HIERARCHICAL GAUSSIAN PROCESS MODELS FOR LOSS RESERVING

ANG ZI QING

UNIVERSITI SAINS MALAYSIA

2021

HIERARCHICAL GAUSSIAN PROCESS MODELS FOR LOSS RESERVING

by

ANG ZI QING

Thesis submitted in fulfilment of the requirements for the degree of

Master of Science

December 2021

ACKNOWLEDGEMENT

I would like to thank my supervisor, Assoc. Prof. Dr. Lee See Keong for his patience, guidance and encouragement thus far. I am grateful to be able to work with a firm believer that the application of mathematical modeling should be backed by a good understanding of the Mathematics behind the machinery. I was given all the freedom I could have wanted, trust, as well as opportunities to sit in Analysis lectures to strengthen my mathematical foundation that was indispensible to study advanced level materials. He has also suggested excellent sources of materials throughout my academic pursuit while I tried to figure out a research direction, which eventually led to this dissertation. For your willingness and open-mindedness to take part in my journey to explore the application of Mathematics in a multidisciplinary project, I owe my thanks to thee.

After gaining the necessary mathematical foundations to study Bayesian modeling, the case studies in this research were started during the early days of the Covid-19 outbreak, which coincidentally I was involved in the development of the CoranaTracker project. My involvement with people from various backgrounds behind the monitor screens in this project has accelerated the process of picking up the skills necessary for data analysis. While studying the SIR model, I have been able to pick up the open source probabilistic programming software, Stan in building a Bayesian statistical model, which I later was able to apply in my research. I am indebted to both the CoronaTracker team as well as the Stan Forum community in this regard.

My appreciation also extends to the staffs in the School of Mathematical Sciences, USM as well as fellow students in the lab, especially when attending lab physically was still possible before the lockdown. They have given me ample support throughout the journey.

Finally, I would like to acknowledge with gratitude for the unwavering support given by my family and friends, especially for the constant reminder and gentle pushes during the final stages of writing.

My study was supported by the Public Service Department of Malaysia under the Yang di-Pertuan Agong scholarship.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Page

LIST OF SYMBOLS

$\mathbf{x}, \mathbf{y}, \mathbf{z}$	Vectors are denoted by bold lowercase.
$\mathbf{A}, \mathbf{B}, \mathbf{C}$	Matrices are denoted by bold uppercase.
$\mathbf{A}^{-1}$	Inverse of a matrix.
$\mathbf{x}^T, \mathbf{A}^T$	Transpose of a vector or a matrix.
$\mathbf{x}^T\mathbf{y}$	Dot product of a vector $\mathbf{x}$ and a vector $\mathbf{y}$.
$\mathbf{A}\backslash\mathbf{b}$	The vector $\mathbf{x}$ that solves $\mathbf{A}\mathbf{x} = \mathbf{b}$.
$\mathbb{N}$	The natural numbers, i.e. positive integers.
$\mathbb{R}$	The real numbers.
$\forall x$	For all x.
$S \to T$	Mapping from S to T.
$\cup, \cap$	Union, intersection of sets.
$\emptyset$	Empty set.
$B \subset A$	B is a subset of A.
$a \in A$	a is an element of A.
$\propto$	Proportional to.
$\sim$	Distributed according to.
$\sum_{i=1}^{N}$	Sum of x_i: $x_1 + \cdots + x_N$.
$\prod_{i=1}^{N}$	Product of x_i: $x_1 \times \cdots \times x_N$.
$P(\cdot)$	Probability measure.
$k(\cdot)$	Kernel function.
$\mathcal{N}(\mathbf{x}\|\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Gaussian (Normal) distributed variable $\mathbf{x}$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.
$\mathcal{GP}$	Gaussian process.

MODEL PROSES GAUSSIAN BERHIERARKI UNTUK RIZAB KERUGIAN

ABSTRAK

Peramalan rizab kerugian merupakan antara aktiviti utama aktuari dalam industri insuran. Amalan ini dilakukan untuk memastikan kesihatan kewangan sesebuah syarikat insuran dalam keadaan baik, dan pada masa yang sama melindungi hak pelanggan syarikat tersebut. Walaupun teknik-teknik yang digunapakai adalah dikawalselia dengan ketat, namun para penyelidik masih giat mencari kaedah yang mampu menambah baik ketepatan dan mengukur ketidakpastian ramalan rizab kerugian ini. Beralih daripada kaedah nisbah kerugian yang merupakan kaedah biasa dalam amalan aktuatri, para penyelidik telah mengkaji kaedah-kaedah yang lain termasuk model berparametrik seperti model lengkung pertumbuhan, model yang berasaskan system dinamik dan juga model tak berparametrik. Wujudnya peningkatan minat para penyelidik dalam bidang ini untuk menggunakan kaedah *Bayesian* dalam pengukuran ketidakpastian ramalan. Perkembangan terkini kajian topik ini melibatkan satu model proses *Gaussian* berhierarki dan bersifat *Bayesian* yang memodelkan hubungan spasial tiga dimensi yang wujud dalam data kerugian. Model ini telah diuji dengan tiga fungsi kernel yang berbeza dan, termasuk penggunaan kaedah ledingan input untuk menguruskan data yang berada dalam keadaan tidak pegun. Pemeriksaan yang lanjut dalam fungsi kernel eksponen berkuasa dua yang bersifat anisotropi dalam kajian ini menunjukkan fungsi kernel tersebut terdiri daripada pendaraban dua fungsi kernel eksponen berkuasa dua yang bersifat pegun. Kedua-dua fungsi kernel dalam pendaraban ini mewakili dimensi input masing-masing. Dalam satu kajian yang lain pula, fungsi kernel yang dihasilkan oleh pendaraban mempunyai kuasa ekstrapolasi yang lebih lemah berbanding dengan fungsi kernel yang dihasilkan melalui tambahan, namun kedua-dua jenis fungsi kernel mempunyai sifat yang berbeza. Kombinasi fungsi kernel telah diperkenalkan dalam kajian tersebut, terutamanya fungsi kernel berdaya tambah penuh yang mengandungi kedua-dua fungsi kernel berdaya darab dan fungsi kernel berdaya tambah. Tesis ini akan menunjukkan kesan menggunakan kombinasi fungsi kernel eksponen berkuasa dua dalam model proses *Gaussian* berhierarki yang bersifat *Bayesian* untuk peramalan rizab kerugian. Walaupun hasil daripada kajian ini tidak menunjukkan kemajuan yang ketara daripada kajian sebelum ini, namun wujud juga bukti yang menunjukkan bahawa kombinasi fungsi kernel mampu memberi lebih struktur untuk model proses *Gaussian* berbanding dengan sebelum ini, apabila digunakan bersama kaedah ledingan input.

