

Bayesian Parameter Estimation for Process Monitoring

by

Marno Basson

Thesis presented in partial fulfilment
of the requirements for the Degree

of

MASTER OF ENGINEERING
(EXTRACTIVE METALLURGICAL ENGINEERING)

in the Faculty of Engineering
at Stellenbosch University

Supervisor

Dr JT Cripwell
Process Engineering
Stellenbosch University

Co-Supervisor/s

Dr L Auret
Process Engineering
Stellenbosch University &
Data Science and Process Manager
Stone Three Digital

Prof RLJ Coetzer
School of Industrial Engineering
North-West University &
Senior Manager Machine Learning and Statistics
SASOL (Central Digital Office)

Prof RS Kroon
Computer Science
Stellenbosch University

March 2020

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: March 2020

Plagiarism Declaration

1. Plagiarism is the use of ideas, material and other intellectual property of another's work and to present it as my own.
2. I agree that plagiarism is a punishable offence because it constitutes theft.
3. I also understand that direct translations are plagiarism.
4. Accordingly all quotations and contributions from any source whatsoever (including the internet) have been cited fully. I understand that the reproduction of text without quotation marks (even when the source is cited) is plagiarism.
5. I declare that the work contained in this document, except where otherwise stated, is my original work and that I have not previously (in its entirety or in part) submitted it for grading in this document or another document.

Initials and surname: M

Basson Date: 17 February 2020

Abstract

The underlying mechanism of many physical systems studied in engineering can be described by algebraic, ordinary differential and auxiliary equations. While these equations stem from engineering expertise, the principles underpinning the model development phase do not always provide sufficient insight into selecting suitable values for all the model parameters. Furthermore, it might not be possible to directly measure all the model parameters (which can be related to several physicochemical system properties) from the system under consideration due to physical, economic and time constraints. As a result, the engineer often has to estimate the model parameters from noise-corrupted, time series data obtained from the physical system, while simultaneously quantifying how reliable these parameter estimates are.

The purpose of the current study is to investigate model parameter estimation, from both the frequentist and Bayesian statistical inference perspectives, and to evaluate the merit of applying Bayesian probabilistic techniques in the chemical engineering setting. Two Bayesian parameter estimation methodologies were developed. The first methodology applies to estimating the parameters of lumped system algebraic dynamic models, while the second methodology is focused on lumped system ordinary differential equation model parameter estimation. Both proposed Bayesian methodologies were benchmarked against the Gauss-Newton nonlinear least squares implementation for which the resulting estimated model parameters have a (frequentist) maximum likelihood interpretation. The results obtained from the proposed Bayesian methodologies were compared to the benchmark approach results based on several performance criteria for a single data set manifestation as well as for multiple independently generated data sets. It was found that the proposed Bayesian methodologies, as well as the benchmark approaches, provide consistent parameter estimation results when compared to the simulation ground truth parameter values, across the multiple independent data sets.

Based on the parameter inference results obtained from the different case studies considered in the current work, it was determined that, from a pragmatic engineering perspective, there is no reason to favour the use of the proposed Bayesian methodologies over the frequentist benchmark approaches and vice versa as both approaches provide comparable results. However, the benefit of the Bayesian approach (which explicitly expresses the model parameter uncertainty) was illustrated by considering a simple cost-benefit analysis for several of the case studies where it was possible to make more informed engineering decisions under uncertainty compared to the traditional frequentist benchmark approach.

In conclusion, even though there is no noteworthy difference between the parameter inference results obtained from the benchmark and proposed Bayesian approaches, the value of the Bayesian approach shows up when one considers the subsequent application of the inferred parameters in day-to-day engineering tasks. Consequently, it is worth further exploring the benefit of applying probabilistic techniques and explicitly modeling with uncertainty, i.e. Bayesian statistical inference, in chemical engineering applications.

Opsomming

Die onderliggende meganisme van baie fisiese stelsels bestudeer in ingenieurswese kan beskryf word deur algebraïese, gewone differensiaal- en hulpvergelings. Terwyl hierdie vergelings uit ingenieurkundigheid stam, gee die beginsels wat die model ontwikkelingsfase ondersteun, nie altyd genoeg insig om gepaste waardes vir al die modelparameters te kies nie. Verder mag dit dalk nie moontlik wees om al die modelparameters (wat verband kan hou met verskeie fisikochemiese stelseienskappe) direk uit die stelsel onder oorweging te meet nie, as gevolg van fisiese, ekonomiese en tydbepenkings. As 'n resultaat moet die ingenieur dikwels die modelparameters uit geraas korrupte, tydreeks data verkry uit die fisiese stelsel, terwyl gelyktydig gekwantifiseer moet word hoe betroubaar hierdie parameter beraminge is.

Die doel van die huidige studie is om modelparameterberaming te ondersoek, uit beide die frekwentis en Bayesiaanse statistiese inferensie perspektiewe, en om die meriete van die toepassing van Bayesiaanse waarskynlikheidstegnieke in die chemiese ingenieursomgewing te evalueer. Twee Bayesiaanse parameterberamingmetodologieë is ontwikkel. Die eerste metodologie is van toepassing op die beraming van die parameters van saamgehoopte stelsel algebraïese dinamiese modelle, terwyl die tweede metodologie gefokus is op saamgehoopte stelsel ordinêre differensiaal vergelyking model parameterberaming. Beide voorgestelde Bayesiaanse metodologieë is genormeer teen die Gauss-Newton nie-liniêre kleinste kwadrate implementasie waarvoor die resulterende beraming modelparameters 'n (frekwentis) maksimum aanneemlikheid interpretasie het. Die resultate verkry uit die voorgestelde Bayesiaanse metodologieë is vergelyk met die normbenaderingresultate op verskeie doeltreffendheidskriteria vir 'n enkel datastel manifestasie sowel as vir veelvoudige onafhanklik gegenereerde datastelle. Dis gevind dat die voorgestelde Bayesiaanse metodologieë, sowel as die normbenaderings, konsekwente parameterbenaderingresultate lewer as vergelyk word met die simulatie grondkontroleparameterwaardes, regoor die veelvoudige onafhanklike datastelle.

Gebaseer op die parameter inferensieresultate verkry uit die verskillende gevallestudies beskou in die huidige werk, is dit bepaal dat, vanuit 'n pragmatiese ingenieursperspektief, daar geen rede is om die gebruik van die voorgestelde Bayesiaanse metodologieë oor die frekwentis normbenaderings en vice versa te gebruik nie, omdat beide benaderings vergelykbare resultate lewer. Die voordeel van die Bayesiaanse benadering (wat duidelik die modelparameter onsekerheid uitdruk) is geïllustreer deur 'n eenvoudige koste-voordeelanalise vir verskeie van die gevallestudies te beskou waar dit moontlik was om meer ingeligte ingenieursbesluite onder onsekerheid te maak, in vergelyking met die tradisionele frekwentis normbenadering.

Ten slotte, selfs al is daar nie merkwaardige verskille tussen die parameter inferensie resultate verkry uit die norm- en voorgestelde Bayesiaanse benaderings nie, kom die waarde van die Bayesiaanse benadering na vore as mens die daaropvolgende toepassing van die afgeleide parameters in dag-tot-dag ingenieurstake in ag neem. Gevolglik is dit die moeite werd om die voordeel van die toepassing van waarskynlikheidstegnieke en uitdruklike modellering met onsekerheid, i.e. Bayesiaanse statistiese inferensie, in chemiese ingenieurswese toepassings, verder te ondersoek.

