	5.99	5.99	6.00
3	2.28	7.23	4.94	5.00	6.00
4	3.11	6.38	3.27	5.00	6.00
5	2.14	5.99	3.85	5.00	5.99
5	1.69	7.68	5.99	5.99	6.00
6	1.12	5.28	4.16	5.00	5.28
6	0.21	5.50	5.28	5.28	5.50
6	2.08	7.58	5.50	5.50	6.00
7	0.71	5.61	4.90	5.00	5.61
7	1.89	7.50	5.61	5.61	6.00
8	2.99	5.72	2.73	5.00	5.72
8	0.39	6.11	5.72	5.72	6.00
9	0.98	5.66	4.68	5.00	5.66
9	0.46	6.12	5.66	5.66	6.00
10	3.07	5.85	2.77	5.00	5.85
10	1.03	6.88	5.85	5.85	6.00

cated samples are known and also with the alternative treatment of the unknown exposure histories, which is to discard all of the left truncated samples. I fit these models with a weakly informative, independent priors

$$\begin{aligned}\hat{F}_{t_1} &\sim N(1.1, 1) \\ \hat{F}_{t_2} &\sim N(1, 1).\end{aligned}$$

The three stan models are available on the GitHub repository. To sample from the posteriors, I use four chains each 1000 iterations long with a burn in of 500. Figure 2.7 shows the three posteriors. Plots (a), (b), and (c) show the joint draws of the two weibull parameters for the fully observed, discarded, and imputed treatment of the left truncated samples respectively. The corresponding CDFs with uncertainty are shown in (d), (e), and (f). The true values of the parameters and the true CDF are plotted in red in the plots. The resulting inference from imputing the installation times of the left truncated samples is almost the exact same as if we had fully observed the left truncated samples. Whereas, discarding the left truncated samples results in a much more diffuse posterior and consequently, more uncertainty around the CDF, particularly for larger exposure times. Next, I refit the models using a slightly more diffuse version of the informative prior shown in Fig. 2.5 (c) and (f) which informs the upper tail of the distribution.

2.4.3 Strongly informative prior

The prior in Fig. 2.5 (c) and (f), which elicits information about the CDF at $t_1 = p_{0.80} = 1.54$ and $t_2 = p_{0.95} = 2.71$ encodes an informative prior that strongly informs the Weibull model in the upper tail of the distribution but is sufficiently vague in the lower tail, where the data are strongly informative. Here I refit the weibull models with a slightly more diffuse version of the prior; i.e. I increase the uncertainty around the estimates of the CDF at each elicitation time. The

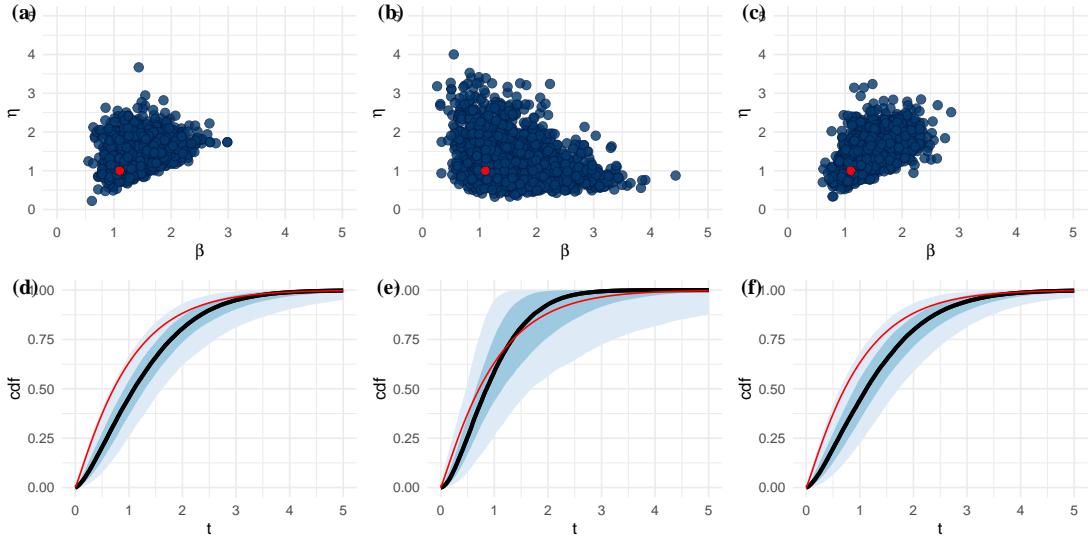


Figure 2.7: The prior distributions.

informative prior is

$$\hat{F}_{t_1} \sim N_0^1(0.8, 0.1)$$

$$\hat{F}_{t_2} \sim N_0^1(0.95, 0.05).$$

Using the informative prior, I refit the three models from Section 2.4.2. Once again, I use four chains each 1000 iterations long with a burn in of 500 to sample from the three posteriors. The resulting draws are plotted in (a), (b), and (c) of Figure 2.8 for the fully observed, discarded, and imputed treatment of the left truncated samples respectively. Comparing the posterior of the imputation method (plots (c)) with the case where the left truncated samples are fully observed (plots (a)), the resulting inference is again very similar. However, now using the informative prior, the posterior of the case where I have discarded the left truncated samples, plot (b), is much closer to the other two posteriors than when a vague prior is used. The corresponding posterior CDFs are shown in (d), (e), and (f) of Fig. 2.8. The uncertainty in all cases has been refined, but most drastically in the upper right part of the CDF. In Fig. 2.8 (e), the effect of combining a likelihood that informs the CDF at lower exposure times with a prior

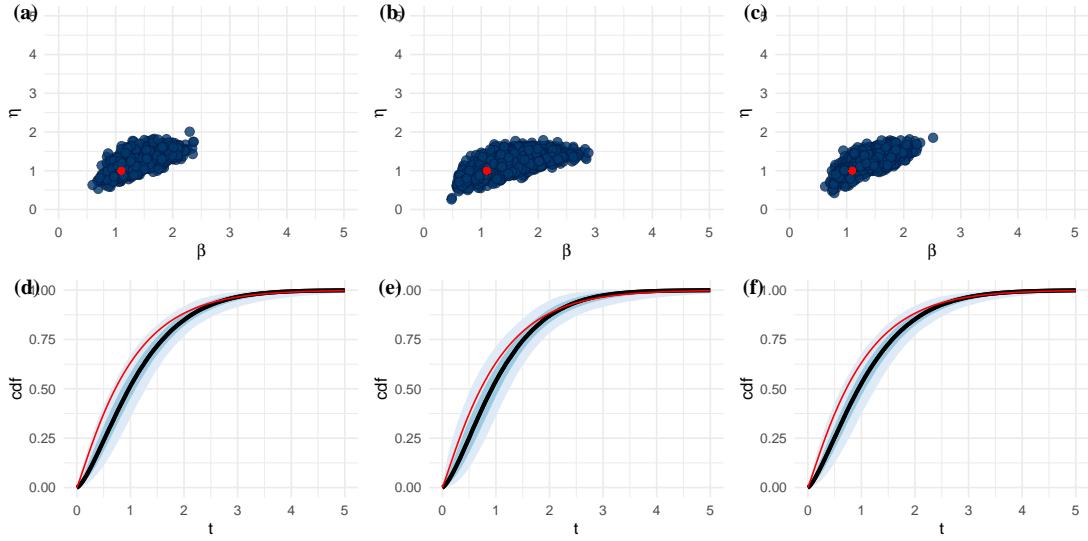


Figure 2.8: The prior distributions.

that informs the CDF at higher exposure times has resulted in a precise estimate of the CDF.

In the implementation of the joint prior that I have used, the prior properly updated. i.e. the observed data have slightly updated our belief about the value of the CDF at both t_1 and t_2 . Figure 2.9 compares the distribution of the posterior draws of $\hat{F}(t_i)$ (gray densities) with the prior (red curves). In the three plots in Fig. 2.9, the posterior distributions are clearly narrower than the prior; more so at t_1 . Fitting the Weibull a models for the data in Table 2.1 has shown that, for small samples sizes, in cases where we do not know the exposure history of the left truncated samples the imputation method provides almost equivalent inference to if we had fully observed the left-truncated samples. Further more, if we alternatively ‘throw away’ the left truncated observations this results in a large loss of information, however, this loss can be compensated for if there is prior information about longer lifetimes and this information is properly encoded into a joint prior. Here I have shown one case for a reasonably small sample. Next, I devise a small simulation experiment to more rigorously evaluate the imputation method for the left truncated samples with unknown exposure history.

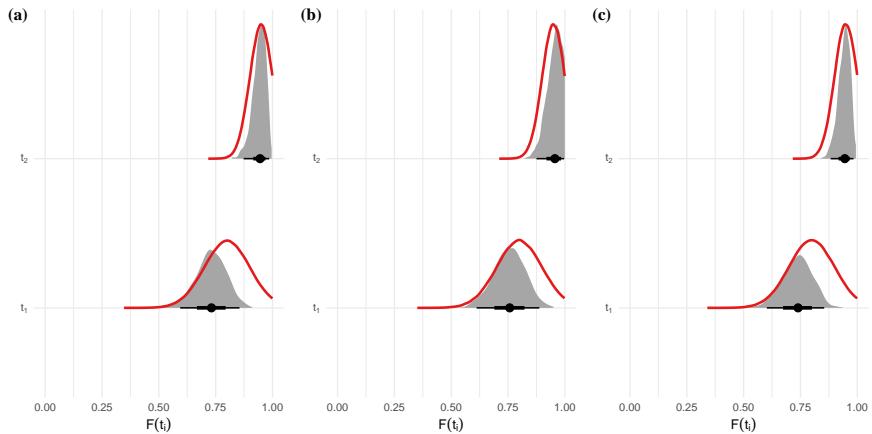


Figure 2.9: Comparison of the prior and posterior for F_{t_1} and F_{t_2} to show how both the elicited distributions have been updated.

2.5 Simulation experiments

To explore the behaviour of the imputation method for left truncated lifetimes with unknown exposure history, I repeat the simulation and model fitting of Section 2.4 for a number of different combinations of the simulation parameters; number of units, start of observation time, and length of observation time. I compare the imputation method—under both a vague and informative prior—with the alternative method of discarding the left truncated observations as well as with the case where the left truncated samples are fully observed (as a ground truth). First I define the factor levels that make up the factor combinations, next I describe the measures of model accuracy that I use to compare the three models, finally, I present the results of the simulation experiments.

2.5.1 Factor levels

I vary three factors when simulating the datasets, each with three levels;

- N : 10, 100, 500 units
- t_{start} : 1, 5, 15 mean lifetimes
- $t_{end} - t_{start}$: 1, 3, 6 mean lifetimes

Increasing the number of units, N , increases the sample size, increasing t_{start} increases the range of possible values the left truncated samples with unobserved truncation time could take—and therefore reduces the information they contribute—and increasing the window size, $t_{end} - t_{start}$, increases the number of fully observed lifetimes, the maximum length of the fully observed lifetimes and the number of samples that are both left truncated and right censored. For each factor combination, I perform one hundred simulations, each time fitting all of the three treatments for the left truncated observations with both a vague and informative prior:

- model: include lt, discard lt, impute lt
- prior: weakly, strongly informative

...

2.5.2 Accuracy measures

To compare how well the Bayesian models reclaim the true data generating mechanism, I calculate the Bayesian p-values of the true parameter values— $\beta = 1.1$ and $\eta = 1$ —under the fitted posteriors and the *elpdd* of a dataset of one hundred fully observed lifetimes generated from the true weibull distribution. The Bayesian p-value for beta is easily calculated using the posterior draws as

$$p_\beta = \frac{\sum_{s=1}^S \beta_{true} \leq \beta_s}{S},$$

similarly the P-value of η can be calculated in the same way. If the model posterior is well calibrated, then the p-values from repeating the same simulation should have a uniform distribution. If the parameter estimates are bias then the P-values should on average sit above or bellow 0.5. If the posterior uncertainty is overly diffuse then the p-values will be clustered around 0.5 and if they the uncertainty is too narrow then the P-values will be pushed towards zero or one. The P-value

provides a indication that the model is over or under predicting the parameter values, but do not express to what degree.

To determine the scale of any discrepancy, I also calculate the *elppd* of a new sample of one hundred fully observed lifetimes under the different posteriors. I do this by simulating one hundred datasets, each with one hundred observations, and calculating the expected log likelihood of each simulated dataset and then taking the average,

$$elppd_{100} = \frac{\sum_{n=1}^{100} \frac{1}{S} \sum_{s=1}^S \sum_{i=1}^{100} f(\tilde{y}_{n,i} | \beta_s, \eta_s)}{100}.$$

2.5.3 Results

The factor combination effects on the p-values.

- Description of the plots
- Bias in the estimate when the window size is small (I think comes the issue that stan tries to update the truncation time).
- There is a small bias in the other cases, these get more visible as the dataset becomes more informative and therefore, the prior more precise.
- Despite the slightly biased estimates, the elppd values are remain close. I show this in more detail in the next figure.

The effects on the elppd score. (elppd under the true model is -95.7)

- The value is small when the data set is small and the window size is short

The observed bias is small relative to the noise of the measures.

2.6 Concluding remarks

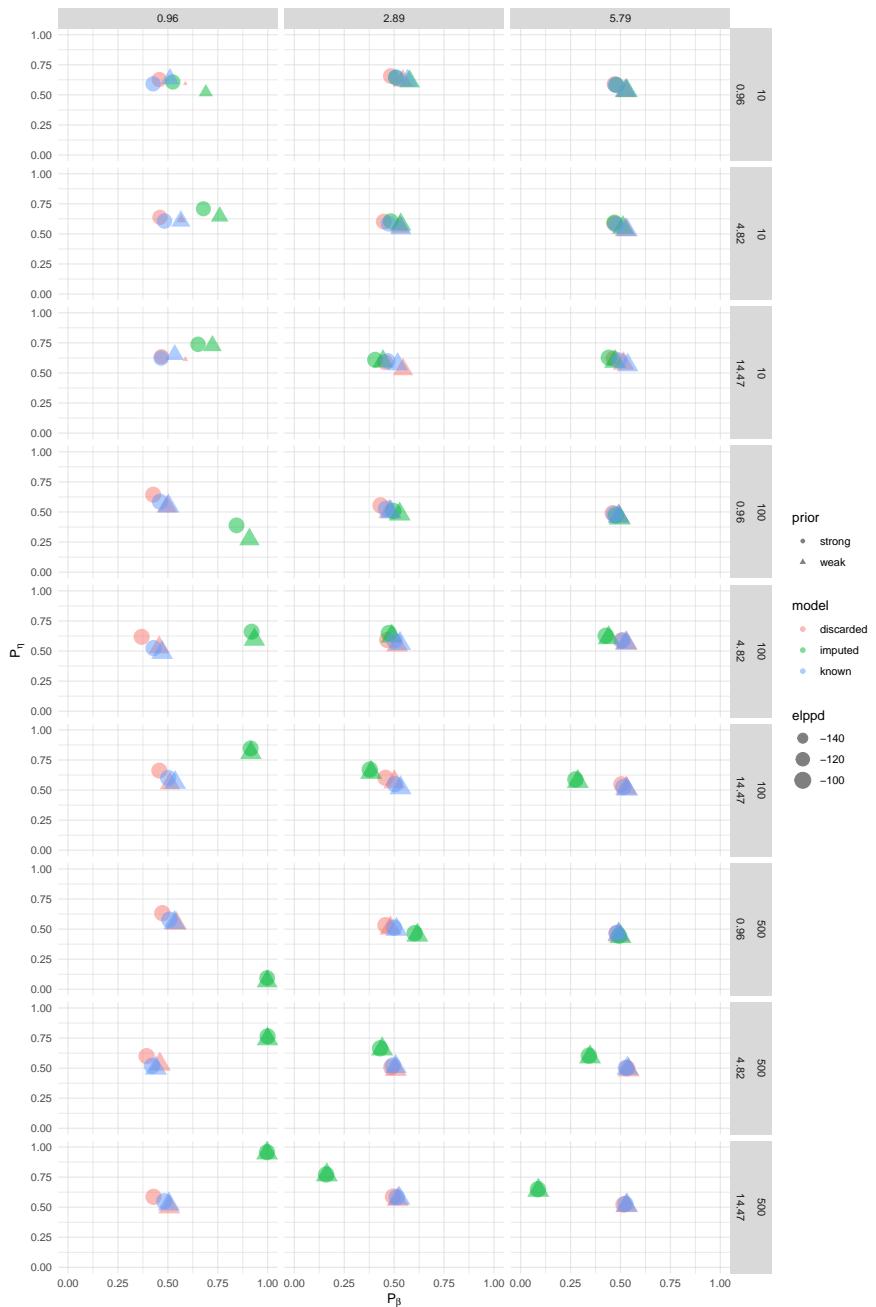


Figure 2.10: The simulation results

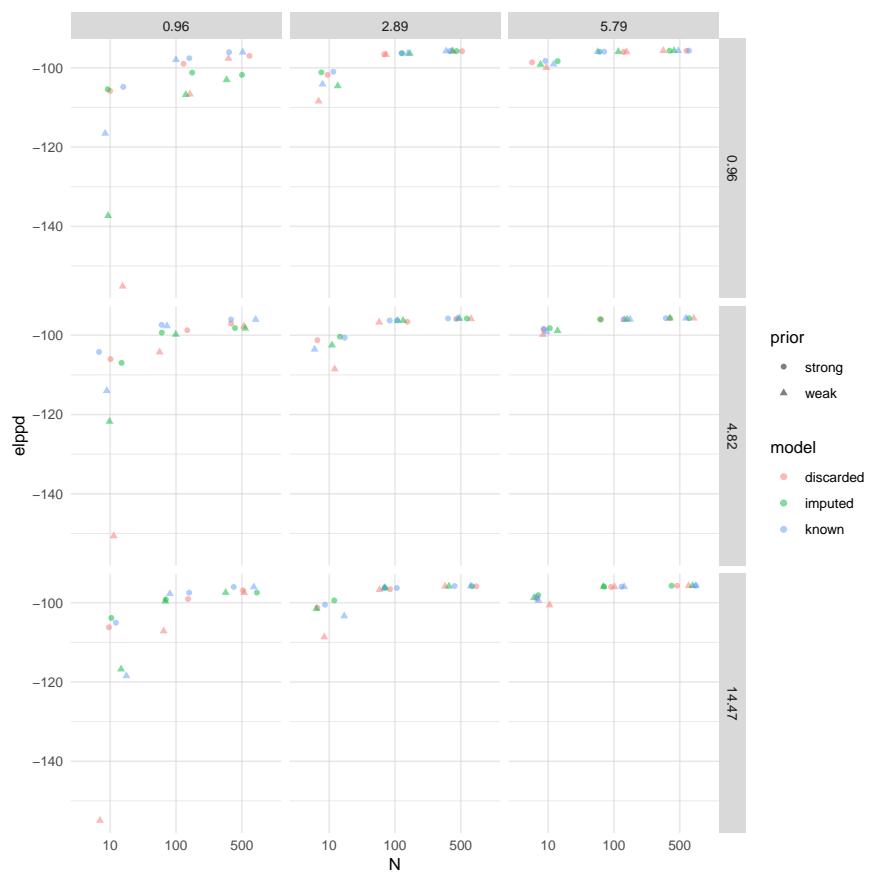


Figure 2.11: The simulation results

Chapter 3

Case study

In this chapter, I apply the work in the previous chapter to the idler frame dataset and show how the draws from the full posterior, obtained through MCMC sampling, can be used to generate useful quantities for decision making, such as expected number of failures in the next small period and expected cost of a maintenance strategy.

3.1 Idler frame data

Summary of the data and relate to findings from previous chapter.

3.2 An informative prior

Justify the prior based on manufacturers information.

3.3 Fitting

Model fitting, joint posterior, and CDF.

3.4 Using the posterior

fit for each unit with uncertainty This naturally comes in the posterior samples

expected number of failures Should be able to determine this for each draw (i.e. the cumulative sum of the imputed failure time of the units still under test.)

cost function Construct my own cost function around the ones presented in Jardine.

Assumptions

- Shuts are every six weeks. t_p is the replacement interval and therefore must be a multiple of six.
- N is the number of idler frames on the belt.
- C_p is the cost of a preventative replacement.
- When an idler frame fails during operation, there is a $P = 0.1$ chance that it will damage the belt. C_f is the cost of unexpectedly replacing the belt.
- The expected number of failures, $E(N_f)$, in a time interval t_p is $N \times F_{\theta|y}(t_p)$.
- The probability that at least one unplanned idler failure damages the belt during the interval is $1 - (1 - P)^{N F_{\theta|y}(t_p)}$.

$$\text{Cost per unit time} = \frac{C_p N + C_f (1 - (1 - P)^{N F_{\theta|y}(t_p)})}{t_p} \quad (3.1)$$

$F_{\theta|y}(t_p)$ is the posterior CDF of the Weibull distribution (the failure time distribution). Hence, we can average over the uncertainty in the posterior draws to get the expected cost per unit time and uncertainty estimates by calculating the cost function for every posterior draw; $F_{\theta^s}(t_p)$.

If we calculate this for the set of possible shut periods we can compare the choices with uncertainty.

In this example, I have not accounted for the possibility that the belt could be damaged, replaced, and then damaged again in a single interval! Jardine talks about some methods for expected number of repeat failures in a time interval in section 2.4.3

This is just a demonstration on how to use the full posterior in the decision making process.

3.5 Discussion and conclusions

Typical results section.

Part II

Part two: Degradation modelling

In this second part of the thesis, I focus on degradation modelling—primarily using the gamma stochastic process. The more foundational work in this second half of the thesis—chapters 4 and 5—arose out of necessity to apply the gamma process to industry data from the Centre’s industry partners. Particularly, the work was motivated by two applied problems: modelling the degradation of many pipe spools in a bauxite processing plant and modelling the wear of an overland iron ore conveyor’s belt—the latter was in one of my industry placements. The pipe degradation project, undertaken by Gabriel González Caceres and Aloke Phatak, looked at the degradation in wall thickness of pipe spools. The gamma process should theoretically be a good model for the erosion of the inside of the spools. However, the degradation signals for each spool are very noisy and consist of only a few measurements; therefore, there was a need to accommodate noise in the gamma process model and to borrow information between similar units in order to estimate the parameters of the model. The belt wear project was an industry placement where I was asked to analyse the belt wear of all conveyors at an industry partner mine, processing facility, and port and devise a better method for modelling belt wear. To analyse the many wear profiles, I used functional data analysis techniques. This led to taking the same approach when modelling the belts’ degradation. The gamma process should also be a good model for the degradation of the belt since the degradation mechanism is wear. However, similar to the pipe spool problem, the signals of the spline coefficients were noisy and sparsely observed.

In the first chapter of this part, chapter 4, I demonstrate how the gamma process can be extended to incorporate noisy degradation measurements through Bayesian hierarchical modelling. Then, in chapter 5, I show how unit-to-unit variability (which facilitates the borrowing of information between units) can be included in the proposed noisy gamma process model. The work in these two chapters has been submitted for publication as the manuscript “*Bayesian hierarchical modelling of noisy gamma processes: model formulation, identifiability,*

model fitting, and extensions to unit-to-unit variability” to Applied Models in Business and Industry. The chapters present an expansion of the manuscript by tying the supplementary material submitted alongside the manuscript into the main analysis and discussion and providing a deeper discussion of the two main topics. There is also a larger focus on the computational details of the analysis.

In chapter 6, I present the belt wear model developed as part of the belt wear industry placement, which utilises the noisy gamma process model. The work in chapter 6 has been submitted for publication to the Annals of Applied Statistics as the manuscript “*Functional degradation modelling of the wearing surface of conveyor belting using Bayesian hierarchical modelling and Gamma processes*”. In addition to the work presented in the manuscript, chapter 6 provides a side-by-side comparison of the gamma process-based belt wear model from the manuscript with a linear general path-based model.

The e-prints for the two papers have been uploaded to arXiv. The full references for the manuscripts are

1. Leadbetter, R., González Cáceres, G., & Phatak, A. (2024). *Bayesian hierarchical modelling of noisy gamma processes: Model formulation, identifiability, model fitting, and extensions to unit-to-unit variability*.
2. Leadbetter, R. & Phatak, A. (2024). *Functional degradation modelling of the wearing surface of conveyor belting using Bayesian hierarchical modelling and Gamma processes*.

Chapter 4

A noisy gamma process for modelling degradation measurements with uncertainty

If there are very few or no failures observed for a particular component or asset then the lifetime methods that we have looked at in Chapters 2 and 3 are not very useful in reliability decision-making. If there is some measure of the degradation process that drives failure, then degradation modelling can be used to forecast the degradation of units and inform reliability decision-making. Gamma stochastic processes are a widely used degradation model for degradation that evolves monotonically (Lawless & Crowder, 2004). However, most degradation data collected in industrial settings is contaminated by noise, or error. This noise can be attributed to different sources, including measurement error, instrument noise, placement of sensors, and other environmental factors (Ye & Xie, 2015). Consequently, models for gamma processes must be extended to account for such noise.

In this chapter, I show that this necessary extension can be facilitated in a straightforward and tractable way using the Bayesian hierarchical modelling framework. I also demonstrate, through simulation, that this noisy gamma pro-

cess model is more difficult to fit than a standard gamma process due to a pre-asymptotic identifiability issue. In the remainder of the chapter, I begin in Section 4.1 with a brief introduction to the gamma process as it is used for degradation modelling. In Section 4.2, I then provide a short overview of current works on modelling noisy gamma processes in the reliability literature and introduce how the model can be implemented using the Bayesian hierarchical modelling framework. In Section 4.3, I discuss the merits of reparameterising the gamma distribution in terms of orthogonal parameters that are interpretable in terms of the average degradation rate and volatility of the gamma process; these parameters allow us to more easily think about how to specify prior distributions for the parameters of a gamma process. I then discuss the equally important step of justifying the prior and performing prior predictive checking in Section 4.4, which allows us to specify sensible prior distributions. In Section 4.5, I fit the noisy gamma process model to simulated data and show that when there are only a few noisy observations, MCMC sampling and posterior inference are poorly behaved. I diagnose that these issues results from the challenges of separating the measurement error from the inherent volatility of the gamma processes using the useful diagnostics of HMC. After identifying and explaining the issue, the section concludes with a demonstration of how the poor behaviour of sampling and inference can be resolved by adding a small amount of supplementary information into the analysis, both through a more informative prior or supplementary data. I end in Section 4.6 by summarising the main results and pointing the way to future work.

4.1 The gamma process

The gamma process is a type of stochastic jump process. It was introduced to the reliability domain by Abdel-Hameed (1975), and since then has been used in many applications including the modelling of the corrosion of steel coatings, wear of brake pads, erosion of breakwaters, thinning of pressure vessels, and

degradation of LED lights (van Noortwijk, 2009).

Consider a sequence $\{z_i\}$ of noise-free measurements of the degradation of a unit observed at times t_i , $i = 0, 1, 2 \dots, I$. Without loss of generality, I assume that $z_0 = 0$ at $t_0 = 0$. A gamma process (Lawless & Crowder, 2004) models the jumps in degradation between measurements, $\Delta z_i = z_i - z_{i-1}$, as independent samples from a gamma distribution. Thus, we can write that

$$\Delta z_i | \eta(\cdot), \xi \sim \text{Ga}\{\eta(t_i) - \eta(t_{i-1}), \xi\}, \quad (4.1)$$

with rate ξ and shape $\eta(t_i) - \eta(t_{i-1})$, where $\eta(\cdot)$ is a given monotone increasing shape function. The simplest gamma process for modelling degradation is a stationary gamma process, which has a linear shape function (Frenk & Nicolai, 2007), for example, $\eta(t_i) = \beta t_i$. Of course, nonlinear shape functions can be used; however, even when the degradation trace appears to be nonlinear, a time transformation can often be applied so that a stationary gamma process can be fitted. Therefore, in what follows, I consider only the stationary gamma process.

When using a linear shape function, we can write eq. 4.1 more simply as

$$\Delta z_i | \beta, \xi \sim \text{Ga}(\beta \Delta t_i, \xi), \quad (4.2)$$

where $\Delta t_i = t_i - t_{i-1}$.

The gamma process described in eqs. 4.1 and 4.2 can be extended to describe situations commonly encountered in practice, namely, the need to account for measurement error and/or unit-to-unit variability when the degradation of several identical or similar units is being measured. I discuss measurement error next and defer discussing unit-to-unit variability until Chapter 5.

4.2 A noisy gamma process

In this section, I give a background to the noisy gamma process and describe how its implementation can be simplified using the BHM framework introduced in Sec. 1.4. In an early paper, Kallen and van Noortwijk (2005) fit a single parameter gamma process to noisy data by using the additive model $y_i = x_i + \epsilon_i$, where y_i represents the noisy observations, x_i represents the underlying gamma process, and ϵ_i is independent and identically distributed Gaussian noise. The gamma process is parameterised in terms of the mean wear rate (β/ξ). They then use the differences of the measured (noisy) jumps, $\Delta y_i = y_i - y_{i-1}$, to formulate the likelihood; consequently, the likelihood is determined by a convolution because the random variable $\Delta Y_i = \Delta X_i + \Delta E_i$ is the sum of the two random variables $\Delta X_i = X_i - X_{i-1}$ and $\Delta E_i = E_i - E_{i-1}$. In addition, calculating the difference of the errors leads to a dependence structure between the $\Delta \epsilon_i$. To carry out inference, Kallen and van Noortwijk (2005) use simulation to approximate the likelihood. D. Lu, Pandey, and Xie (2013) extended their work by developing a faster method for approximating the likelihood using the Genz transform and a quasi-Monte Carlo method. Their method also allows both of the parameters of the gamma process, β and ξ , in 4.2 to be estimated.

Building on the work of Kallen and van Noortwijk (2005) and D. Lu et al. (2013), Pulcini (2016) proposed a way to include degradation-dependent measurement error. Other researchers focused on improving computational efficiency by alternative methods such as deconvolution (Rodríguez-Picón et al., 2021) or by using faster algorithms to approximate the likelihood, for example, approximate Bayesian computing (Hazra, Bhadra, & Pandey, 2022; Hazra, Pandey, & Manzana, 2020). Common to all of these works, however, is a convolution-based likelihood based on a *marginal* model that requires the evaluation of, or approximations to, a complicated multidimensional integral. By contrast, hierarchical modelling based on *conditional* models provides a more straightforward, tractable, and flexible alternative when it is combined with an efficient inferen-

tial method. We describe hierarchical modelling in a Bayesian framework in the next section, but first note in passing that Giorgio, Mele, and Pulcini (2019) and Esposito, Mele, Castanier, and Giorgio (2022) also formulate a conditional likelihood to model a complex noisy gamma process and use maximum likelihood estimation combined with an EM algorithm and particle filtering for estimation and inference.

To demonstrate how a noisy GP can be postulated under the BHM framework described in Sec. 1.4, consider the noisy degradation trace in Fig. 4.1. Figure 4.1 shows a degradation trace generated from a stationary gamma process (the solid line) and the noisy observations of this degradation trace (the red points). Let, y_i refer to the measured, noisy, degradation data at time t_i , $i = 0, 1, 2, \dots, I$ and, using the same notation as in Section 4.1, $\{z_i\}$ refer to the values of the underlying gamma degradation path at times t_i .

To specify the Bayesian hierarchical model, in the data model, I assume that *given the value of the underlying gamma process*, the noisy observations are normally distributed and independent of each other; in other words, the y_i are *conditionally* independent. That is

$$y_i|z_i, \sigma \sim N(z_i, \sigma) \quad \text{data model}$$

where σ is the standard deviation of the Gaussian distribution. I then assume in the next level of the model that the underlying degradation, the z_i , follow a gamma degradation process. As a consequence of the independence of the increments and eq. 4.2, $z_i = \sum_{j=0}^i \Delta z_j$ has a gamma distribution given by $\text{Ga}(\beta t_i, \xi)$. Therefore, we write the process model as

$$z_i = \sum_{j=0}^i \Delta z_j$$

$$\Delta z_i|\beta, \xi \sim \text{Ga}(\beta \Delta t_i, \xi) \quad \text{process model}$$

In the final level of the hierarchy, I specify a distribution for the parameters β, ξ and σ , but for the moment, I write the distribution in its most general form, as the joint distribution

$$\beta, \xi, \sigma | \theta \sim \pi(\theta) \quad \text{parameter model}$$

where $\pi(\theta)$ represents the parameters of the joint distribution. In the next section, I show how a reparametrisation of the process model results in more interpretable parameters than the shape and rate, and I explain how this simplifies the last step of specifying the parameter model. Then, in Section 4.4 I use simulation to choose suitable distributions for these parameters.