HIERARCHICAL GAUSSIAN PROCESS MODELS FOR LOSS RESERVING

ABSTRACT

Loss reserving is one of the main activities of actuaries in the insurance industry and is done to ensure the financial health of companies as well as protecting consumers' interest. Techniques applied by the practitioners are highly regulated, but researchers are still ongoing in the pursuit of finding methods to improve predictive accuracy and to establish a measure of predictive uncertainties. Diverting from the link ratio methods, researchers have experimented with parametric models such as growth-curve models and models involving dynamical systems, as well as nonparametric models. Researchers in this field have increasingly shown interests in utilizing Bayesian methods to measure predictive uncertainties. The latest development in the loss reserving literature involves a hierarchical Bayesian Gaussian Process model that models the three-dimensional spatial relationship in the loss observation, experimented with three stationary kernels with input warping to manage the nonstationarities present in the data. Upon inspection of the anisotropic squared exponential kernel used in the study, we have shown that the kernel consists of a multiplication of two stationary squared-exponential kernel applied on each of the two input spaces, i.e. a multiplicative kernel. In another independent study, it was found that multiplicative kernels have poorer extrapolation abilities compared to additive kernels, but both types of kernels have different properties. In that study, combinations of kernels were introduced, especially the full additive kernel that includes both additive and multiplicative kernels. This thesis presents the case of using various combinations of squared exponential kernels into the hierarchical Bayesian Gaussian Process model for loss reserving. While the results obtained from the simulations do not show a great improvement from the original study that has proposed this model, there are evidence that suggest that the combination of kernels have provided more structure to the Gaussian Process model, along with the usage of input warping.

CHAPTER 1

INTRODUCTION

Loss reserving is one of the main activities of actuaries other than pricing products in the insurance industry to ensure an insurer's financial health as well as to protect consumers' interest. There already exist some typical workflow for this highly regulated practice but practitioners constantly seek for improvements in obtaining better predictions for loss reserves as well as fulfilling the need for measure of uncertainties. While improving predictive accuracy has always been the main focus not only in the insurance industry, researchers in this field have been working on Bayesian models in order to capture predictive uncertainties, whether by parametric or by nonparametric models. Gaussian process regression has been recently introduced into the loss reserving literature and further study was needed to be applied in practice. This thesis focuses on investigating the effect of various combinations of kernels in the loss reserving Gaussian process models.

1.1 The loss reserving problem

The insurance business exists as a result of pooling individuals exposed to certain risks, whether a loss of life or nonlife, and this group of individuals (insured) are consumers of a risk transfer service whereby the insurer receives some return in the form of premiums on its willingness to take on the risks of its customers. It is essentially a sharing of risks since the event may not occur to every individual in the pool. General (nonlife) insurers cover property and casualty risks of their customers (insured), that is, the insurer is liable to any losses suffered by the insured in an event of occurrence of a covered risk during the policy effective period. However, if the covered risk does not occur in an insured's policy effective period, the premium paid by every insured in the pool at the beginning of the policy effective period will be earned by the insurer. In order to make a profit out of this business, the insurer has to charge a premium to its customers that is sufficient to cover the losses when an event happens, to cover operational costs of the company as well as to make a profit for the shareholders. Despite the usual logic of profiting from a business, the insurance business differs slightly from any typical business since the costs of the product is unknown at sale, hence the accounting equation

$$\text{Profit} = \text{Income} - \text{Costs} \tag{1.1}$$

cannot be applied directly. The similar equation in the context of the general insurance industry is as follows:

$$\begin{aligned}
\text{Total Profit} &= \text{Investment income} + \text{Underwriting Profit} \qquad (1.2)\\
&= \Big(\text{Investment income on capital}\\
&\qquad + \text{Investment income earned on policyholder-supplied funds}\Big)\\
&\quad + \Big(\text{Premium} - \text{Losses} - \text{Loss Adjusting Expenses}\\
&\qquad - \text{Underwriting Expenses}\Big).
\end{aligned}$$

From 1.2, the costs of an insurance company generally include underwriting expenses to keep a company operational as well as the losses and loss adjusting expenses that may only be realised in the future. To report the profitability of the company in its financial papers, the unknown future liabilities at a point of reporting has to be estimated. The actuaries' responsibility in a general insurance company are typically separated into pricing of products and valuation of future liabilities. This research focuses on the latter.

The valuation of future liabilities typically involves forecasting future liabilities and assessing the amount of assets to address them. Insurance companies are required to allocate a provision for the future liabilities known as a loss reserve to pay for the unknown claims that may occur in the future. The loss reserve is recorded in the financial statements of the insurers as a form of recognition of future liabilities, usually being the largest portion of a general insurer's liability. If the reserve is set too low, there is a risk of inadequate funds for claims in the future and an overestimation of profit; if it is set too high, there is a loss of chance to invest in the funds and an underestimation of profit. An inadequate of funds to pay for liabilities puts a company at risk of bankruptcy as well. In fact, the insurance business is a heavily regulated business, especially after a compensation crisis in the 1980s and 1990s (American Academy of Actuaries, 2000) that have brought the actuarial industry into the spotlight. Regulations were set and tough reporting standards have been imposed to improve actuaries' credibility. Hence a proper estimation of this ultimate liability is vital, but the task is challenged by underlying uncertainties which normally arise due to the unknown ultimate number and size (amount) of reported claims (Zhang et al., 2012). Future claims may occur randomly, and may or may not be made known to the insurer at the time of occurrence. If claims are reported and make known to the company, case reserves can be allocated for the particular claim but loss payments can stretch out to months or even years for certain line of businesses before the claim is complete, causing lengthy claim settlement periods. In some cases, especially when huge amount of losses is involved, it may involve litigation process and so may take even more time to reach settlement. On the other hand, the insurer needs to also take into account claims that are unknown to them at a particular evaluation time and thus the need to also set up a provision for these unknown claims. Therefore, the very existence of the actuarial profession has been primarily regarded to predict these underlying uncertainties to safeguard company and consumer interests.

1.1.1 The claims process

When an event covered in an insured's policy occurs, the claims process begins when a claim is reported and made known to the insurer. This process in particular is called the opening of claims. An amount known as a *case reserve* or also commonly known as a *case outstanding* will be allocated for each open claim that is being reported to pay for losses of the particular claim. If the case outstanding is insufficient as a result of claims development, the amount can be increased (as incurred but not enough reported, IBNER), else the remainder returns to the company funds. As long as the claim is not settled (closed claims), the case outstanding remains allocated for future incoming losses. For situations where an event has happened but not made known to the company up until before the policy effective period is over, the incurred but not yet reported (IBNYR) or pure IBNR will be allocated. In general, IBNR is the sum of IBNER and IBNYR. Hence, as an acknowledgement of the insurer's future liabilities, the loss reserve is the sum of IBNR and the case outstanding at a particular evaluation time. Having known the paid losses at a particular evaluation point, together with the provision for development of known losses and provision for unknown losses, we have the amount of ultimate value of losses. Alternatively, we can also say that by estimating the ultimate value of losses, the amount of IBNR can be estimated by subtracting the paid amount and the case outstanding amount at an evaluation point from the ultimate value of losses, which can be seen from the third line in the equation below, where t represents a point of evaluation:

$$\begin{aligned}
\text{Ultimate Loss} &\approx \text{Paid Losses}(t) + \text{Provision for Known Losses}(t) \\
&\qquad + \text{Provision for Unknown Losses}(t) \\
&\approx \text{Paid Losses}(t) + \Big(\text{Case Outstanding}(t) + \text{IBNER}(t)\Big) \\
&\qquad + \text{IBNYR}(t) \\
&\approx \text{Paid Losses}(t) + \text{Case Outstanding}(t) + \text{IBNR}(t) \\
&\approx \text{Paid Losses}(t) + \text{Reserve}(t).
\end{aligned} \tag{1.3}$$

1.1.2 The loss triangle

Loss reserving techniques have traditionally evolved around loss triangles which show the development of aggregated claims of the insureds by accident years. Table 1.1 shows a typical loss triangle for *cumulative paid claims* of the worker's compensation line of business from a single company. The loss data was extracted from statutory annual statements of selected general insurance companies in the United States, which are required to be submitted to the National Association of Insurance Commisioners (NAIC) periodically. The extracted data was made available from the website of the Casualty Actuarial Society (Meyers, 2011). The link to the dataset is provided in the references and is still accessible at the time of submission of this thesis.