Acknowledgements

The author would like to acknowledge the following individuals for their contributions towards the completion of this project:

1. The multitude of supervisors associated with this project, Dr Lidia Auret, Prof Steve Kroon, Prof Roelof Coetzer and Dr Jamie Cripwell for their guidance, management, support, motivation, patience and advice during the execution of this project.
2. Professor Hans Eggers (Department of Theoretical Physics, Stellenbosch University) for teaching me the fundamentals of Bayesian statistical inference.
3. My family for their support, love and encouragement. A special thanks to my mother for all the food she made me!
4. My friends Chanté du Toit, Ryan Pottinger, Grant Albertus, Sagal Rabikoosun, Candice Fritz, Kirsten la Vita, and Arné du Toit for their support and motivation.

Table of Contents

| | |
|--|-----------|
| Chapter 1: Introduction | 1 |
| 1.1. The Modeling Problem and Physical Systems..... | 2 |
| 1.2. Probability and Uncertainty | 3 |
| 1.2.1. Frequentist Statistical Inference..... | 3 |
| 1.2.2. Bayesian Statistical Inference | 4 |
| 1.3. Parameter Tracking..... | 5 |
| 1.3.1. Fault Detection and Isolation..... | 5 |
| 1.3.2. Condition-based Maintenance | 6 |
| 1.4. Why Parameter Estimation? | 6 |
| 1.5. Research Outcomes..... | 7 |
| 1.5.1. Aims..... | 7 |
| 1.5.2. Objectives | 7 |
| 1.5.3. Research Scope | 8 |
| 1.5.4. Contributions..... | 9 |
| 1.6. Thesis Overview | 10 |
| Chapter 2: Theoretical Background | 11 |
| 2.1. Introduction to Probability Theory | 11 |
| 2.1.1. Fundamental Rules of Probability Theory..... | 11 |
| 2.1.2. Bayes' Rule..... | 12 |
| 2.1.3. Bayes' Rule for Parameter Estimation..... | 13 |
| 2.1.4. Expected Values..... | 13 |
| 2.2. Concepts from Information Theory | 15 |
| 2.2.1. Entropy..... | 15 |
| 2.2.2. Kullback-Leibler Divergence..... | 15 |
| 2.3. Useful Distributions | 15 |
| 2.3.1. The Gaussian Distribution | 16 |
| 2.3.2. The Gamma Distribution | 20 |
| 2.3.3. Conjugacy | 21 |
| 2.4. Bayesian Linear Regression..... | 22 |
| 2.4.1. Bayesian Linear Regression - Probabilistic Model I | 22 |
| 2.4.2. Connection to Simple Least Squares Regression..... | 24 |
| 2.4.3. Bayesian Linear Regression - Probabilistic Model II | 28 |

| | |
|---|-----------|
| 2.5. Gaussian Processes | 29 |
| 2.5.1. Bayesian Linear Regression – Gaussian Process Motivation | 29 |
| 2.5.2. Gaussian Processes for Non-Parametric Regression | 31 |
| 2.6. Variational Inference | 36 |
| 2.7. Nonlinear Deterministic Functions | 37 |
| 2.8. Confidence vs Credibility Intervals | 38 |
| 2.9. Summary | 39 |
| Chapter 3: Literature Review | 41 |
| 3.1. Overview | 41 |
| 3.1.1. Previous Chapters | 41 |
| 3.1.2. Present Chapter | 42 |
| 3.2. Parameter Estimation Techniques | 42 |
| 3.3. Parameter Estimation Methods for Algebraic Dynamic Models | 43 |
| 3.4. Parameter Estimation Methods for ODE Dynamic Models | 48 |
| 3.5. Remark: Suitability of the Gaussian Process Regression Approach | 53 |
| 3.6. Summary | 53 |
| Chapter 4: Proposed Approaches and Methodology | 55 |
| 4.1. Overview | 55 |
| 4.2. Simulation Case Studies | 55 |
| 4.2.1. Case Study 1 & 2: Continuous Stirred Tank Reactor | 55 |
| 4.2.2. Case Study 3: Liquid Draining Tank | 58 |
| 4.3. Parameter Estimates Required | 60 |
| 4.4. Benchmark Method | 60 |
| 4.5. Bayesian Parameter Estimation Methods | 64 |
| 4.5.1. Variational Bayesian Nonlinear Regression | 64 |
| 4.5.2. ODE Parameter Estimation via Gaussian Process Gradient Matching | 70 |
| 4.6. Parameter Tracking Applications | 80 |
| 4.6.1. Extended Case Study 2: CSTR with Catalyst Decay | 80 |
| 4.6.2. Extended Case Study 3: Liquid Draining Tank with Valve Degradation | 82 |
| 4.7. Data Generation Process | 83 |
| 4.7.1. Case Study 1: Isothermal CSTR – Perfect Input Step Disturbance | 83 |
| 4.7.2. Case Study 2: Isothermal CSTR – Random Exogenous Input Disturbance | 84 |
| 4.7.3. Case Study 3: Draining Tank – Random Exogenous Input Disturbance | 85 |
| 4.8. Performance Criteria | 86 |
| 4.9. Summary | 87 |

| | |
|--|------------|
| Chapter 5: Results and Discussions | 89 |
| 5.1. Overview..... | 89 |
| 5.2. Case Study Parameter Inference Results | 89 |
| 5.2.1. Case Study 1: Isothermal CSTR – Perfect Step Input Disturbance | 89 |
| 5.2.2. Case Study 2: Isothermal CSTR – Exogenous Input Disturbance..... | 94 |
| 5.2.3. Case Study 3: Liquid Draining Tank – Exogenous Input Disturbance..... | 99 |
| 5.3. Parameter Tracking Application Results | 103 |
| 5.3.1. Extended Case Study 2: Isothermal CSTR with Catalyst Decay..... | 103 |
| 5.3.2. Extended Case Study 3: Draining Tank –Valve Degradation | 110 |
| 5.4. Comment on ‘Conservative’ Estimates and Coverage Frequencies | 117 |
| 5.5. Summary | 118 |
| Chapter 6: Conclusions | 121 |
| 6.1. Frequentist vs Bayesian | 121 |
| 6.2. Review of Objectives..... | 123 |
| 6.3. Novelty and Contribution | 124 |
| Chapter 7: Recommendations | 125 |
| 7.1. Considering more Complex Systems..... | 125 |
| 7.2. Bayesian Practicalities | 126 |
| 7.3. Gaussian Process Considerations..... | 127 |
| 7.4. Algorithm 4 Practicalities | 128 |
| References | 131 |
| Appendix A: Illustrative Example..... | 135 |
| A.1. Frequentist Viewpoint..... | 137 |
| A.2. Bayesian Viewpoint | 138 |
| A.3. On-Line Learning/Sequential Estimation | 143 |

Acronyms

| | |
|----------------|---|
| CAVI | Coordinate ascent variational inference |
| CSTR | Continuous stirred tank reactor |
| ELBO | Evidence lower bound |
| FDI | Fault detection and isolation |
| \mathcal{GP} | Gaussian process |
| HOT | Higher order terms |
| \mathcal{KL} | Kullback-Leibler divergence |
| MAP | Maximum <i>a posteriori</i> |
| MCMC | Markov chain Monte Carlo |
| ML | Maximum likelihood |
| NOC | Normal operating conditions |
| ODE | Ordinary differential equation |
| RUL | Remaining useful life |

Greek Symbols

| | |
|------------------------|--|
| β | Precision parameter of univariate Gaussian distribution |
| β_{ML} | Maximum likelihood precision parameter estimate (frequentist) |
| $\Gamma(a)$ | Gamma function |
| ϵ_n | Zero-mean Gaussian distributed noise |
| θ | Dynamic model parameters (pertains to any arbitrary model) |
| μ | Mean parameter of univariate Gaussian distribution |
| $\boldsymbol{\mu}$ | Mean vector of multivariate Gaussian distribution |
| ρ_{step} | Stepping parameter size |
| σ | Standard deviation parameter of univariate Gaussian distribution |
| σ^2 | Variance parameter of univariate Gaussian distribution |
| σ_{noise}^2 | Sensor variance parameter (follow the subscript notation used) |
| $\boldsymbol{\Sigma}$ | Covariance matrix of multivariate Gaussian distribution |
| τ | Process time constant |
| $\boldsymbol{\phi}(x)$ | Vector containing model basis functions |
| $\boldsymbol{\Phi}$ | Regression design matrix |
| $\psi(a)$ | Digamma function |
| $\boldsymbol{\Omega}$ | Parameter vector (reserved for algebraic dynamic models) |