4.3 Reparametrisation

The gamma process described in eq. 4.2 has density function

$$f(z_j; \beta t_i, \xi) = \frac{\xi^{\beta t_i}}{\Gamma(\beta)} e^{-\xi z} z^{\beta t_i - 1}, \quad (4.3)$$

and the mean and variance, which I denote by μ and σ^2 , are given by

$$\mu = \frac{\beta}{\xi} t_i \quad \text{and} \quad \sigma^2 = \frac{\beta}{\xi^2} t_i. \quad (4.4)$$

Both the average degradation rate and the variability of the gamma process depend on the parameters β and ξ . Hence, it is challenging to specify prior distributions of β and ξ so as to separate their effects on the stochastic process. From the perspective of the user, it is desirable to reparameterise the gamma process so that the new parameters have clear interpretations and effects. In addition, if they are *orthogonal* (Cox & Reid, 1987), there are several desirable statistical consequences for estimation, inference, and computation.

One such parameterisation is in terms of the mean μ and coefficient of variation $\nu = \sigma/\mu = 1/\sqrt{\beta}$: the mean represents the average degradation rate per

unit time, whereas the coefficient of variation describes the volatility of the degradation process; or how much heterogeneity there is in the wear rate over time. For the user, therefore, μ and ν have a more intuitive interpretation than the shape and the rate. Furthermore, using a result due to Huzurbazar (1956), it is straightforward to show that these parameters are also orthogonal. (We note in passing that orthogonal parameterisations are not unique; the mean μ and shape β are also orthogonal (Huzurbazar, 1956).)

Substituting μ and ν in the expression for the distribution of the increments in the process model in Section 4.1 yields

$$\Delta z_i | \mu, \nu \sim \text{Ga} \left(\frac{\Delta t_i}{\nu^2}, \frac{1}{\mu \nu^2} \right). \quad (4.5)$$

I use this reparameterisation in the remainder of this thesis. Kallen and van Noortwijk (2005) also use the shape and coefficient of variation, pointing out that it can be easier for a plant engineer to interpret them. They do not, however, exploit their orthogonality, preferring to fix the value of ν in their analysis instead of estimating it.

4.4 Constructing the prior

The prior distribution in the parameter model summarises our beliefs about the parameters. There are two ways this information is encoded: the choice of distribution and the values of the hyperparameters. Before the advent of contemporary sampling algorithms, Bayesian analysis relied on conjugate prior distributions, or convenient prior distributions that facilitated the use of Gibbs samplers or conventional Metropolis-Hastings algorithms (Gilks et al., 1996). However, with the development of more efficient sampling algorithms such as Hamiltonian Monte Carlo (M. Betancourt, 2017), we are no longer limited by such requirements and can select priors that reflect our state of knowledge, facilitate efficient computation, and that can be justified and evaluated in a principled way. In this section,

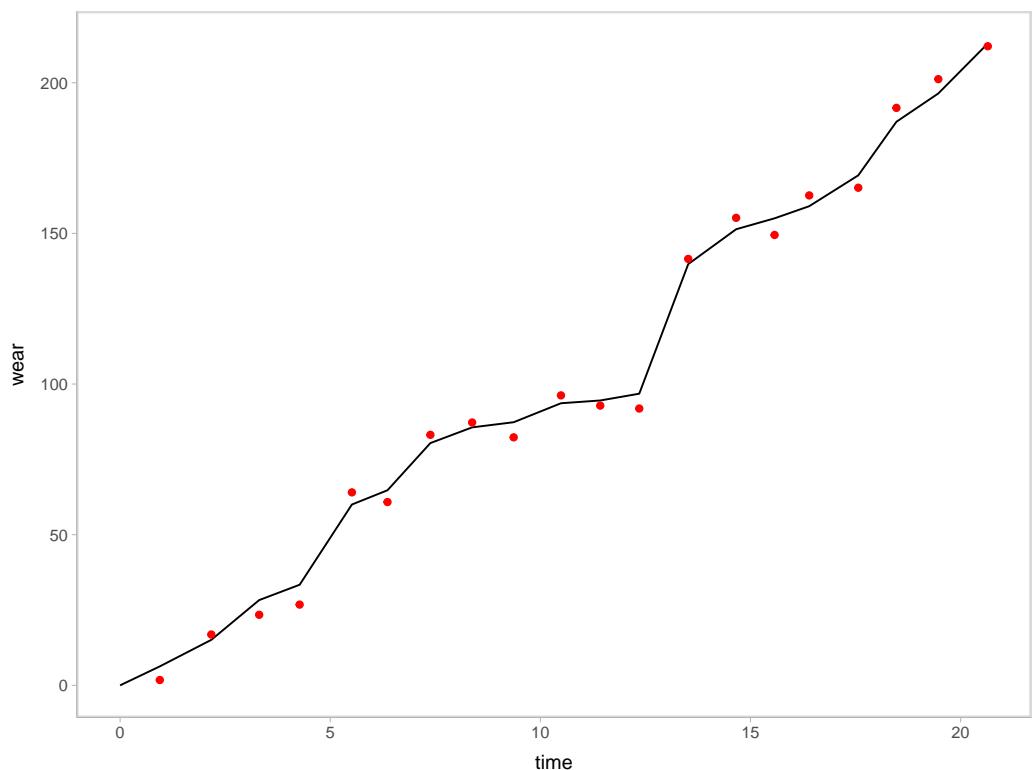


Figure 4.1: A simulated degradation trace generated from a gamma process with parameters $\mu = 10$ and $\nu = 1.119$ and the simulated noisy observations of the trace. The black line is the underlying degradation path and the points are the noisy observations.

I compare some commonly used ‘default’ priors for the GP from the reliability literature with a set of well-thought-out priors for the new parameters through prior predictive simulation: along the way, I provide justification for the new choice of priors for the alternative parameters μ and ν .

In the degradation modelling literature, a gamma distribution is often used as the prior distribution for the rate parameter ξ of the gamma process (Lawless & Crowder, 2004) and also for the shape parameter (Rodríguez-Picón et al., 2018). It is well known that a gamma prior on the rate parameter is conditionally conjugate (Pradhan & Kundu, 2011), and its use leads to analytically tractable results, as Lawless and Crowder (2004) show. Nevertheless, little work has been done to assess whether other prior distributions might be more appropriate. The gamma distribution has a heavy tail, and its use can lead to MCMC chains that converge very slowly or that are highly autocorrelated; moreover, it can lead to physically implausible realisations of the gamma process, as we demonstrate below. Using the new parameterisation of the GP in terms of μ and ν , conditional conjugacy no longer exists, and so there is even less motivation for a gamma prior.

In Figure 4.2, I illustrate prior predictive checking of a noise-free gamma process using three sets of priors for its parameters: first, ‘conventional’ priors, $\text{Ga}(1, 0.001)$ and $\text{Ga}(0.001, 0.001)$, that are widely used in the literature for both the shape and rate parameters of the usual parameterisation of a noise-free gamma process in eq. (4.2), and second, priors on μ and ν in the alternative parameterisation of eq. (4.5) with carefully thought out weakly-informative priors. All three sets of priors yield an average degradation rate of 1 unit per unit time.

The distribution $\text{Ga}(\epsilon, \tilde{\epsilon})$, where $\epsilon, \tilde{\epsilon} \rightarrow 0$, is often used as a noninformative prior distribution, especially in mixed linear models, where it is a conditionally conjugate prior for the precision (Hodges, 2014, p. 33). In addition, as we pointed out above, the gamma distribution is conditionally conjugate for the rate parameter: if $\{z_i\}$, $i = 1, 2, \dots, n$, represents an independent sample from $\text{Ga}(\beta, \xi)$, then the conditional distribution of ξ given β and the data is $\text{Ga}(n\beta + \epsilon, \sum_{i=1}^n z_i + \tilde{\epsilon})$

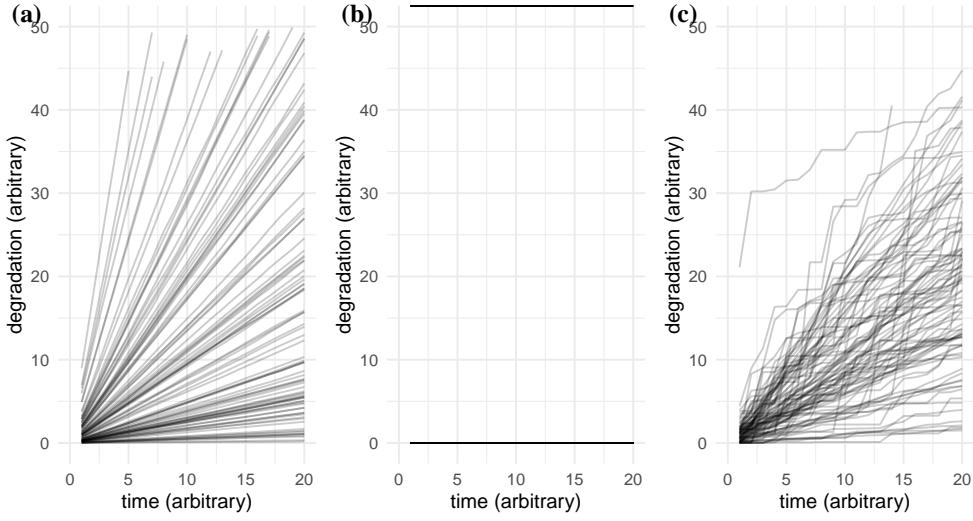


Figure 4.2: One hundred realizations from prior predictive distributions of a noise-free gamma process with mean degradation rate of 1 unit per unit time and different prior distributions for parameters: (a) shape, rate $\sim \text{Ga}(1, 0.001)$, (b) shape, rate $\sim \text{Ga}(0.001, 0.001)$, and (c) parameterization mean/coefficient of variation—see text for details.

when the prior distribution of ξ is $\text{Ga}(\epsilon, \tilde{\epsilon})$. Hence, when ϵ and $\tilde{\epsilon}$ are both small, the prior adds very little information, but it is noninformative *with respect to the rate parameter only*; furthermore, inferences about ξ may be sensitive to the values of ϵ and $\tilde{\epsilon}$ in data sets where small values of ξ may be possible (Gelman, Vehtari, et al., 2020, p. 130). When $\text{Ga}(\epsilon, \tilde{\epsilon})$ is used for *both* parameters, we can no longer assume that the joint prior will be noninformative and, therefore must evaluate it to determine whether it is indeed diffuse. For further discussion on the consequences of using $\text{Ga}(\epsilon, \tilde{\epsilon})$ as a prior distribution and guidance on using more sensible alternatives, see Hedges (2014) and Gelman, Vehtari, et al. (2020).

Figure 4.2 (a) and (b) show 100 draws from the prior predictive distribution of a noise-free gamma process when both the shape and rate parameters are assigned the prior distribution $\text{Ga}(1, 0.001)$ (Fig. 4.2 (a)) or $\text{Ga}(0.001, 0.001)$ (Fig. 4.2 (b)). In Fig. 4.2 (a), we can clearly see that the degradation traces resulting from a $\text{Ga}(1, 0.001)$ prior distribution are all nearly linear, without the jumps expected of gamma processes; furthermore, many of the rates of degradation are unrealistically high and unrealistically low. In Fig. 4.2 (b), where a $\text{Ga}(0.001, 0.001)$ prior

is used, most of the prior predictive distribution has mass around implausibly low values of the average rate, and there is one unrealistically steep degradation trace. As we pointed out earlier, the gamma distribution is highly skewed and has heavy tails; consequently, depending on the values of the shape and rate, the prior can place mass on high, low, or both high and low values, resulting in simulated data that simply could not be observed in practice. By contrast, prior simulations generated according to the weekly-informative priors constructed with respect to the alternative parameters μ and ν in Fig. 4.2 (b) look much more plausible.

To specify independent prior distributions of the parameters μ and ν in the GP model in (4.5), we adopt the approach introduced by Simpson, Rue, Riebler, Martins, and Sørbye (2017): design priors that favour simpler models over more complex ones and that are consistent with domain knowledge. The mean μ controls the average degradation rate, similar to the action of the slope parameter in a linear degradation path model. We have no reason to believe that the variability about the mean degradation rate would be asymmetric, so a Gaussian distribution with a small standard deviation is both appropriate and convenient

$$\mu \sim N(1, 0.5).$$

The coefficient of variation ν is a measure of the volatility of the degradation process, and although we might expect some heterogeneity in the wear rate as degradation progresses, we do not expect the wear rate to be extremely volatile. Hence, we use a truncated Student t -distribution with 3 degrees of freedom as a prior for ν ;

$$\nu \sim t_3^+(0, 1),$$

where the superscript (+) denotes a truncated distribution whose lower bound is zero; furthermore, I use the location-scale form of a t distribution with n degrees of freedom, written as $t_n(\text{location}, \text{scale})$. This prior places a large mass near zero but still allows the posterior distribution to move away from zero. In addition,

it has lighter tails than a gamma distribution and consequently does not give too much weight to extremely volatile degradation paths. Figure 4.2 (c) shows 100 draws from this parameter model for the reparameterised gamma process. The degradation traces have the appearance of paths expected from a gamma process; that is, there are discrete jumps between time points, in contrast to Fig. 4.2 (a), where all the traces are straight lines. Furthermore, more than half the degradation values at the end (eleventh time point) are between 6 and 16, as would be expected when the degradation varies around one. Finally, although there are some extreme realisations, there are only one or two that are completely implausible.

To fully specify the model, I also need to specify a prior for the standard deviation of the measurement error, σ . Following the recommendations of Gelman, Vehtari, et al. (2020, Chapter 17), I use a vague Uniform(0, A) prior for σ , where A is chosen to be large relative to the expected scale of σ . We use such a vague prior for demonstration purposes. However, in practice, an analyst should have a reasonable grasp of the scale of the measurement error and should be able to specify a weakly informative prior; we do exactly this in Section 4.5.4. Because our initial prior on σ is so vague, we do not include the measurement error in the prior predictive checking in Fig. 4.2.

Of course, had I used different values of the hyperparameters for the N and t_3^+ priors, the appearance of the degradation traces in Figure 4.2 (c) would have been different. For example, if we had specified a much more diffuse prior for the mean wear rate, then no doubt unrealistically fast and slow wearing degradation traces would have been generated. It is only through the use of prior predictive checking to choose sensible values of the hyper-parameters in conjunction with suitable distributional forms of the priors that a well-justified prior is obtained.

4.5 Fitting the noisy gamma process

To improve our understanding of the noisy GP model, I fit the Bayesian hierarchical model outlined in Sec. 4.2 with the priors defined in Sec. 4.4 to the single simulated degradation trace in Fig. 4.1 as well as to a subset of the simulated degradation measurements. For clarity, the full model is

$$\begin{aligned}
 y_i | z_i, \sigma &\sim N(z_i, \sigma) && \text{data model} \\
 z_i &= \sum_{j=0}^i \Delta z_j && \text{process model} \\
 \Delta z_i | \mu, \nu &\sim Ga\left(\frac{\Delta t_i}{\nu^2}, \frac{1}{\mu\nu^2}\right) \\
 \mu &\sim N^+(10, 10) && \text{parameter model} \\
 \nu &\sim t_2^+(0, 1) \\
 \sigma &\sim \text{Unif}(0, 100).
 \end{aligned}$$

The single path example shows that the noisy GP is more difficult to fit than a noise-free GP model. When the sample size is small, the model struggles to separate the parameters describing the variance of the measurement error and the volatility of the underlying gamma process because there is not enough information in the data to do so. To demonstrate this problem with identifiability, I fit the BHM of two data sets: one ‘large’ data set consisting of all 20 simulated noisy degradation measurements in Fig. 4.1, and another ‘small’ data set that is a subset of 10 points. I fit the BHM of the noisy GP outlined in Sections 4.2 and 4.4 to these two data sets and evaluate how well the true parameter values and underlying degradation path is reclaimed in the two resulting posterior distributions. I also investigate the efficiency of the No-U-Turn sampler for the two cases.

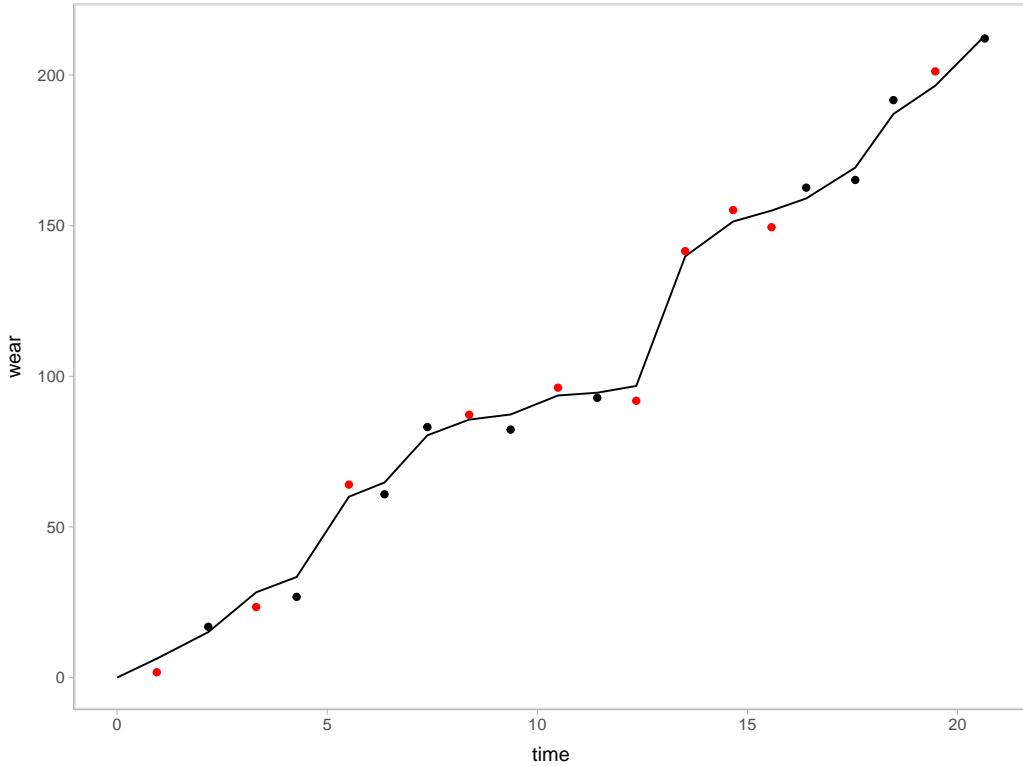


Figure 4.3: The simulated degradation trace with the subset of observations selected for the small dataset highlighted in Red.

4.5.1 Data simulation

To generate the degradation trace in Fig. 4.1—which I coin the ‘large’ dataset—by sampling twenty time increments from a $\text{Unif}(0.8, 1.3)$ distribution. Then next, I sampled twenty jumps in degradation from $\text{Ga}(\Delta t_i/\nu^2, 1/\mu\nu^2)$, using $\mu = 10$ and $\nu = 1.119$. I then calculated the cumulative sum of the jumps to obtain the underlying, noise-free degradation trace z_i , where $z_0 = 0$ at $t_0 = 0$. Finally, I add Gaussian noise with standard deviation $\sigma = 4$ to the underlying degradation path to get the noisy observations. The big dataset is described in Table 4.1. To create the second smaller data set, I randomly select ten of the twenty noisy observations. The degradation observations selected for the small data set are highlighted in Fig. 4.3, and the small data set is displayed in Table 4.2.

Table 4.1: The twenty simulated noisy degradation observation that make up the ‘big’ data set.

t	Δz	z	y
0.00	NA	0.00	NA
0.95	6.29	6.29	1.77
2.17	8.81	15.10	16.87
3.31	13.20	28.30	23.42
4.27	5.09	33.39	26.80
5.52	26.63	60.02	64.04
6.36	4.73	64.75	60.84
7.38	15.66	80.41	83.15
8.38	5.21	85.62	87.24
9.36	1.71	87.33	82.31
10.49	6.30	93.63	96.23
11.43	0.92	94.54	92.85
12.35	2.25	96.79	91.88
13.52	43.01	139.80	141.53
14.66	11.57	151.37	155.16
15.57	3.63	155.00	149.48
16.40	4.02	159.02	162.60
17.56	10.23	169.26	165.16
18.47	17.79	187.05	191.65
19.47	9.40	196.44	201.20
20.65	16.57	213.01	212.11

Table 4.2: The subset of ten simulated noisy degradation observation from the ‘big’ data set which make up the ‘small’ data set.

t	Δz	z	y
0.00	NA	0.00	NA
0.95	6.29	6.29	1.77
3.31	13.20	28.30	23.42
5.52	26.63	60.02	64.04
8.38	5.21	85.62	87.24
10.49	6.30	93.63	96.23
12.35	2.25	96.79	91.88
13.52	43.01	139.80	141.53
14.66	11.57	151.37	155.16
15.57	3.63	155.00	149.48
19.47	9.40	196.44	201.20

4.5.2 Computation

To sample from the posteriors of the noisy GP model conditioned on the two different datasets I use the No-U-Turn sampler implemented in the probabilistic programming language *Stan* (Stan Development Team, 2022). I generated 88,000 samples from each posterior distribution using four chains of 25,000 iterations each, with a burn-in of 3,000 and no thinning. To ensure a detailed exploration of the posterior, I also change the sampling parameters *adapt delta* and *maximum tree depth* to 0.99 and 13 respectively. Raising *adapt delta* results in a more aggressive (smaller) choice of the step size used for the leapfrog algorithm that approximates the Hamiltonian trajectories, and raising the *maximum tree depth* allows each leapfrog algorithm to run for longer. Increasing these two sampling parameters results in a slower sampler but ensures a more detailed exploration of the posterior. All of the code to define the model in Stan, simulate the data in R, and sample from the posterior using RStan is available on a GitHub repository.

During sampling—despite increasing *adapt delta* and *maximum tree depth*—80 divergent transitions occur while fitting the model to the small data set, whereas only four occur when sampling from the posterior conditioned on the larger data

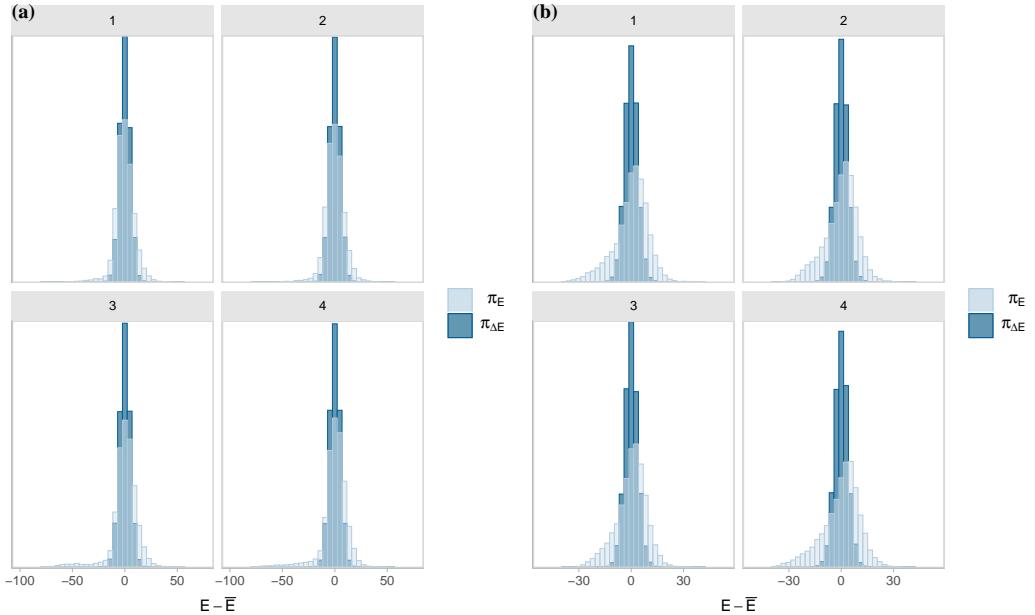


Figure 4.4: The cain energy diagnostics for each chain when sampling from the posterior of the big data set (left) and small data set. Each plot compares the marginal energy distribution, π_E , with the first differenced distribution, $\pi_{\Delta E}$.

set. The divergencies signify incomplete exploration of the target distribution, which I explore in the next section. In addition to evaluating sampling through divergent transitions, energy diagnostics quantify the heaviness of the tails of the posterior distribution and thus can identify inefficient sampling (Gabry & Mahr, 2024). The chain energy plots in Fig. 4.4 compare the marginal energy distribution, π_E , with the first differenced distribution, $\pi_{\Delta E}$, for each chain. These plots are similar to those in M. Betancourt (2017) that compare the energy transition distribution (equivalent to $\pi_{\Delta E}$) with the marginal energy distribution (equivalent to π_E). Ideally, these two overlaid distributions should look the same. However, if the distribution of $\pi_{\Delta E}$ is much narrower than that of π_E , then this indicates slow exploration of the target distribution. The energy diagnostics from fitting the model to the small data set in Fig. 4.4 (b) show that exploration of the posterior is very inefficient, whereas when fitting the model to the big dataset, Fig. 4.4 (a), sampling is much more efficient; although it is not perfect.

4.5.3 Results and diagnostics

My objective here is to investigate how the size of the dataset affects inference from the BHM for a noisy GP. To do so, I assess how well the parameters and underlying degradation trace are reclaimed in the two posterior distributions. Visualising the two posteriors shows that when the model is fitted to all twenty degradation observations, it is able to recover the parameter values and underlying degradation path; when only a subset of ten noisy observations is used, the model fails to do so because it is unable to disentangle the observation noise from the volatility of the gamma process. I come to this conclusion by exploring the degenerate behaviours in the posterior that are flagged by divergent trajectories that occur during sampling.

Marginal densities Figure 4.5 shows the marginal distributions of the parameters μ , ν , and σ conditioned on the small and big data sets as well as the true values of the parameters. For each marginal density, the median and 66% and 95% credible intervals are shown. It is clear that when the model is fit to the small data set, it fails to reclaim the true parameter values, but when it is fit to the bigger dataset, it successfully reclaims the true values. The marginal posterior densities of the parameters conditioned on all twenty degradation observations are centred around the true values of the parameters, whereas the marginal posteriors of σ and ν conditioned on the subset of ten observations are centred around fifteen and zero, respectively. Furthermore, the marginal posterior of σ conditioned on the smaller subset of the data appears to have some multimodality. To understand the implication of these parameter estimates as well as the effect of how they covary with one another in the posterior I look at their joint effect on the outcome variables, which in this case is the predictive distribution of the filtered degradation path.

Posterior predictive density Figure 4.6 shows the posterior predictive distribution of the underlying degradation trace, the z_i , for the two posteriors with

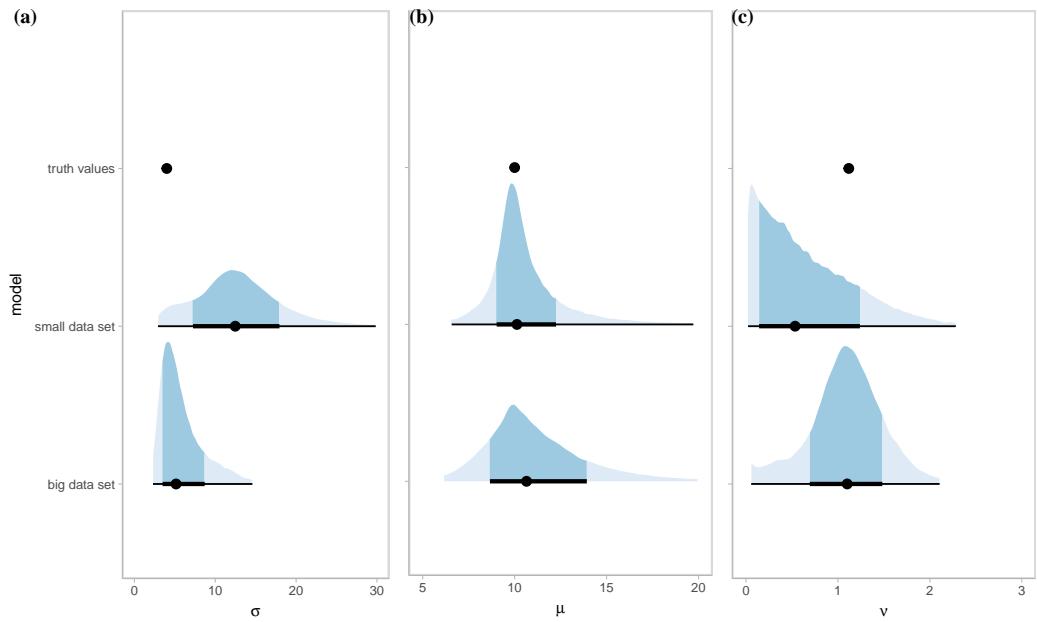


Figure 4.5: The marginal posterior distributions of the parameters σ , μ , and ν when the BHM for the noisy gamma process is fitted to the ‘large’ and ‘small’ simulated data in Fig. 4.1. The points and intervals shown in each distribution represent, respectively, the median and 95% and 66% credible intervals. The values used to simulate the data are shown in the top row.

the true degradation trace and noisy observations overlaid. The thick grey line in each plot is the median of the posterior predictive distribution; additional quantiles are shown in different shades of blue. Clearly, in Figure 4.6 (a), the model has been able to reclaim the underlying degradation from the noisy degradation observations when fit to all twenty observations: the median path follows the actual path almost exactly, with uncertainty bands that are narrow enough to be useful. However, as was the case with parameter values, the median path derived from the posterior distribution conditioned on the subset of the data has not recovered the true path (Fig. 4.6 (b)). In Figure 4.6 (b), the median path is a nearly straight line through the data points. In addition, the uncertainty intervals are much wider.

Pairs plots Clearly, some issues are occurring in the posterior distribution of the model conditioned on the smaller subset of the data. These issues were pre-

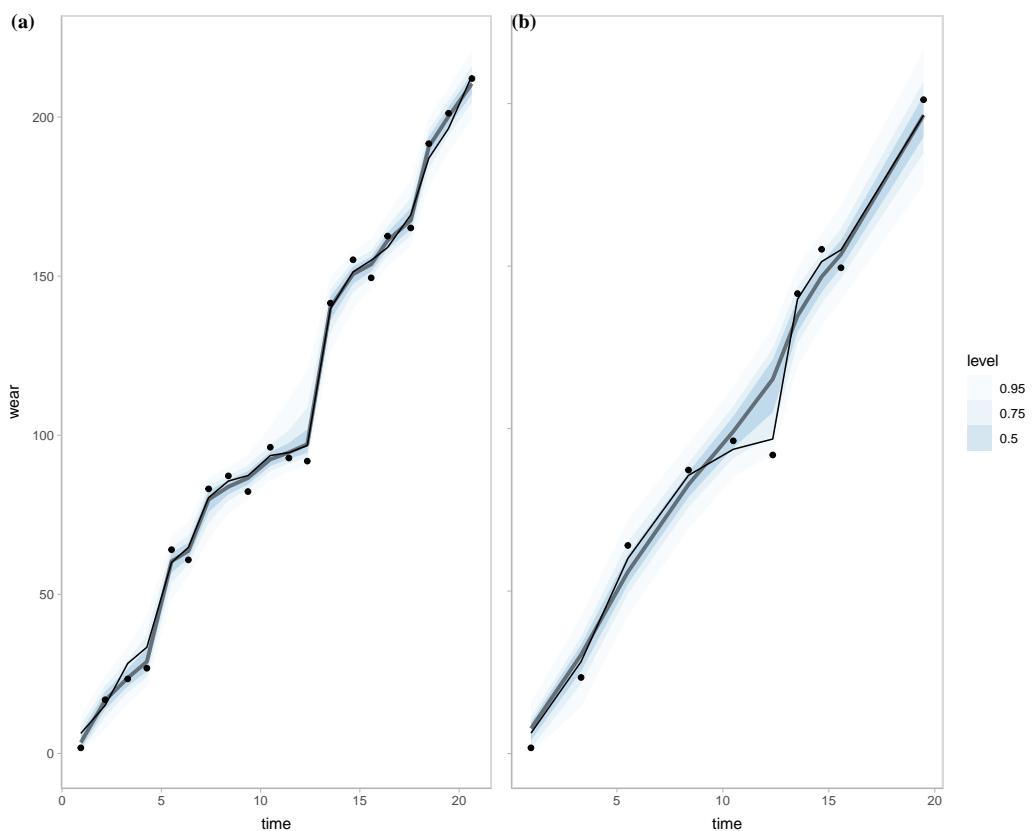


Figure 4.6: The posterior predictive distribution of the filtered degradation path compared to the true degradation path when the model is fitted using the big (left) and small (right) data set. The median of the predictive distribution is shown as a grey line and the corresponding uncertainty intervals as blue ribbons.

eluded to by the divergent trajectories that occurred during sampling (Sec. 4.7). In the case of fitting the model to simulated data, where we can be sure that the model is properly specified and implemented, poorly behaved sampling is often a sign of a deeper issue with the model. As discussed in Sec. 1.4, the divergent transitions can point to the problematic areas in the posterior. Figure 4.7 shows a pairs plot of the parameters μ , ν , and σ and the first degradation jump, Δz_1 . The divergent trajectories are shown in red. In the bivariate scatter plots, there are strong funnel shapes between μ and $\log(\nu)$ and between $\log(\nu)$ and the first degradation jump. The divergent trajectories are concentrated at the entrance to these funnels, suggesting that they are the cause of the sampling issues. The funnel shapes occur because as ν shrinks towards zero μ and Δz_1 approach very particular values; $\mu = 10$ and $\Delta z_1 = 10 \times \Delta t_1$.