This loss triangle was built with information from years 1988 to 1997. The claims were aggregated into accident years, that is, years when the events occurred, not necessarily the time when claims get reported since reporting can be delayed, and the development of

Table 1.1: Loss triangle of cumulative paid claims: Company GRCODE = 337.

Year	Premium	Development Lags									
		1	2	3	4	5	6	7	8	9	10
1988	104437	9558	22778	33298	40348	45146	48048	49782	50623	51812	51939
1989	88883	7913	19472	29622	36816	40975	43302	44707	45871	46229	
1990	85956	8744	24302	35406	43412	48057	50897	52879	53956		
1991	99339	13301	32950	47201	56394	61650	65039	66566			
1992	104897	11424	29086	42034	50910	56406	59437				
1993	119427	11792	27161	38229	46722	50742					
1994	110784	11194	26893	38488	45580						
1995	77731	12550	31604	44045							
1996	63646	13194	31474								
1997	48052	9372									

each cohort across time is observed. The 1988 row is ‘complete’ since the development of the 1988 accident year cohort for the full ten years would have been known by the end of 1997; whereas for the 1997 row, only one year worth of information is known by then. For example, the value 51939 at the tenth development lag for accident year 1988 is accumulated from the first development lag at 9558 throughout the span of 9 years. Suppose x_j represents a development lag such that $j \in \mathbb{N}$, the increment of the cumulative paid claims for each row decreases as it develops toward development year 10, that is, x_{10}. The development will stop at development year x_∞, that is, when the ratio between two consecutive development reaches 1.0, and that is when the ultimate paid amount for each cohort can be obtained. Therefore, the prediction of the ultimate paid amount for each cohort is done by ‘completing’ the lower half of the loss triangle, which is essentially an extrapolation problem. The representation of the loss table is generalized in Table 1.2:

Table 1.2: Generalization of the cumulative loss triangle. Source: Zhang et al. (2012)

Year	Premium	Development Lag					Tail
		x_1	x_2	$\cdots$	x_{I-1}	x_I	$\cdots x_\infty$
1	p_1	$y_1(x_1)$	$y_1(x_2)$	$\cdots$	$y_1(x_{I-1})$	$y_1(x_I)$	
2	p_2	$y_2(x_1)$	$y_2(x_2)$	$\cdots$	$y_2(x_{I-1})$		
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$			
i	p_i	$y_i(x_1)$	$\cdots$	$y_i(x_{I+1-i})$			
$\vdots$	$\vdots$	$\vdots$	$\vdots$				
$I-1$	p_{I-1}	$y_{I-1}(x_1)$	$y_{I-1}(x_2)$				
I	p_I	$y_I(x_1)$					

Here

$$
\begin{aligned}
y_i(x_j) &= \text{cumulative paid loss amount for a cohort of insurance claims that} \\
&\quad\ \text{occur in year } i, \text{ evaluated at } x_j \text{ years since inception year } i \\
I &= \text{Total number of rows or columns} \\
\mathcal{D}_I &= \{y_i(x_j)|i+j \leq I+1\} \\
&= \text{Observed data} \\
\mathcal{D}_I^c &= \{y_i(x_j)|I+1 < i+j \leq 2I\}, t_0 = 0 \\
&= \text{Unobserved data.}
\end{aligned}
$$

The problem lies in the fact that we do not know x_∞, thus the value $y_i(x_\infty)$ is uncertain at a point of evaluation. In general, the goal of insurance loss reserving is to estimate $y_i(x_\infty)$ for $i = 1, ..., I$ given $\mathcal{D}_I$, which are the ultimate values for each cohort and then determine the unpaid amount for the provision of both known and unknown claims. Actuaries around the world have been equipped with methods recommended by professional bodies in tackling this problem but researchers continue to devise methods that may improve predictive accuracy and ability to measure predictive uncertainties more effectively. While the trend of devising new methods based upon Bayesian inference has been going on for some time now, recent inclusion of Gaussian Process models into the loss reserving literature has opened up opportunities to look into yet another interesting method to solve the same problem. Since the method is still relatively new, there is a need to look into details of Gaussian Process Models in fitting various types of data.

1.2 Problem Statement

The Gaussian process regression model has provided a framework to detect spatial relationships among the loss observations and then complete the lower half of the loss triangle as a ‘surface’. The Gaussian Process regression model is specified mainly by a kernel function that is responsible to carry most of the assumptions in the model. A common kernel function applied in practice is the squared exponential kernel, which is also a stationary kernel function. However, using a stationary kernel function in a Gaussian process model to fit the cumulative loss observations may be insufficient as nonstationarities clearly exist in cumulative data. Hence, input warping was also proposed along with the model to handle the nonstationarities. Just like any existing models, the proposed Gaussian process model is not problem-free. Upon inspection, it can be found that there exist some over-smoothing issue in the completed surface as well as a huge predictive uncertainties associated to it.

1.3 Research Objectives

The primary objective of this research is to improve the predictive accuracy and to reduce predictive uncertainties of the loss reserve estimates generated by the hierarchical Bayesian Gaussian process regression model for loss reserving. In order to achieve that, the objectives are broken down as follows:

1. To investigate the causes of the over-smoothing problem.
2. To investigate the effect of specifying the kernel functions by adding and multiplying squared exponential kernel functions in the Gaussian process regression model for loss reserving.

1.4 Research Methodology

The research begins by establishing a mathematical foundation critical to understand not only the theory behind Bayesian modelling but also to briefly understand the reason behind using Markov Chain Monte Carlo (MCMC) in estimating sophisticated integrals associated to Bayesian inference. In this research, the MCMC sampling in the case studies are handled by the Stan probabilistic programming software and hence will not be elaborated in the thesis, although the related materials will be recommended to an interested reader. Since this research covers a multiple discipline of studies from actuarial science to machine learning, all mathematical objects mentioned in the thesis such as functions are treated as defined in classical mathematics to avoid ambiguities. For example, $f(\cdot)$ denotes a function that does not necessarily strictly represent a probability density function, while the notation $P(\cdot)$ is treated as a probability measure as defined in measure-theoretic probability theory and not a ‘frozen notation’ for a prior distribution. Some background in undergraduate linear algebra as well as measure-theoretic probability theory which are especially critical in understanding Gaussian Process regression are expected from a reader. The writing in the thesis do not cater specifically to any targeted group of readers from any specified field.