Other Symbols

| | |
|-------------------------------------|--|
| $\mathbf{0}$ | Zero vector |
| A | Cross-sectional area |
| \mathbf{B} | Relevant Gaussian process matrix (follow the subscript notation used) |
| C_A | Outlet concentration |
| C_{A0} | Inlet concentration |
| \mathbf{C}_A | Outlet concentration - function values |
| \mathbf{C}'_A | Rate of change of outlet concentration - function values |
| \mathbf{C}_{A0} | Inlet concentration - function values |
| $cov[x, y]$ | Covariance between random variables x and y |
| $dom(y)$ | Domain of random variable y |
| \mathcal{D}, \mathbf{d} | Data set (related to specific variable via the appropriate subscript notation) |
| $\mathbb{E}_{p(x)}[f(x)]$ | Expected value of $f(x)$ under the distribution $p(x)$ |
| $E(\mathbf{w}), E(\mathbf{\Omega})$ | Simple least squares error/objective function |
| F_0 | Inlet flow rate - function values |
| F | Flow rate |
| $f(x)$ | Function of the random variable x |
| $\mathbb{H}[p(x)]$ | Entropy of distribution $p(x)$ |
| $\mathbf{I}_{N \times N}$ | N-by-N identity matrix |
| \mathbf{J}_R | Reduced Jacobian matrix |
| \mathbf{J} | Jacobian matrix |
| K_p | Process gain |
| k_v | Flow restriction coefficient |
| k | Reaction rate constant |
| $k(t_i, t_m)$ | Kernel function |
| \mathbf{K} | Gram matrix (reserved for Gaussian process regression in Chapter 2) or diagonal matrix used to compute \mathbf{J}_R (reserved for <i>Algorithm 2</i> in Chapter 4) |
| \mathbf{L}' | Rate of change of liquid level - function values |
| \mathbf{L}_s | Vector containing liquid level steady state sensor measurements |

| | |
|---|--|
| \mathcal{L} | Likelihood function (used in Appendix A) or ELBO (used in Chapter 4) |
| L | Liquid level |
| \mathbf{L} | Liquid level - function values |
| \mathbf{m}_0 | Gaussian prior distribution mean vector |
| \mathbf{m}_N | Gaussian posterior distribution mean vector |
| \mathcal{M} | Relevant model under consideration |
| M | Number of regression model parameters |
| N | Number of sensor measurements/observations |
| $P_{CSTR}(t)$ | Isothermal, constant volume CSTR with catalyst decay - profit function |
| $P_{Tank}(t)$ | Liquid draining tank with valve degradation - profit function |
| $p(x, y)$ | Joint probability distribution over random variables x and y |
| $p(x), p(y)$ | Probability distribution over random variables x or y |
| $p(\mathbf{x})$ | Probability distribution over the vector random variable \mathbf{x} |
| $q(\mathbf{w}), q(\beta)$ | Variational approximating posterior distributions (mean-field family) |
| \mathbf{S}_0 | Gaussian prior distribution covariance matrix |
| \mathbf{S}_N | Gaussian posterior distribution covariance matrix |
| \mathcal{S}_ρ | Physical system description under consideration |
| t | Time |
| V | Volume |
| $\text{Var}[f(x)]$ | Variance of function $f(x)$ |
| $\text{Var}[x]$ | Variance of random variable x |
| \mathbf{w} | Parameter vector (reserved for ordinary differential equation models) |
| $\mathbf{w}_{MAP}, \mathbf{\Omega}_{MAP}$ | Maximum <i>a posteriori</i> parameter estimates (Bayesian) |
| $\mathbf{w}_{ML}, \mathbf{\Omega}_{ML}$ | Maximum likelihood parameter estimates (frequentist) |
| x^* | Test/prediction point (reserved for Gaussian process regression) |
| x, y | Random variables x and y |
| \mathbf{x} | Vector random variable |
| \mathbf{z} | Latent variable vector |

Chapter 1

Introduction

“All models are wrong but some are useful.”

- George E.P. Box, 1976

Scientific discovery and exploration are based on hypothesising models from experimental data and studying these models' properties to obtain an understanding of the underlying system phenomena generating the data. The problem of discovering patterns in systems is fundamental to our understanding of dynamical systems. Models, whether in the form of hypotheses, physical laws or empirical equations, all attempt to link data to an underlying system. This allows the experimentalist to identify useful interpretable properties which can subsequently be used to make informed decisions and to take appropriate actions (Ljung, 1999; Bishop, 2006). From an industrial process control perspective, informed decision making is of particular importance when considering (1) smooth plant operating conditions, (2) annual production rate, (3) product quality, and (4) profit optimisation. Furthermore, depending on the decision-making process, executing the appropriate actions can play a vital role in (1) safety and environmental considerations, (2) process equipment protection and (3) monitoring and diagnosis of process irregularities (Marlin, 2000).

System identification predominantly deals with constructing a mathematical model of a dynamical system from data, and, is thus a realisation of the scientific methodology. Since dynamical systems are widespread across a multitude of scientific environments, system identification is widely applicable (Ljung, 1999). Dynamic models - a particular type of mathematical description of a dynamical system - are of particular interest in the chemical engineering context as these models can be used to establish relationships between process variables (Marlin, 2000). Lumped system dynamic models derived from fundamental principles, such as the conservation of mass or energy principle, employ nonlinear ordinary differential equations (ODEs) to relate process variables. (Marlin, 2000; Calderhead, Girolami and Lawrence, 2009).

This thesis assumes from the outset that a set of candidate lumped system dynamic models parameterised by a vector θ are available and considers the problem of finding the best model, i.e. the model that best describes the observational data, by identifying the most suitable parameter vector θ^* . A multitude of methods exist for finding θ^* , each of which organises the procedure for selecting θ^* in a different way following a set of rules and heuristics (Ljung, 1999).

The problem faced within the context of the current study, refer to Figure 1.1, is to decide on how to use a dataset \mathcal{D} to select θ^* , and implicitly, the relevant model member $\mathcal{M}(\theta^*)$ in the parameterised candidate model set. Such procedures are known in statistics as *parameter estimation methods* (Ljung, 1999; Migal, 2008)