Parallel coordinate plot The divergent trajectories can help to further explore how the degenerate behaviour manifests in the multidimensional posterior. Figure 4.8 shows a parallel coordinate plot of the posterior draws for all of the parameters in the model. The divergent trajectories are highlighted in red. The divergencies draw a clear structure through parameter space. They all pass through the values $\mu = 10$, $\nu = 0$, and $\Delta z_i = 10 \times \Delta t_i$ and have an inflated value of σ that is much larger than the true value $\sigma = 4$. This structure equates to a linear degradation trace with large uncertainty. To emphasise this point, in Fig. 4.9, I plot the posterior predictive distribution of the underlying degradation path from the small data set and overlay the divergent trajectories. From this, it is clear that the areas of tight curvature in the posterior occur around the models where the degradation trace is effectively linear.

Comparison of the two posteriors In comparison, this degenerate behaviour in the posterior is almost completely washed out by extra information in the large dataset. Figure 4.10 shows the joint distributions of the intermediate quantities Δz_{15} from the model conditioned on the big dataset and Δz_9 for the small with

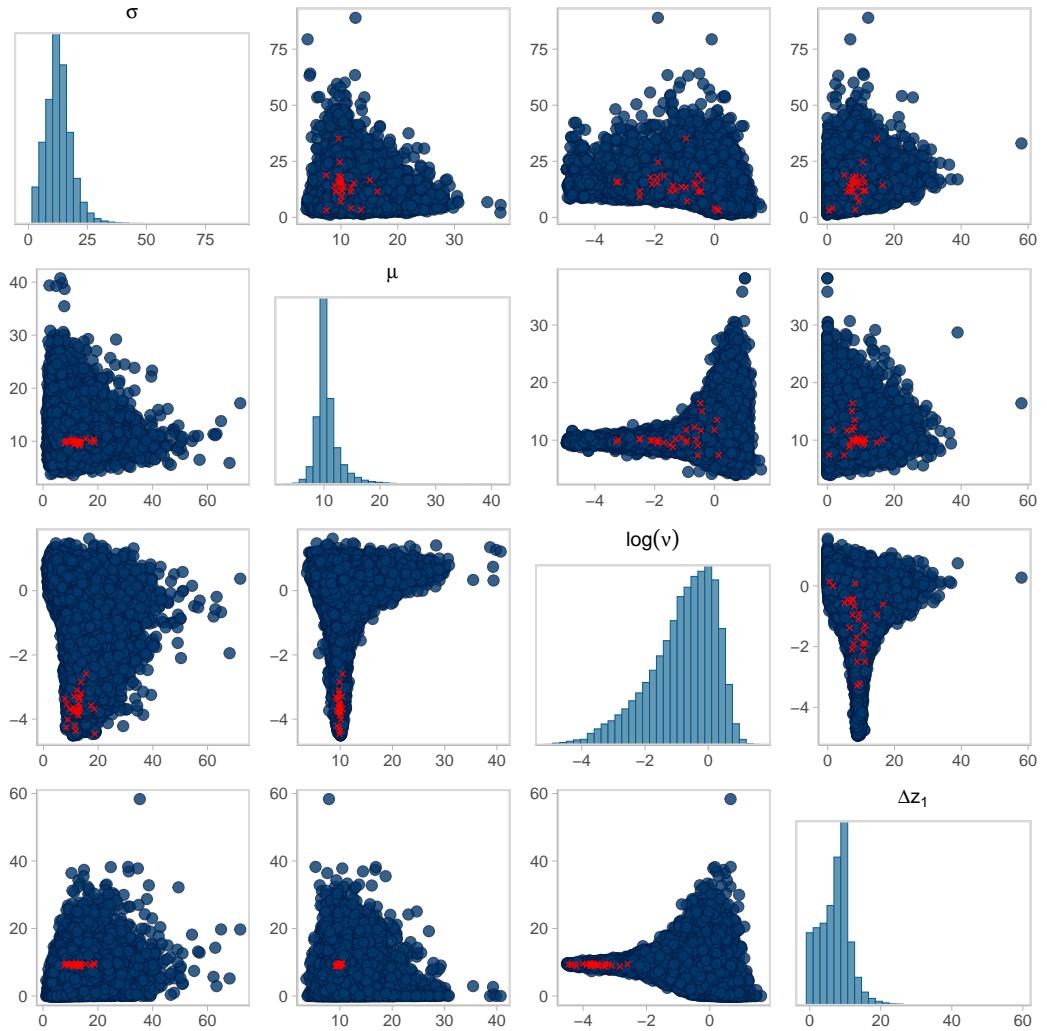


Figure 4.7: Pairs plot showing the MCMC draws from the posterior distribution of the parameters σ , μ , $\log \nu$, and the filtered value Δz_1 when the BHM of the noisy gamma process is fitted to the small dataset. The red points indicate divergences, which congregate at the end of the funnel in the pairwise plots of $\log \nu/\Delta z_1$ and $\log \nu/\mu$.

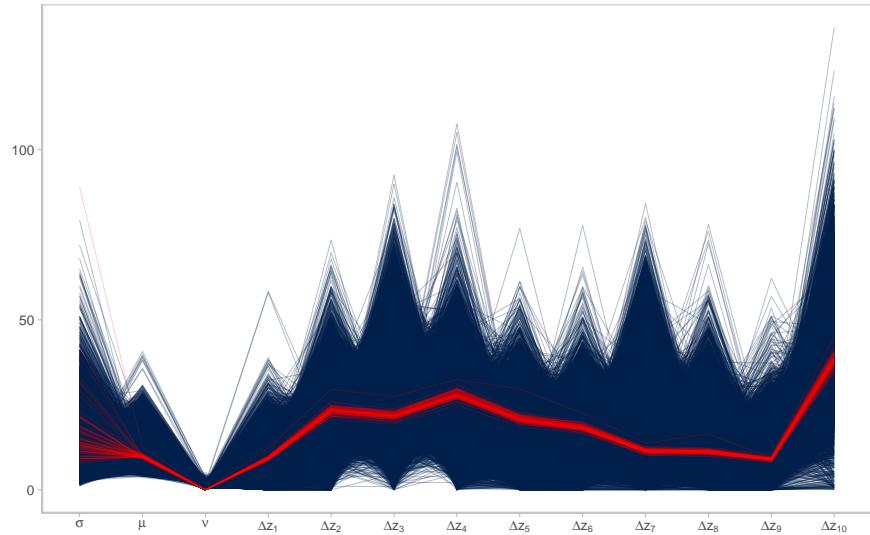


Figure 4.8: The parallel coordinate plot of the draws from the posterior conditioned on the small dataset with the divergent traces plotted in red.

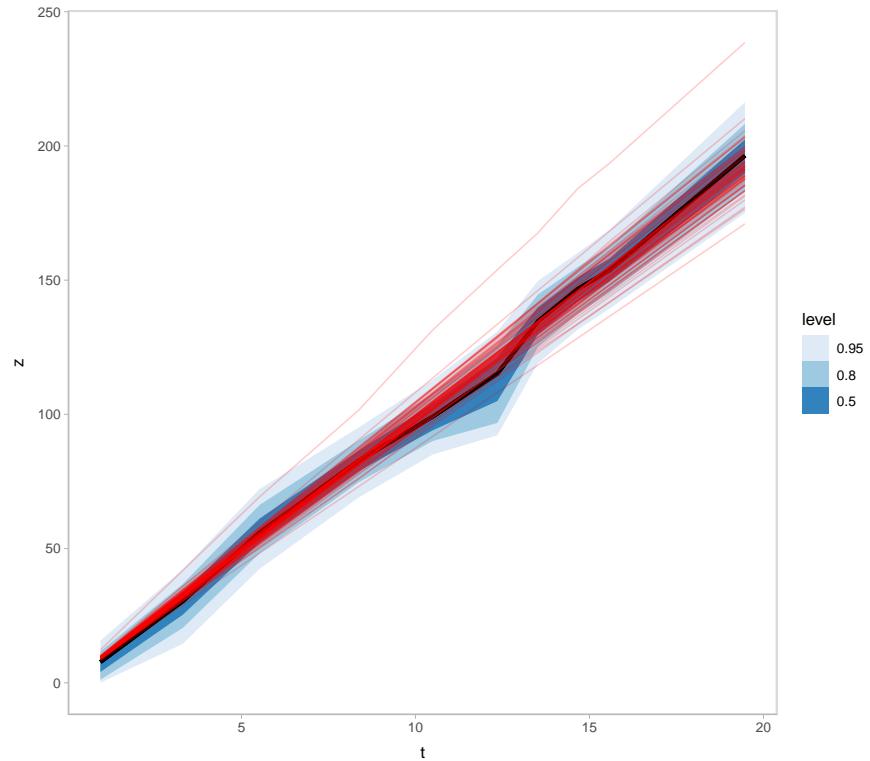


Figure 4.9: The divergent traces overlaid (in red) on the posterior predictive distribution of the degradation path for the small data set.

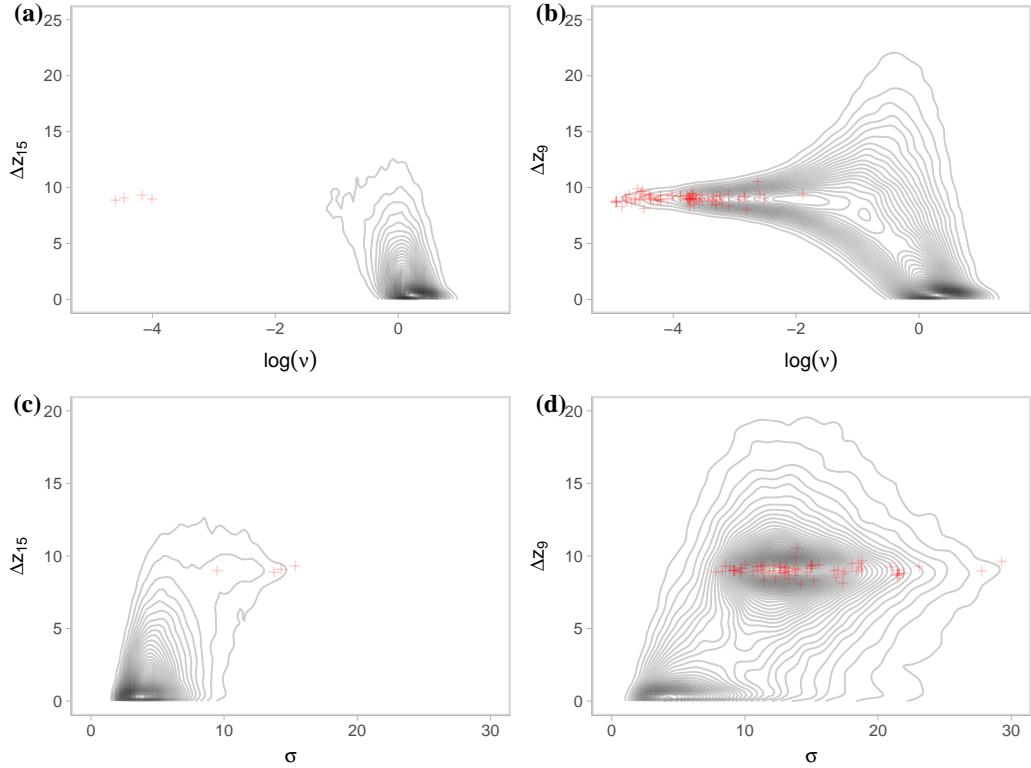


Figure 4.10: .

$\log(\nu)$ and σ . In the two datasets, Δz_{15} and Δz_9 are the same jump in degradation. In the joint posterior of Δz_9 and $\log(\nu)$, there is the deep funnel shape around $\Delta z_9 = 9$ and $\log(\nu) = -\infty$, and there is a second mode in the joint distribution of Δz_9 and σ . However, in the joint distribution of Δz_{15} and $\log(\nu)$ there is very little mass around $\Delta z_{15} = 9$ and no second mode in the joint distribution of Δz_{15} and σ . This leads to the conclusion that nonidentifiability only exists when there are few observations.

4.5.4 Solutions to computational issues

The identifiability issue for the small data set can also be solved by injecting more information into the analysis that helps to disentangle σ and ν . This information can come in the form of either supplementary data or prior information that informs one of the nonidentifiable parameters. Getting extra information about

the measurement error would typically be much easier than the coefficient of variation of the gamma process. Here, I show that adding a small amount of supplementary information using either supplementary data or a stronger prior helps to identify σ and, therefore, ν , resulting in much smoother geometries in the posterior and, therefore, much more efficient sampling. The results on the inference are arguably better than when I fit the model to all twenty degradation observations.

Prior information In Section 4.5.3, I have used an effectively non-informative prior for the standard deviation of the measurement error. Typically, a technician would have some understanding of the variability in the measurement process. To emulate this, I place a Gaussian prior on the standard deviation of the measurement error

$$\sigma \sim N(4, 1).$$

This prior is centred around the true value of σ and places 95% of the mass between $\sigma = 2$ and $\sigma = 6$. Sampling from the posterior of this model with this stronger prior conditioned on the small dataset is much quicker, and no divergent transitions occur. Figure 4.11 shows the chain energies of the sampler when a stronger prior is used on sigma. The marginal energy distribution and the first differenced distribution now match closely, showing that the chains have efficiently explored the posterior.

The pairs plot of the MCMC samples in Fig. 4.12 look much smoother, and there is little remanence of the deep funnel-shaped degeneracies between $\log \nu$ and μ and Δz_1 . In Figure 4.13, I compare the marginal distribution for σ , μ , and ν from this posterior with the true values and the models fit in Sec. 4.5.3. The marginal distributions for the model with the stronger prior are much smoother, and there is now no mass around zero in the posterior of ν .

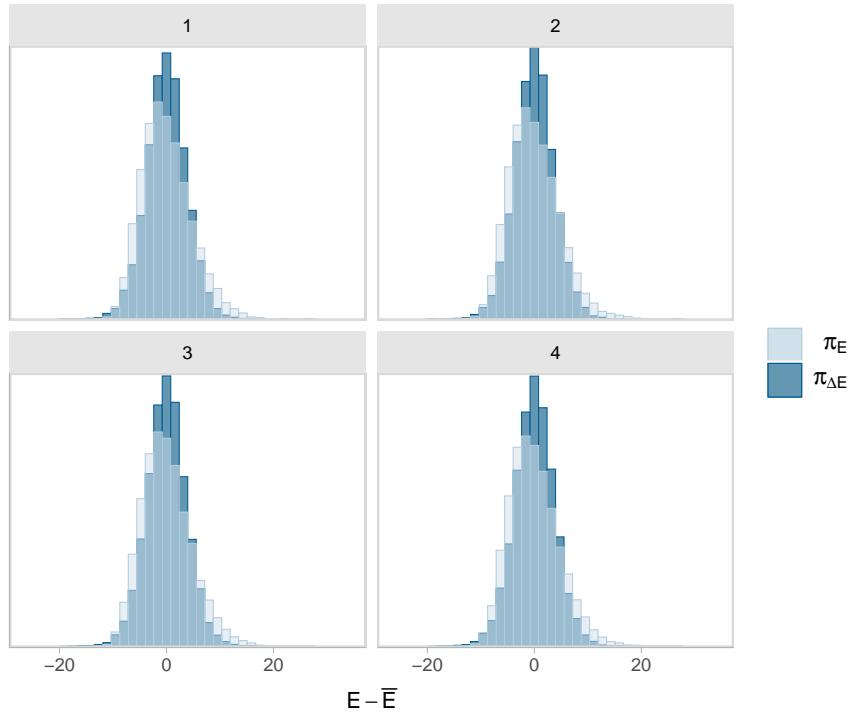


Figure 4.11: The chain energy diagnostics when the model is fit with a stronger prior on σ .

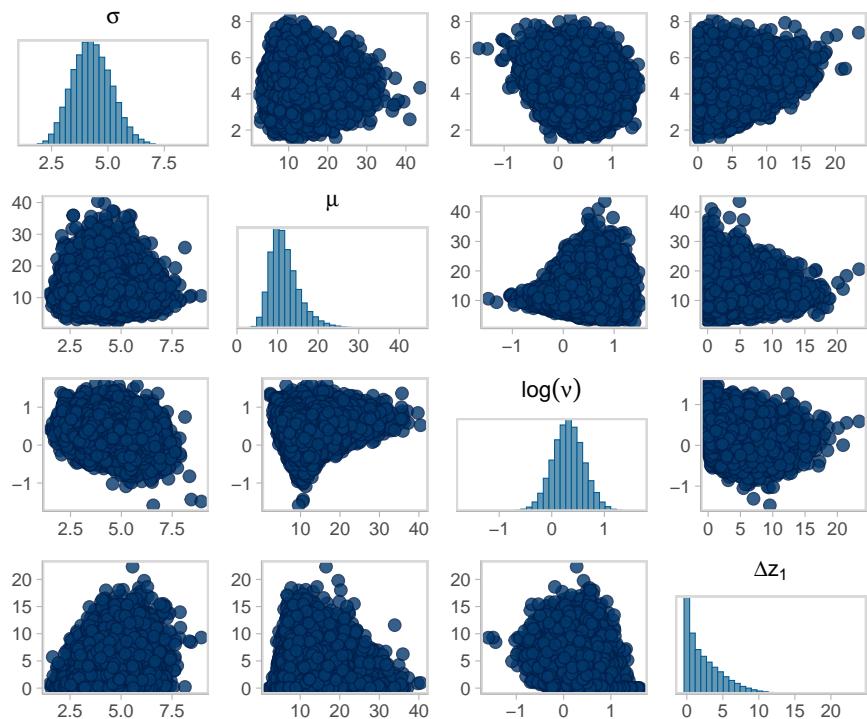


Figure 4.12: The pairs plot of the MCMC draws when using a stronger prior on σ .

Supplementary data Alternative to adding a stronger prior, I sample five supplementary observations of the measurement error from the distribution

$$y_{\text{sup}} \sim N(0, 4).$$

Extra supplementary observations such as these could be obtained by taking multiple measurements at time $t = 0$, when the degradation is known to be zero, just before decommissioning the component—after which detailed non-noisy measurements can be obtained—or by performing a small experiment. The supplementary observations can be easily incorporated into the Hierarchical model through the data model

$$\begin{aligned} y_i | z_i, \sigma &\sim N(z_i, \sigma) && \text{data model} \\ y_{\text{sup}} &\sim N(0, \sigma). \end{aligned}$$

Similar to the more informative prior, the sampler is much more efficient, no divergencies occur during sampling, and the bivariate posterior distributions in the pairs plots look much smoother. The resulting marginal posterior distributions of σ , μ , and ν for the supplementary data are also compared in Fig. 4.13. The marginal distributions of the parameters are very much the same as when a stronger prior is used, and the model successfully reclaims the true parameter values.

4.6 Discussion

The main focus of this chapter was to show that using the Bayesian hierarchical formalism allows us to frame a model for a noisy gamma stochastic process in a tractable and transparent manner. Decomposing the noisy gamma process into a sequence of conditional models—the data, process, and parameter models—removes the need for complex deconvolutions that require the evaluation of, or

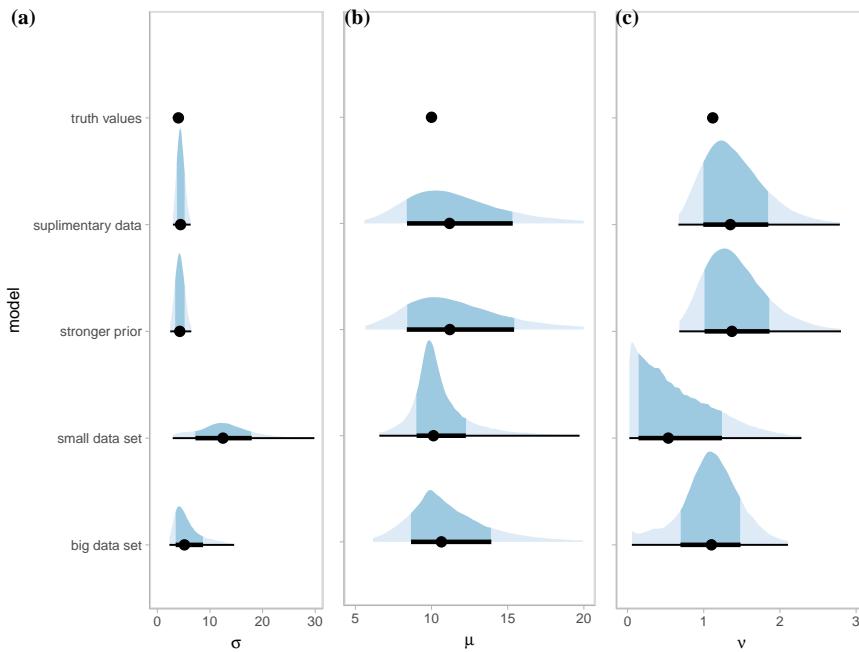


Figure 4.13: The marginal posterior distributions of the parameters σ , μ , and ν when information is added into the analysis either through a stronger prior or supplementary data compared to the original two posteriors.

approximations to, multidimensional integrals. Making this connection that allows the simple extension of the GP to noisy degradation observation and showing its implementation in the contemporary Bayesian computational environment Stan, is a step towards making stochastic degradation models applicable to use in industry and also accessible to practitioners. Below, I summarise the main element of this chapter and highlight the important findings and contributions, as well as areas for future work.

Reparameterising the gamma process in terms of the mean μ and coefficient of variation ν results in more interpretable parameters than the shape β and rate ξ : μ is the mean wear rate, and ν is the inverse of the ‘signal-to-noise’ ratio, and hence is a measure of the volatility of the gamma process. The interpretability of μ and ν simplifies specifying prior distributions because they are easier to elicit domain information about. Additionally, reparameterising the gamma process in this way also helps clarify how extensions of the model, such as unit-to-unit variability (which I show in the next chapter) or covariates, can be incorporated into

the model. Finally, the parameters μ and ν are orthogonal, which has desirable computational benefits.

Under this new parameterisation, I re-assessed good choice of priors and shown a principled way of assessing these priors. I draw away from the conventional gamma priors used for the gamma process and think deeply about the justification of priors to construct a weekly informative set of prior distributions that conform with my understanding of the underlying data-generating process. I then evaluate if the priors are, in fact, weekly informative through prior predictive checking. This is particularly important in the case of a noisy gamma process, since the noisy observation of the degradation trace means that the data do not strongly inform the underlying degradation model and in such cases using non-informative prior distributions can put large amounts of mass in unrealistic parts of parameter space (Tian et al., 2024).

In fitting the noisy gamma process to simulated data, I identify issues with sampling from the posterior when there are only a few observations. Investigating the poor sampling uncovers an identifiability issue between the volatility of the gamma process (expressed by ν) and the measurement error (σ), which is only present when the sample size is small. The observed degenerate behaviour of inference from the small-data posterior results from what M. Betancourt (2020) refers to as ‘pre-asymptotic non-identifiability’.

Because variation in the noisy degradation signal can be a result of both the randomness of jumps of the GP and the randomness of the measurement error, it can be difficult to separate these two sources when there is only a small number of observations. In the small dataset, the data do not strongly inform the parameters σ and ν , and it is therefore difficult to distinguish between competing models—the noisy gamma process and one where ν approaches zero—as reflected in their multi-modal posterior distributions. Using the terminology of M. Betancourt (2020), we can say that these parameters are pre-asymptotically non-identifiable.

I further confirm this in Section 4.5.4 by showing that the computational is-

sues and pathological behaviour in the posterior are completely resolved by adding additional information that specifically informs one of the pre-asymptotically non-identifiable parameters. Although fitting the model to all twenty noisy observations results in a much better behaved posterior, there is still remanence of the degenerate areas seen in the posterior conditioned on only ten observations; in addition, there are still a few divergences and the chain energy plots (Fig. 4.4) show that sampling is still slow. In contrast, when I use a stronger prior for σ or add a small amount of supplementary data that informs σ , there is no sign of degenerate behaviour in the posteriors and sampling becomes very efficient. With enough noisy observations, the model eventually becomes identifiable from the noisy degradation data alone. However, in a typical reliability application, the analysis will have small sample sizes, in which case adding additional information to the analysis can help to identify the model.

The issue of pre-asymptotically non-identifiability is not unique to the noisy gamma process. In an early paper on noisy Wiener processes, Whitmore (1995) also remarked on the difficulty in estimating the measurement error variance of a noisy Wiener process. In a Bayesian reanalysis of the same data, Hamada, Wilson, Reese, and Martz (2008b) imposed strong prior distributions on the measurement error variance and the variance of the Wiener process in order to ensure identifiability, although they do not explicitly justify their reasons for doing so. More work should be done to understand the interplay between the scale of the measurement error, the volatility of the underlying stochastic degradation process, and these small sample identifiability issues.

In the context of real noisy degradation data, there is no way of checking if there is enough information in the data to properly identify the model. Therefore, practitioners applying the noisy GP model should use as much information as they have available to them. This includes encoding their domain-expert knowledge into the prior distributions of the parameters, rather than choosing a default non-informative prior; incorporating supplementary data that informs one of the

pre-asymptotically non-identifiable parameters; and, if available, modelling the degradation of groups of similar units jointly as to ‘borrow’ information. With respect to the latter, in the next chapter, I extend the noisy gamma process for a single degradation trace to model noisy degradation paths from multiple units while assuming that the measurement error is the same for all units. Doing so drastically improves the problems with identifiability and MCMC sampling that we have seen in this chapter.

Chapter 5

Noisy gamma process with unit-to-unit variability

In Chapter 4, I discussed how to model a single degradation path using a noisy gamma process postulated through the Bayesian hierarchical modelling framework. I concluded that there is an identifiability issue between the noise and the volatility of the underlying gamma process when there are only a few degradation measurements. A resolution to this preasymptotic nonidentifiability is to add extra information into the analysis of the degradation trace, which can be done by modelling the degradation of a population of m nominally identical units simultaneously. This raises the question of how the degradation traces of each unit are related to one another. For example, we may assume that all of the units are realisations from the same underlying gamma degradation process.

However, this assumption may be too restrictive in practice since there may be additional variability in their degradation resulting from slight variations in operating conditions or their manufacture. The most common approach to modelling this extra layer of heterogeneity between units beyond what can be explained by the volatility of the gamma degradation process and any covariates is to use a ‘mixed effects’ model, in which some of the parameters of the model—so-called ‘random effects’—vary between units or individuals, whereas others, the ‘fixed

effects', do not¹. Early examples in the degradation literature include C. J. Lu and Meeker (1993) and Lawless and Crowder (2004), who incorporated random effects into a general path model and gamma process, respectively. A more recent example is Rodríguez-Picón et al. (2018), who modelled the GaAs laser dataset that was also analysed by Meeker and Escobar (1998). To model the heterogeneity in the degradation paths, Rodríguez-Picón et al. (2018) incorporate random effects into a noise-free gamma process by specifying the effect in either the mean or variance of the gamma process. By contrast, Peng and Ai (2018) follow the methodology of Lawless and Crowder (2004) and specify random effects in the scale parameter of a gamma process.

In this chapter, I show how the hierarchical model for the noisy gamma process can be neatly extended to incorporate unit-to-unit variability through the same BHM formalism and show the advantages of using the mean/coefficient of variation parameterisation in this context. Before going any further, however, it is worth clarifying the terminology that I use. As pointed out above, the terms random and fixed effects are used when mixed effects models are used to describe unit-to-unit variability. However, as Gelman (2005) and Gelman and Hill (2006) point out, all parameters in a Bayesian analysis are random variables; furthermore, because there is a multiplicity of definitions of fixed and random, such terms can engender considerable confusion [Section 6](Gelman, 2005). Consequently, Gelman (2005) and Gelman and Hill (2006) make a plea for abandoning these long-used terms in place of more descriptive ones: *varying*, for parameters that differ between groups or units, and *constant*, for parameters that are identical for all groups or units. In this chapter, I simply identify which parameters are common across units, those that are unique to each unit, and, most importantly, the specification of the prior distribution(s) for parameters that vary from unit-to-unit.

To demonstrate the models for multiple units, I use a data set from an exper-

¹This definition is just one of five that Gelman (2005) lists.

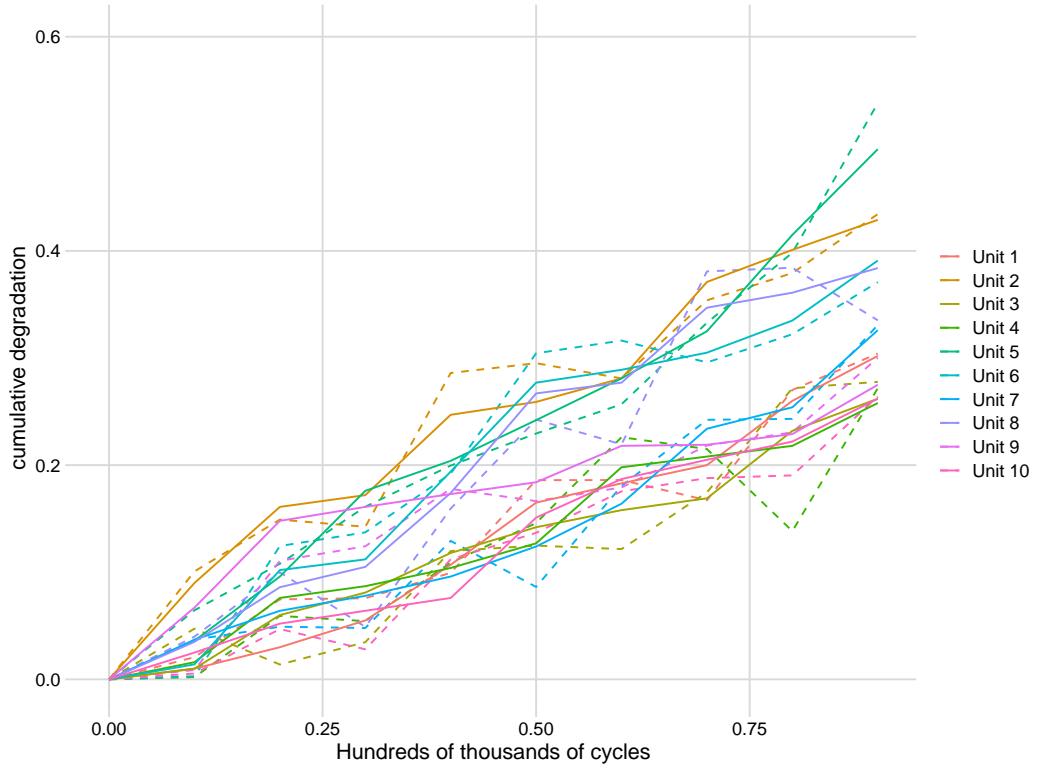


Figure 5.1: Crack-growth propagation data of Rodríguez-Picón et al. (2018): original data (solid lines); noisy data (dashed lines).

iment to measure and then model crack-propagation in the terminal of nominally identical electronic devices originally published by Rodríguez-Picón et al. (2018). The non-noisy data are shown in Figure 5.1 as solid lines. To emulate measurement error, I add a small amount of $N(0, 0.025)$ noise. These new noisy degradation traces are shown as dashed lines in Fig. 5.1. The soft failure of the terminals is considered to be when the crack length reaches 0.4 mm, and we can see from the figure that by the end of the experiment, several units have yet to fail. There are two reasons why such data may be collected (Robinson & Crowder, 2000): to estimate the remaining useful life or failure time distributions of units that have yet to fail during operation, or the corresponding quantities for new units. In the analysis that follows, I model the noisy data and show how to estimate such failure time distributions along with uncertainty intervals.