With the mathematical foundation in place, the loss reserving literature is reviewed where focus is given to the Bayesian models, from the most common Link Ratio method to the Bayesian parametric and nonparametric models. The focus is then narrowed down to the hierarchical Bayesian Gaussian Process regression model. To achieve the aforementioned research objectives, the Gaussian process models will be tested with kernel functions that are constructed from combinations of additive and multiplicative squared exponential kernel functions. This is done to investigate the structural support that may be provided by such specification to the Gaussian process model. These Gaussian process models defined by the various kernel functions will be fitted to the NAIC data to evaluate their predictive abilities and conclusions will be made from the results of the case studies. The entire workflow is summarized and linked to the respective chapters of the thesis in Figure 1.1.

1.5 Outline and contribution

The contribution of this thesis is to investigate Bayesian Gaussian process models for loss reserving that may be built from various combinations of kernels, thus testifying these kernels and make conclusions from the experiments carried out upon chosen data sets from the NAIC.

Chapter 2 lays out the bare minimal mathematical background for the rest of this thesis, especially Bayesian inference for both parametric and nonparametric models.

Chapter 3 discusses some methods used by practitioners in determining loss reserves as well as some Bayesian models available in the loss reserving literature leading up to this research.

Chapter 4 introduces the modifications that can be made to the Gaussian Process

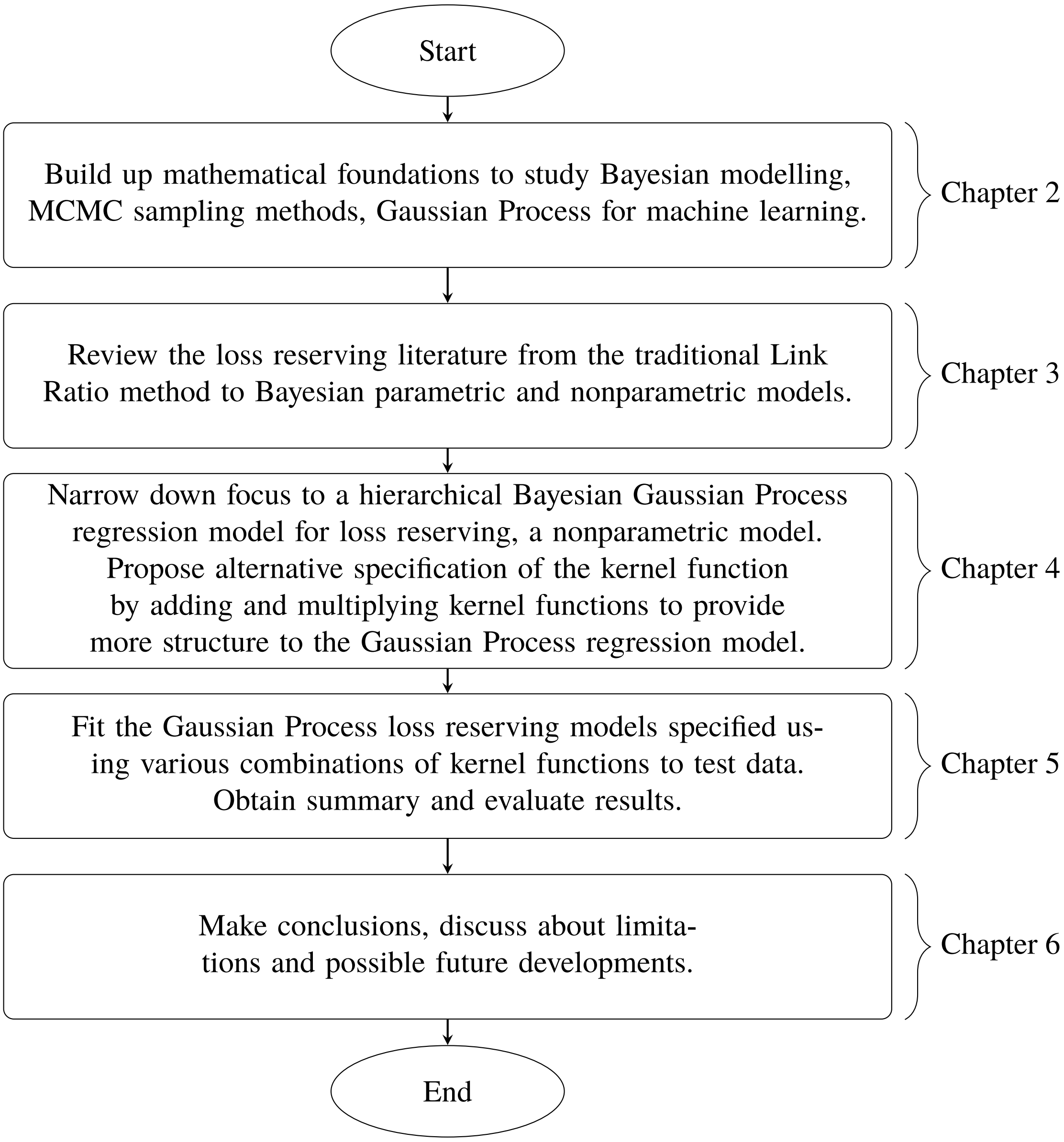
Start
Build up mathematical foundations to study Bayesian modelling, MCMC sampling methods, Gaussian Process for machine learning.
Chapter 2
Review the loss reserving literature from the traditional Link Ratio method to Bayesian parametric and nonparametric models.
Chapter 3
Narrow down focus to a hierarchical Bayesian Gaussian Process regression model for loss reserving, a nonparametric model. Propose alternative specification of the kernel function by adding and multiplying kernel functions to provide more structure to the Gaussian Process regression model.
Chapter 4
Fit the Gaussian Process loss reserving models specified using various combinations of kernel functions to test data. Obtain summary and evaluate results.
Chapter 5
Make conclusions, discuss about limitations and possible future developments.
Chapter 6
End

Figure 1.1: A block diagram summarizing the research methodology.

loss reserving models by adding and multiplying kernels. The case study setup and the data sets used for simulations will also be laid out in this chapter.

Chapter 5 presents the results from the case studies for Gaussian Process loss reserving models using various combinations of additive and multiplicative squared exponential kernels, with and without input warping.

Finally Chapter 6 concludes the research, highlights some limitations that has arisen throughout the research and discusses some potential developments that can be undertaken or looked into in the future.

CHAPTER 2

BAYESIAN INFERENCE

This research involves predicting into the future as well as estimating the uncertainties underlying the predictions. In order to do so, Bayesian inference will be employed to represent uncertainties with probability. Since Bayesian inference is the pillar of this research, the mechanism of Bayesian inference will be very briefly introduced in this chapter before we proceed into the details of its application in predicting loss reserves. We will begin by emphasising the utilization of probability in measuring beliefs in this statistical inference method, and later provide examples of Bayesian inferences with a parametric model and a nonparametric model.

2.1 Random variable and Probability

Given a countable sample space Ω, let $\mathcal{G} = \sigma(\Omega)$ be a σ-field on Ω, that is, a family of subsets of Ω that fulfils the following closure properties:

(i) $\mathcal{G}$ is a collection of subsets of Ω that includes Ω itself and is closed under its complements and countable unions.

(ii) $\mathcal{G}$ includes the empty subset and is closed under countable intersections.

Then by the De Morgan Laws, $\mathcal{G}$ is closed under the intersection and difference operations. The pair $(\Omega, \mathcal{G})$ is called a measurable space.