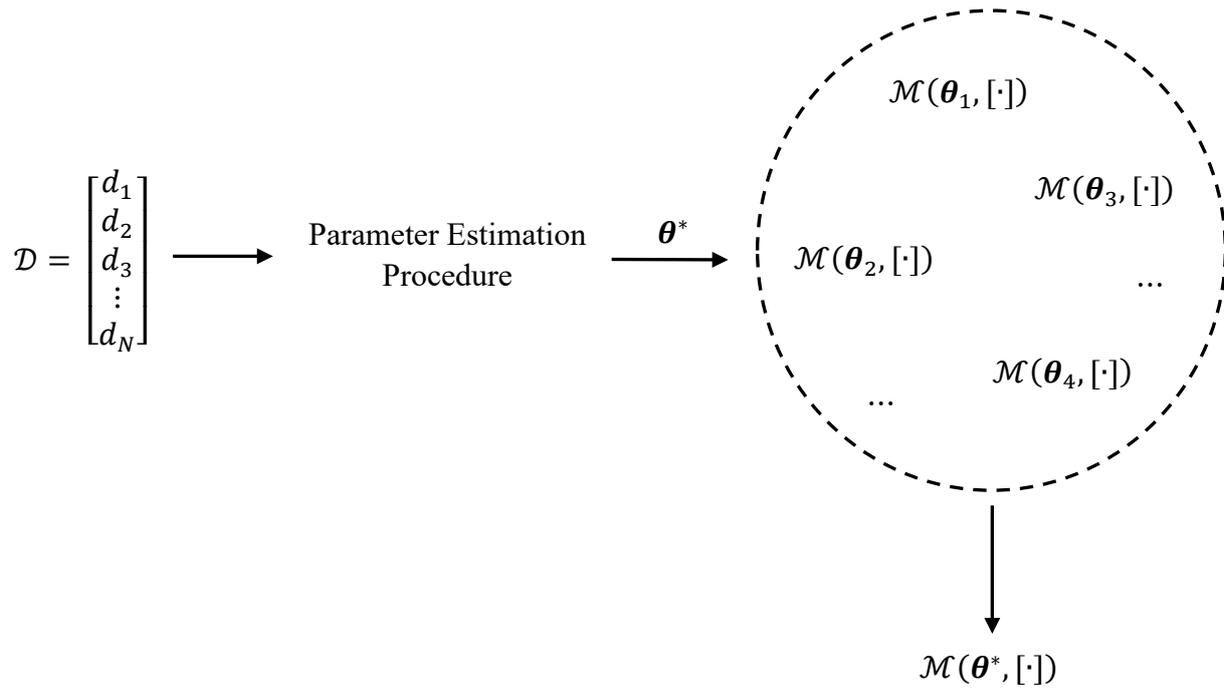


Figure 1.1: General parameter estimation procedure. Given a data set \mathcal{D} and a parameter estimation procedure, find the most suitable parameter vector θ^* such that one can select the relevant model member $\mathcal{M}(\theta^*)$ from the candidate model set

The notation $[\cdot]$ makes explicit that the model can have various inputs depending on the physical process considered. Standard parameter estimation techniques typically provide a point estimate for θ , however, a more general description of θ is obtained by defining a probability distribution over the space of possible parameter values. The argument follows that given a set of experimental data, an *a priori* description of the model parameters, and a dynamic model of the physical process under consideration, one can obtain an *a posteriori* description of θ . A natural question that might arise is “Why should the engineer be interested in the posterior description of θ ?” One possible reason stems from the fact that the posterior distribution explicitly expresses the uncertainty associated with the model parameters which can subsequently be used for more informed engineering decision-making (Section 5.3) (Ljung, 1999; Bishop, 2006; Murphy, 2012).

1.1. The Modeling Problem and Physical Systems

Let \mathcal{S}_p be the description of the physical system under consideration. Two examples of \mathcal{S}_p considered in this thesis are the descriptions of (1) a liquid draining tank with a flow restriction, and (2) an isothermal, constant volume, continuous stirred tank reactor (CSTR) with first-order reaction kinetics. It is assumed that \mathcal{S}_p includes all the necessary information to completely characterise the physical system under consideration, i.e. the set of equations developed during the model development phase are known exactly and sufficiently describe the physical process. However, the principles underpinning the model development phase do not always provide insight into selecting suitable values for all the model parameters and these parameters cannot always be measured directly from the physical system. As a result, the model parameters have to be estimated from noise-corrupted time series data. Tarantola (2005) outlines the scientific procedure for the study of physical systems into the following three steps:

CHAPTER 1: INTRODUCTION

- (1) **System parameterisation** – this step focuses on discovering the minimal set of model parameters that can completely characterise the dynamical system. Within the chemical engineering context, one can achieve this through various modelling considerations, with the most significant consideration being the final application of the model. In certain modeling situations, a more complex model is essential, and this typically coincides with a model that requires more parameters to explain the experimental data. (Marlin, 2000).
- (2) **Forward modeling** – within the context of the thesis, this step focuses on using dynamic models derived from physical principles, such as the conservation of mass and energy principle, to make deterministic predictions, i.e. given a sufficiently detailed description of a physical system with the model parameters *known exactly*, one can predict the outcome of a measurement. The task of predicting the results of measurements exactly from a system description is known as the *forward modeling procedure*. The forward modelling procedure, concerning deterministic physics, has a unique solution. One can also refer to the forward modelling procedure as the *simulation procedure*. In other words, given an assumed ground truth, one can generate noise-free measurements for in a deterministic manner.
- (3) **Inverse modeling** – the task of inverse modeling reverses the direction of the forward modeling procedure. Given experimental data and a dynamic model with *unknown* model parameters, one would like to make conclusions about the actual values of the dynamic model parameters. However, due to noise-corrupted sensor measurement data, there is uncertainty associated with the estimated model parameters. As a result, no unique solution is typically available (cf. *Forward modelling*), and one would ideally like a modeling framework that explicitly incorporates the parameter uncertainty (Bishop, 2006).

1.2. Probability and Uncertainty

One can view the concept of probability from two different paradigms. The first views it in terms of infinitely repeated sampling, and is called the frequentist interpretation of probability. The second view of probability is the Bayesian interpretation, also referred to as the subjective probability paradigm, where probability provides a quantification of uncertainty (Bishop, 2006).

1.2.1. Frequentist Statistical Inference

The frequentist approach to statistical inference treats the model parameters as unknown but fixed values and relies on several heuristics to find *point estimates* for these parameters from data. Two popular heuristics from the frequentist statistical literature includes the Maximum Likelihood and Unbiasedness heuristics. Once the experimentalist has decided on which heuristic to use, the uncertainty associated with the model parameter estimates is obtained from a sampling distribution, which is the distribution of the estimates generated by the proposed estimator applied to multiple data sets. From the sampling distribution, *confidence intervals* (Section 2.8) can be constructed to reflect the experimentalist's uncertainty in the estimated model parameters (Hastie et al., 2006; Hastie, Tibshirani and Friedman, 2009).

1.2.2. Bayesian Statistical Inference

The Bayesian interpretation of statistical inference treats the model parameters as unknown and thus random variables themselves. This is because the Bayesian interpretation uses probability theory to quantify uncertainty. Using Bayes' rule, the experimentalist can obtain a probability distribution over the unknown model parameters in contrast to the frequentist parameter *point estimates*. Once the experimentalist has obtained the probability distribution, the uncertainty associated with the model parameters is described by the probability distribution itself. Plausible values can be extracted from this distribution to construct *credibility intervals* (Section 2.8) (cf. *confidence intervals*).

Thus, one observes that Bayesian statistical inference explicitly incorporates uncertainty into the inference procedure avoiding the need to (1) estimate the model parameters by choosing between several heuristics, and, (2) does not require a sampling distribution to quantify uncertainty (Jaynes, 2003; MacKay, 2004; Bishop, 2006; Barber, 2012; Murphy, 2012; Sivia and Skilling, 2012). The current thesis draws on Bayesian statistical inference, thus, in a preview of subsequent chapters and the illustrative examples, the author would like to explicitly point out several advantages (and disadvantages) of Bayesian analysis that might appear *ad hoc* for the moment but will become clear as the thesis progresses:

- (i) The Bayesian methodology is grounded on fundamental rules (*sum and product rule*) and relies on operations such as *marginalisation* and *conditioning* to perform inference. It allows one to avoid choosing between *ad hoc* heuristics to estimate model parameters.
- (ii) Most Bayesian models of practical interest are intractable and does not scale well to big data settings. However, due to recent advances in approximate inference techniques and computational performance, the practical applicability of Bayesian methods has been greatly enhanced and become mainstream.
- (iii) Bayesian inference provides a principled way to incorporate both prior engineering knowledge (about the unknown dynamic model parameters) and experimental data. However, the Bayesian methodology does not tell the experimentalist how to select the prior distribution. In fact, there is no 'correct' way to select a prior, and Bayesian inference requires the experimentalist to translate their prior beliefs into a mathematical formulation.
- (iv) Furthermore, the posterior distribution results can be heavily influenced by the choice of prior. From a practical engineering perspective, it might be difficult to convince other experimentalists of the selected prior distribution and the attained posterior distribution results. Consequently, it is beneficial to compare the Bayesian inference results to the results of a traditionally practised and well-established benchmark technique.
- (v) Bayes' rule is inherently online. As new observations become available, the current posterior distribution can be used as the prior distribution for future analysis.
- (vi) One of the most desirable benefits of Bayesian inference is the principled way in which uncertainty is incorporated into the inference procedure and the ease of interpreting the credibility interval (Section 2.8).
- (vii) Another major benefit of the Bayesian methodology is the ease of extensibility. From the frequentist statistical inference perspective, it is not always obvious how to extend the inference procedure to more complex inference scenarios. However, from the Bayesian perspective, more complex inference scenarios typically correspond to dealing with

CHAPTER 1: INTRODUCTION

additional distributions where Bayes' rule guides the experimentalist on how to logically perform inference with the added complexity. Thus, all inference procedures, whether simplistic or complex, logically follow from Bayes' rule.