In the rest of this chapter, I first propose several noisy gamma process models

for the multiple noisy degradation traces in Fig. 5.1 in Section 5.1; a complete pooling model and three where either one or both parameters of the gamma process vary across units. In Section 5.2, I then go on to sample from and evaluate the posterior distributions of these models conditioned on the noisy crack-growth data set. I also compare the models using elppd and cross-validation methods. I then show how to construct failure time distributions for a new unit and a unit that is currently under test but yet to fail in Section 5.3 using both a complete and a partial pooling model. Finally, I discuss the findings and conclusions of the analysis in Section 5.4.

5.1 Models for multiple units

There are three ways in which we might consider allowing a model for the degradation data in Fig. 5.1 to vary, each of which leads to a different form of *pooling*, or, alternatively, of how information is shared or not among the units (Johnson, Ott, & Dogucu, 2022). The added advantage of the mean/coefficient of variation parameterisation is that it makes it explicit which characteristics of the model we are sharing between units. For example, first, we might make the assumption that for a particular characteristic of the model, such as the mean degradation rate (described by the parameter μ), the units do not contain information that may be relevant to each other and therefore estimate completely separate values of μ for each unit; this corresponds to *no pooling*. Secondly, we might assume that all of the units have the same mean wear rate, and the variation that we observe in Fig. 5.1 is only due to the volatility of the gamma process and the fact that we have only observed them over such a short period, and hence estimate the parameter μ by averaging the data from all units—this is *complete pooling*. Finally, although the units are different from each other, they have the same specifications, so we might expect their average degradation rates share similar characteristics: this supposition can be modelled by allowing the parameter μ to vary from unit-to-unit yet arise from a common distribution. Doing so results in

partial pooling of information, which is especially useful when sample sizes are small, but we do not want to make the assumption that all units are identical (McElreath, 2020, Section 13.1).

Different forms of pooling can be applied to each parameter. For example, Lawless and Crowder (2004) allow the rate parameter of the gamma process to vary from unit-to-unit and assume these unit-specific rate parameters arise from a common distribution whose parameters are estimated from the data (partial pooling) while also assuming that the shape parameter is the same for all units; i.e. completely pooled. The multitude of possible pooling combinations is one reason why it is useful to specify a hierarchical model so that the parameters have separate effects on the outcome, like μ , ν , and σ do. By doing so, we can use our understanding of the data-generating process to select sensible models.

In the analysis that follows, I confine my exploratory modelling of the crack growth data to models where σ is completely pooled, and μ and ν are either completely or partially pooled across the units. Since the crack growth data is from a formal experiment, it is reasonable to assume that measurement error in the degradation measurements is constant across the units and observation times and hence to completely pool σ . Moreover, because the degradation traces are from nominally identical units, it makes sense to assume that their mean wear rate and volatility are in some way related; i.e. either partially or completely pool μ and ν . Furthermore, if there is no pooling of either μ or ν , there is no way to make statements about new units without using heuristics and one of the motivations for modelling the crack growth data is to produce reliability estimates for new units. In the next section, I define the complete and partial pooling models that I explore in the rest of the chapter. If it was suspected that the measurement error varied between the units, then the same methods that I describe of varying μ and ν in section 5.1.2 below could be used for σ .

5.1.1 The complete pooling model

I denote by y_{ij} , $j = 1, 2, \dots, J$, the measured degradation of J identical units, and without loss of generality, assume that they are measured at the same times t_i , $i = 1, 2, \dots, I$. In a complete pooling model for the crack growth data, both parameters μ and ν are completely pooled between the ten units. In other words, each unit is a realisation from the same underlying gamma degradation process.

I specify the complete pooling model

$$\begin{aligned}
y_{ij}|z_{ij}, \sigma &\sim N(z_{ij}, \sigma) && \text{data model} \\
\Delta z_{ij}|\mu, \nu &\sim Ga\left(\frac{\Delta t_i}{\nu^2}, \frac{1}{\mu\nu^2}\right) && \text{process model} \\
\sigma &\sim \text{Unif}(0, 10) && \text{parameter model} \\
\mu &\sim N^+(0.5, 0.2) \\
\nu &\sim t_3^+(0, 0.5)
\end{aligned}$$

for the crack growth data. Note that this model is essentially the same model as I explored in Chap 4 except there are now multiple realisations from the gamma process for each Δt_i corresponding to the jump in degradation from each unit. I also use new values of the hyperparameters that are adjusted to the scale of the crack growth data. I chose these new priors using prior predictive checking.

Figure 5.2 shows four prior predictive simulations from the noisy crack growth model with complete pooling. In the figure, each simulated dataset contains the same number of units and observations as the true data set in Fig. 5.1. Clearly, the simulations are noticeably different from the true data; in the first simulation, the units wear much faster; in the second and third, the degradation traces are much more volatile; and in the fourth, there is almost no variability between the pathways. However, as I stressed in Sec. 1.4, the point of performing prior predictive checks is not to tune the prior until it matches the observed data but rather to ensure that the model produces plausible realisations of the data,

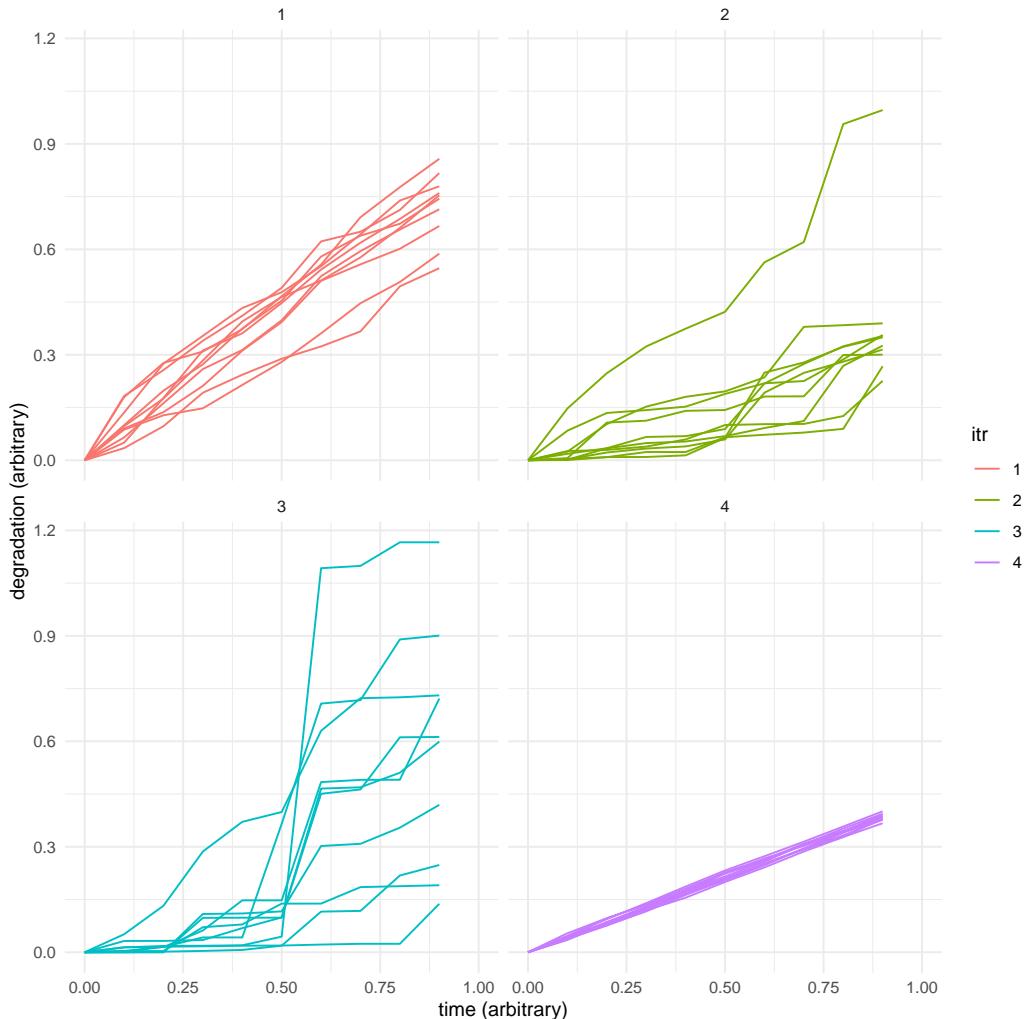


Figure 5.2: Four non-noisy prior predictive simulations from the complete pooling model. Each simulation contains the same number of units and observations as the crack growth data in Fig. 5.1.

which appears to be the case here. Interestingly, through the prior predictive simulations, we can see that the volatility of the gamma process with no varying parameters already allows for a reasonable amount of variation in the degradation traces of the units. Next, I explain how I extended the complete pooling model to allow the parameters to vary between units.

5.1.2 Partial pooling models

To incorporate unit-to-unit variability into the complete pooling model above, I assign a hierarchical prior on μ , ν , or both μ and ν and then estimate the hyperparameters from the data. I use the general structure

$$\begin{aligned}\theta_j | \mu_\theta, \sigma_\theta &\sim N^+(\mu_\theta, \sigma_\theta) \\ \mu_\theta &\sim {}_\theta\pi \\ \sigma_\theta &\sim \text{Cauchy}^+(0, 1)\end{aligned}$$

for the hierarchical prior, where θ represents the parameter we allow to vary between units, θ_j are the unit-specific parameters, and ${}_\theta\pi$ the corresponding prior for that parameter in the complete pooling model. In this hierarchical prior distribution, I assume that the θ_j arise from a Gaussian distribution whose hyperparameters μ_θ and σ_θ are to be estimated. In this way, information is shared across units since a unit-specific estimate of the parameter influences the hyperparameters and, hence, the other unit-specific parameters. I use the same prior for μ_θ as the prior used for the completely pooled case of θ in the complete pooling model since μ_θ now expresses the expected value of the θ_j . Finally, I use a vague truncated Cauchy hyperprior for the standard deviation of the hierarchical prior following the recommendations of Gelman, Carlin, et al. (2020, chap. 17). These choices are just a general starting point; the more mass close to zero in the hyper prior for σ_θ , the more information is pooled between the units (McElreath, 2020), and if I were to use a distribution with heavier tails than a Gaussian, such as the student t, then inference about the hyperparameters would be more robust to outlying units (Gelman, Carlin, et al., 2020, chap. 17). Using this general structure of a hierarchical prior, I explore a varying μ model, varying ν model and a model where both μ and ν vary from unit-to-unit.

varying μ model In the varying μ model, I am assuming that each of the degradation traces in Fig.5.1 arise from different gamma processes where these processes have the same volatility (described by ν) and similar but not the same the average degradation rates. The varying μ model is specified as

$$\begin{aligned}
y_{ij}|z_{ij}, \sigma &\sim N(z_{ij}, \sigma) && \text{data model} \\
\Delta z_{ij}|\mu_j, \nu &\sim Ga\left(\frac{\Delta t_i}{\nu^2}, \frac{1}{\mu_j \nu^2}\right) && \text{process model} \\
\sigma &\sim \text{Unif}(0, 10) && \text{parameter model} \\
\mu_j &\sim N^+(\mu_\mu, \sigma_\mu) \\
\nu &\sim t_3^+(0, 0.5) \\
\mu_\mu &\sim N^+(1, 0.2) \\
\sigma_\mu &\sim \text{Cauchy}^+(0, 1).
\end{aligned}$$

varying ν model In the varying ν model, I once again assume that each degradation trace is a realisation from a different gamma process. However, this time, I assume that all of these processes share the same average degradation rate μ but have varying degrees of volatility ν . I specify the varying ν model as

$$\begin{aligned}
y_{ij}|z_{ij}, \sigma &\sim N(z_{ij}, \sigma) && \text{data model} \\
\Delta z_{ij}|\mu, \nu_j &\sim Ga\left(\frac{\Delta t_i}{\nu_j^2}, \frac{1}{\mu \nu_j^2}\right) && \text{process model} \\
\sigma &\sim \text{Unif}(0, 10) && \text{parameter model} \\
\mu &\sim N^+(1, 0.2) \\
\nu_j &\sim N^+(\mu_\nu, \sigma_\nu) \\
\mu_\nu &\sim t_3^+(0, 0.5) \\
\sigma_\nu &\sim \text{Cauchy}^+(0, 1).
\end{aligned}$$

varying μ and ν model In the final and most flexible model, I assume that the gamma processes that each degradation trace arises from have unique values

of μ and ν . This model where both μ and ν are partially pooled is

$$\begin{aligned}
 y_{ij}|z_{ij}, \sigma &\sim N(z_{ij}, \sigma) && \text{data model} \\
 \Delta z_{ij}|\mu_j, \nu_j &\sim Ga\left(\frac{\Delta t_i}{\nu_j^2}, \frac{1}{\mu_j \nu_j^2}\right) && \text{process model} \\
 \sigma &\sim \text{Unif}(0, 10) && \text{parameter model} \\
 \mu_j &\sim N^+(\mu_\mu, \sigma_\mu) \\
 \nu_j &\sim N^+(\mu_\nu, \sigma_\nu) \\
 \mu_\mu &\sim t_3^+(0, 0.5) \\
 \sigma_\mu &\sim \text{Cauchy}^+(0, 1) \\
 \mu_\nu &\sim t_3^+(0, 0.5) \\
 \sigma_\nu &\sim \text{Cauchy}^+(0, 1).
 \end{aligned}$$

These models can be seen as a set of nested models where the varying μ , varying ν , and complete pooling models are special cases of the model where both μ and ν vary. In models where either μ , ν , or both (as in the complete pooling model) are constant across the different units, the complete pooling is equivalent to a model in which the hyperparameters σ_μ or $\sigma_\nu \rightarrow 0$ and hence the unit specific parameters are forced to be equal to the mean hyperparameters μ_μ or μ_ν .

5.2 Computation, posteriors, and predictive distributions

I fit all four models using the probabilistic programming language Stan. The code is available on a GitHub repository. The HMC algorithm is remarkably efficient, particularly for the complete pooling model, and exploring the posterior distributions requires only six chains of length 1000 after a burn-in period of 1000 iterations. The n_{eff} and \hat{R} statistics for the parameters of interest for all four

Table 5.1: The number of divergent transitions that occur during sampling.

model	number of divergent transitions
complete pooling	0
partial pooling mu	23
partial pooling nu	110
partial pooling mu and nu	202

models indicate that in all cases chains have mixed well (Vehtari, Gelman, Simpson, Carpenter, & Bürkner, 2021), these statistics are shown in tables 5.2, 5.3, 5.4, and 5.5 for the different models. During the sampling from the posteriors of the hierarchical models, divergent transitions occur, particularly in models where ν varies from unit-to-unit. Table 5.1 lists the number of divergent transitions that occur while sampling from the posterior of each model. In the first part of this section, I summarise the inference from each model and show that all models have been able to reclaim the scale of the measurement error and the true underlying degradation paths from the noisy data. I also investigate the cause of the divergent transitions for the hierarchical models, identifying the cause to be the tight curvature in the posterior distributions where the partial pooling collapses towards the complete pooling case. I then show some brief posterior predictive checking to understand the practical differences between the fitted models. In the last part of the section, I use the $\text{elpd}_{\text{LOO-CV}}$ scoring method described in Chap. 1 to compare the four models.

Complete pooling The posterior samples from the complete pooling model for the parameters σ , μ , and ν are summarised in Table 5.2. The model has done a reasonable job at reclaiming the standard deviation of the measurement error (which is 0.025) since the expected value of σ is 0.030 and the majority of the posterior mass sits between $\sigma = 0.020$ and $\sigma = 0.040$. The BHM also provides posterior distributions of the underlying degradation. These are shown in Fig. 5.3 as 95% credible intervals, along with the noisy data and the true underlying degradation traces from which they were generated. As we can see,

Table 5.2: Output from fitting a model with complete pooling to the noisy data of Fig. 5.1. We assume that the data from all units is a manifestation of a single underlying gamma process, and hence the mean and coefficient of variation of the process do not vary from unit-to-unit.

Parameter	Mean	2.5%	50%	97.5%	n_{eff}	\hat{R}
σ	0.03	0.02	0.03	0.04	2003	1.00
μ	0.38	0.33	0.38	0.44	7889	1.00
ν	0.21	0.15	0.21	0.29	652	1.01

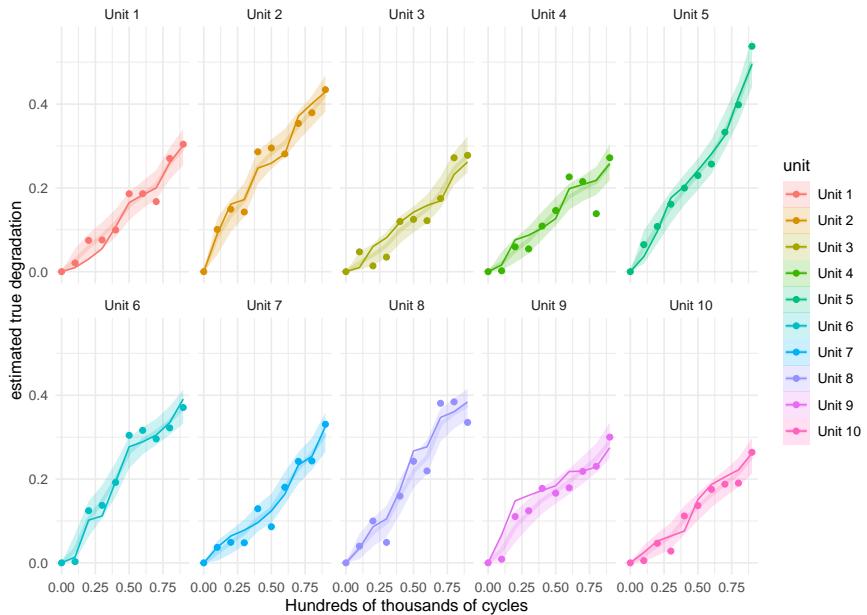


Figure 5.3: Marginal posterior distributions of the underlying gamma process from a BHM where all parameters are completely pooled.

the credible intervals contain the underlying true degradation over the entire time span for each of the ten units, with few exceptions.

Varying μ Table 5.3 shows some summary statistics of the marginal posterior distributions from the varying μ model for select model parameters, and Fig. 5.4 shows the marginal posterior distributions of the parameters σ , ν , μ_1, \dots, μ_{10} , μ_μ and σ_μ . As Table 5.3 and Fig. 5.4 show, a posteriori, the mean degradation rates of the units arise from the distribution $N^+(0.38, 0.07)$. The small expected standard deviation of 0.07 indicates that the unit-specific mean degradation rates

Table 5.3: Partial output from fitting a BHM to the noisy data of Fig. 5.1 where mean degradation μ_j varies between units. Only statistics for Units 1–4 are shown.

Parameter	Mean	2.5%	50%	97.5%	n_{eff}	\hat{R}
σ	0.03	0.02	0.03	0.04	747	1.01
μ_1	0.36	0.26	0.36	0.47	2150	1.00
μ_2	0.42	0.33	0.42	0.56	954	1.00
μ_3	0.35	0.25	0.35	0.46	1001	1.00
μ_4	0.34	0.23	0.34	0.45	746	1.01
ν	0.18	0.09	0.18	0.28	248	1.03
μ_μ	0.38	0.32	0.38	0.46	3079	1.00
σ_μ	0.07	0.01	0.06	0.17	343	1.01

vary in a relatively narrow range, as the posterior distributions in Fig. 5.4 indicate. In addition, the lower tail of the marginal posterior of σ_μ has a considerable mass near zero, and hence, there is reasonable evidence that the average degradation rate is constant across units.

Interestingly, the marginal posterior of μ_μ is wider relative to μ in the CP model; however, both have the same mean, and the uncertainty intervals of the unit-specific μ_j are even wider still. A possible reason is that because I am not making the simplifying assumption that all of the μ_j are equal, the data do not inform the parameters as strongly since there are now even more parameters to estimate in the model, and hence the uncertainty is larger. Additionally, the estimate of ν has shrunk slightly (particularly in the lower tail). From the shrinkage of ν , it appears that because more of the variability between the traces is being attributed to the variation of the μ_j , the resulting traces are less volatile. Despite these slight changes in inference regarding the mean wear rates and coefficient of variation, the varying μ model reclaims the true value of σ to effectively the same degree as the complete pooling model.

During the sampling from the posterior of the varying μ model, a small number (23) of divergent transitions occur. Figure 5.5 show the pairs plots of the samples from the posterior for the parameters μ_1 , μ_2 , μ_μ , and σ_μ . The 23 divergent

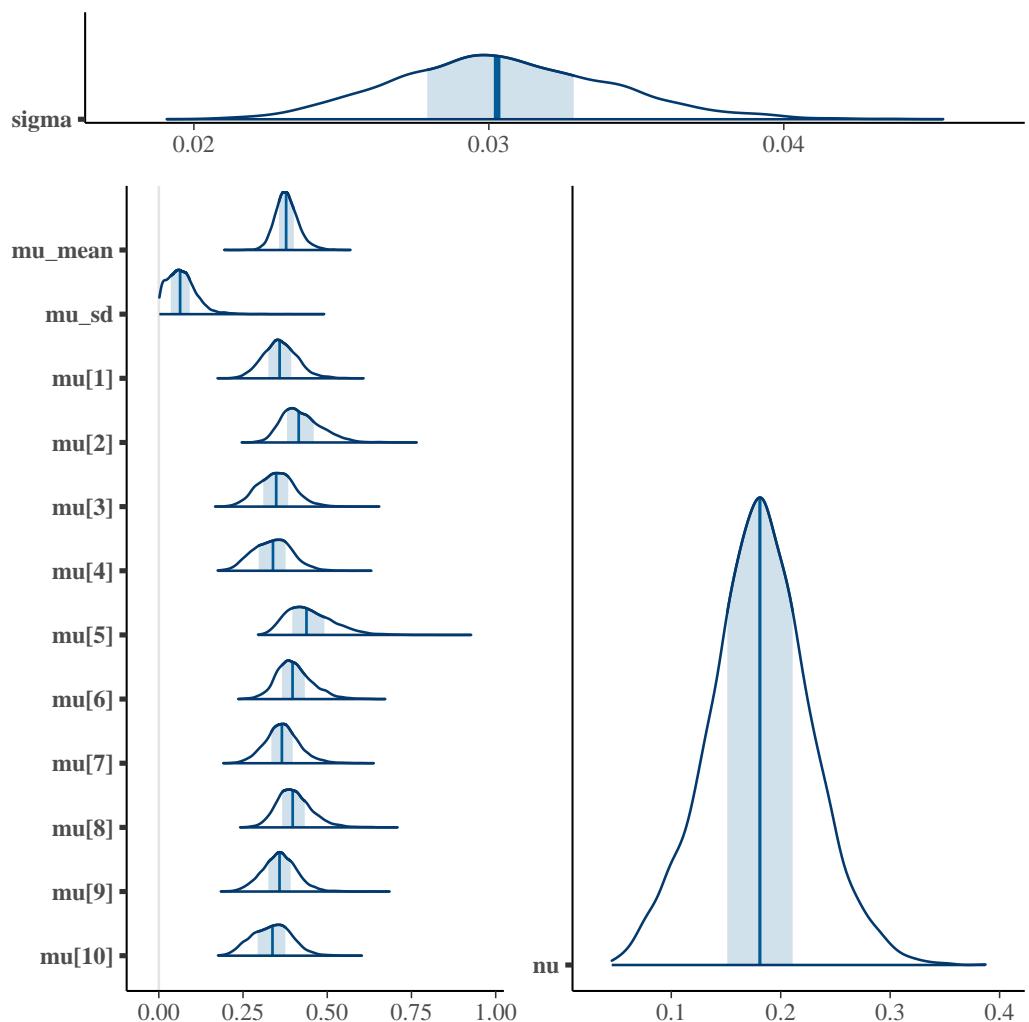


Figure 5.4: Marginal posterior distributions of parameters from the model where mean degradation rate μ_j varies from unit-to-unit.

transitions are plotted in red in each of the bivariate scatter plots. From the plots (particularly those in the lower off-diagonal), we can see that the divergent trajectories tend to occur for very small values of σ_μ . This pattern suggests that in the area of the posterior where the partial pooling model collapses towards the simpler complete pooling case—when $\mu_\mu = \mu_1 = \dots = \mu_{10}$ and $\sigma_\mu = 0$ —there is tight curvature, and hence the sampler starts to misbehave. To confirm this hypothesis, Figure 5.6 show the parallel coordinate plot for σ , μ_μ , μ_1 , \dots , μ_{10} , and σ_μ with the non-divergent traces plotted in blue and the divergent traces plotted in red. Tracking the divergent traces through the parameter space, it is clear that the divergencies occur when all the unit specific μ_j are very close to the mean hyperparameter μ_μ and the standard deviation hyperparameter is very close to zero.

Despite the small issues with sampling, in the varying μ model, the posterior predictive distributions for the reclaimed ‘non-noisy’ degradation traces match the true degradation traces of the units very closely. Figure 5.7 shows the posterior predictive distributions of each unit’s degradation using the varying μ model. Like with the complete pooling case, the 95% posterior predictive intervals from the varying μ model contain the true underlying degradation traces most of the time.

Varying ν When ν is allowed to vary between units instead of μ , all of the ν_j are effectively equal to the mean μ_ν , resulting in most of the mass of σ_ν being very close to zero; Table 5.4. In other words, there is little evidence that ν varies unit-to-unit. Like in the posterior of the varying μ model, the expected value of the hierarchical prior, in this case, μ_ν , is effectively the same as the expected value of the parameter in the complete pooling posterior, ν , and the uncertainty intervals are slightly wider in the partial pooling case. As for the completely pooled parameters, the marginal posteriors of μ in Table 5.4 and Table 5.2 are almost identical, as are the marginal posteriors of σ . The posterior predictive distributions of the degradation traces (not shown) also look similar

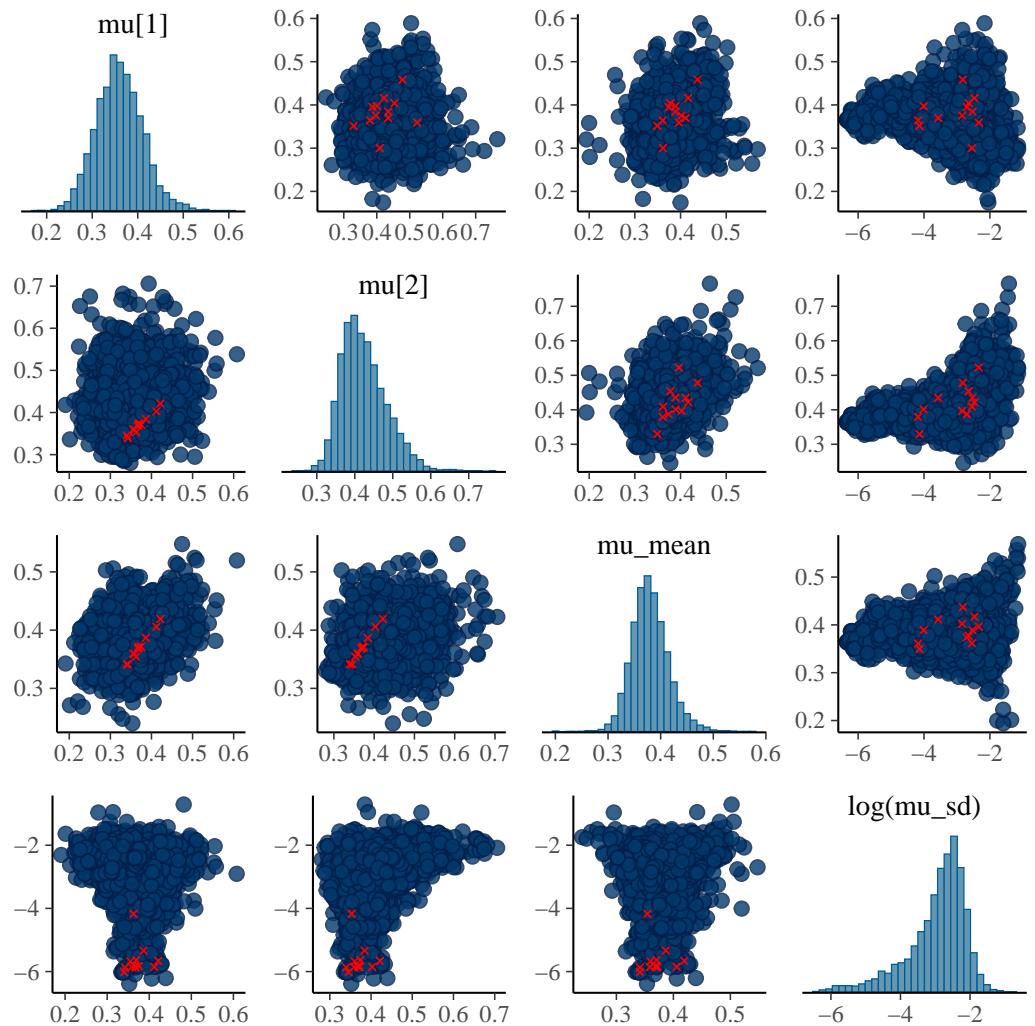


Figure 5.5: A pairs plot of the posterior samples of μ_1 , μ_2 , μ_{μ} , and $\log(\sigma_{\mu})$ from the varying μ model. In each bivariate plot, divergencies are plotted in red.

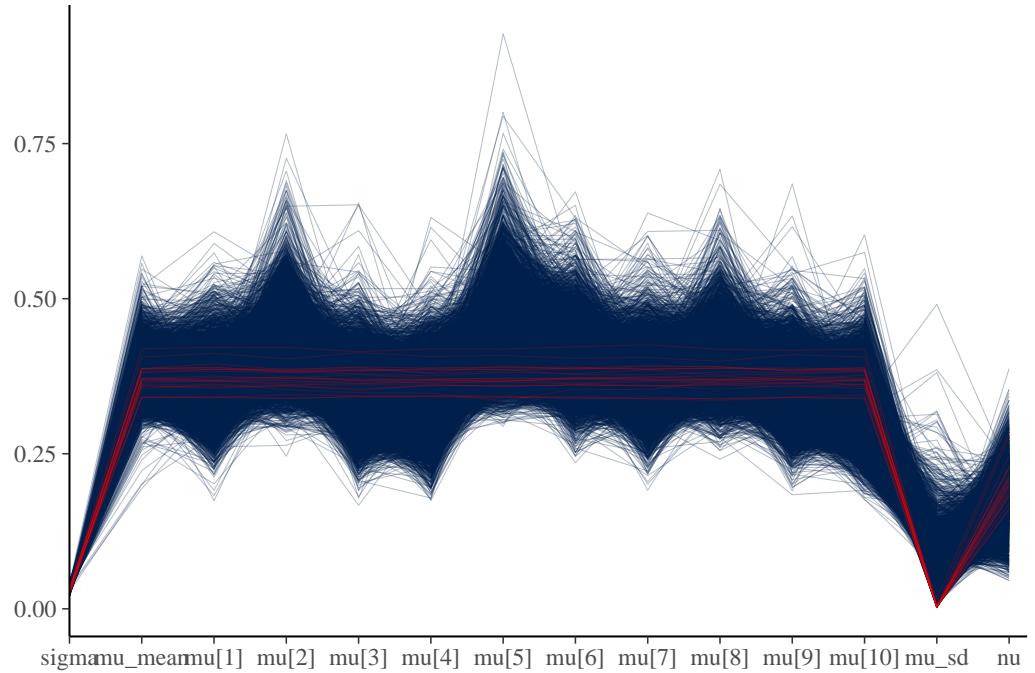


Figure 5.6: Parallel coordinate plot for the parameters and hyper parameters of the varying μ model. The divergent traces are plotted in red.