Definition 2.1.1 (Probability measure)**.** A *probability measure* P is a function of $\mathcal{G}$ satisfying the properties

(i) For every $A \subset \Omega$, $P(A) \geq 0$;

(ii) For any disjoint subsets A and B of Ω,

$$P(A \cup B) = P(A) + P(B);$$

(iii) $P(\Omega) = 1$.

When such a probability measure P is defined on $\mathcal{G}$, a probability triple $(\Omega, \mathcal{G}, P)$ is formed, that is, sets in $\mathcal{G}$ are measurable and each set has probability.

Definition 2.1.2 (Random variable for countable Ω)**.** Let X be a numerically valued function defined on elements of Ω. Then X is called a *random variable* on Ω if

$$\begin{aligned} X : \Omega &\to \mathbb{R} \\ \omega &\longmapsto X(\omega). \end{aligned} \tag{2.1}$$

Suppose Ω is finite or countable, and let $P(\omega)$ be the 'weight' of each ω in Ω. Then the probability of A such that $A \subset \Omega$ is given by

$$P(A) = \frac{\sum_{\omega \in A} P(\omega)}{\sum_{\omega \in \Omega} P(\omega)}, \tag{2.2}$$

and $P(A)$ reduces to the proportion of the cardinality of A relative to the cardinality of Ω when the weights of all sample points in Ω are equal.

The set A in Equation (2.2) is determined by values of random variables. For instance, let A be a subset of Ω such that

$$A = \{\omega \in \Omega \mid a \leq X(\omega) \leq b\}, \tag{2.3}$$

where a and b are constants and X is a random variable. Then

$$P(A) = P(a \leq X \leq b) = \sum_{a \leq v_n \leq b} p_n \tag{2.4}$$

is the probability of set A, where p_n denotes the weight $P(v_n)$ of v_n in a set of the range of $X(\omega)$, $V_X = \cup_{\omega \in \Omega} X(\omega)$ since the mapping in Equation (2.1) can be many-to-one. Similarly, for the set $\{X(\omega) \leq x\}$,

$$F_X(x) = P(X \leq x) = \sum_{v_n \leq x} p_n \tag{2.5}$$

is a function that collects all the probabilities of random variable X in the set up to and including x and is usually known as the *cumulative probability distribution function* of X. An average value, also known as the *expectation value* of the set determined by the random variable X can be obtained by

$$E(X) = \sum_{\omega \in \Omega} X(\omega)P(\omega). \tag{2.6}$$

In the case where Ω is an uncountably infinite set such as $\mathbb{R}$, it is impossible to define a probability measure P on Ω since there are infinitely many subsets and the probability of each subset will be reduced to 0. Hence, the condition $P(\Omega) = 1$ cannot hold. In order to solve this measurability problem, the uncountable set needs to be prepped into a measurable space by restricting Ω to a Borel field. A Borel field $\mathcal{F}$ is constructed by defining a σ-field on a set A of all finite open intervals in Ω, giving the smallest σ-field containing A. Every set in $\mathcal{F}$ is measurable and when a probability measure P is assigned to $\mathcal{F}$, a probability triple $(\Omega, \mathcal{F}, P)$ is formed and every set in $\mathcal{F}$ has probability. Therefore, as compared to defining a probability measure on a countable Ω where each element (point) has a probability, defining a probability measure on an uncountable Ω assigns probability to intervals.

Definition 2.1.3 (Random variable for uncountable Ω)**.** Let X be a real-valued function defined on an uncountable Ω. Then X is a *random variable* on Ω if for any $x \in \mathbb{R}$,

$$\{\omega \in \Omega \mid X(\omega) \leq x\} \in \mathcal{F}.$$

One property of random variable worth mentioning is that, if X is a discrete random variable such that X is defined on a countable Ω, then any function g of X is also a random variable; if X is a continuous random variable such that X is defined on an uncountable Ω, a function f of X is a random variable if $f : \mathbb{R} \to \mathbb{R}$ is a measurable function.

Definition 2.1.4 (Stochastic process)**.** If we consider a set of random variables, the collection of these random variables $\{X_i\}_{i \in \mathbb{I}}$ over the same measure space where $\mathbb{I}$ is an index set, is a *stochastic process*.

Since an uncountable Ω is an infinite set and it is impossible to assign probability to every single point, the interval defined by a random variable X is instead given a probability density such that probability is associated to area under the curve f and the total area equals to 1. A *density function* f is a function defined on $\mathbb{R}$ satisfying two conditions:

(i) $f(u) \geq 0, \forall u \in \mathbb{R}$,

(ii) $\int_{-\infty}^{\infty} f(u)du = 1$.

The probability density of any given subset defined by a random variable is obtained by summing up the area under the density function that is partitioned into intervals by integration. So for an interval $[a, b]$,

$$P(a \leq X \leq b) = \int_a^b f(u)du \tag{2.7}$$

and the distribution function is then

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)du. \tag{2.8}$$

Note that f is the derivative of F if f is continuous, by the fundamental theorem of calculus. The expectation associated to a continuous random variable is

$$E(X) = \int_{-\infty}^{\infty} uf(u)du. \tag{2.9}$$

This research involves inferring parameters from the real space. For readers exposed only to elementary probability theory, one may regard it as focusing on continuous random variables in the following sections. It is crucial to emphasise here that with only elementary probability theory, one may run into trouble when it is required to think about discrete and continuous random variables at the same time, which are especially prevalent in Bayesian inferences. Measure-theoretic probability theory fixes the issue of having to discuss discrete and continuous random variables separately and deals with all probability distributions in a unified way. However, as measure-theory is a huge subject on its own, it will not be discussed in this thesis but rather we will just benefit from its results.

2.2 Conditional Probability and Bayes' Theorem

In Section 2.1, probability of a set determined by a random variable is measured with respect to the entire sample space Ω as shown in Equation (2.2). However in practical usage, we may have greater interest in the proportional weight of a set A with respect to a set S such that $A \subset S \subset \Omega$, especially when more information is available or when it is certain that some of the elements in Ω are no longer relevant in an inquiry.

Definition 2.2.1 (Conditional probabililty)**.** The relationship between set A and set S when attention is focused into S as a new universe instead of Ω is expressed as

$$P(A|S) = \frac{P(A \cap S)}{P(S)}, P(S) \neq 0 \tag{2.10}$$

and is known as the *conditional probability* of A relative to S.

The numerator is the weight of the part of A in S relative to Ω and the denominator is the weight of S relative to Ω. This gives a different probability than when Ω is set as the 'universe'. The conditional probability is often useful when the mutual or joint relation of several random variables are to be inspected.

Definition 2.2.2 (Partition of a set)**.** Let S be a sample space of a union of mutually exclusive subsets $B_1, B_2, \cdots, B_k$ for some positive integer k such that

(i) $S = B_1 \cup B_2 \cup \cdots \cup B_k$

(ii) $B_i \cap B_j = \emptyset$, for $i \neq j$.

The collection of sets $B_1, B_2, \cdots, B_k$ is called a *partition* of S.