Van de Schoot et al. (2014) provide an overview of some of the similarities and differences between the frequentist and Bayesian view of statistics. Refer to Table 1.1.

Table 1.1: Similarities and differences between the Bayesian and frequentist view of statistics [Table adapted from Van de Schoot et al. (2014)].

| | Frequentist | Bayesian |
|---|------------------------------------|--|
| Possible to include prior knowledge? | No | Yes |
| Treatment of model parameters | Unknown and treated as fixed | Unknown and therefore treated as random |
| Nature of parameter estimates | Point estimates | A distribution over possible parameter values |
| Definition of the uncertainty | Based on the sampling distribution | Captured by the probability distribution over the model parameters |
| Expression of uncertainty intervals (Section 2.8) | Confidence intervals | Credibility intervals |

The similarities and differences (as well as the associated implications thereof) outlined in Table 1.1 will become clear in Sections 4.4 (frequentist benchmark), 4.5 (proposed Bayesian methodologies) and 5.3 (parameter tracking application results).

1.3. Parameter Tracking

Parameter tracking techniques attempt to monitor processes directly based on the use of process model parameters that have a physical interpretation while state variable techniques typically assume the process model parameters are known and monitor the process signals (Isermann, 1984, 1985). While these methods are complementary, the current thesis focuses on parameter estimation techniques with parameter tracking applications. The parameter tracking methodology is as follows: Dynamic model parameters relate to several physical system parameters such as area, reaction rate coefficient and density, among others. Faults that manifest themselves as changes in physical system parameters also manifest as changes in the process model parameters. Based on this methodology, if unexpected changes in the process model parameters are detected, this triggers an alarm of a possible fault condition within the process (Isermann, 1984, 1985; Jardine, Lin and Banjevic, 2006).

1.3.1. Fault Detection and Isolation

Fault detection and isolation (FDI) plays a vital role in industrial settings, especially in the context of equipment protection, environmental considerations, and employee safety. Process models (as derived from first principles), parameter estimation techniques, and decision rules make it possible to monitor physical processes. However, FDI is also possible from a purely

CHAPTER 1: INTRODUCTION

data-driven perspective with methods such as principal component and factor analysis. An essential characteristic of any automatic supervision system is early fault detection and isolation such that one can instigate corrective action to avoid undesirable plant operating behaviour. A possible way of incorporating such preventative action is through the use of process models that relate process variables to one another. However, these process models require suitable values for the model parameters. As a result, the inverse modeling problem, i.e. determining the model parameters θ from noise-corrupted data, plays a crucial role (Isermann, 1984, 1985; Poshtan, Doraiswami and Stevenson, 1997; Marlin, 2000; Migal, 2008; Zhiqiang and Zhihuan, 2013).

Isermann (1984) defines a *fault* as a non-permitted deviation of some dynamical system property which leads to the inability of this property to fulfil an intended purpose. If such non-permitted deviations occur, it should be detected as early as possible. *Fault detection* commences by checking if particular measurable or unmeasurable estimated variables are within pre-specified tolerance thresholds. Typically, these tolerance thresholds rely on the notion of nominal operating conditions (NOC), and deviation from NOC triggers an alarm of a possible fault condition in the process. Once a process fault is detected, *fault diagnosis (isolation)* commences (Isermann, 1984; Marlin, 2000; Seborg et al., 2011).

1.3.2. Condition-based Maintenance

Condition-based maintenance is concerned with establishing a maintenance policy which can inform the engineer when maintenance should be performed and what associated actions should be taken. Hence, Condition-based maintenance is concerned with preventing system failure by performing preventative system equipment replacement (Ghasemi, Yacout, and Ouali 2009, 2010). Condition monitoring entails observing and gathering information about the degradation condition of system equipment to prevent future failure and to determine the appropriate maintenance actions (Jardine, Lin and Banjevic, 2006).

When one subjects a piece of system equipment to condition monitoring, data about degradation indicators are periodically collected to diagnose the equipment condition and to establish a prognosis of future performance. This diagnosis and prognosis procedure is based on parameterised mathematical models with several unknown model parameters. In order to perform Condition-based maintenance, i.e. performing preventative system equipment replacement, the unknown model parameters must be estimated from data (Martin, 1994; Ghasemi, Yacout and Ouali, 2010).

1.4. Why Parameter Estimation?

A central theme that arises in both the *Fault Detection and Isolation* and *Condition-based Maintenance* settings is the idea of using parameters to perform practical engineering tasks on a day-to-day basis. Given the important role parameters play in engineering tasks, one would ideally like a methodology that can ‘reliably’ estimate the unknown dynamic model parameters from noise-corrupted time series data. For the purposes of this thesis, a ‘reliable’ estimate of the dynamic model parameters is defined as an estimate that is *good in quality/accuracy with a pragmatic interpretation of the associated parameter uncertainty*. The practical engineering importance of parameters form the basis for the current research.

CHAPTER 1: INTRODUCTION

Several techniques exist for estimating the parameters of dynamic models from noise-corrupted time series data. However, in recent years, Gaussian process regression (which is a nonparametric Bayesian approach) has established itself as a successful tool in the field of system identification (Gorbach, Bauer and Buhmann, 2017; Wenk et al., 2018). Gaussian processes have been used in conjunction with nonlinear finite-impulse response (NFIR) (Ackermann, De Villiers and Cilliers, 2011) as well as nonlinear autoregressive (NARX) (Kocijan et al., 2005) models to construct time series prediction models. Further applications include extensions to state-space and nonlinear Box Jenkins models making Gaussian processes a versatile tools in the field of system identification (Kocijan, 2016; Särkkä, 2019).

1.5. Research Outcomes

The underlying mechanism of many physical systems in engineering can be described by algebraic, ordinary differential and auxiliary equations. These equations stem from expert knowledge, however, the principles underpinning the model development phase do not always provide insight into selecting suitable values for all the model parameters. Consequently, one has to estimate the model parameters from noise-corrupted observational data using the inverse modeling approach, i.e. the experimentalist has to explicitly deal with uncertainty.

1.5.1. Aims

The current study investigates a Gaussian process based approach for system identification, with a specific focus on parameter learning of chemical engineering systems. It is based on the work of Calderhead, Girolami and Lawrence (2009), Dondelinger et al. (2013), Gorbach, Bauer and Buhmann (2017) and Wenk et al. (2018). The Gaussian process based approach, otherwise referred to as gradient matching in the relevant literature, enables Bayesian inference of parameters for ordinary differential equations without explicitly solving the system of equations. The idea behind gradient matching is to minimise the difference between two calculations of time derivatives: one provided by the ordinary differential equations describing the physical system state variables, and another from a Gaussian process interpolating the dynamics of the state variables (Wenk et al., 2018).