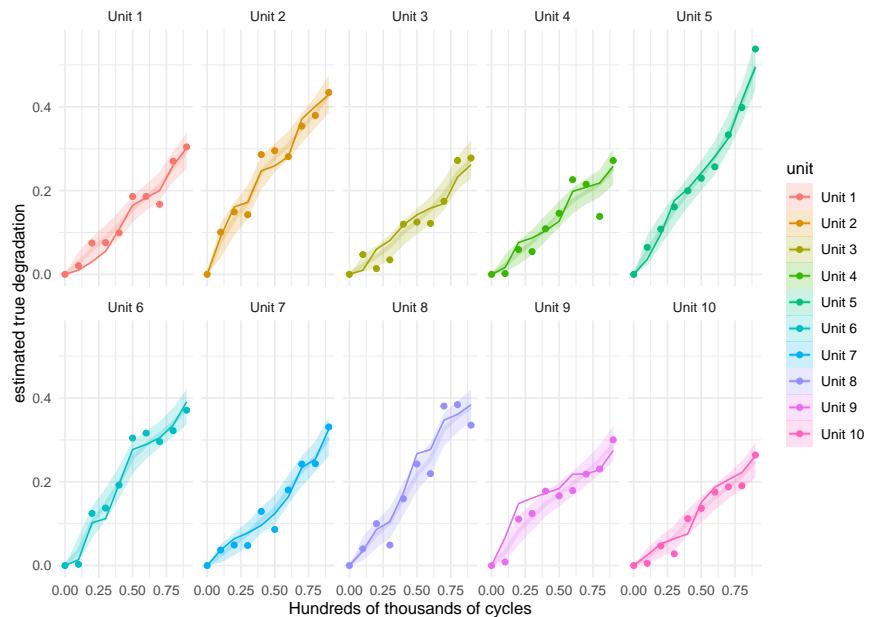


Figure 5.7: Marginal posterior distributions of the underlying gamma process from a BHM where mean degradation rate μ_j varies from unit-to-unit.

Table 5.4: Partial output from fitting a BHM to the noisy data of Fig. 5.1 where the coefficient of variation ν_j varies between units. Only statistics for Units 1–4 are shown.

Parameter	Mean	2.5%	50%	97.5%	n_{eff}	\hat{R}
σ	0.03	0.02	0.03	0.04	6072	1.00
ν_1	0.21	0.10	0.21	0.32	2501	1.00
ν_2	0.22	0.14	0.22	0.34	3092	1.00
ν_3	0.22	0.13	0.22	0.34	2767	1.00
ν_4	0.22	0.13	0.22	0.34	2821	1.00
μ	0.38	0.32	0.37	0.44	6808	1.00
μ_ν	0.22	0.15	0.22	0.31	2127	1.00
σ_ν	0.03	0.00	0.03	0.11	915	1.01

to the complete pooling and varying μ cases.

Like the varying μ model, divergent transitions occur while sampling from the posterior of the varying ν model. However, there are almost five times more divergencies when fitting the varying ν model than the varying μ model; Table 5.1. Figure 5.8 shows the parallel coordinate plot of the MCMC draws for the parameters σ , μ_ν , ν_1 , \dots , ν_{10} , and σ_ν . Like with varying μ model, the divergent transitions diagnose a degenerate area in the posterior around $\mu_\nu = \nu_1 = \dots = \nu_{10}$ and $\sigma_\nu = 0$. In the case of the varying ν model, there is even more posterior mass around this area since there is little evidence that ν should vary from unit to unit, and so a higher number of divergent transitions occur.

varying μ and ν The posterior of the varying μ and ν model is summarised in Table 5.5. When both parameters are allowed to vary unit-to-unit, the results are almost a synthesis of the two models where only one of the parameters varies from unit-to-unit. In Table 5.5, the summaries of the unit specific μ_j match the results of the varying μ model almost exactly (Table 5.3), and like for the varying ν model, there is very little variation between ν_1 , ν_2 , and μ_ν and σ_ν has considerable mas near zero. In Table 5.5, the marginal posterior distributions of the unit specific ν_j and their mean μ_ν match the completely pooled estimate

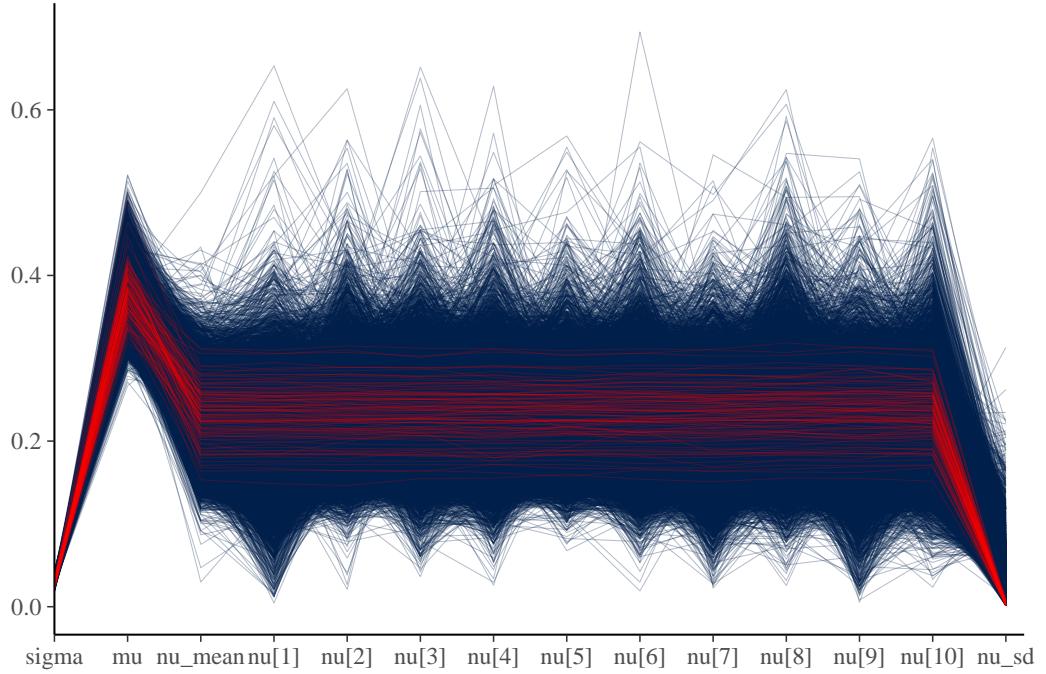


Figure 5.8: Parallel coordinate plot for the parameters and hyper parameters of the varying ν model. The divergent traces are plotted in red.

of ν from the varying μ model. While obtaining the samples from the posterior of the varying μ and ν model, the poor behaviour of the sampler is even more severe than the models where only μ or ν vary; roughly 3% of the transitions are divergent. Still, the varying μ and ν model is able to reclaim the true scale of the measurement error and underlying degradation traces of the units.

Posterior predictive checks Besides the sampling issues, it is hard to differentiate one model from the others since in all of the posteriors σ and the underlying degradation paths of the units are reclaimed to roughly the same level. The easiest way to understand the practical differences between the four models is to look at posterior simulations of new datasets. Figure 5.9 shows three posterior predictive simulations generated from the posteriors of the different models. In each subplot, each of the different colours/line types indicates a posterior predictive simulation with the same number of units and observations as the original data. These simulations from the posterior now look a lot more like the observed

Table 5.5: Partial output from fitting a BHM to the noisy data of Fig. 5.1 where both the coefficient of variation ν_j and the mean wear rate μ_j varies between units. Only statistics for Units 1–4 are shown.

Parameter	Mean	2.5%	50%	97.5%	n_{eff}	\hat{R}
σ	0.03	0.02	0.03	0.04	2700	1.00
ν_1	0.18	0.06	0.18	0.31	1376	1.00
ν_2	0.19	0.08	0.19	0.32	1358	1.00
ν_3	0.19	0.08	0.19	0.32	1275	1.00
ν_4	0.19	0.07	0.19	0.33	1254	1.00
μ_1	0.36	0.26	0.36	0.48	7195	1.00
μ_2	0.42	0.32	0.42	0.56	4363	1.00
μ_3	0.35	0.24	0.35	0.47	4168	1.00
μ_4	0.34	0.23	0.34	0.46	3206	1.00
μ_μ	0.38	0.31	0.38	0.46	13484	1.00
σ_μ	0.07	0.01	0.07	0.17	1468	1.01
μ_ν	0.19	0.09	0.19	0.29	915	1.00
σ_ν	0.04	0.00	0.03	0.11	1102	1.01

data than the prior predictive simulations in Fig. 5.2 did. In Fig 5.9, there is little difference between plot (a) and (c) as well as between (b) and (d) since there is so little variation between the unit-specific ν_j when we allow ν to vary unit to unit. The main difference between the plots is when μ is allowed to vary between units. In (b) and (d), where μ varies from unit-to-unit, the spread of the degradation traces is wider, and the paths are slightly straighter than in (a) and (c), where μ is constant. The posterior simulations from the models where μ is completely pooled look more like the true data. However, I could not say that the data in Fig. 5.1 did not come from one of the varying μ models; some of the simulated datasets in the two models where μ varies from unit-to-unit also look very similar to the observed data. With no obvious choice just by visually evaluating the posteriors, I quantitatively compare the models using cross-validation in the next section.

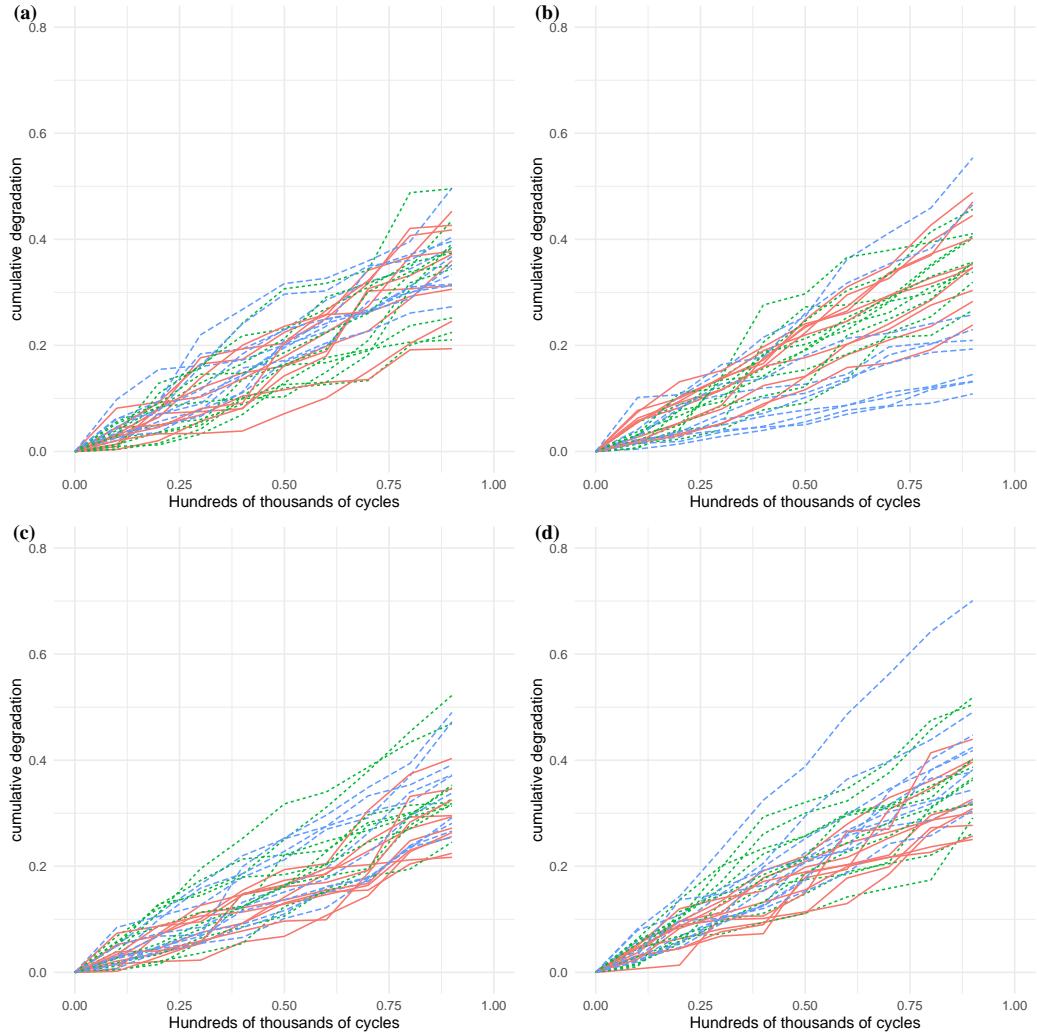


Figure 5.9: Posterior predictive simulations from each of the four models. Each simulated dataset, is indicated by a different colour and line type. Plot (a) shows three simulated datasets from the fitted complete pooling model, (b) from the model with varying μ , (c) from the varying ν model, and (d) from the model where both μ and ν vary from unit-to-unit.

5.2.1 Model comparison

Having chosen four suitable models, fitted them to the data, and checked that the sampling and inference appear reasonable, the next step is to compare the models. To do so, I use the *elppd* and cross-validation methods introduced in Sec. 1.4. In the case of the crack growth data, as mentioned previously, the motivation for analysis is to predict the failure time of new units and the units currently under test but yet to fail. Doing so requires that the models predict the future degradation of the units under test and any new, unobserved units. Hence, it makes sense to compare the four models with respect to these two things.

The ability to predict the degradation of new units can be estimated from eq. (1.12) by treating the I observations from each unit j collectively as a single new observation, i.e., $y_j = (y_{1j}, \dots, y_{Ij})$. We call this method leave-one-unit-out cross-validation (LOUO-CV) and the elppd score calculated in this way the $\text{elppd}_{\text{LOUO-CV}}$. By contrast, the elppd for new observations from the units under test can be approximated using eq. (1.12) by sequentially withholding the final observation, y_{Ij} , from each of the J units; we refer to this method as step-ahead cross-validation (SA-CV) and the corresponding score as $\text{elppd}_{\text{SA-CV}}$. In both cases, we construct the likelihood of the withheld observations in the same way since we have assumed in the data model that the noisy observations y_{ij} from the different units are independent and normally distributed conditional on the true underlying degradation, $\tilde{z}_{i,j}$, and the measurement error, σ . However, the definition of the posterior predictive distribution of \tilde{z} that should be used depends on both the cross-validation method (LOUO-CV or SA-CV) and the model structure (i.e. complete pooling or partial pooling). The details of constructing the posterior predictive distribution of \tilde{z} for LOUO-CV or SA-CV and the results are outlined below, and the code may be found on the GitHub repository.

Leave-one-unit-out cross-validation To calculate $\text{elppd}_{\text{LOUO-CV}}$, we iteratively withhold the data from each unit j , condition on the data from the re-

Table 5.6: Leave-one-out cross-validation statistics for the models fitted in Section ??.

	elppd _{LOOO-CV}	elppd _{SA-CV}
complete pooling	153.8700	15.17704
varying μ	154.1152	14.00906
varying ν	153.3352	15.12410
varying μ and ν	154.2274	15.07771

maining $J - 1$ units and then calculate the log-likelihood of the withheld unit's observations, $y_j = (y_{1j}, \dots, y_{Ij})$. This calculation is based on posterior predictive draws of a new unit's (the withheld unit) filtered degradation path under the fitted model, $\tilde{z}_j = (\tilde{z}_{1j}, \dots, \tilde{z}_{Ij})$, and the posterior draws of σ . Thus, we can write elppd_{LOOO-CV} as

$$\text{elppd}_{\text{LOOO-CV}} = \sum_{j=1}^J \sum_{i=1}^I \log \frac{1}{S} \sum_{s=1}^S p(y_{ij} | [\tilde{z}_{ij}, \sigma]_{-j}^s). \quad (5.1)$$

To generate posterior predictive draws of the non-noisy degradation path of a new unit, we sample $I - 1$ jumps $\Delta \tilde{z}_{ij}^s$ in degradation from $\text{Ga}([\tilde{\mu}_j, \tilde{\nu}_j]_{-i}^s)$ and then calculate their cumulative sum to generate the degradation path. If μ is completely pooled, the $\tilde{\mu}_j^s$ are taken from posterior draws μ^s ; similarly, if ν is completely pooled, the $\tilde{\nu}_j^s$ are posteriors draws ν^s . If, however, the mean degradation varies across units, $\tilde{\mu}_j^s$ is sampled from the (hierarchical prior) $N^+(\mu_\mu^s, \sigma_\mu^s)$; in the same way, $\tilde{\nu}_j^s$ would be also sampled from $N^+(\mu_\nu, \sigma_\nu)$ if the coefficient of variation varied across units. For the models discussed in Sec. 5.1.1 and 5.1.2, the first column of Table 5.6 shows the elppd_{LOOO-CV} scores calculated in this way.

Step-ahead cross-validation Step-ahead cross-validation is carried out by iteratively withholding the most recent observation from each of the units under test, and the SA-CV estimate of elppd is calculated as

$$\text{elppd}_{\text{SA-CV}} = \sum_{j=1}^J \log \frac{1}{S} \sum_{s=1}^S p(y_{Ij} | [\tilde{z}_{Ij}, \sigma]_{-[Ij]}^s). \quad (5.2)$$

To generate the posterior predictive draws in this case, we sample the jump in degradation for unit j from $\text{Ga}(\mu_j, \nu_j)$ and then add this jump to the posterior draws of $\tilde{z}_{I-1,j}$. Where either μ or ν are completely pooled, μ_j and ν_j are posterior draws μ^s and ν^s , respectively; otherwise, we use the draws from the posterior distributions of the unit-specific parameters of the gamma process. The $\text{elppd}_{\text{SA-CV}}$ scores for each of the different models are shown in the right-hand column of Table 5.6.

As the elppd results in Table 5.6 show, for the LOOU case, the models where μ is allowed to vary between units perform slightly better, whereas, for the SA scores, the units where μ is constant across units perform slightly better. However, the difference in both cases is marginal. The similarity of inference from the complete pooling model to that from the partial pooling models—as well as the fact that it has the highest $\text{elppd}_{\text{SA-CV}}$ score—suggests that the completely pooled gamma process with measurement error is sufficient to explain the variability in the degradation traces. However, the added variability of the unit-specific μ_j helps slightly when generalising to new units. Although, it is worth noting that the closeness of the models means that the elppd results are sensitive to the priors. In Leadbetter, Caceres, and Phatak (2024) we used a slightly miss-specified prior for μ , and the result is that the complete pooling model has the highest elppd scores for both LOOU-CV and SA-CV.

In this section, I have shown how to fit different noisy gamma process models for multiple units and demonstrated a principled way of evaluating and comparing them to identify the most appropriate models for the data set being analysed. In the case of the crack growth data, there is a marginal difference between all of the models I explore; this may not be the case for other datasets. The elppd scores seem to suggest that both the completely pooled and a model where μ varies between units are useful for predicting the degradation of the current units under test and new units, respectively. From both posterior predictive checking and elppd scores, there is little motivation to allow ν to vary between units

compared to the simpler constant ν alternative. Holding ν constant also results in better-behaved sampling. Therefore, in the next section, I construct failure time distributions for the units under test as well as for new units using only the complete pooling and varying μ model.

5.3 Failure time distributions

As noted in the introduction, degradation measurements are collected to estimate the failure time distribution of individual units that are in-service but have not failed and/or of new units with the same nominal specifications as the experimental units. The soft failure of the terminals is considered to be when the crack length exceeds $z_f = 0.4\text{mm}$. For degradation models and a soft definition of failure, the failure time T can be defined as the first passage time when the true degradation path crosses the failure threshold z_f (Balakrishnan, Tsai, & C.-T., 2017), that is,

$$T = \inf [t | Z_t \geq z_f].$$

Note that it depends on the true degradation path, not on the observed one, and hence, it does not involve the measurement error (Hamada et al., 2008b). In the Bayesian context, we write the failure time distribution as $F_{T|\Theta}(t)$, where Θ denotes the (hyper) parameters in the model. It can, therefore, be written as

$$F_{T|\Theta}(t) = p(T < t | \Theta) = p(Z_t > z_f | \Theta).$$

One of the advantages of using a fully Bayesian treatment is that we can use the posterior predictive distribution of the underlying degradation Z_t to calculate the failure time distribution, thereby incorporating uncertainty in the parameters, which will be reflected in credible intervals for $F_{T|\Theta}(t)$. Alternatives to a fully Bayesian treatment include, for example, bootstrapping (Peng & Ai, 2018).

Although there is no explicit expression for $F_{T|\Theta}(t)$ it is straightforward to

Complete pooling	Varying μ
<ol style="list-style-type: none"> 1. Draw a sample from the posterior distribution of (μ, ν); 2. Given that $Z_t \mu, \nu \sim \text{Ga}(t/\nu^2, 1/\mu\nu^2)$, calculate $p(Z_t > z_f)$ numerically for a range of values of t to generate one draw of $F_T(t)$; 3. Repeat Steps 1. and 2. n_{sim} times. 	<ol style="list-style-type: none"> 1. Draw a sample from the posterior distribution of $(\mu_\mu, \sigma_\mu, \nu)$; 2. Generate μ_j from $N^+(\mu_\mu, \sigma_\mu)$; 3. Using the μ_j from Step 2. and the corresponding value of ν in Step 1., generate a draw of $F_T(t)$ for a range of values of t as in Step 2. for complete pooling; 4. Repeat Steps 1.–3. n_{sim} times.

Table 5.7: Algorithms for calculating the posterior distribution of the failure time distribution $F_T(t)$ for the complete pooling and varying μ models.

obtain the posterior distribution of $F_{T|\Theta}(t)$ by simulation and numerical evaluation of the distribution function of a gamma distribution (e.g., by using the R function `pgamma`), using a modified version of the procedure outlined by Hamada et al. (2008b, Sec. 8.2.1). For the complete pooling and varying μ models, the algorithms are shown in Table 5.7.

Fig. 5.10 shows the posterior of the failure time distributions for new units, calculated from the complete pooling and varying μ models. There is considerably greater uncertainty in $F_T(t)$ from the partial pooling model, but this is not surprising: in addition to the inherent variability of the gamma process, the partial pooling model also includes the variability in the μ_j .

Using slight modifications of the algorithms shown in Table 5.7, we can also calculate the failure time distribution for a unit that is currently under test and that has yet to fail, for example, Unit 3. This distribution, also known as the predictive failure time distribution (Lawless & Crowder, 2004), is conditional on the unit not having failed by t_I and having attained a degradation level z_I . Since z_I is a latent variable in the model, the posterior distribution contains samples of z_I from which we can calculate the jump in degradation that corresponds to a soft failure. Figure 5.11 shows the posterior predictive failure time distributions

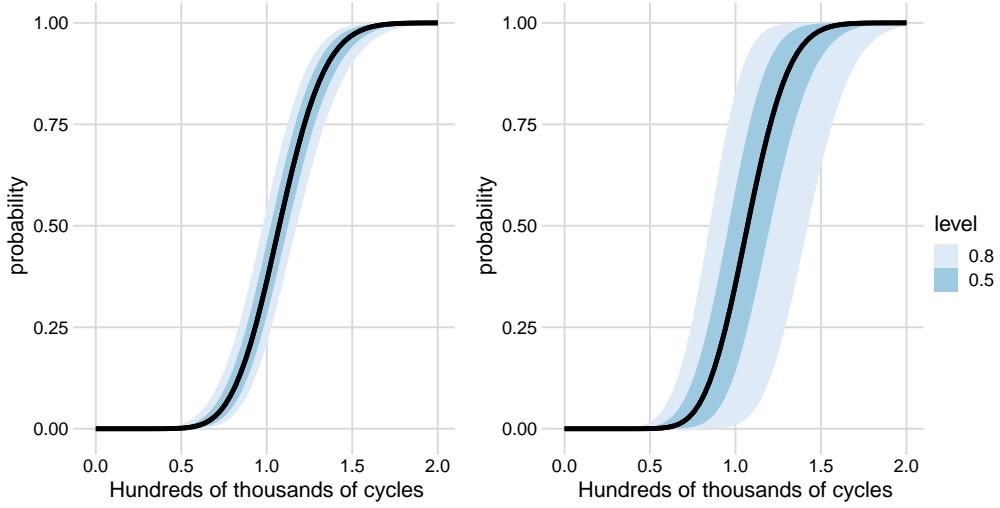


Figure 5.10: Posterior distributions of the failure time distributions from the complete pooling model (left) and the varying μ model (right).

for Unit 3, which has not failed; again, we see that because of the additional layer of uncertainty, the credible intervals for the varying μ model are wider than those from the complete pooling model, although the difference is not as extreme compared to the failure time distributions for new units. Comparing Fig. 5.10 with Figure 5.11, the unit-specific failure time distributions have tighter uncertainty intervals than their new unit counterparts. This difference is more significant for the varying μ model since the unit-specific estimate does not average over the variability in the μ_j and so is a much more precise estimate.

5.4 Discussions

In this chapter, I showed how the noisy gamma process model for a single degradation path from Chap. 4 could be extended—using the same hierarchical modelling framework—to incorporate unit-to-unit variability when modelling the noisy degradation traces of multiple nominally-identical units, and how allowing some parameters to vary determines how information is shared between observational units. I then demonstrated the fitting, evaluation, and comparison of these models on an experimental crack growth dataset with added measure-

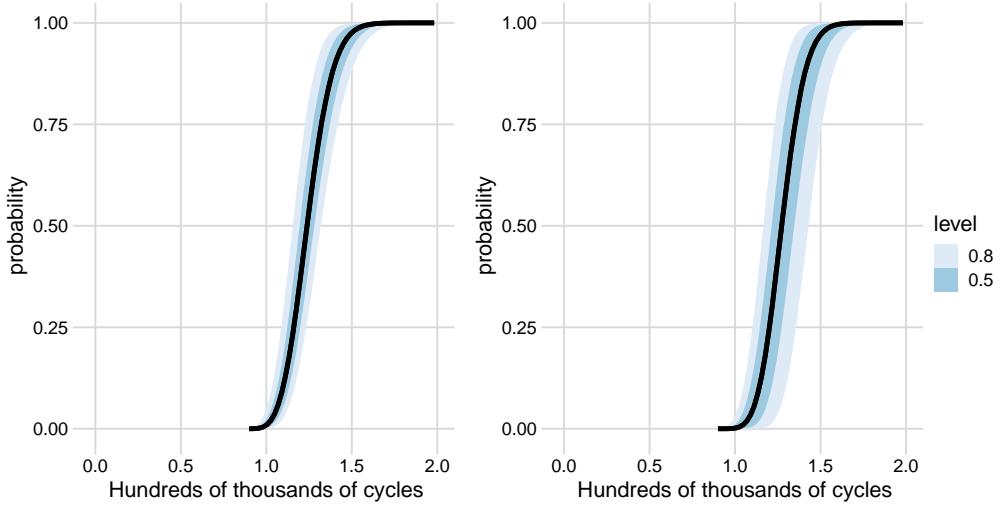


Figure 5.11: Posterior distributions of the predictive failure time distributions from the complete pooling model (left) and the varying μ model (right) for Unit 3.

ment error. Lastly, I showed how failure time distributions (with uncertainty bands) can be constructed for both new units and units under test using the posterior distributions of either a complete pooling or partial pooling gamma process model. In this final section, I review the chapter’s main points and discuss the key findings, contributions, and areas for future work.

To model the ten units’ noisy degradation traces in Fig. 5.1 simultaneously, I explored several models: one where all of the degradation traces arise from the same underlying gamma process (complete pooling) and others where either μ , ν , or both are allowed to vary between units (partial pooling). The parameterisation in terms of μ and ν clarifies how unit-to-unit variability can be incorporated into the model and forces the analyst to be explicit in how they expect the units to vary—do we expect the mean wear rate to vary among units, or the volatility? In the same way, the new parameterisation also clarifies how we might model the effect of covariates on the degradation: would we expect, for example, environmental variables such as varying temperature or humidity to affect the mean wear or the volatility? There are no covariates for the crack growth data, but extending the BHM to include covariate information would be interesting future work if data were available to do so. The fact that parameters μ and

ν have clear and separate effects also helps diagnose the sampling issues and interpret the posterior of the parameters and hyperparameters directly.

Based on the elppd criterion, the models where μ varies between units perform best for predicting out-of-sample. In contrast, models where μ is constant across observational units predict the future observations of the units under test best. Although, the differences in the elppd scores are relatively small. All of the models explored fit the data well since the predictive distribution of each unit's underlying degradation path recovers the true underlying degradation path; in addition, the marginal posterior distributions of σ include the true value of $\sigma = 0.025$ that I used to simulate the measurement error. In the posterior distributions of the partial pooling models, there is some evidence that we should allow μ to vary between units since there is some variability in the modes of the marginal posteriors of the unit-specific μ_j ; however, these distributions are wide enough to encompass the mean μ_μ and the posterior of σ_μ has mass near zero and therefore it could well be that all units share the same value, even under a model where we allow them to vary. In the models where ν varies, the marginal posterior distributions of the unit specific ν_j are very similar, and the marginal posterior of σ_ν has considerable mass near zero, showing that there is little evidence that the coefficient of variation varies among the units.

Given the weak evidence in the hierarchical models' posteriors that μ varies amongst the different units and even weaker evidence that ν varies, it is understandable that the complete pooling case performs at the same level as the partial pooling cases for the crack growth data. The complete pooling model performs best in the case of $\text{elppd}_{\text{SA-CV}}$, possibly because assuming the simpler model structure results in the data more strongly informing the three parameters and hence more precisely estimating the volatility of the gamma process (ν), which is important for accurately forecasting future degradation. Because of the marginal difference between models, it is not surprising that in Leadbetter et al. (2024), where we used a slightly different prior for μ , the complete pooling model

performs best with respect to both $\text{elppd}_{\text{LOUO-CV}}$ and $\text{elppd}_{\text{SA-CV}}$; showing that the ordering is sensitive to the model specification and prior. Rodríguez-Picón et al. (2018) analyses the same data using gamma processes that incorporate unit-to-unit variability but without measurement error. They find that, when the data do not include measurement error, one of the partial-pooling models they explore outperforms the complete pooling model according to information criteria methods. However, there is also very little difference among the models they explore.

The struggle to clearly identify the best model could be caused by the nested structure of the models. When the observed variability of the degradation traces can easily be explained by the complete pooling model (which appears to be the case for the crack growth data), all of the models I explored here contain this ‘true’ model. Future work exploring how well the elppd methods identify the true model from complete and partial pooling models when the data are generated from one of the models would be interesting. The presence of measurement error adds an additional ‘degree of freedom’ to the models, making it even more difficult to clearly identify the best ‘underlying’ model candidate. To this end, it would also be helpful if future work exploring elppd through simulation looked at the effect of sample size and noise level on how well true models are reclaimed. In an early work on unit-to-unit variability, Lawless and Crowder (2004) devise a statistical test to determine whether or not a random effect should be included in a degradation model when working in a non-Bayesian framework. Similar guidance for Bayesian models would be helpful.

The crack growth data that I have analysed do not show strong signs that there is variability among the units outside of the usual ‘jumpiness’ of a gamma process. For other data, identifying a suitable partial pooling model may be much more obvious. Nevertheless, I have shown how analysts can propose, fit, check, and then, finally, choose the most suitable model candidate for a set of Bayesian models using a fully Bayesian framework.