If A is any subset of S and $B_1, B_2, \cdots, B_k$ is a partition of S, then A can be decomposed as:

$$A = (A \cap B_1) \cup (A \cap B_2) \cup \cdots \cup (A \cap B_k). \tag{2.11}$$

By the Law of Total Probability (Chapter 2 ,Wackerly et al. (2014)), assuming that $B_1, B_2, \cdots, B_k$ is a partition of S such that $P(B_i) > 0$, for $i = 1, 2, \cdots, k$, then for any event A of the same probability space,

$$P(A) = \sum_{i=1}^{k} P(A \cap B_i) = \sum_{i=1}^{k} P(A|B_i)P(B_i), \tag{2.12}$$

where any $P(A|B_i) = 0$ is omitted from the summation since $P(A|B_i)$ is finite. Then from Definition 2.2.1 and the Law of Total Probability, the result known as *Bayes' rule* can be derived as follows:

$$P(B_j|A) = \frac{P(A \cap B_j)}{P(A)} = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{k} P(A|B_i)P(B_i)}. \tag{2.13}$$

If one understands a random variable as defined in classical mathematics, it is straightforward to replace sets A, B and S from Equation (2.10) to Equation (2.13) with random variables, which are measurable functions that maps subsets in Ω to its range that is usually $\mathbb{R}$.

2.3 Bayesian inference: Probability at work

Probability is commonly used to measure frequencies of outcomes, but it can also be used to measure degree of belief. Bayes' rule comes into picture when the conditional probability of unobserved variables given observed variables is to be calculated, and this kind of problem is known as the *inverse probability problem* (MacKay, 2003).

Bayesian inference is a statistical inference method that employs probability to measure degree of belief of inferences as compared to common statistical methods that make predictions by assuming a selected most plausible hypothesis to be true ahead of an inference. It first assumes a range of possible values of a hypothesis that are associated with uncertainties, represented by probabilities. Then prediction is made by marginalizing over all possible values of the hypothesis, which helps to avoid extreme predictions. As more information or evidence is made known, the probabilities of the values of the hypothesis will be updated accordingly.

A formal definition for statistical models from Bickel and Doksum (2015) is as follows.

Definition 2.3.1 (Model)**.** Let a sample space Ω represents a random experiment and $\omega \in \Omega$ represents the outcome of the experiment. Then let $\mathbf{y}(\omega)$ be the observations that we obtain from the experiment such that $\mathbf{y} = (y_1, \cdots, y_n)$ is a random vector. Since there would be different realizations of $\mathbf{y}$ if the experiment is repeated, there is a probability distribution associated to $\mathbf{y}$ that is assumed to be part of another family of probability distributions $\mathcal{M}$ on $\mathbb{R}^n$, where $\mathcal{M}$ is the *model*.

Models are largely categorized into classification or regression. *Classification* models predict labels or assign input variables that are often binary or fall into any one of n discrete classes $\mathcal{C}_1, \cdots, \mathcal{C}_n$. On the other hand, *regression* models map input variables to the real space $\mathbb{R}$, that is, to predict a quantity. This research falls under the later category.

Definition 2.3.2 (Parameterization)**.** $Parameterization$ is a mapping that describes $\mathcal{M}$ such that $\mathcal{M} = \{M_\theta : \theta \in \Theta\}$, where Θ represents the parameter space.

When the parameter space Θ is of fixed finite dimensional, the model is $parametric$, otherwise the model is $nonparametric$. For example, the parametric Normal family is parameterized by $\theta = (\mu, \sigma^2)$ so Θ is finite dimensional with dimension two. We are making it clear at this point that θ can be more than one-dimensional but for the following discussion, the model parameter(s) to be estimated will be represented by a one-dimensional θ to simplify the equations.

The model becomes a Bayesian model when we consider assumptions upon θ prior executing the experiment via the *prior distribution* $P(\theta)$. This would mean that θ is being treated as a random variable to generate a joint distribution for (y, θ) such that

$$f(\theta, y) = P(y|\theta)P(\theta), \tag{2.14}$$

where we let $P(y|\theta) = M_\theta$ in order to observe the application of Bayes' Rule. The integral of the joint distribution across all possible values of θ is the complete Bayesian model.

Bayesian statistical inference involves choosing a model that summarizes the data generating process represented by the conditional distribution $P(y|\theta)$, and then requires an input of an assumed prior distribution for the model parameter(s) θ that includes judgement of how plausible the parameter(s) could have certain values in various regions of the parameter space before any measurements are taken. Combining the prior distribution and the conditional distribution $P(y|\theta)$, we get a joint distribution for all quantities related to the problem by Bayes' rule

$$\begin{aligned} P(\theta, y) &= P(\theta, y) \qquad (2.15) \\ &= P(\theta)P(y|\theta) \end{aligned}$$

It is possible to make inferences before considering observed data, which is usually done to check the model configurations before fitting data in order to make sure the specifications are aligned to prior knowledge of a problem. This is done by integrating the joint distribution across all possible values of θ to obtain the *prior predictive distribution*, which is a marginal distribution $P(y)$ of y, where

$$\textit{Prior predictive distribution} = P(y) = \int P(\theta)P(y|\theta)d\theta. \tag{2.16}$$

The likelihood function $P(y|\theta)$ from the joint distribution above is a function of θ, and encapsulates the relative abilities of the parameter values to describe the observed data. Thus the function can be considered as a measure of plausibility of the parameter values, so the *likelihood principle* (Chapter 2, MacKay (2003)) requires all inferences and predictions to be depending only on the likelihood function given some observed outcomes. By defining a joint distribution for both the model parameter(s) θ and the observed data, the Bayesian approach reduces a statistical inference to a probabilistic inference. When observed data $\mathbf{y} = (y_1, \cdots, y_n)$ is considered, a posterior probability distribution for the parameters that updates the prior distribution can be obtained as follows:

$$\begin{aligned} Posterior\ distribution &= P(\theta|y_1, ..., y_n) \\ &= \frac{P(\theta, y_1, ..., y_n)}{P(y_1, ..., y_n)} \\ &= \frac{P(\theta)\prod_{i=1}^{n} P(y_i|\theta)}{\int P(\theta)\prod_{i=1}^{n} P(y_i|\theta)d\theta}. \end{aligned} \tag{2.17}$$

All information about a system is quantified by the posterior distribution after the measurements are taken. Hence, a new joint distribution can be formed with a likelihood function for the unknown y_{n+1} and the posterior distribution from Equation (2.17). Then to generate a prediction distribution for an unknown y_{n+1} given observed data, a *posterior predictive distribution* $P(y_{n+1}|\mathbf{y})$ is formed by integrating over values of θ from the new joint distribution:

$$\begin{aligned} \textit{Posterior predictive distribution} &= P(y_{n+1}|\mathbf{y}) \\ &= P(y_{n+1}|y_1, ..., y_n) \\ &= \int P(y_{n+1}|\theta)P(\theta|y_1, ..., y_n)d\theta. \end{aligned} \tag{2.18}$$

Comparing to Equation (2.9), the posterior predictive distribution can also be viewed as the expectation of the likelihood function $P(y_{n+1}|\theta)$ with respect to the posterior distribution for θ,

$$E[P(y_{n+1}|\theta)].$$

Hence, any statistical queries are answered by computing expectations with respect to the posterior distribution.