The previously mentioned work focuses on simultaneously inferring posterior distributions over the system state variables and the model parameters for coupled ordinary differential equations. This thesis aims to apply and evaluate a similar gradient matching Gaussian process based approach but restricts attention to lumped system dynamic model parameter inference only for a single continuous-time model with time-invariant parameters.

1.5.2. Objectives

The current study outlines the following objectives to achieve the aim provided in section 1.5.1, successfully:

- (1) Dynamic modeling and simulation of process unit case studies using the forward modeling approach with the addition of sensor measurement noise.

CHAPTER 1: INTRODUCTION

Motivation: By using a dynamic model (with an assumed ground truth) in combination with a simulation environment – for example, MATLAB[®]/Simulink – one can generate data which captures important properties of many real-world data sets, namely that these data sets have an underlying regularity, which one would like to learn, but individual observations are noise-corrupted.

- (2) The proposal and application of various Bayesian inference techniques for estimating the parameters of lumped system dynamic models.

Motivation: Over the last few decades, Bayesian techniques have become increasingly mainstream. Bayesian methods are used for various practical applications since these techniques allow one to model uncertainty explicitly. Due to recent advances in approximate inference algorithms and computational performance, Bayesian techniques can scale to large industrial settings which allow practitioners to exploit the incremental learning characteristic of these methods. Both modelling uncertainty explicitly and real-time learning are of particular importance in an industrial chemical engineering setting, thus, the author intends to adopt the Bayesian methodology within the current research framework (Bishop, 2006; Fox and Roberts, 2012; Murphy, 2012; Blei, Kucukelbir and McAuliffe, 2018).

- (3) Benchmark the proposed Bayesian techniques against traditional parameter estimation methods.

Motivation: In order to establish whether the proposed Bayesian approaches provide sensible results, i.e. parameter estimates that are *good in quality/accuracy*, it is necessary to compare the Bayesian parameter results to parameter estimates obtained from traditional parameter estimation methods. Additionally, since the data is generated using the forward modeling approach, the results from both the proposed Bayesian approaches and traditional parameter estimation methods can be compared to the simulation ground truth.

- (4) Illustrate the application of the proposed Bayesian and benchmark parameter estimation techniques to parameter tracking applications.

Motivation: Due to the principled way in which Bayesian inference incorporates prior knowledge and experimental data, in conjunction with the practical importance of parameters in engineering tasks, it is worth investigating how the results obtained from the proposed Bayesian approaches measure up against the results from traditional parameter estimation methods in the context of parameter tracking applications. If similar results are obtained, then this can serve as motivation to recommend the use of Bayesian inference, which is grounded on the fundamental rules of probability theory (*sum and product rule*) where all inference procedures logically follow from Bayes' rule.

1.5.3. Research Scope

The primary research scope of the current study comprises the following aspects:

- (1) Attention is given to a single lumped system continuous-time dynamic model with time-invariant parameters. Furthermore, the current study restricts attention to continuous-time models that are linear in the model parameters. In scenarios where the dynamic

CHAPTER 1: INTRODUCTION

model is nonlinear in the model parameters, the nonlinear model is approximated by a first-order Taylor expansion. While there is no linearity restriction imposed on the model state variables, the Bayesian methods developed in this thesis only apply to systems with the above-mentioned restrictions.

- (2) Any form of automated control is excluded from the case study process units. The author does not intend to consider further scalability of any of the outlined approaches.
- (3) Two case studies, inspired by Marlin (2000), are used for illustrative purposes throughout this thesis: The first is an isothermal, constant volume CSTR with first-order reaction kinetics. This particular choice of process unit results in a linear (in the state) ordinary differential equation model that is also linear in the model parameters and can be solved explicitly under certain conditions. The second case study is a single liquid draining tank with a flow restriction.

This tank unit is considered for two reasons. From a modelling perspective, the draining tank model results in a nonlinear (in the state) ordinary differential equation that is linear in the model parameters. From a pragmatic perspective, draining tanks are common process units encountered in everyday life, e.g. tanks used to collect rainwater or potable water systems installed for recycling and household usage. Thus, familiarity with single tank systems may aid in understanding some of the advanced concepts used throughout this thesis.

- (4) Both case studies contain two adjustable model parameters that are related to physical process parameters. Having only two adjustable model parameters allows visual interpretation of the results, which makes the results more accessible to readers that might not necessarily be familiar with the mathematics behind the parameter inference procedures.
- (5) It is intended to motivate and show proof of concept for the Bayesian approach to inference not only from a purist theoretical stance by explicitly modeling uncertainty but also from a pragmatic engineering viewpoint.
- (6) The proposed thesis audience is graduate and undergraduate chemical engineers with a working knowledge of linear algebra, multivariate calculus, and some familiarity with probability theory. However, the thesis includes a self-contained overview of the basic concepts of probability theory. Furthermore, the author assumes that the target audience is familiar with the concept of mathematical modelling and the simulation of chemical engineering processes.

1.5.4. Contributions

The chief contributions of this thesis are:

- (1) An introduction to probabilistic techniques and modelling with uncertainty within the chemical engineering framework, with illustrative case studies.

CHAPTER 1: INTRODUCTION

- (2) The implementation and evaluation of a gradient matching Gaussian process based technique for estimating the parameters of the dynamic model case studies to serve as proof of concept for further application.
- (3) An extension of the current Gaussian process based parameter inference procedures (Table 3.2) to include ordinary differential equations with an arbitrary exogenous input disturbance structure (Sections 3.5 and 4.5.2).
- (4) Open-source software for each simulation case study with the corresponding code illustrating the proposed method's implementation, and the obtained results.

The MATLAB[®]/Simulink code used in the current work for implementing *Algorithms 1* through *4* (introduced in Sections 4.4 and 4.5), as applied to the the various case studies (introduced in Sections 4.2 and 4.6), is freely available at (MATLAB R2018b and higher required):

Open-source software URL:

<https://gitlab.com/pleased/bayesian-ode-parameter-estimation>

1.6. Thesis Overview

The remainder of the thesis is structured as follows: Chapter 2 provides a brief theoretical background with a self-contained introduction to concepts of probability theory – an example of the application of Bayes' rule is presented in Appendix A. Chapter 3 reviews the relevant frequentist and Bayesian parameter estimation literature before Chapter 4 outlines the proposed Bayesian approaches and benchmark techniques used in the current work. Chapter 5 presents and discusses the results followed by the main conclusions in Chapter 6. Lastly, Chapter 7 outlines potential pitfalls, limitations, and recommendations for further research

Chapter 2

Theoretical Background

“Probability made sense, but was just a game; statistics was important, but it was a bewildering collection of tests with little obvious rhyme or reason.”

- Devinder S. Sivia, 2012

2.1. Introduction to Probability Theory

Given a lumped system dynamic model of a physical process, the primary goal is to estimate the dynamic model parameters θ . In a practical setting, engineers obtain data of a physical system by measuring physical properties of the system using sensors. However, since the measurement equipment is not perfect, the engineer only observes a noise-corrupted version of the true physical properties. It is precisely this imperfect state of knowledge about the true underlying system physical properties that gives rise to the inverse modeling procedure (Section 1.1). In other words, the engineer would like to make conclusions about the model parameter values θ from the measured physical system properties, however, the measurements are noise-corrupted and, as a result, the engineer is uncertain about the values of θ .

2.1.1. Fundamental Rules of Probability Theory

The two fundamental rules of probability theory, which form the basis for all the probabilistic machinery used throughout this thesis, are the sum and product rule. Consider two discrete random variables x and y , i.e. random variables that can take on a finite set of values.