Chapter 6

Conveyor belt wear forecasting

In the past two chapters, I have looked at expanding the gamma stochastic process model to account for common situations encountered in practice, namely, the need to account for measurement error and the ability to borrow information across similar processes. In this chapter, I look at the practical example of modelling the degradation of a conveyor belts wearing surface using gamma processes. In doing so, I use both the noisy gamma process construction from chapter 4 and the partial pooling structures in chapter 5.

Conveyors are critical to the productivity of iron ore mines and other mining operations. As such, their unplanned failure can cause a significant loss of production and, subsequently, a substantial loss of profits. On the conveyor, one of the main components that can fail is the belt, and a major failure mode of the belt is wear (Bortnowski, Kawalec, Król, & Ozdoba, 2022). To manage the risk of failure due to wear, reliability engineers monitor the thickness of the belt's protective topcoat using ultrasonic thickness (UT) measurements. An example of this data is shown in *Figure 6.1*. Engineers then use this condition monitoring data to estimate the failure time of the belt and plan when to replace it. However, at each observation time, the UT data only provide a detailed description of the wear profile across the belt's width at one random location along its length. Furthermore, the observation times are sparse. These factors result in

lots of uncertainty around the underlying degradation of the belt. Therefore, estimating the failure time of the belt from the UT measurement data to inform maintenance decisions requires robust statistical modelling and the quantification of uncertainty. In this chapter, I show how the Bayesian hierarchical approach can be used to extend functional data analysis (FDA) of the wear profiles to different degradation models in order to forecast the belt’s wear and predict the remaining useful life. Particularly I compare a gamma stochastic process and a linear general path model for modelling the underlying degradation process of the belt.

Although there are papers that address the condition monitoring of conveyor belts, for example, identification of damage from puncture or modelling cord damage (Bortnowski et al., 2022), very few academic works focus on wear from abrasion. This is surprising considering that wear is a major failure mode of the belt (Bortnowski et al., 2022), especially for shorter, highly-utilised belts like stackers and reclaimers (which are also highly critical and difficult to maintain). Webb et al. (2020) demonstrate one typical method that an engineer would use to estimate when the failure of the belt will occur due to wear. For each measurement location along the belt’s width, the engineer fits a linear relationship to the UT measurement at that location using cumulative tonnes as the predictor variable. Next, at the location with the most aggressive wear rate—the steepest gradient—they extrapolate the line up to some predetermined soft failure threshold, indicated by a dashed line at 25mm of wear in *Figure 6.1*. The time at which the extrapolated line intersects the soft failure threshold is the predicted failure time. An alternative but similar approach is to take the maximum wear measurement from each profile and fit a linear relationship to these maximum wear measurements. This second approach is used by some conveyor condition monitoring software. *Figure 6.2* demonstrates these two methods using the data in *Figure 6.1*. Unfortunately, these method neglects the many sources of uncertainty in the data-generating and observation processes. Firstly, trend fitting of

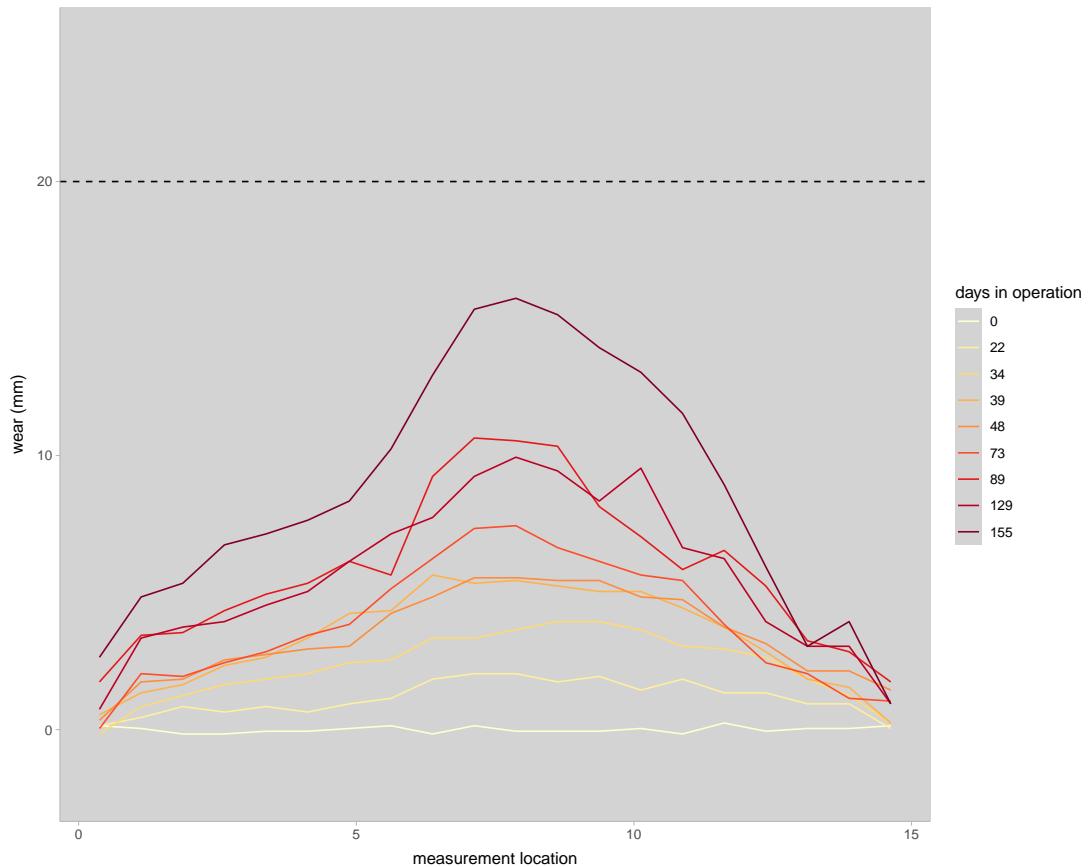


Figure 6.1: The growing wear profile of a conveyor's belt over time. Ultrasonic thickness measurements are taken at $N = 20$ measurement locations across the width of the belt at repeated times. The vertical axis is wear in mm , calculated by subtracting the measured thickness from the original thickness of the belt. The horizontal dashed line at $25mm$ of wear indicates the maximum allowable wear before the belt needs to be replaced, i.e. the soft failure threshold. The colour gradient indicates the time the belt has been in operation. Measurements taken at the beginning of the belt's life are shown in light yellow, and the most recent set of measurements are plotted in dark red

the raw UT measurements is sensitive to noise in the data, especially early on in the belt's life when there are few observations. Secondly, there is no formal structure for managing the different sources of uncertainty, for example, uncertainty in the UT measurements due to measurement error, uncertainty in the wear profile because of spatial variation along the length of the belt, uncertainty in the wear rate due to variations in operating conditions, and uncertainty in the parameters of the degradation process. Finally, the prediction is based solely on the forecast from only a few measurements on that particular belt with no uncertainty quantification. Therefore, the engineer cannot quickly assess the future wear of the entire belt surface, nor can they interpret risk, which limits their ability to justify and defend their maintenance decisions.

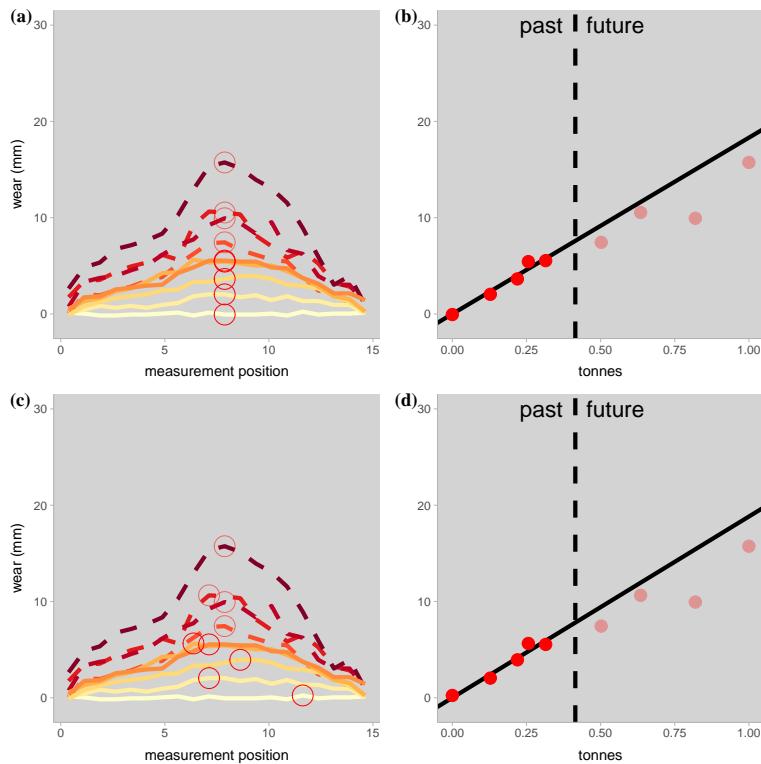


Figure 6.2: A demonstration of some typical approaches used to trend belt wear. (a) and (b) demonstrate the method of (Webb et al., 2020), where the measurements at the fastest wearing location which are circled in (a) are trended into the future in figure (b). (c) and (d) demonstrate an alternative but similar approach used by some reliability software, where the maximum wear measurement in each profile is trended instead.

Rather than forecasting the wear at a single measurement location or only for the maximum wear measurement, I propose a Bayesian hierarchical method for forecasting the evolution of the entire wear profile of the belt through time. The method I use expands degradation models, such as the previous work in chapters 4, and 5, to functional time series. In the first level of the hierarchical model, I apply a functional data analysis (FDA) approach to modelling the condition monitoring data in Figure 6.1. The FDA interpretation smooths the observations, which helps account for measurement error in the UT testing process; it also reduces the dimension of the data, making it easier to model the degradation at all the measurement locations simultaneously without too much of a computational burden. This functional data model, when paired with a suitable process model for the underlying degradation—either general path or stochastic process—and a suitable parameter model, formally manages the different sources of uncertainty. The proposed method produces forecasts of the belt’s wear that:

1. properly quantifies the uncertainties in the prediction, and
2. produces an intuitive forecast of the entire wear profile.

The method is not limited to conveyor belt wear; it can be used for any degrading surface monitored over a grid of locations. Here, I show an application for a one-dimensional profile, but the method could be expanded to two.

I begin in Section 6.1 by providing a brief overview of functional data analysis and showing how it can form the data model in a Bayesian hierarchical model for degradation. I then define two process models for the degradation of the belt in Section 6.2: a noisy gamma process and a linear general path. In Section 6.3, I define the prior distributions for the two models, show how an informative prior can be constructed from historic belt wear datasets, and check the plausibility of the process model through prior predictive simulation. Section 6.4 describes sampling and investigates the posterior draws from each model in terms of the marginal posterior distributions of the parameters, the posterior distribution of

the intermediate quantities in each model that describe the underlying wear process, and the prior predictive distributions for replications of the data in Fig. 6.1. In Section 6.5, I describe how to generate forecasts for the belt’s wear profile using the two different models and demonstrate with the ninth wear profile observation that I withhold when fitting the two models. I then evaluate and compare the two models based on their ability to predict the degradation of the belt at future time points in Section 6.6. Lastly, in Section 6.7, I demonstrate how to construct failure time distributions for the two models conditioned on the current condition of the belt. I finish in Section 6.8 by revisiting the main results and pointing out areas of useful future work.

6.1 Functional data analysis (the data model)

In functional data analysis, the data in each observation are considered as coming from a smooth underlying random function rather than being a scalar or vector-valued random variable (Gelman, Carlin, et al., 2020, p. 512). That is, we interpret the set of UT measurements at each observation time as noisy observations at discrete locations of some smooth underlying random function

$$z_{i,n}|f_i(n), \sigma \sim N(f_i(n), \sigma). \quad (6.1)$$

In this way, the standard deviation σ describes the noise in the UT measurement process—which we assume to be normally distributed—and the function $f_i(n)$ describes a smooth wear profile across the width of the belt at time t_i . The aim of a functional data analysis is, therefore, to model the collection of functions $\{f_i(\cdot)\}_{i=1}^I$. To do so requires the analyst to choose a functional form for $f_i(\cdot)$, a decision on which the model is implicitly conditioned. In this analysis, I use a B-spline to model the wear profiles of the conveyor belt.

B-splines B-splines are piecewise continuous functions (Ramsay, Hooker, & Graves, 2009, p. 33-38) that are constructed as the weighted sum of a set of M locally defined polynomial B-spline basis functions,

$$f_i(n) = \sum_{m=1}^M y_{i,m} b_m(n). \quad (6.2)$$

Here, $y_{i,m}$ is the weight of the m^{th} basis function, $b_m(\cdot)$, for the i^{th} observation. The number of basis functions and their shapes and locations are defined by a set of knots and the order of the basis functions. To describe the wear profiles, I use eight evenly spaced knots and third-order basis functions. However, I also drop the outer two sets of basis functions to constrain how flexible the spline can be towards the edges of the belt and to ensure that the wear profile is fixed at zero at the boundaries. These choices were made by measuring the goodness of fit for many different conveyor wear profiles from the industry partner's condition monitoring system and weighing up simplicity and flexibility. Figure 6.3 (c) shows the set of un-weighted basis functions on which we condition our model. The spline is fit to the UT data by estimating the weights each basis function. As an example, Fig. 6.3 (a) shows the fitted spline for the fifth observation and Fig. 6.3 (b) shows the weighted set of basis functions that make up the wear profile. Fitting a B-spline to each set of UT measurements using the set of basis functions in Fig. 6.3 (c) yields the set of spline coefficients $\{y_{i,m}\}_{m=1}^M$ that fully describes the wear profile at time t_i . The next level of the BHM, the process model, models how these spline coefficients evolve through time.

Functional time series Before moving on, I note that modelling the evolution of spline coefficients through time is not new. Functional time series analysis is an area of statistics that models how time-ordered functional observations evolve (Hormann & Kokoszka, 2012). It has also been done in a Bayesian context (Kowal, 2017). Usually, these methods use functional PCA (Ramsay et al., 2009, p. 16) and, for the most part, autoregressive processes to model the evolution

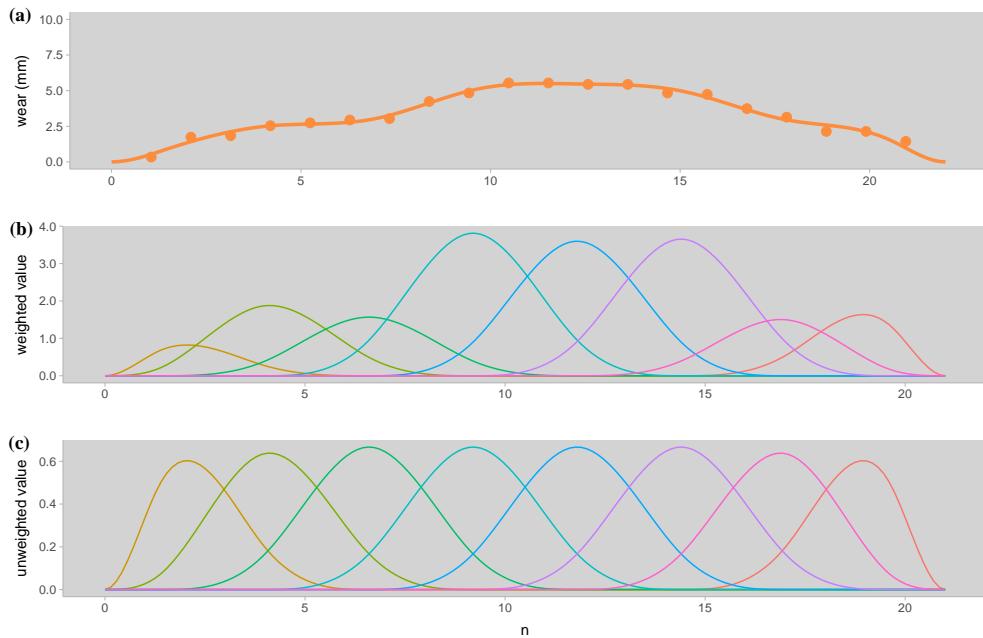


Figure 6.3: (a) shows a fitted wear profile for the fifth observation; (b) shows the weighted set of basis functions that make up the profile in (a); and (c) shows the unweighted set of basis functions

of the eigen weights (the coefficients of the eigenfunctions). Using functional PCA first reduces the number of coefficients to model. Furthermore, because the eigenfunctions are orthonormal, the eigen weights can be modelled independently. This dimension reduction is helpful for cases where many basis functions are needed to fit the data because, in these cases, modelling the basis coefficients directly would require a very large covariance matrix. However, in our case, the number of functional observations of the belt is too small to perform PCA reliably, and the degradation process is not stationary. Rather, we use a degradation model to model the spline coefficients directly and, for the moment, assume that the coefficients are independent of one another. In this way, our approach may be more down the path of the spatio-temporal models of Wikle et al. (2019, p. 218-224).

6.2 Process models

As potential process models for the underlying degradation process of the belt, I explore both a noisy gamma process and a linear general path model. In the noisy gamma process model, I model the spline coefficients with effectively the same model that I presented in chapters 4, and 5, except with some minor adjustments to make the model more robust to outlying observations. In the linear path model, I use a much simpler linear model for the spline coefficients. The major difference between the two process models is that the noisy stochastic process model explicitly breaks down the variation in the degradation signal into variation attributed to changes in degradation rate through time—which could arise from changes in operation or environmental conditions—and the uncertainty of the measurement process—which in this case arises because the cross-section is not always measured in the same location. In contrast, the general path model assumes that the degradation path is deterministic conditional on the model’s parameters and the time t and lumps the variability from both the variation in signal through time and measurement error into a single error term. When using a general path model, additional structure in the noisy degradation measurements can be accounted for by adjusting the error distribution, for example, adding autoregressive error terms. In some cases, the two approaches are directly comparable; for example, Whitmore (1995) formulates a Weiner degradation process with measurement error as a multiple linear regression with added covariance structure. Some discussions of stochastic process and general path models can be found in Ye and Xie (2015). The benefit of using the noisy gamma process, in this case, is that by modelling these two sources of variation separately, we can assign them different distributions (Gaussian-distributed measurement error and gamma-distributed jumps), which explicitly splits up the two sources of uncertainty. However, there may be a limit of measurement error and/or volatility of the degradation process where it is simpler and better to use the general path model even though it is less physically motivated. Below, I outline the two process

models.

6.2.1 Noisy gamma process

In the gamma process version of the process model, I model each spline coefficient as coming from a t distribution with 10 degrees of freedom where the location depends on the estimated ‘average’ value of the spline coefficient along the length of the belt at that time and the scale depends on the estimated average value and scale parameter ϕ . Assuming a t distribution for the spline coefficients makes the model more robust to outlying observations (Gelman, Carlin, et al., 2020, Chapter 17). I then model the progression of the ‘average’ degradation of the spline coefficient through the gamma process parametrised in terms of the mean and coefficient of variation

$$\begin{aligned} y_{m,i}|y_{m,i}^*, \phi &\sim t_{10}(y_{m,i}^*, \phi y_{m,i}^*) \\ \Delta y_{m,i}^* &= y_{m,i}^* - y_{m,i-1}^* \\ \Delta y_{m,i}^* | \nu_m, \mu_m &\sim \text{Ga}\left(\frac{\Delta t_i}{\nu_m^2}, \frac{1}{\mu_m \nu_m^2}\right). \end{aligned}$$

In this process model, each of the m spline coefficients has specific μ_m and ν_m .

6.2.2 Linear model

In the linear general path version of the process model, I use the same t_{10} structure to model the noisy spline coefficients conditional on their estimated average value and ϕ , but now model the progression of the average value of each spline coefficient as a deterministic function of t that is specified in terms of the average wear rate μ_m

$$\begin{aligned} y_{m,i}|y_{m,i}^*, \phi &\sim t_{10}(y_{m,i}^*, \phi y_{m,i}^*) \\ y_{m,i}^* &= \mu_m t_i. \end{aligned}$$

6.3 Parameter model

Because I have used the mean/coefficient of variation parameterisation of the gamma process and a conditional structure for both the models, they share many of the same parameters— σ , ϕ , and the μ_m . Hence, I use the same priors in both models. For the variance parameters σ and ϕ I use the priors

$$\begin{aligned}\sigma &\sim U(0, 100) \\ \phi &\sim \text{Cauchy}^+(0, 25)\end{aligned}$$

based on the same justifications presented in Section 4.4 (Gelman, Carlin, et al., 2020, chap. 17). For the mean wear rate of the different spline coefficients, I now use the prior

$$\mu_m \sim N(\hat{a}, \hat{b})$$

where \hat{a} and \hat{b} are estimated from historic data. I go into more detail on how this is done below, but first, define the remaining priors for the gamma process.

The gamma process model has slightly more parameters than the simpler linear path model. For the coefficient of variation parameters ν_m , I use the hierarchical prior

$$\begin{aligned}\nu_m &\sim N(\mu_\nu, \sigma_\nu) \\ \mu_\nu &\sim t_3(0, 0.5) \\ \sigma_\nu &\sim \text{Cauchy}^+(0, 0.25)\end{aligned}$$

defined in the same way as the unit-to-unit variability model in Section 5.1.2. The hierarchical prior partially pools information between the ν_m of the gamma processes for each spline coefficient (because they all belong to the same belt we expect them to be similar) while still allowing them to vary.

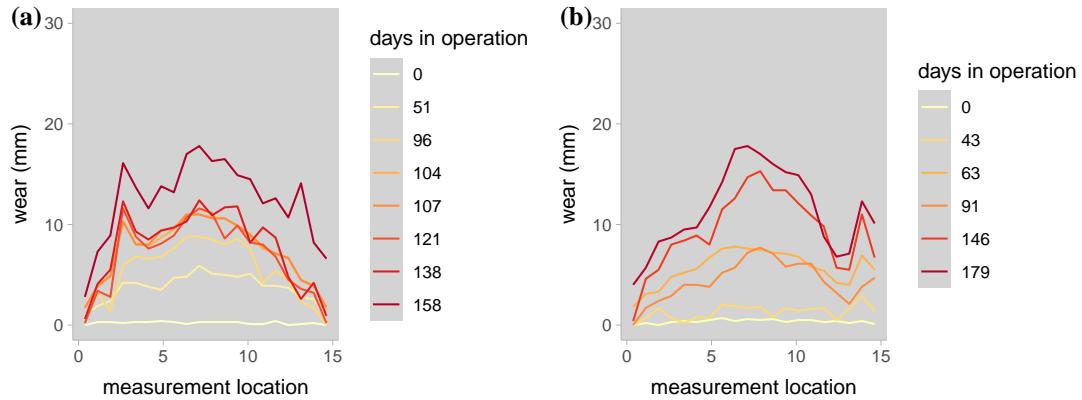


Figure 6.4: Two historic belt wear data sets from the same conveyor

Choosing informative hyperparameters for μ There are also historic belt wear datasets for the particular conveyor that the data in Fig. 6.1 come from. Two of these historical datasets are shown in Figure 6.4. According to the industry partner from whom we got the data, the future wear behaviour of the belt is expected to be different from the historic wear datasets because of variations in ore composition, belt manufacturer, and operational strategies. However, the historic wear behaviour of the belt should indicate future behaviour. Although the historical data cannot be directly used in the analysis, in the Bayesian framework, the historical data can be used to formulate an informative prior and, therefore, still supplement the analysis.

I encode the prior information in the model through the parameter μ . To do so, I fit B-splines to each of the historic wear profiles using the same set of basis functions in Fig. 6.3 (c) to calculate the spline coefficients and then estimate the average wear of each coefficient. Figure 6.5 shows the linear regressions for each spline coefficient. Based on the estimates, I set \hat{a}_m in the prior for μ_m as the estimated slope of the coefficient based on the historical data and \hat{b}_m as five times the standard error of the estimate.

Prior predictive checking To check the plausibility of the proposed priors in the context of the two process models, I perform prior predictive checking.

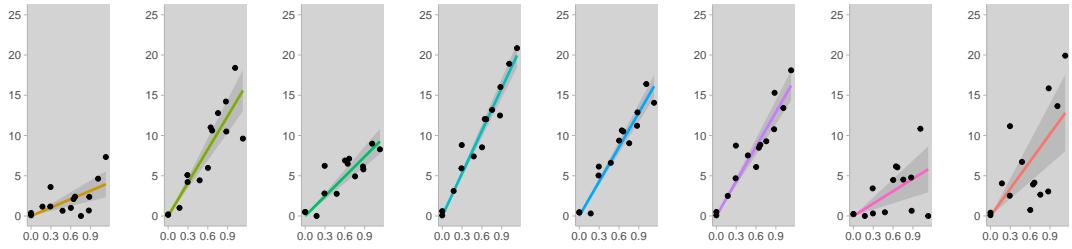


Figure 6.5: The estimated average wear rate for each spline coefficients calculated from the datasets in Fig. 6.4. The colour of each fit corresponds to the basis functions in Fig. 6.3 (c). The horizontal scale of each plots is in tonnes and the vertical scale is the value of the spline coefficient.

Figure 6.6 shows sixteen prior predictive simulations from the gamma process (solid lines) and linear model (dashed lines). Because I have used very vague priors for the parameters σ and ϕ , simulating the noisy wear profiles and UT measurements will not make sense until after conditioning on the observed data. Therefore, the simulations shown in Fig. 6.6 are for the belt's non-noisy (average) wear profile.

To generate each fictitious data set, I sample a value of each μ_m , μ_ν , and σ_ν from their prior. I then sample the values of ν_m from the hierarchical prior using the realisation of μ_ν and σ_ν . Next, I generate values of the smoothed spline coefficients for each observation time t_i from each of the two process models and apply the coefficients to the set of basis functions to calculate the smooth wear profiles. For the linear general path model, this means simply multiplying each of the μ_m by each t_i to get the values of the spline coefficients. Whereas, for the gamma process, I simulate the sets of jumps in degradation using the distributions $\text{Ga}(\Delta t_i / \nu_m^2, 1 / (\mu_m \nu_m^2))$ and take the cumulative sum to calculate the values of the filtered spline coefficients.

The fictitious data resulting from each model, for the most part, look somewhat realistic in scale and shape. Some of the gamma process simulations look unrealistic—there are some where the wear jumps 60mm between observation times and others where the belt does not wear at all—but many look sensible. The shapes of most of the linear simulations look reasonable, but their growth

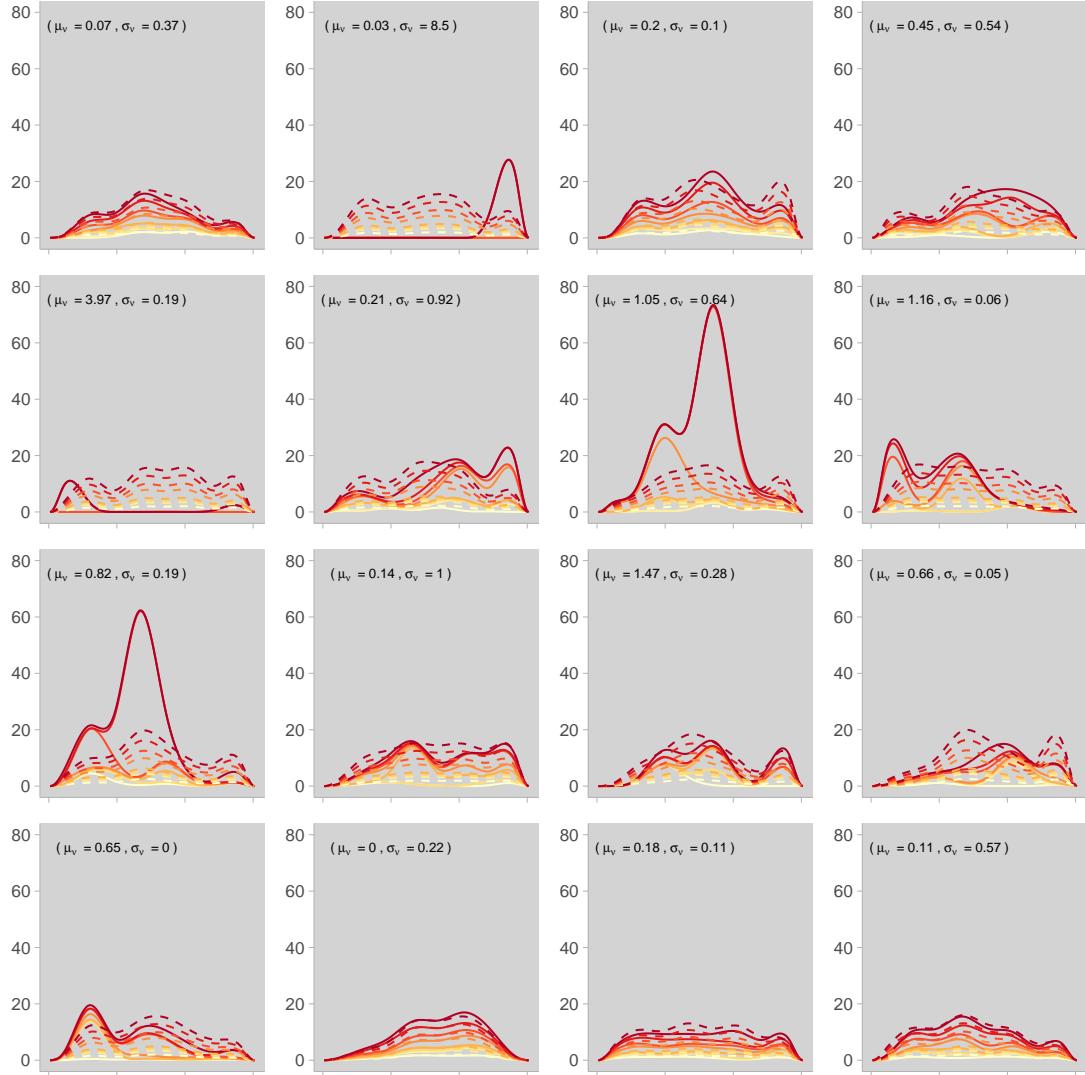


Figure 6.6: Sixteen prior predictive simulations of the average wear profile generated from the linear path (dashed line) and gamma process (solid line) process models with the priors defined in Sec. 6.3. The values of the hyper parameters for ν in the gamma process that were used to generate each simulation are displayed on each plot.

looks synthetic and not natural enough—this should change when I add in the noise layer. In each plot, I have used the same set of realisations of the μ_m to simulate from both the gamma process and linear general path model. The realised values of the hyperparameters of the hierarchical prior for the ν_m used in the gamma process models are displayed in each subplot. Notice that when μ_ν and $\sigma_\nu \rightarrow 0$ the profiles from the gamma process match those from the linear model very closely and when the values of the hyperparameters allow for large values of the ν_m —i.e. when either μ_ν , σ_ν , or both are large—the simulated data from the gamma process model look unrealistic. However, in these simulations, the gamma process model looks to be able to generate the most natural-looking datasets. One thing of note is that many of the simulations, particularly for the gamma process, appear unrealistically wiggly. This ‘wiggliness’ is due to a lack of large-scale spatial structure in the model and is a feature of the postulated model, not the prior. Incorporating large-scale spatial structure could be addressed in future works, but I do not address it here since this behaviour is ‘smoothed out’ when all the possible curves are interpreted as a distribution. All in all, the simulations from both models are in the realms of plausibility, which is what is desired of a weakly informative prior (Gabry et al., 2019b).