However, computing these expectations is equivalent to computing some difficult integrals. Specifically, the denominator of the posterior distribution often involve some calculations of intractable high dimensional integrals. The problem seems to be solvable by employing Monte Carlo sampling method to get independent samples in order to approximate the integrals, but getting independent samples from the posterior distribution is easier said than done. So, one way to tackle this problem is to get a sequence of dependent samples by using Monte Carlo Markov Chains (MCMC) to explore the high dimensional space such that each step of the sampler samples from the numerator of the posterior distribution in Equation (2.17), known as un-normalised posterior

$$P(\theta|y_1,...,y_n) \propto P(\theta)\prod_{i=1}^{n} P(y_i|\theta) \tag{2.19}$$

and the next sample is dependent only on the previous sample. Since the acceptance of the next sample is determined by ratio of the samples from the un-normalised posterior, the problem of determining the denominator of the posterior distribution can be avoided. This research employs a probabilistic programming language Stan (Stan Development Team, 2020) to implement the MCMC algorithms in order to obtain posterior predictive samples. A comprehensive guide to using Stan in Bayesian Modeling and understanding the basics of MCMC has been laid out by Lambert (2018). To understand the mechanism of the MCMC algorithm employed by Stan in which the performance has been optimized by Hamiltonian Monte Carlo and No-U-Turn sampler, see Betancourt (2018) for a simpler introduction and Betancourt et al. (2017) for a higher level understanding.

2.3.1 Bayesian Parametric model: An example

We are often interested to unravel information from observations obtained from an experiment. If the 'true' mechanism that generates the observations can be recovered, we can use the mechanism to answer questions that can only be answered if the experiment is being carried out continuously (interpolation) or never stopping the experiment (extrapolation). Some well studied experiments that produce observations about certain phenomenon of interest may already have some theoretical model established by researchers. These models are often proven to be effective in interacting with the latent phenomenon in question and hence are able to answer questions that we have related to the said phenomenon. In fact, some models being built on one phenomenon may be also suitable for another phenomenon. During a study, each observation on a phenomenon may differ depending on its magnitude or methods of observation, hence a parametric model with some specified functional form governed by parameters gives a suitable guideline as a starting point and we only have to approximate as close as possible the parameter values that when applied into the model provides values as close to the observations as possible. This section demonstrates the mechanism of approximating parameter values by fitting an example Bayesian parametric regression model to a set of time series data.

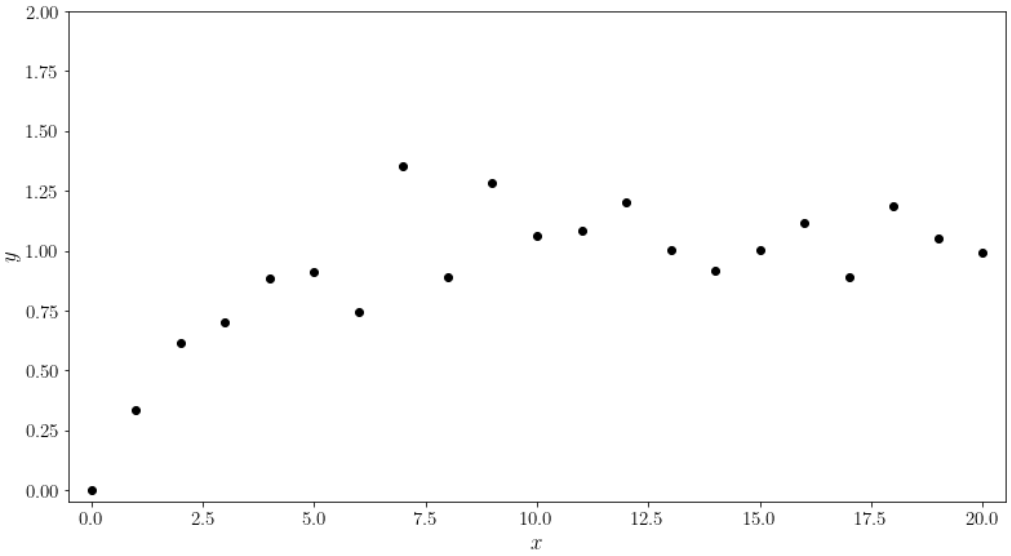

Figure 2.1: A set of time series data to be fitted with a model.

We begin with a set of time series data visualized in Figure 2.1, where x represents time and y denotes the observation at each x. Assuming that a modeler has no knowledge of the ‘true’ data-generating process but has information such that the nonnegative outputs from the source experience initial positive growth at smaller input values x_i such that the growth rate decreases and approaches zero as the input values increases. The modeler wishes to model the underlying phenomenon that generates these observations and infer the parameters of the model to obtain more insights about this phenomenon.

A possible parametric model to be chosen as a potential data-generating process for the time series data would be an exponential function to capture the growth pattern. On top of that, it is obvious that the observations contain some form of errors, deviating from a supposedly smooth exponential curve. Hence, we can consider an error component in the data-generating process.

A common candidate for such component would usually be a Normal distribution, but using a Normal distribution for the error component may result in negative valued samples, which might not produce relevant samples given the information that observations are nonnegative. For ordinary linear regression, a Normal distribution is usually assumed to relate the variability in the observations to changes in the mean. However, if the volatility of the observations follows the Normal distribution, that means the uncertainty that grows with the mean can be greater at higher input values x , which contradicts the prior knowledge that the output values should stabilize at higher input values. Another possible candidate would be the Lognormal distribution. This is equivalent to applying a log-transform to the constant variance in Normal distribution that results in a constant standard deviation over mean ratio, so that the variance will not grow as the mean increases.

Considering the above information, suppose that the modeler decides on choosing the exponential function to model the latent phenomenon of the data $\mathbf{y} = (y_1, \cdots, y_{21})$ and addresses the error with the Lognormal distribution. The model can be written as

$$y_i \sim \text{Lognormal}\left(\log\left(f\left(x_i|\lambda\right)\right), \sigma\right), \tag{2.20}$$

where

$$f(x_i|\lambda) = 1 - e^{-\lambda x_i},\ \lambda > 0$$

is the candidate function chosen to model the latent phenomenon in the time series parameterized by λ. The mean is wrapped with a log function so that the mean of the Lognormal distribution centers around $f(x_i|\lambda)$. As mentioned in Section 2.3, θ can be more than one-dimensional. Thus in this case, the likelihood function or the sampling distribution is

$$P(y_i|\theta) = P(y_i|\lambda, \sigma, x_i),$$

but since inputs x_i are considered fixed, the conditioning on x_i is suppressed in subsequent notation.