The Sum Rule

$$p(x) = \sum_y p(x, y) \quad \text{Equation 1.1}$$

The sum rule, given by Equation 1.1, for discrete random variables, states that the *marginal probability* $p(x)$ can be obtained from the *joint probability* $p(x, y)$, verbalised as *the probability of x and y* , by summing over all possible values of the random variable y . The marginal probability $p(x)$ is verbalised as *the probability of x* . The sum rule can be extended to continuous random variables by replacing the summation with integration to obtain Equation 1.2.

$$p(x) = \int_{dom(y)} p(x, y) dy \quad \text{Equation 1.2}$$

The notation $dom(y)$ refers to integrating over the domain of the random variable y . A justification for the extension of the sum rule to continuous random variables requires a branch of mathematics called Measure Theory and falls outside the scope of the current thesis (Jaynes, 2003; MacKay, 2004; Bishop, 2006; Barber, 2012).

CHAPTER 2: THEORETICAL BACKGROUND

The Product Rule

$$p(x, y) = p(x|y)p(y) \text{ where } p(y) \neq 0 \quad \text{Equation 1.3}$$

Equation 1.3 states that the *joint probability* $p(x, y)$ over x and y can be factored into a product of a *conditional probability* $p(x|y)$ and a marginal probability $p(y)$. The *conditional probability* $p(x|y)$ is verbalised as *the probability of x given y* (Jaynes, 2003; MacKay, 2004; Bishop, 2006; Barber, 2012).

An important assumption often encountered in the statistics and machine learning literature is *independence* (Equation 1.4).

Independence

$$p(x, y) = p(x)p(y) \quad \text{Equation 1.4}$$

Random variables x and y are independent if knowledge about the state of one variable provides no additional information about the other variable. That is, the joint probability of x and y factorises into the product of marginal probabilities (Jaynes, 2003; MacKay, 2004; Bishop, 2006; Barber, 2012).

2.1.2. Bayes' Rule

The specific factorisation expressed in Equation 1.3 is not unique. An alternative and equally valid factorisation would be to state that the joint probability $p(x, y)$ can factorise as:

$$p(x, y) = p(y|x)p(x) \quad \text{Equation 1.5}$$

From Equation 1.3 and 2.5, one observes that both factorisations express the same joint probability $p(x, y)$. Equating Equation 1.3 and 2.5, one can immediately obtain a relationship between conditional probabilities, known as Bayes' rule (Bishop, 2006):

Bayes' Rule

$$p(y|x)p(x) = p(x|y)p(y) \quad \text{Equation 1.6}$$

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} \text{ where } p(x) \neq 0 \quad \text{Equation 1.7}$$

Using the sum rule (Equation 1.1), the denominator $p(x)$ in Bayes' rule can be expressed as,

$$p(x) = \sum_y p(x, y) = \sum_y p(x|y)p(y) \quad \text{Equation 1.8}$$

For the case of continuous random variables, the denominator in Bayes' rule is obtained by replacing the summation with integration such that,

$$p(x) = \int_{\text{dom}(y)} p(x, y)dy = \int_{\text{dom}(y)} p(x|y)p(y)dy \quad \text{Equation 1.9}$$

CHAPTER 2: THEORETICAL BACKGROUND

2.1.3. Bayes' Rule for Parameter Estimation

Recall that the primary goal of this thesis is to estimate the dynamic model parameters θ , given a lumped system dynamic model description of a physical process and data \mathcal{D} . Bayes' rule for parameters allows the engineer to make inferences about quantities of interest, such as the model parameters θ , from observed data. In this case, the observed data corresponds to sensor measurements of physical system properties.

Bayes' Rule for Parameters

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} \quad \text{Equation 1.10}$$

The denominator of Equation 1.10 can be expressed, using the sum and product rule of probability theory for continuous random variables, as:

$$p(\mathcal{D}) = \int_{\text{dom}(\theta)} p(\theta, \mathcal{D}) d\theta \quad \text{Equation 1.11}$$

$$p(\mathcal{D}) = \int_{\text{dom}(\theta)} p(\mathcal{D}|\theta)p(\theta) d\theta \quad \text{Equation 1.12}$$

Here the notation \mathcal{D} represents a data set $\mathcal{D} = \{d_i\}_{i=1}^N$ with N sensor measurements. It is standard practice to assume that the sensor measurements are independent given θ . Given that the independence assumption holds (Section 2.1.1, Equation 1.4), the quantity $p(\mathcal{D}|\theta)$ can be expressed as $p(\mathcal{D}|\theta) = p(d_1, d_2, \dots, d_N|\theta) = \prod_{i=1}^N p(d_i|\theta)$. As a result, Bayes' rule for parameter estimation from independent sensor measurements can typically be expressed as:

$$p(\theta|\mathcal{D}) = \frac{\prod_{i=1}^N p(d_i|\theta) p(\theta)}{\int_{\text{dom}(\theta)} \prod_{i=1}^N p(d_i|\theta) p(\theta) d\theta} \quad \text{Equation 1.13}$$

Refer to Appendix A for an illustrative example of the application of Bayes' rule to a parameter estimation problem.

2.1.4. Expected Values

Calculating the expected value of a function is one of the central operations involving probabilities, and it entails finding the weighted average of function values. The average value of $f(x)$, where x is a continuous scalar quantity, under a probability distribution $p(x)$ is called the linear expectation of function $f(x)$ and is denoted by $\mathbb{E}_{p(x)}[f(x)]$ (MacKay, 2004; Bishop, 2006; Barber, 2012; Murphy, 2012). For continuous random variables, the expected value of function $f(x)$ is expressed as:

Expected Value

$$\mathbb{E}_{p(x)}[f(x)] = \int_{\text{dom}(x)} p(x)f(x) dx \quad \text{Equation 1.14}$$

CHAPTER 2: THEORETICAL BACKGROUND

The variance of function $f(x)$ is defined by Bishop (2006) as:

Function Variance

$$\text{Var}[f(x)] = \mathbb{E}_{p(x)} \left[(f(x) - \mathbb{E}_{p(x)}[f(x)])^2 \right] \quad \text{Equation 1.15}$$

This quantity provides a measure of how much variability one expects there to be in $f(x)$ around its mean value (given by $\mathbb{E}_{p(x)}[f(x)]$). By expanding the square in Equation 1.15, one observes that the variance of $f(x)$ can also be written as:

$$\text{Var}[f(x)] = \mathbb{E}_{p(x)} \left[(f(x))^2 \right] - (\mathbb{E}_{p(x)}[f(x)])^2 \quad \text{Equation 1.16}$$

One can consider the variance of the continuous random variable x itself which is given by:

Variable Variance

$$\text{Var}[x] = \mathbb{E}_{p(x)}[(x - \mathbb{E}[x])^2] \quad \text{Equation 1.17}$$

Similar to that of the function variance, one can expand out the square of Equation 1.17 and rewrite the variable variance as:

$$\text{Var}[x] = \mathbb{E}_{p(x)}[(x)^2] - (\mathbb{E}_{p(x)}[x])^2 \quad \text{Equation 1.18}$$

The notation $\mathbb{E}_{p(x)}[x]$ requires evaluating the expected value of the random variable x under the distribution $p(x)$. The expectation is readily obtained by evaluating Equation 1.19.

$$\mathbb{E}_{p(x)}[x] = \int_{\text{dom}(x)} p(x)x dx \quad \text{Equation 1.19}$$

Similarly, $\mathbb{E}_{p(x)}[(x)^2]$ can be obtained by evaluating Equation 1.20.

$$\mathbb{E}_{p(x)}[(x)^2] = \int_{\text{dom}(x)} p(x)x^2 dx \quad \text{Equation 1.20}$$

In the case of two continuous scalar random variables x and y , their covariance is defined by Bishop (2006) as:

Variable Covariance

$$\text{cov}[x, y] = \mathbb{E}_{p(x,y)}[(x - \mathbb{E}_{p(x)}[x])(y - \mathbb{E}_{p(y)}[y])] \quad \text{Equation 1.21}$$

The covariance expresses the extent to which random variables x and y vary together. If the random variables x and y are independent (Section 2.1.1, Equation 2.4), then the covariance is zero. The notation $\mathbb{E}_{p(x,y)}$ makes explicit that the expectation is with respect to both random variables x and y under the joint probability distribution $p(x, y)$.