6.4 Posterior sampling and inference

Conditioning on the first eight observations, I generate samples from the posterior distributions of each model using the No-U-Turn sampler implemented in *Stan* (Stan Development Team, 2022). In total, I generated 12000 samples from each posterior, using four chains that are 4000 iterations in length with no thinning and a burn-in of 1000 iterations. For the case of the gamma process, sampling results in a small number of divergent transitions—roughly 0.9%—which appear to arise in the hierarchical prior for the same reasons as previously discussed in Section 5.2; however, because of the very limited number of divergent transitions, I do not try and rectify the issue. The linear general path model, on the other

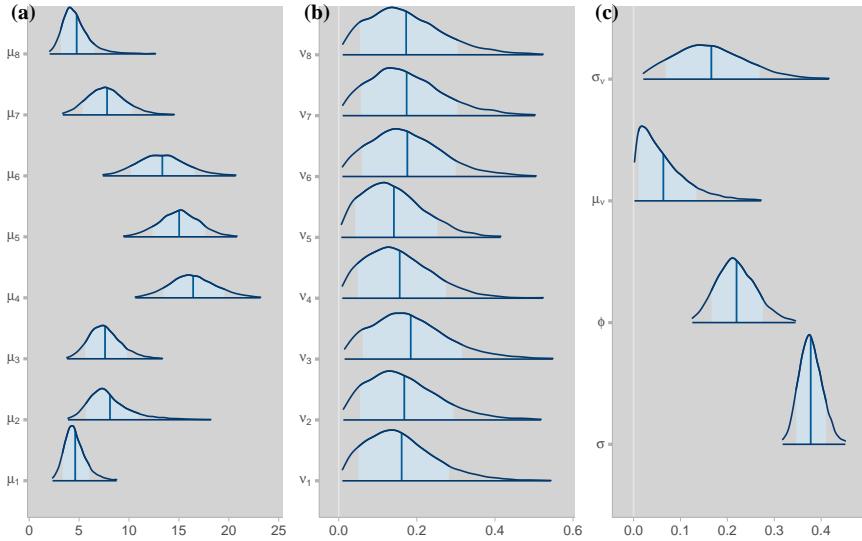


Figure 6.7: Marginal posterior distributions of the gamma process model parameters.

hand, fits very efficiently without any divergent transitions, and both models show good \hat{R} statistics and effective sample sizes for the parameters. In this section, I visually analyse and compare the posterior samples from each model through the marginal posterior distributions of the parameters, the posterior distributions of the intermediate quantities in the models, and the posterior predictive distributions of the UT measurements and underlying wear profiles.

Marginal posteriors of parameters The marginal posteriors of the parameters show our updated belief after conditioning on the data. Essentially, compared with the prior, they show what we have learnt from the data. Additionally, the posterior draws of the parameters are important because we rely on them when forecasting the underlying process into the future in Section 6.5. Figure 6.7 and 6.8 show the marginal posterior distributions of the parameters in the gamma process and linear general path model, respectively.

In the marginal distributions of the parameters of the Gamma process model, shown in Fig. 6.7, we can see three main things. Firstly, by looking at the distribution of the different μ_m , Fig. 6.7 (a), we can see that the model has

successfully captured the belt’s general ‘dishing out’ behaviour since the mean wear rates of the spline coefficients closer to the centre of the belt are higher. Secondly, the general wear of the belt appears to be lop-sided since there is a lack of symmetry in the μ_m —the posterior median of μ_6 is much greater than that of μ_3 . These first two points may be obvious to the reader when looking at the raw data in Fig. 6.1, but it is important that the model has identified the general behaviour as chronic and not just noise, especially when our main goal is to forecast the wear profile through time. The third main observation is how little variability there is amongst the ν_m in Fig. 6.7 (b). Their expected values are very similar, and the marginal posterior of σ_ν , shown in Fig. 6.7 (c), has a lot of mass near zero. However, it is hard to say whether or not the variability is negligible.

In addition to these three main points, in Fig. 6.7 (c), it looks as though the marginal posteriors of the two variance parameters σ and ϕ and the two hyperparameters of ν encode similar levels of uncertainty into the model. However, this is not the case, even though the marginal posteriors have similar scales. σ can be interpreted directly in mm , but the effect of ϕ is scaled by the values of the filtered spline coefficient and should rather be interpreted as a proportion of y^* . The influence of the posterior distributions of σ and ϕ is much more obvious in the posterior predictive distributions. On the other hand, the uncertainty in the degradation path that results from the underlying gamma process and the posterior distribution of its parameters and hyperparameters— μ_ν , σ_ν , ν_m , and μ_m —is almost impossible to interpret from the marginal distributions in Fig. 6.7. We gain a better intuition of the uncertainty in the underlying gamma degradation process by looking at the marginal posterior of the filtered spline coefficients $y_{i,m}^*$. I look at the posterior distributions of these intermediate quantities and the posterior predictive distributions next, but first, I compare the marginal posteriors from the fitted general path model to the marginal distributions from the gamma process that I have looked at here.

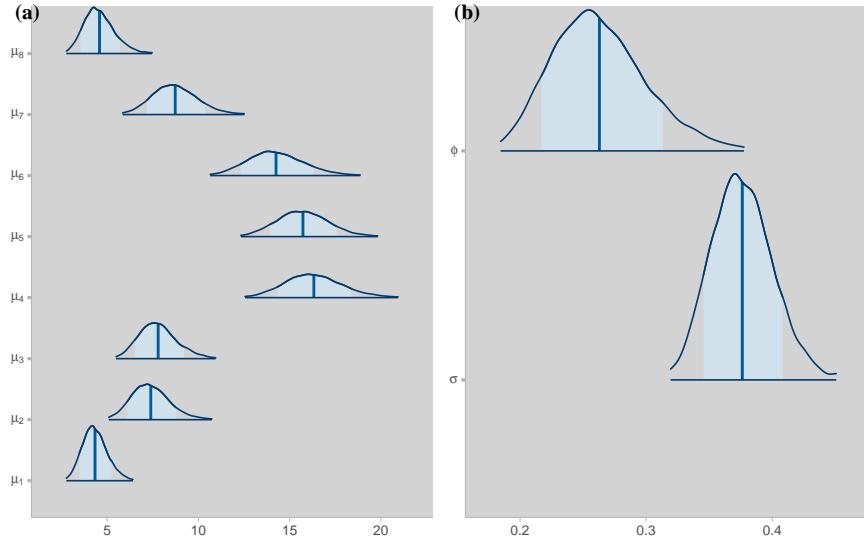


Figure 6.8: Marginal posterior distributions of the linear general path model parameters.

Figure 6.8 shows the marginal posteriors of parameters in the general path model. The marginal posterior distributions of the μ_m from the general path model in Fig. 6.8 (a) show the same general dishing out behaviour of the belt and asymmetric wear pattern. In fact, comparing Fig. 6.8 (a) with Fig. 6.7 (a), it appears that the posterior distribution of the μ_m are almost identical in the two models. The only observable difference is that the marginal distributions have heavier upper tails in the posterior of the gamma process model. In the marginal posteriors of the parameters σ and ϕ in Fig. 6.8 (b), the estimated value of sigma and corresponding uncertainty is also very similar to the gamma process model; however, the estimated value of ϕ is clearly higher. As expected, the scale of the t_{10} distribution in the parameter model has inflated in the case of the general path model to account for the variability in the degradation rate that would otherwise be accounted for by the jumpy gamma process. The distinction between the two process models is clearest in the posteriors of the intermediate quantities in the two models.

Intermediate quantities The $y_{i,m}^*$ are the filtered degradation paths of each spline coefficient, which essentially describe the ‘average’ wear along the length of the belt at each time. The intermediate quantities in the model are treated similarly to parameters in the Bayesian framework; as such, we also obtain posterior draws of the $y_{i,m}^*$ during MCMC sampling. Figure 6.9 shows the mean posterior value of the filtered degradation of each coefficient—plotted as coloured lines. Fig. 6.9 (a) shows the posterior of the gamma process model, and Fig. 6.9 (b) shows the posterior of the general path model. The colours of the mean path correspond to the basis functions in Fig. 6.3 (c). One hundred individual draws from each of the joint posteriors of the y^* are also plotted in Fig. 6.9 as dark grey lines.

The draws of the spline coefficient from the gamma process mode in Fig. 6.9 (a) are noticeably ‘jumpy’ while the mean paths of the filtered spline coefficients look reasonably linear. For the general path model in Fig. 6.9 (b), on the other hand, all of the posterior draws are perfectly straight lines. Despite this difference, the spread of the draws of the pathways in each equivalent subplot in Fig. 6.9 (a) and (b) are very similar. The fact that most of the draws from the gamma process appear as jumpy processes—even though we specified a prior which favours straight processes—seems to suggest that the wear rate varies through time, and as such, we should allow the model to do so too. The general path model has accounted for this added variability in the data by inflating the value of ϕ , as we previously noted. This difference is why some argue that the gamma process is a more realistic and physically motivated model (Ye & Xie, 2015). However, the simplification of reality that linear models apply is notoriously successful. To understand if the simplified linear model is flexible enough to sufficiently describe the data and if the data contain enough information and a clear enough signal to identify the gamma process with its extra parameters, I look at the posterior predictive distributions for the two different models and compare them with the observed data.

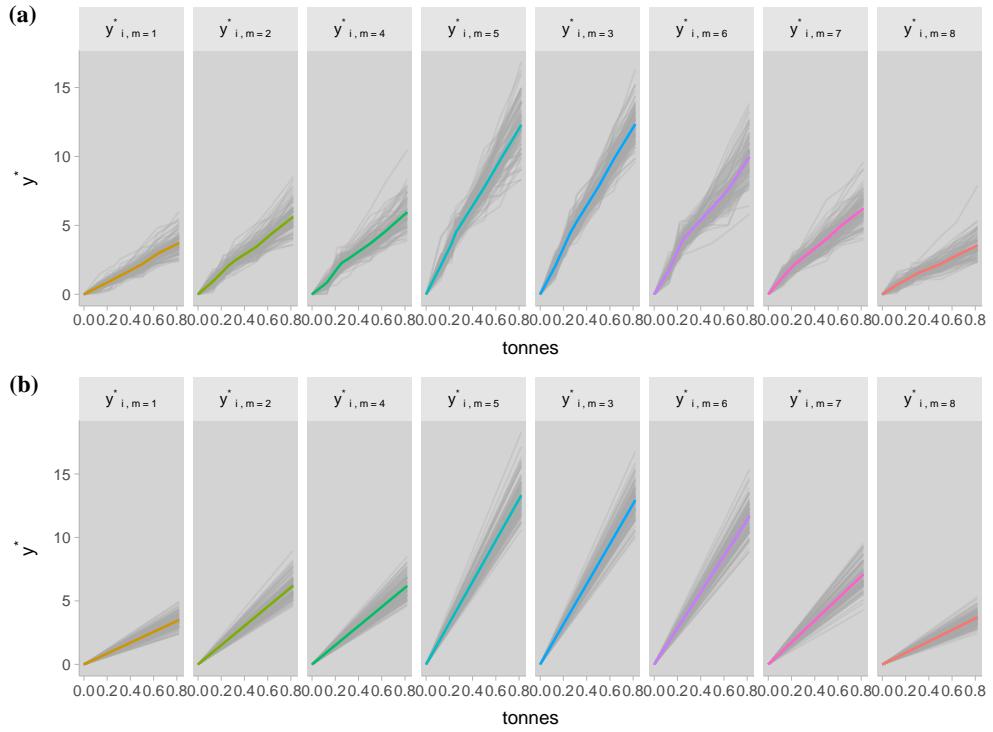


Figure 6.9: The posterior draws of the filtered spline coefficients for (a) the gamma process model and (b) the linear general path model.

Posterior predictive distributions As discussed in section 1.4, a method for checking the fit of Bayesian models is to check if the data look plausible under posterior predictive distribution. In hierarchical models, this comparison can be performed at different levels of the model to check the fit of the data model and process models. Here, I generate a compare the posterior predictive distributions of replications of the UT measurements at the same times and locations along the belt as the data in Fig. 6.1 as well as the predictive distribution of replications of the wear profile at different locations along the belt’s length at the same observation times.

To generate posterior predictive distributions for replications of the UT measurements, \tilde{z} , I sample from the data model conditional on the draws of the spline coefficients and σ

$$\tilde{z}_{i,n} | \underline{y}_i^s, \sigma^s \sim N(f_i^s(n), \sigma^s). \quad (6.3)$$

The posterior predictive distribution is generated in the same way for both models. The joint posterior predictive distribution of \tilde{z} from the gamma process model is shown in Fig. 6.10 (a), and from the general path model in Fig. 6.10 (c). The observed data are also plotted in both sub-plots for comparison. At the noisy UT observation levels, the prior predictive distributions of \tilde{z} for both models are very similar and fit the observed data well. In both cases, all of the observed UT data sit very close to the median wear profiles (showing that the models are flexible enough to fit the data) and sit neatly within the 95% uncertainty intervals.

To generate the posterior predictive distribution for replications of the measured wear profiles at each observation time, I sample new values of the noisy spline coefficients from the first level of the process model conditioned on the posterior draws of the filtered (mean) values of the spline coefficients and ϕ ,

$$\tilde{y}_{i,m} | y_{i,m}^{*s}, \phi^s \sim t_{10}(y_{i,m}^{*s}, \phi^s y_{i,m}^{*s}), \quad (6.4)$$

and then calculate the values of the spline functions $f_i(n) = \sum_{m=1}^M b_m(n) \tilde{y}_{i,m}$. This process is also the same for both models. The posterior predictive distribution of the wear profiles at each observation time generated from the posterior of the gamma process is shown in Fig. 6.10 (b); Fig. 6.10 (d) shows their posterior predictive distribution generated from the general path model. In Fig. 6.10 (b) and (d), I plot each observation time separately so that the 95% uncertainty intervals are clear. In all sub-plots, the observed data are included for comparison.

In the case of the posterior predictive distribution of \tilde{y} , the average wear profiles of the belt at each time are now clearly monotonic increasing and the uncertainty around the average wear profile also grows with time. In both figures 6.10 (b) and (d), the observed noisy UT wear profiles sit inside the uncertainty intervals, showing that the observed data are also plausible under both Bayesian hierarchical models at the process model level, but comparing Fig. 6.10 (b) and (d), the linear model appears to predict a greater wear rate and larger uncertainty than the gamma process model.

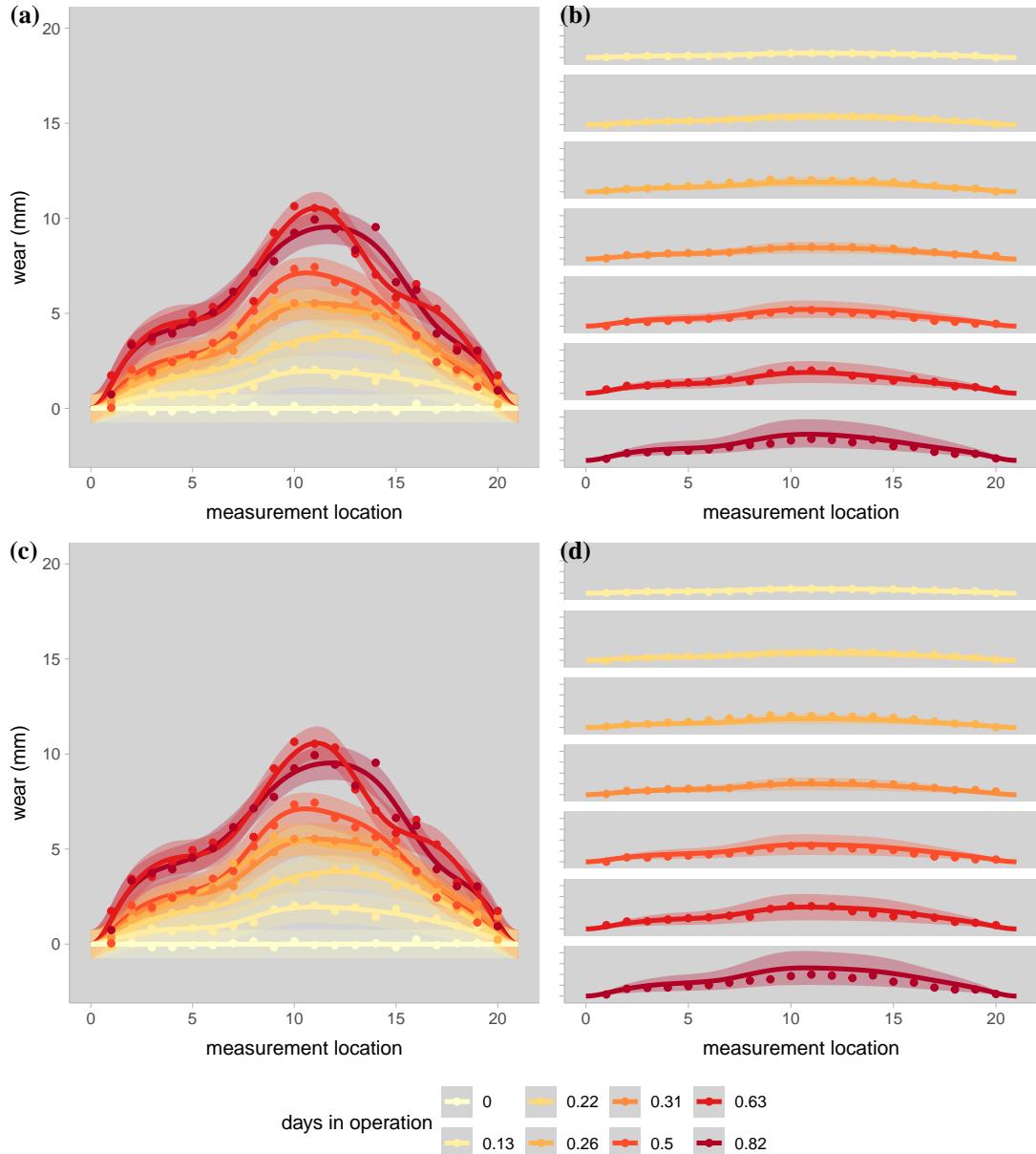


Figure 6.10: Four posterior predictive distributions, two form each model. (a) and (c) shows the predicted smooth functional observation of the wear profiles underlying the sets of UT measurements from the gamma process and general path model, respectively. The UT measurements are also plotted as points as a reference. (b) and (d) shows the posterior predictive distribution of a new observation of the wear profile along the length of the belt at each observation time for the gamma process and general path model, respectively. In (b) and (d), the observed wear profiles are also plotted for comparison.

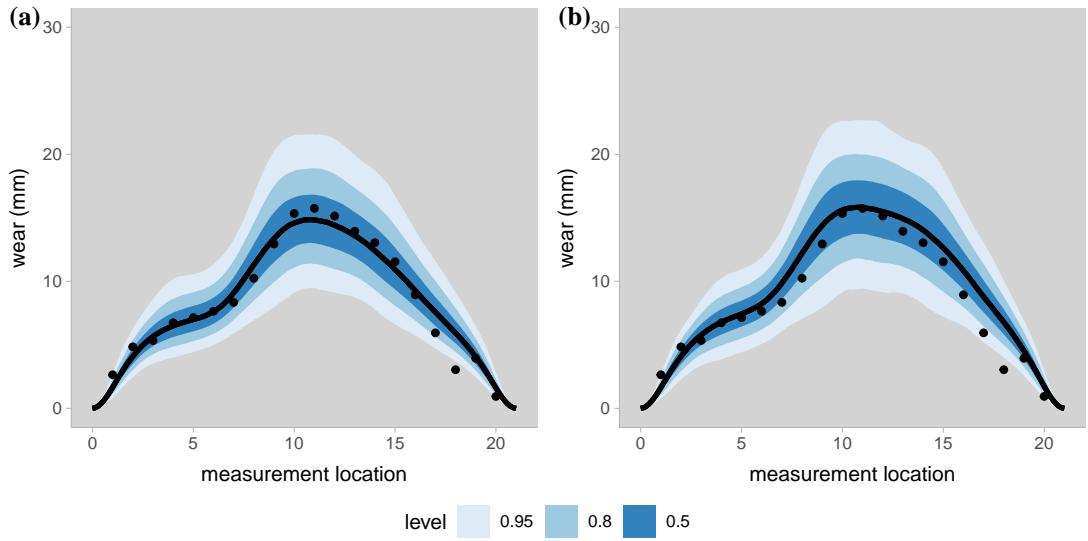


Figure 6.11: Forecasts of the ninth, withheld, observation from (a) the gamma process model and (b) the linear general path model. The true, observed, measurements are plotted for reference.

6.5 Forecasting degradation curves

Using the posterior draws from the two posteriors, we can forecast the degradation process of the spline coefficients using the process model and predict the belt’s wear profile at any time, t_{I+1} , in the future, conditioned on the belts current state of degradation. When producing the forecast, we should do so along the entire belt’s length since we want to predict when any part of the belt will exceed the soft failure threshold. In other words, we want to predict the distribution of the noisy spline coefficients at the forecast time t_{I+1} . The forecasts to the time $t_9 = 1$ of the ninth (withheld) belt wear observation are shown in Fig. 6.11 for the gamma process model (a) and general path model (b) (approximately 0.18 tonnes or 3.7 weeks into the future). In each sub-plot, the median forecast is shown as a solid black curve, and the 0.50, 0.80, and 0.95 uncertainty intervals are shown in various shades of blue. The observed UT measurements at t_9 —withheld from the model—are also plotted in each figure for comparison.

The process of generating the two forecasts in Fig. 6.11 are quite different. For the gamma process model, I first sample jumps of the spline coefficients from

the gamma distribution

$$\Delta y_{I+1,m}^* | \nu_m, \nu_m, z \sim Ga \left(\frac{t_{I+1} - t_I}{\nu_m^2}, \frac{1}{\mu_m, \nu_m^2} \right) \quad (6.5)$$

using the posterior draws of the gamma process parameters. I then add these jumps to the filtered values of the spline coefficients at time t_I , $\{y_{I,m}^*\}_{m=1}^M$, to predict the values of the filtered spline coefficients at time t_{I+1} . Lastly, I average over the spatial variability in the wear profiles along the length of the belt by sampling values from

$$y_{I+1,m} | y_{I+1,m}^*, \phi, z \sim t_{10}(y_{I+1,m}^*, \phi y_{I+1,m}^*). \quad (6.6)$$

For the general path model, I simply calculate the values of the filtered splice coefficients using the deterministic wear function $y_{I+1,m}^* = \mu_m \times t_{I+1}$ and then average over the spatial variability in the wear profiles along the length of the belt in the same way as I did for the last step of the gamma process forecast.

Despite the difference in how the two forecasts are produced, they both appear to reasonably predict the observed data at $t_9 = 1$. For both distributions, the observed data sit close to the median and comfortably within the uncertainty intervals. However, like for the posterior predictive distributions, the linear general path results in a slightly higher prediction of the wear and slightly wider uncertainty intervals. In the next section, I compare both models based on their predictive performance in terms of the whole wear profile and in terms of the maximum wear observation.

6.6 Comparison of methods

To distinguish between the two methods, I compare them based on both their ability to predict the whole wear profile at N steps ahead using the expected log scores and also visually on their ability to predict the maximum wear measure-

ment using a resampling technique similar to bootstrapping and cross validation. In the latter, I also compare the predictions of the method described in Webb et al. (2020). The first of these comparisons is mechanistically the same as elppd_{cv} , except that in this case it is unclear what predictive density I am trying to approximate and I reuse the withheld wear profiles as test sets more than once. Therefore, I refer to the method as the expected log score rather than elppd_{cv} , but for the purpose of model comparisons, the method is the same as that described in Sec 1.4.

Expected log score First I compare the two models based on the expected log probability of the spline coefficients of a withheld observation under the forecasted distribution. I fit the models to a portion of the data (i.e. observations 1:5, 1:6, 1:7, and 1:8) and then evaluate the expected log score of the withheld future wear profiles. To calculate the log score, I estimate the underlying spline coefficients of the withheld observation using maximum likelihood, assuming that $\sigma = 0.38$ (the median estimate from both models). I then calculate the expected log score of the set of spline coefficients under the distribution $t_{10}(\tilde{y}_{m,I}, \phi \tilde{y}_{m,I})$, using the eq. 1.12, as I would to calculate $\text{elppd}_{\text{LOO-cv}}$. The expected log scores of the N-step-ahead predictions are presented in table 6.1. The summations of the expected log score (similar to an elppd_{cv} measure) are presented in the final row of table 6.1.

In the expected log scores, the linear general path model outperforms the gamma process model in every scenario, even for the case of the forecasts shown in the previous section (Fig. 6.11). So, it appears that the linear general path model is a better model for predicting the overall wear profile at future times. However, since the main purpose for modelling the belt wear data is to predict the soft failure of the belt, reliability practitioners may be willing to forgo some accuracy in the prediction around the edges of the belt—where the wear rate is slow—for a more accurate prediction of the maximum wear.

Table 6.1: The expected log score (ELS) for each model when fitting to a portion of the data and predicting n-steps ahead. I is the maximum observation that the model was fit to and $I+1$ is the withheld observation that the forecast is generated for. The summation of the elppd scores are displayed at the bottom of the table.

I	$I + 1$	$ELS_{gamma\ process}$	$ELS_{linear\ path}$
5	6	-15.304	-14.543
5	7	-14.275	-13.268
5	8	-16.124	-15.156
5	9	-14.369	-13.724
6	7	-12.919	-12.207
6	8	-13.172	-13.142
6	9	-12.177	-11.928
7	8	-13.412	-13.287
7	9	-11.767	-11.480
8	9	-11.084	-10.582
-134.603		-129.317	

Test quantity An additional way of scrutinising the predictive performance of the two models is to choose a test quantity $T(z)$ (Gelman, Carlin, et al., 2020, p. 145) to evaluate the predictions solely on an aspect of the forecast that is important for the decision we are trying to inform; which in this case is the maximum wear measurement. To do so, I re-fit the models to all possible combinations of five, six, seven, and eight observations and predict from the most recent observations to the withheld future observations. For example, in one combination, I fit the models to the observations at $(t_1, t_3, t_4, t_5, t_6, t_7)$ —leaving out t_2 , t_8 and t_9 —and then predict the wear profile at t_8 and t_9 . I drop out observations in this way to remove the effect of individual observations; a similar concept to resampling techniques such as bootstrapping. I then calculate the predictive distributions of the maximum wear measurement from the two forecasts and compare them with the observed maximum wear measurement. I also compare the point estimate method of Webb et al. (2020) demonstrated in Fig. 6.2.

To generate the predictive distributions for the maximum wear measurement,

I first average over the UT measurement error by sampling from

$$\tilde{z}_{n,I+1} \sim N \left(\sum_{m=1}^M b_m(n) y_{m,I+1}, \sigma \right) \quad (6.7)$$

using the forecasted joint distribution of the $\{y_{m,I+1}\}_{m=1}^M$ and the posterior draws of σ . I then calculate $\text{Max}(\{z_{n,I+1}\}_{n=1}^N)$ for each set of draws. I average over the UT measurement error now because, in this case, we are directly comparing the prediction with the noisy observations. The predictive distributions are compared with the observed maximum UT measurements and the method of Webb et al. (2020) in Figure 6.12. Each subplot shows the the predictive distribution of the gamma process in blue and the general path model in red as well as the observed max wear measurement as a black vertical line and the point estimate of the max wear generated according to Webb et al. (2020) as a vertical red line. The title of each plot shows a vector indicating which functional observations the models were fit to and a number indicating the forecast time. For example, '[1, 0, 1, 1, 1, 1, 1, 0, 0] → 8' indicates that the models were fit to the observations at $(t_1, t_3, t_4, t_5, t_6, t_7)$ and that the forecasts are generated for t_8 .

The comparisons of Figure 6.12 show that the point estimates of Webb et al. (2020) and the maximum a posteriori (MAP) estimates of the Bayesian models are similar, and both reasonably predict the maximum wear measurement. However, the Bayesian MAP estimates are more robust when fewer observations are used to produce the forecasts. Interestingly, the point estimate of Webb et al. (2020) is typically closer to the MAP of the gamma process model than the linear general path model. Comparing the two predictive densities, the gamma process degradation model is much more optimistic than the linear general path. It also looks like it is generally as good or better than the linear general path model at predicting the maximum wear. In general, the observed maximum wear measurement is always contained in both the predictive distributions; with the exception being observation 8. A closer look at the eighth observation shows

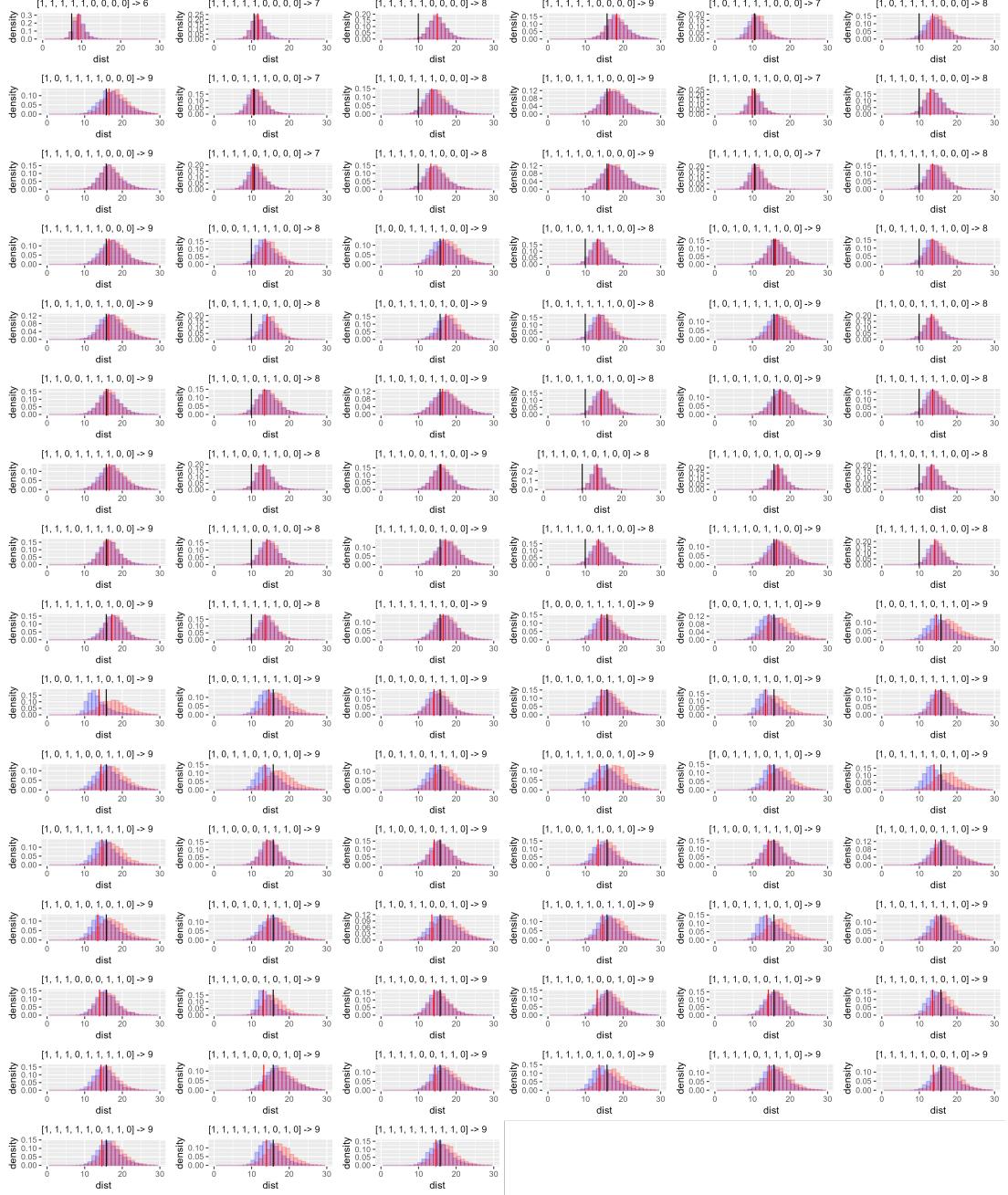


Figure 6.12: Comparisons of the forecasted maximum wear according to the two Bayesian models (the blue histogram shows the posterior predictive distribution from the gamma process and the red from the linear general path) and the method of Webb et al. (2020) (red line) with the observed maximum wear (black line). In each plot, the different methods are fit to a subset of the data and forecast a withheld future observation. For example, the plot title ' $[1, 0, 1, 1, 1, 1, 1, 0, 0] \rightarrow 9$ ' indicates that the model was fit to the observations at $(t_1, t_3, t_4, t_5, t_6, t_7)$ —leaving out t_2 , t_8 , and t_9 —and that the forecast is generated for t_8 .

that it sits below the seventh observation in many places along the profile, even though it is far from the seventh observation in tonnage. Hence, the eighth observation may well be an extreme outlier. As a reliability practitioner, if you were being conservative, you would use the predictions of the linear general path model. However, if reducing maintenance cost was important, then effort should be spent to properly validate the two methods since if the gamma process model is in fact a better representation of the true data generating mechanism then making decisions based on the linear general path model's predictions would result in the belt being replaced while it still has remaining useful life and, hence, overspending on maintenance.