In order to define a complete Bayesian model, we need to define a model configuration space made up of the prior distributions of the two parameters $P(\lambda, \sigma) = P(\lambda)P(\sigma)$. Choosing priors in Bayesian inference is an art itself, but the rule of thumb is to ensure that the prior generates reasonable samples that we would expect to see (Gelman et al., 2017). In practice, we can apply our judgement on the parameters when setting up priors and usually the assumption should constrain the system especially when dealing with likelihood functions that are not well-behaved to avoid extreme predictions. Assuming that the parameters are nonnegative ($\lambda > 0$, $\sigma > 0$), we can start with the Lognormal priors and try the following values in the Lognormal distributions of the priors :

$$\lambda \sim \text{Lognormal } (\log(0.3), 0.4)$$

$$\sigma \sim \text{Lognormal } (\log(0.15), 0.4).$$

To assess if these priors are suitable, we can generate samples from the prior predictive distribution which can be compared to the one-dimensional Equation (2.21) as follows:

$$P(y_i) = \iint P(y_i|\lambda, \sigma)P(\lambda, \sigma)\ d\lambda d\sigma\ . \tag{2.21}$$

Thus, we can observe the effect of the priors when combined with the data-generating process before fitting the model to the data. As mentioned before, the integrand in Equation (2.21) is the complete Bayesian model and the entire prior predictive distribution is a marginal of the complete Bayesian joint distribution. Because the model configuration in the prior predictive distribution is entirely dependent on prior information, the prior predictive distribution is able to quantify the best understanding of the the phenomenon before observations are taken. Thus, we are able to make predictions before considering the observations, which at the same time provides the opportunity to evaluate the priors. The prior predictive samples shown in the left panel of Figure 2.2 and the prior predictive distribution shown in the left panel of Figure 2.3 suggest that the prior assumptions are reasonable, so we may attempt to fit the model to the data. If the prior samples to do not reflect the modeler's expectation, the values in the Lognormal distributions can be altered according to the modeler's knowledge or redefine the prior distributions until samples deemed satisfactory by the modeler is generated.

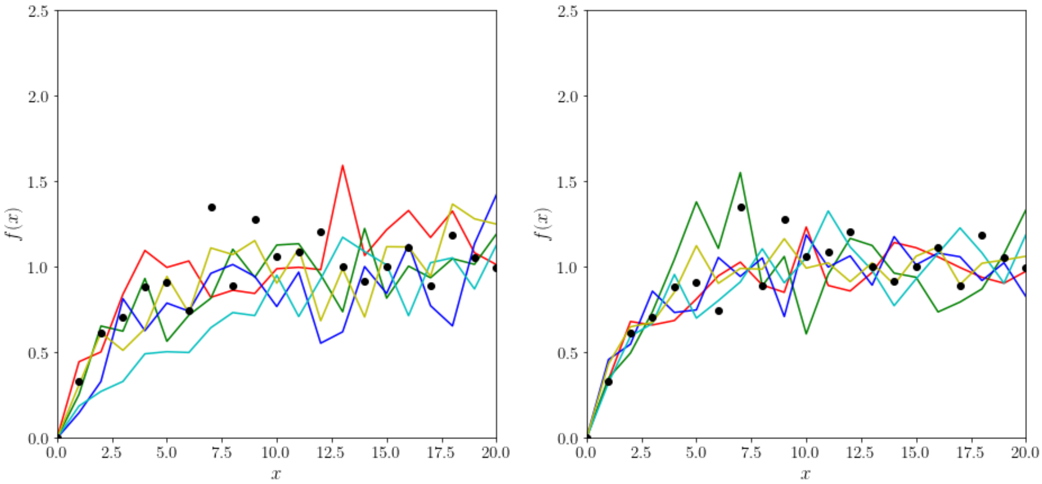

Figure 2.2: Five random samples each drawn from the prior (left) and posterior (right) predictive distributions via MCMC superimposed on the original time series data.

The role of data in Bayesian inference is to update the prior distribution that results in the posterior distribution $P(\lambda, \sigma|\mathbf{y})$ such that the posterior distribution is the compromise between data and prior information. Then with the posterior distribution, we can construct the posterior predictive distribution like the one in Equation (2.18) for a future observable $\tilde{\mathbf{y}}$ such that:

$$P(\tilde{\mathbf{y}}|\mathbf{y}) = \iint P(\tilde{\mathbf{y}}|\lambda, \sigma)P(\lambda, \sigma|\mathbf{y})\ d\lambda d\sigma. \tag{2.22}$$

The model configuration in the posterior predictive distribution now contains information prior considering the observations as well as information given by the data, hence it is ideal to make predictions regarding the phenomenon of interest with it. The integral in Equation (2.22) is approximated by MCMC, which the sampling process produces posterior distributions for the parameters as well as the posterior predictive distribution for $\tilde{\mathbf{y}}$. For the sake of this example, the data shown in Figure 2.1 was actually generated by the exact data-generating process in Equation (2.20) with true values being $(\lambda, \sigma) = (0.5, 0.125)$. The posterior predictive samples are shown in the right panel of

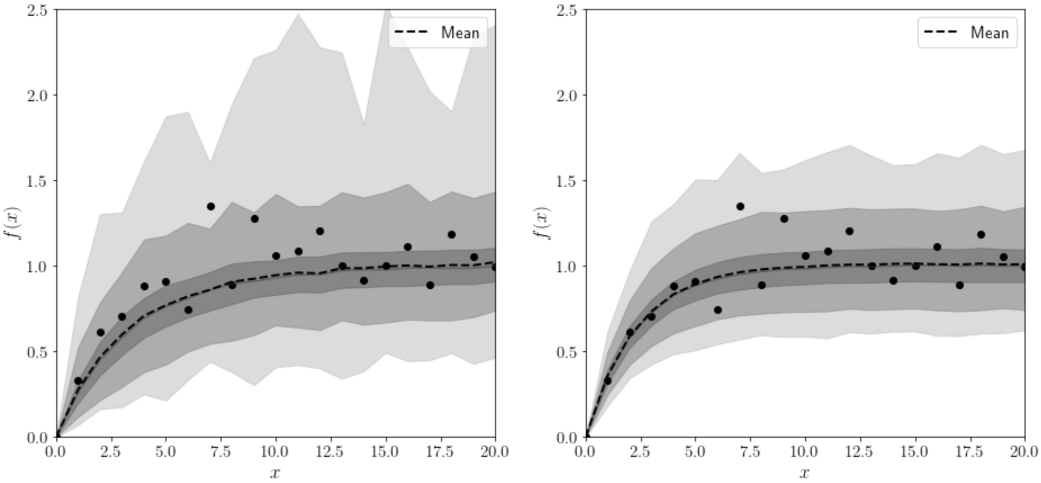

Figure 2.3: Prior (left) and posterior (right) predictive distributions of a parametric model where the shades denote the 75%, 97.5%, and 99.9% intervals of the MCMC samples that approximate the integrals; Dotted line represents the mean of the samples.

Figure 2.2 and the posterior predictive distribution is shown in the right panel of Figure 2.3. The comparison between the predictive distributions and observed data in Figure 2.3 shows how well the Bayesian model approximates the phenomenon of interest before and after considering the observations. The modeler might also be interested to make inferences about the parameter λ that governs the growth rate of the exponential function and the error parameter σ. From Figure 2.4, the concentration of the joint distribution of the parameters towards the region closer to the 'true value' may indicate that the modeler is now less unsure about the latent phenomenon that has produced the data that she was working with. These figures of joint distributions are constructed simply using the prior and posterior distributions of each parameter λ and σ obtained as a result of samples drawn with MCMC.

2.3.2 Gaussian Process model: A Bayesian nonparametric model

Despite the name, a Bayesian nonparametric model is not the case where the model has no parameters at all but is instead a really large Bayesian parametric model, where the dimension of the parameter space can be arbitrarily large. From Section 2.3.1, we have seen that a parametric model for regression is specified by a functional form governed by its respective fixed number of parameters. Nonparametric models for regression on