6.7 Failure time distributions

The original motivation for forecasting belt wear was to inform practitioners decisions about when to replace the belt. To facilitate this decision, I use the fitted models to generate failure time distributions for the belt conditioned on its current state of degradation. I demonstrate calculating the failure time distributions from the seventh and eighth observation since at the time of the ninth observation there is a non-negligible probability that some point along the belt has already exceeded the soft failure threshold of $25mm$ —this is reflected in the max wear predictions for the ninth observation time in Figure 6.12—and so the failure time distribution is not very helpful. The failure time distribution describes our uncertainty about when the belt will reach the soft failure threshold (the first passage time). Note that unlike the failure time distributions in Chapter 5, in this case, the noise in the degradation model for the spline coefficients is important. For some situations, the failure time distribution $F(t)$ has an analytical solution, however, there may be no analytical solution for more complicated processes, and so the failure time distribution must be simulated using Monte Carlo evaluation (Meeker et al., 2022, p. 504-506). In the case at hand, multiple processes are driving the degradation of each spline coefficient and we must average over the spatial variability along

the length of the belt; hence, there is no straightforward analytical solution, and I must simulate the failure time distribution.

Algorithm 1 Numerical procedure for calculating the failure time distribution conditional on the fitted gamma process model and current state of degradation.

```

1: for each posterior draw do
2:   for  $j = 1$  to 1000 do
3:     Starting from most recent filtered spline coefficients.
4:      $\Delta t \leftarrow 0.001$ ,  $t \leftarrow t_I$ ,  $\underline{y}^* \leftarrow \{y_{n,I}^*\}_{m=1}^M$ 
5:     Average over the variability along the length of belt.
6:      $\underline{y} \sim t_{10}(\underline{y}^*, \underline{y}^* \phi)$ 
7:     Calculate the value of B-spline by multiplying the design matrix by
       column vector of coefficients.
8:      $\underline{z} \leftarrow B \cdot \underline{y}$ 
9:     while  $\text{Max}(\underline{z}) < 25$  do
10:     $\Delta y_m^* \sim Ga\left(\frac{\Delta t}{\nu_m^2}, \frac{1}{\nu_m^2 \mu_m}\right)$ 
11:     $t \leftarrow t + \Delta t$ ,  $\underline{y}^* \leftarrow \underline{y}^* + \{\Delta y_m^*\}_{m=1}^M$ 
12:     $\underline{y} \sim t_{10}(\underline{y}^*, \underline{y}^* \phi)$ 
13:     $\underline{z} \leftarrow B \cdot \underline{y}$ 
14:  end while
15:   $FT[j] \leftarrow t$ 
16: end for
17: Calculate the empirical cdf from  $FT$ .
18: end for

```

To calculate the failure time numerically from the gamma process model I use the procedure in *Algorithm 1*. For each of the 12000 posterior draws, I simulate 1000 pathways from the most recent observation until soft failure. Each pathway is simulated by incrementing the GP forward by time steps of 0.001 and averaging over the variability along the length of the belt until the degradation at some point exceeds the soft failure threshold. Using the times at which each of 1000 simulated pathways cross the soft failure threshold, I construct an empirical failure time CDF. The result of *Algorithm 1* is 12000 empirical CDFs—one for each of the posterior draws. I construct the failure time distribution from the general path model in the same way, except that when I calculate the filtered values of the spline coefficients, I do so using the deterministic degradation function $\underline{y}^* = \underline{\mu}t$. The distributions of the CDFs are shown in Figure 6.13 for the gamma

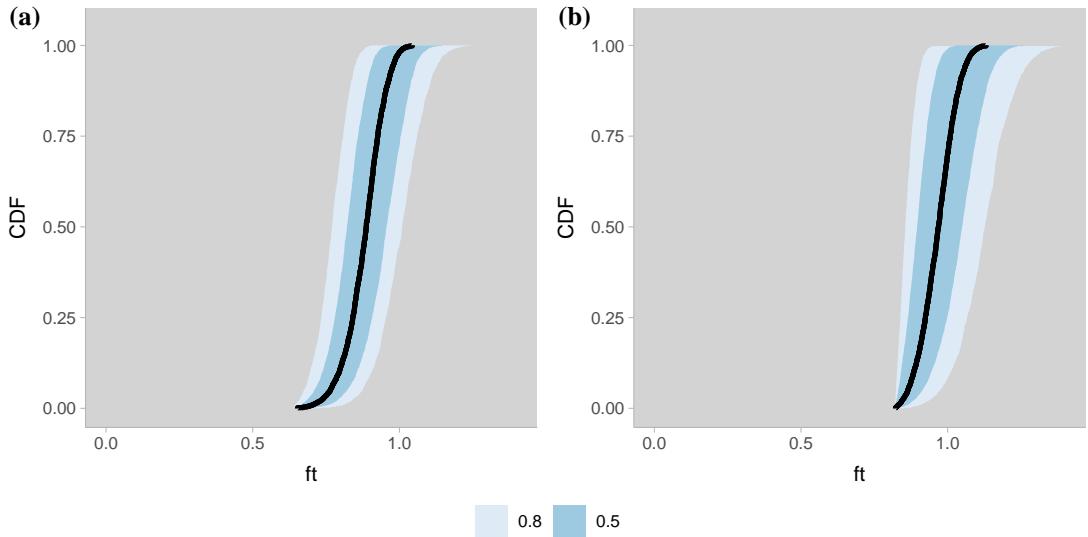


Figure 6.13: The failure time CDFs generated from the posterior of the gamma process model conditioned on the first (a) seven and (b) eight observations. The median CDF is indicated by the black line and the uncertainty intervals are shown by the different shades of blue ribbons.

process and Figure 6.14 for the general path model fitted to the first seven observations (Fig. (a) and 6.14 (a) respectively) and for the first eight observations (Fig. 6.13 (b) and 6.14 (b) respectively). The average CDF is shown as a black line and the 0.5, 0.8, and 0.95 uncertainty intervals in different shades of blue. From the distributions of the failure time CDFs, we can quickly interpret the risk of the soft failure of the belt as we delay the replacement time.

Remembering that maintenance decisions are made in the context of the whole business, not just the specific asset; a reliability engineer may plan to replace belts when there is a 25% chance that they have exceeded the soft failure threshold—remembering that soft failure means that the belt is still technically operational. At the time of the seventh observation, for this particular belt that I have analysed, the predicted time from the gamma process (linear general path) model that there is a 25% chance of soft failure is at 0.82(0.81) tonnes with a lower and upper bound of 0.75 and 0.95(0.74 and 0.90) respectively. Say there are two planned maintenance shutdowns for the particular conveyor coming up; a reliability practitioner can use this distribution to inform which shutdown they

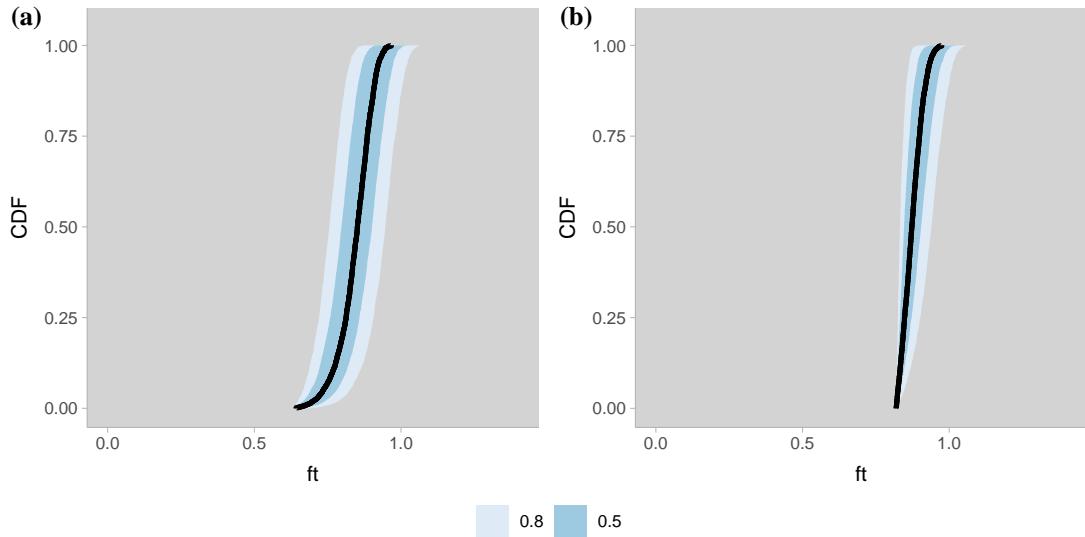


Figure 6.14: The failure time CDFs generated from the posterior of the linear general path model conditioned on the first (a) seven and (b) eight observations. The median CDF is indicated by the black line and the uncertainty intervals are shown by the different shades of blue ribbons.

should replace the belt in. The uncertainty intervals can also help to select one belt over another. Furthermore, since the failure time prediction is with respect to tonnes, the reliability practitioner could look at reducing (or even increasing) the planned operation of the belt to manage the probability of soft failure at the calendar time of the maintenance shutdown. Once we observed the eighth set of measurements, our belief is then updated, and the estimate for the time at which there is a 25% chance of soft failure becomes 0.95 (0.85) tonnes, and the lower and upper uncertainty bounds become 0.83 and 1.09 (0.83 and 0.91), respectively. Using this updated belief, the reliability practitioner could re-evaluate their decision and make last-minute changes to a shutdown plan while formally managing risk.

6.8 Discussions

In this chapter, I've constructed, fit, evaluated, and compared two BHM's for conveyor belt wear, and, in doing so, demonstrated an end-to-end example of the

Bayesian workflow for an applied problem in the mining industry. In the data models for the two BHMs, I extend FDA to degradation modelling in order to model a degrading surface. In the two process models, I compared the noisy gamma process model from chapters 4 and 5 with a linear general path model. Lastly, in the parameter models, I show how historical belt wear data can be used to inform the analysis of the current belt through an informative prior. I compare the two models with one another through the expected log score of N step ahead predictions and also compare them alongside the method of Webb et al. (2020) in their ability to predict the maximum wear measurement. Lastly, I've shown how to construct failure time distributions based on the current degradation of the belt for the two BHMs. In this last section, I distil the main points of the chapter, discuss the advantages and limitations of the BHM models I have explored and point to areas of future work.

The comparison of the two Bayesian models shows that while both models appear to have reasonably fit the data, the linear general path model is a better predictive model for the overall belt's wear. However, when the two models are compared based on their ability to predict the maximum wear observation, which defines the soft failure of the belt, both models' predictions appear reasonable when compared with the observed data and the gamma process model's forecasts are far more optimistic. Consequently, this is also reflected in the failure time distributions generated from the two models. If these models were to be implemented in practice, it would be worth investing in collecting more detailed data for a short period of time to properly validate the models—i.e. collecting wear profiles more frequently and measuring more than one location along the belt's length at each time.

An added advantage of the BHM structure applied to belt wear is that the structure can easily incorporate additional observations into the model without the model becoming over-parameterised and can take full advantage of such extra information to reduce the uncertainty of parameter estimates and predictions.

This is particularly true for the noisy gamma process model. For example, if we were to have a much finer grid of measurements across the width of the belt’s surface at each observation time, these measurements would still be summarised by the same number of spline coefficients. So, the result would be better uncertainty quantification of σ with no additional parameters in the lower levels of the hierarchical model. Alternatively, if more than one functional observation was recorded at each observation time, this could be incorporated by drawing more than one realisation of the noisy spline coefficients. To elaborate, instead of a single set of noisy spline coefficients at each time, we could use $y_{j,i,m} \sim N(y_{i,m}^*, y_{i,m}^* \phi)$, where the new subscript j identifies the different functional observation at time t_i . The result would be better identification of ϕ —the spatial variation in wear profiles along the length of the belt—and subsequently a more precise filtered estimate of the $y_{i,m}^*$. Lastly, if observations were collected more frequently, then the finer temporal resolution would better identify the ‘jumpiness’ of the gamma process—i.e. ν would be estimated more precisely—which refines the uncertainty quantification of the forecasts from the gamma process.

An alternative extension of the data would be if the set of UT measurements at each time were two-dimensional. Here, we have shown an application of this model to a one-dimensional surface; however, the method could easily be extended to a two-dimensional wearing surface by using two-dimensional spatial basis functions such as in Wikle et al. (2019, p. 84). For the case of belt wear, we could extend the model in this way if there were multiple functional observations at each time, and we knew the location along the length of the belt of each observation. By expanding the model to two dimensions, the method could be applied to the degradation of many other assets—for example, the wear liners in transfer station shoots or haul truck beds.

As I touched on briefly when discussing the prior predictive checks in Section 6.3, I have made the simplifying assumption that the spline coefficients are independent of one another. In doing so, we neglect any large-scale spatial struc-

ture in the wear profiles. For both prior and posterior predictive checks, if we simulate a single belt profile, it will look unrealistically ‘wiggly’. Although the B-spline accounts for small-scale spatial structure in the UT measurements, the assumed independence of the spline coefficients means that there is no way to capture any large-scale spacial structure. In reality, we would expect neighbouring coefficients to behave somewhat similarly. A possible way of accounting for this large-scale spatial correlation would be to include a spatial random effect, which would most likely result in better uncertainty quantification in the parameter estimates and forecasts since we are adding information about the underlying process through the model’s structure. One hurdle to implementing a spatial random effect in the gamma process model is there is no straightforward way of coercing correlation in the jumps of multiple gamma processes. This would be an interesting area for future work. Because of this hurdle, it may be simpler to implement large-scale spatial structure in the linear model. This could be done using a conditional autoregressive structure. Nevertheless, the simplifying assumption of independence appears acceptable since the ‘wiggliness’ of the individual realisations is ‘washed out’ when I average over all of the posterior draws, making any predictions look smooth, such as in Fig. 6.11.

When confronted with analysing complicated, small, and messy datasets—something very common in reliability and condition monitoring—a very natural approach is to simplify the data and apply methods we are familiar with, such as regression. However, here we have demonstrated that if we instead take the time to think deeply about how the data arise and what extra knowledge we possess about the data-generating process, we can construct statistical models that take full advantage of all the information available in both the data and our understanding of the problem. In doing so, we get more detailed predictions and defensible uncertainty qualification for the reliability predictions and accompanying maintenance decisions.

Chapter 7

Discussions

7.1 Overview and discussions

7.2 Strengths and limitations

7.3 Future directions

7.4 Implications for industry practitioners

Appendices

Appendix A

Appendix Title

Appendix B

Copyright Information

References

- Abdel-Hameed, M. (1975). A gamma wear process. *IEEE transactions on Reliability*, 24(2), 152–153. (Publisher: IEEE)
- Bae, S. J., Kuo, W., & Kvam, P. H. (2007). Degradation models and implied lifetime distributions. *Reliability Engineering & System Safety*, 92(5), 601–608. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0951832006000536> doi: <https://doi.org/10.1016/j.ress.2006.02.002>
- Balakrishnan, N., Tsai, C.-C., & C.-T., L. (2017). Gamma degradation models: Inference and optimal design. In D.-G. Chen (Ed.), *Statistical modeling for degradation data* (pp. 171–191). Singapore: Springer Nature Singapore Pte. Ltd.
- Berliner, L. M. (1996). Hierarchical Bayesian time series models. In K. M. Hanson & R. N. Silver (Eds.), *Maximum entropy and bayesian methods* (pp. 15–22). Dordrecht: Springer Netherlands.
- Betancourt, M. (2017). A Conceptual Introduction to Hamiltonian Monte Carlo. *arXiv*. Retrieved from <https://arxiv.org/abs/1701.02434> doi: 10.48550/ARXIV.1701.02434
- Betancourt, M. (2020, June). *Identity Crisis*. Retrieved from https://betanalpha.github.io/assets/case_studies/identifiability.html (Accessed: 31-05-2023)
- Betancourt, M., & Girolami, M. (2015). Hamiltonian monte carlo for hierarchical models. In S. K. Upadhyay, U. Singh, D. K. Dey, & A. Loganathan (Eds.),

Current trends in bayesian methodology with applications (p. 24). Chapman and Hall/CRC. doi: 10.1201/b18502

Betancourt, M. J. (2015). *Adiabatic monte carlo*.

Bortnowski, P., Kawalec, W., Król, R., & Ozdoba, M. (2022). Types and causes of damage to the conveyor belt - review, classification and mutual relations. *Engineering Failure Analysis*, 140, 106520. Retrieved from <https://www.sciencedirect.com/science/article/pii/S1350630722004940> doi: <https://doi.org/10.1016/j.engfailanal.2022.106520>

Cox, D. R., & Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society, Series B*, 49(1), 1–39.

Creagh, B. (2020). *Bridgestone: A partner in products, service and technology*. <https://www.australianmining.com.au/wp-content/uploads/2020/04/bridgestone-Material-size-web-1024x768.jpg>.

Cressie, N. A. C., & Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. Hoboken, N.J: Wiley.

Esposito, N., Mele, A., Castanier, B., & Giorgio, M. (2022, January). A new gamma degradation process with random effect and state-dependent measurement error. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, 1748006X2110672. Retrieved 2022-10-25, from <http://journals.sagepub.com/doi/10.1177/1748006X211067299> doi: 10.1177/1748006X211067299

Frenk, J., & Nicolai, R. (2007, May). *Approximating the randomized hitting time distribution of a non-stationary gamma process* (Tech. Rep. No. 2007-18). Erasmus University Rotterdam, Rotterdam, The Netherlands: Econometric Institute.

Gabry, J., & Mahr, T. (2024). *bayesplot: Plotting for bayesian models*. Retrieved from <https://mc-stan.org/bayesplot/> (R package version 1.11.1)

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (2019a).

- Discussion on the meeting on ‘Data Visualization’. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 182(2), 419–441.
- Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (2019b). Visualization in Bayesian Workflow. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 182(2), 389-402.
- Gelman, A. (2005). Analysis of variance—why it is more important than ever. *The Annals of Statistics*, 33(1), 1 – 53.
- Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2020). *Bayesian data analysis, third edition*. Boca Raton, FL: Taylor & Francis.
- Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.
- Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4), 457 – 472. Retrieved from <https://doi.org/10.1214/ss/1177011136> doi: 10.1214/ss/1177011136
- Gelman, A., Simpson, D., & Betancourt, M. (2017). The prior can often only be understood in the context of the likelihood. *Entropy*, 19(10). Retrieved from <https://www.mdpi.com/1099-4300/19/10/555> doi: 10.3390/e19100555
- Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., ... Modrák, M. (2020, November). Bayesian Workflow. *arXiv:2011.01808 [stat]*. Retrieved 2022-03-02, from <http://arxiv.org/abs/2011.01808> (arXiv: 2011.01808)
- Geraerds, W. (1985). The cost of downtime for maintenance: preliminary considerations. *Maint. Manage. Int.*, 5(1), 13–21.
- Gilks, W., Richardson, S., & Spiegelhalter, D. (1996). *Markov chain monte carlo in practice*. Boca Raton, FL: Chapman & Hall/CRC.
- Giorgio, M., Mele, A., & Pulcini, G. (2019, March). A perturbed gamma degradation process with degradation dependent non-Gaussian measurement errors. *Applied Stochastic Models in Business and Industry*, 35(2), 198–210. Retrieved 2022-05-10, from <https://onlinelibrary.wiley.com/>

[doi/10.1002/asmb.2377](https://doi.org/10.1002/asmb.2377) doi: 10.1002/asmb.2377

- Guo, G. (1993). Event-history analysis for left-truncated data. *Sociological Methodology*, 23, 217–243. Retrieved 2024-05-01, from <http://www.jstor.org/stable/271011>
- Hamada, M. S., Wilson, A. G., Reese, C. S., & Martz, H. F. (2008a). *Bayesian reliability*. Springer New York, NY.
- Hamada, M. S., Wilson, A. G., Reese, C. S., & Martz, H. F. (2008b). *Bayesian reliability*. New York, NY: Springer Science+Business Media, LLC.
- Hazra, I., Bhadra, R., & Pandey, M. D. (2022, December). Likelihood-free Hamiltonian Monte Carlo for modeling piping degradation and remaining useful life prediction using the mixed gamma process. *International Journal of Pressure Vessels and Piping*, 200, 104834. Retrieved 2022-11-23, from <https://www.sciencedirect.com/science/article/pii/S0308016122002198> doi: 10.1016/j.ijpvp.2022.104834
- Hazra, I., Pandey, M. D., & Manzana, N. (2020, June). Approximate Bayesian computation (ABC) method for estimating parameters of the gamma process using noisy data. *Reliability Engineering & System Safety*, 198, 106780. Retrieved 2023-02-16, from <https://linkinghub.elsevier.com/retrieve/pii/S0951832019309901> doi: 10.1016/j.ress.2019.106780
- Heng, A., Zhang, S., Tan, A. C. C., & Mathew, J. (2009). Rotating machinery prognostics: State of the art , challenges and opportunities. *Mechanical Systems and Signal Processing*, 23, 724–739. doi: 10.1016/j.ymssp.2008.06.009
- Hodges, J. S. (2014). *Richly Parameterized Linear Models. Additive, Time Series, and Spatial Models Using Random Effects*. Boca Raton, FL: Chapman & Hall/CRC Press.
- Hoffman, M. D., & Gelman, A. (2014). The No U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15, 1351–1381.

- Hong, Y., Meeker, W. Q., & McCalley, J. D. (2009). Prediction of remaining life of power transformers based on left truncated and right censored lifetime data. *The Annals of Applied Statistics*, 3(2), 857–879. Retrieved from <https://doi.org/10.1214/00-AOAS231> doi: 10.1214/00-AOAS231
- Hormann, S., & Kokoszka, P. (2012, 12). Functional time series. *Handb. Stat.*, 30. doi: 10.1007/978-1-4614-3655-3_16
- Huzurbazar, V. S. (1956). Sufficient statistics and orthogonal parameters. *Sankhyā: The Indian Journal of Statistics*, 17(3).
- ISO. (2013). *ISO 12489:2013 petroleum, petrochemical and natural gas industries — reliability modelling and calculation of safety systems*. ISO International Standard. Retrieved from <https://www.iso.org/obp/ui/en/#iso:std:iso:tr:12489:ed-1:v1:en:sec:3.1.8>
- ISO. (2016). *ISO 14224:2016 petroleum, petrochemical and natural gas industries — collection and exchange of reliability and maintenance data for equipment*. ISO International Standard. Retrieved from <https://www.iso.org/obp/ui/en/#iso:std:iso:14224:ed-3:v2:en>
- Jardine, A. K., & Tsang, A. H. (2013). *Maintenance, replacement, and reliability: Theory and applications (2nd edition)*. CRC Press.
- Johnson, A. A., Ott, M. Q., & Dogucu, M. (2022). *Bayes rules! an introduction to applied bayesian modeling*. Boca Raton, FL: CRC Press.
- Kallen, M., & van Noortwijk, J. (2005, November). Optimal maintenance decisions under imperfect inspection. *Reliability Engineering & System Safety*, 90(2-3), 177–185. Retrieved 2022-03-02, from <https://linkinghub.elsevier.com/retrieve/pii/S0951832004002443> doi: 10.1016/j.ress.2004.10.004
- Kaminskiy, M., & Krivtsov, V. (2005, 01). A simple procedure for bayesian estimation of the weibull distribution. *IEEE Transactions on Reliability*, 54, 612-616.
- Kowal, D. R. (2017). *Bayesian methods for functional and time se-*

- ries data (Doctoral dissertation, Cornell University). Retrieved from <https://higherlogicdownload.s3.amazonaws.com/AMSTAT/7328eb46-c57e-49af-8d4f-0460ca4a6a2c/UploadedImages/Kowal.pdf>
- Kundu, D., & Mitra, D. (2016). Bayesian inference of weibull distribution based on left truncated and right censored data. *Computational Statistics & Data Analysis*, 99. doi: <https://doi.org/10.1016/j.csda.2016.01.001>
- Lawless, J., & Crowder, M. (2004, 09). Covariates and random effects in a gamma process model with application to degradation and failure. *Lifetime Data Analysis*, 10, 213–227.
- Leadbetter, R., Caceres, G. G., & Phatak, A. (2024). *Bayesian hierarchical modelling of noisy gamma processes: Model formulation, identifiability, model fitting, and extensions to unit-to-unit variability*.
- Lee, M.-L. T., & Whitmore, G. (2006). Threshold regression for survival analysis: Modeling event times by a stochastic process reaching a boundary [Article]. *Statistical Science*, 21(4), 501 - 513. (Cited by: 215; All Open Access, Bronze Open Access, Green Open Access) doi: 10.1214/088342306000000330
- Lu, C., Meeker, W., & Escobar, L. (1996, 01). A comparison of degradation and failure-time methods for estimating a time-to-failure distribution. *Statistica Sinica*, 6.
- Lu, C. J., & Meeker, W. (1993). Using degradation measures to estimate a time-to-failure distribution. *Technometrics*.
- Lu, D., Pandey, M. D., & Xie, W.-C. (2013, August). An efficient method for the estimation of parameters of stochastic gamma process from noisy degradation measurements. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, 227(4), 425–433. Retrieved 2022-06-08, from <http://journals.sagepub.com/doi/10.1177/1748006X13477008> doi: 10.1177/1748006X13477008
- Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2013). *The*

- bugs book: A practical introduction to bayesian analysis.* Boca Raton, FL: CRC Press.
- McElreath, R. (2020). *Statistical rethinking* (2nd. ed.). Boca Raton, FL: Chapman and Hall/CRC.
- Meeker, W. Q., & Escobar, L. A. (1998). *Statistical methods for reliability data* (1st. ed.). New York, NY: Wiley.
- Meeker, W. Q., Escobar, L. A., & Pascual, F. G. (2022). *Statistical methods for reliability data* (2nd. ed.). Hoboken, NJ: Wiley.
- Mitra, D. (2013). *Likelihood inference for left truncated and right censored lifetime data* (PhD thesis). McMaster University, Hamilton, ON, Canada. (Available at <https://macsphere.mcmaster.ca/handle/11375/12738>)
- Mittman, E. T. (2018). *Applications of bayesian hierarchical models in gene expression and product reliability* (PhD thesis). Iowa State University, Ames, IA, United States. (Available at <https://dr.lib.iastate.edu/entities/publication/7c167cf8-a345-476f-a77d-4fc829e4b2c9>)
- Moore, D. F. (2016). *Applied survival analysis using r* (R. Gentleman, K. Hornik, & G. Parmigiani, Eds.). Springer Cham.
- Neal, R. M. (2011). Mcmc using hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, & X.-L. Meng (Eds.), *Handbook of markov chain monte carlo*. Chapman and Hall/CRC. doi: 10.1201/b10905
- Pandey, M., & Yuan, X.-X. (2006, January). A comparison of probabilistic models of deterioration for life cycle management of structures. In (pp. 735–746). doi: 10.1007/1-4020-4891-2_62
- Peng, C.-Y., & Ai, H.-F. (2018). Degradation analysis with measurement errors. In D. C. et al. (Ed.), *Proceedings of the pacific rim statistical conference for production engineering* (pp. 95–122). Singapore: Springer Nature Singapore Pte. Ltd.
- Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceed-*

ings of the 3rd international workshop on distributed statistical computing (dsc 2003) (pp. 1–10). Vienna.

Pradhan, B., & Kundu, D. (2011). Bayes estimation and prediction of the two-parameter gamma distribution. *Journal of Statistical Computation and Simulation*, 81(9), 1187–1198.

Pulcini, G. (2016, August). A perturbed gamma process with statistically dependent measurement errors. *Reliability Engineering & System Safety*, 152, 296–306. Retrieved 2023-02-02, from <https://linkinghub.elsevier.com/retrieve/pii/S095183201630014X> doi: 10.1016/j.ress.2016.03.024

Ramsay, J., Hooker, G., & Graves, S. (2009). *Functional data analysis with r and matlab*. New York, NY: Springer New York. Retrieved from <https://doi.org/10.1007/978-0-387-98185-7> doi: 10.1007/978-0-387-98185-7

Reich, B., & Ghosh, S. (2019). *Bayesian statistical methods*. Boca Raton, FL: CRC Press.

Robinson, M. E., & Crowder, M. J. (2000, 12). Bayesian methods for a growth-curve degradation model with repeated measures. *Lifetime Data Analysis*, 6.

Rodríguez-Picón, L. A., Méndez-González, L. C., Romero-López, R., Pérez-Olgún, I. J. C., Rodríguez-Borbón, M. I., & Valles-Rosales, D. J. (2021, January). Reliability Analysis Based on a Gamma-Gaussian Deconvolution Degradation Modeling with Measurement Error. *Applied Sciences*, 11(9), 4133. Retrieved 2022-11-23, from <https://www.mdpi.com/2076-3417/11/9/4133> doi: 10.3390/app11094133

Rodríguez-Picón, L. A., Rodríguez-Picón, A. P., Méndez-González, L. C., Rodríguez-Borbón, M. I., & Alvarado-Iniesta, A. (2018, July). Degradation modeling based on gamma process models with random effects. *Communications in Statistics - Simulation and Computation*, 47(6), 1796–1810. Retrieved 2022-11-03, from <https://doi.org/10.1080/03610918>

- .2017.1324981 doi: 10.1080/03610918.2017.1324981
- SAP SE. (2023). *Sap software suite*. <https://www.sap.com>.
- Si, X.-S., Wang, W., Hu, C.-H., & Zhou, D.-H. (2011). Remaining useful life estimation - a review on the statistical data driven approaches. *European Journal of Operational Research*, 213(1), 1-14. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0377221710007903> doi: <https://doi.org/10.1016/j.ejor.2010.11.018>
- Simpson, D., Rue, H., Riebler, A., Martins, T. G., & Sørbye, S. H. (2017). Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors. *Statistical Science*, 32(1), 1 – 28.
- Stan Development Team. (2022). Stan Modeling Language Users Guide and Reference Manual [Computer software manual]. <https://mc-stan.org>.
- Stan Development Team. (2024). Stan User's Guide, Version 2.34 [Computer software manual]. Retrieved from https://mc-stan.org/docs/2_34/stan-users-guide/truncation-censoring.html#censored-data (Chapter: Censored Data)
- Tian, Q., Lewis-Beck, C., Niemi, J. B., & Meeker, W. Q. (2024). Specifying prior distributions in reliability applications. *Applied Stochastic Models in Business and Industry*, 40(1), 5-62. Retrieved from <https://onlinelibrary.wiley.com/doi/abs/10.1002/asmb.2752> doi: <https://doi.org/10.1002/asmb.2752>
- van Noortwijk, J. (2009, January). A survey of the application of gamma processes in maintenance. *Reliability Engineering & System Safety*, 94(1), 2–21. Retrieved 2022-04-18, from <https://linkinghub.elsevier.com/retrieve/pii/S0951832007001111> doi: 10.1016/j.ress.2007.03.019
- Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27(5), 1413–1432.
- Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021).

Rank-Normalization, Folding, and Localization: An Improved \widehat{R} for Assessing Convergence of MCMC (with Discussion). *Bayesian Analysis*, 16(2), 667 – 718.

Webb, C., Sikorska, J., Khan, R. N., & Hodkiewicz, M. (2020). Developing and evaluating predictive conveyor belt wear models. *Data-Centric Engineering*, 1, e3. doi: 10.1017/dce.2020.1

Whitmore, G. (1995). Estimating degradation by a Wiener diffusion process subject to measurement error. *Lifetime Data Analysis*, 1, 307–319.

Wikle, C. K., Zammit-Mangion, A., & Cressie, N. (2019). *Spatio-Temporal Statistics with R*. New York: Chapman and Hall/CRC. doi: 10.1201/9781351769723

Ye, Z.-S., & Xie, M. (2015). Stochastic modelling and analysis of degradation for highly reliable products. *Applied Stochastic Models in Business and Industry*, 31(1), 16-32. Retrieved from <https://onlinelibrary.wiley.com/doi/abs/10.1002/asmb.2063> doi: <https://doi.org/10.1002/asmb.2063>

Every reasonable effort has been made to acknowledge the owners of copyright material. I would be pleased to hear from any copyright owner who has been omitted or incorrectly acknowledged.