In situations where no ambiguity arises as to which variable, concerning some probability distribution, is being averaged over, short-hand notation omitting the subscript is used, i.e. $\mathbb{E}_{p(x)}[x] = \mathbb{E}[x]$.

2.2. Concepts from Information Theory

This section introduces basic concepts from the field of information theory which will prove useful in the development of techniques in this thesis.

2.2.1. Entropy

Entropy is a measure of the uncertainty in a distribution and is related to equilibrium thermodynamics as a measure of disorder. The entropy for a distribution defined over a continuous scalar random variable x , otherwise referred to as the *differential entropy* of x , is evaluated as follows (MacKay, 2004; Bishop, 2006; Barber, 2012):

Differential Entropy

$$\mathbb{H}[p(x)] = - \int_{\text{dom}(x)} p(x) \ln p(x) dx = - \mathbb{E}_{p(x)}[\ln p(x)] \quad \text{Equation 1.22}$$

2.2.2. Kullback-Leibler Divergence

Suppose one has some unknown but true distribution $p(x)$ over a continuous scalar quantity x and wishes to model the distribution $p(x)$ with an approximating distribution $q(x)$. If $q(x)$ is to be used for inference purposes, then the expected additional amount of information (Shannon, 1948) required to specify the value of x resulting from the use of the distribution $q(x)$, instead of the true distribution $p(x)$, is given by the Kullback-Leibler divergence $\mathcal{KL}(p(x)||q(x))$ from $p(x)$ to $q(x)$ (Kullback and Leibler, 1951; Bishop, 2006.).

Kullback-Leibler Divergence

$$\mathcal{KL}(p(x)||q(x)) = - \int_{\text{dom}(x)} p(x) \ln \frac{q(x)}{p(x)} dx = - \mathbb{E}_{p(x)} \left[\ln \left\{ \frac{q(x)}{p(x)} \right\} \right] \quad \text{Equation 1.23}$$

The quantity given by Equation 1.23 is also known as the relative entropy or the \mathcal{KL} divergence between $p(x)$ and $q(x)$. Note that the \mathcal{KL} divergence is not a symmetric functional - for the two distributions $p(x)$ and $q(x)$, $\mathcal{KL}(p(x)||q(x)) \neq \mathcal{KL}(q(x)||p(x))$. The \mathcal{KL} divergence between two distributions $p(x)$ and $q(x)$ satisfies $\mathcal{KL}(p(x)||q(x)) \geq 0$ with equality precisely when $p(x) = q(x)$.

2.3. Useful Distributions

Section 2.3 is devoted to the introduction of various probability distributions which will form the basis for most of the work presented in this thesis. Although many different types of distributions exist and are interesting in their own right, Section 2.3 provides an overview of the Gaussian and Gamma distributions which form the building blocks for the probabilistic models used in subsequent sections.

2.3.1. The Gaussian Distribution

One of the most important and well-known probability distributions for continuous random variables is the Gaussian, also referred to as Normal, distribution. For a real-valued scalar quantity x , the properties of the Gaussian distribution are outlined by Bishop (2006) as follows:

Univariate Gaussian Distribution

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2} (x - \mu)^2\right\} \quad \text{Equation 1.24}$$

The Gaussian distribution defined by Equation 1.24 has two parameters: μ (called the *mean*), and σ^2 (called the *variance*). The square root of σ^2 is the *standard deviation* σ , and the reciprocal of σ^2 is called the *precision* β . Equation 1.24 has support for $x \in \mathbb{R}$ with $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. It is straightforward to show that Equation 1.24 is normalised by noting that:

$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) dx = 1$$

The average value of x under the Gaussian distribution $\mathcal{N}(x|\mu, \sigma^2)$, i.e. the expected value of x , is given by (refer to Equation 1.19):

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) x dx = \mu \quad \text{Equation 1.25}$$

The second order moment can be calculated as (refer to Equation 1.20):

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) x^2 dx = \mu^2 + \sigma^2 \quad \text{Equation 1.26}$$

It follows that the variance of random variable x is calculated as (refer to Equation 1.18):

$$\text{Var}[x] = \mathbb{E}[(x)^2] - (\mathbb{E}[x])^2 = \sigma^2 \quad \text{Equation 1.27}$$

The maximum of a distribution is its mode. For a Gaussian distribution, the mode corresponds to the mean μ . Thus, the mode for $\mathcal{N}(x|\mu, \sigma^2)$ is:

$$\text{mode}[x] = \mu \quad \text{Equation 1.28}$$

Similarly, the entropy for the Gaussian distribution $\mathcal{N}(x|\mu, \sigma^2)$ is calculated, using Equation 2.22, as follows:

$$\mathbb{H}[x] = \frac{1}{2} (1 + \ln(2\pi\sigma^2)) \quad \text{Equation 1.29}$$

Figure 1.1 shows a graphical depiction of three univariate Gaussian distributions over the scalar random variable x for different settings of the parameters μ and σ^2 .

CHAPTER 2: THEORETICAL BACKGROUND

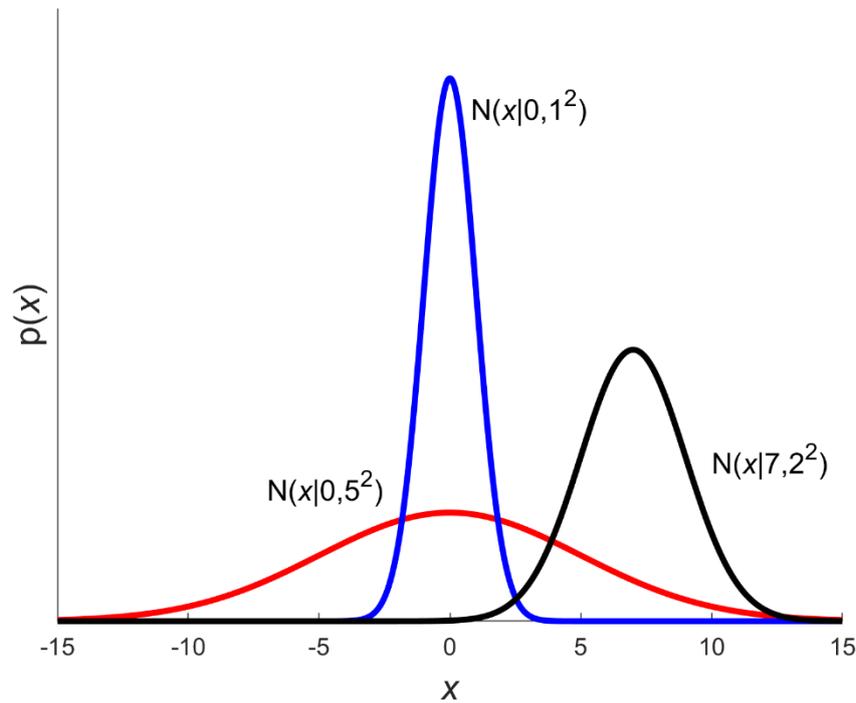


Figure 1.1: Example univariate Gaussian distributions with different mean μ and variance σ^2 parameters. Figure adapted from Koller and Friedman (Figure 2.2, 2009).

From Figure 1.1 and Equation 1.24, one observes that the Gaussian distribution takes on a bell-like curve where the parameter μ controls the peak location; at this location, the Gaussian distribution takes on its maximum value, and the variance σ^2 determines how peaked the distribution is. The smaller the distribution variance σ^2 (refer to the blue and red curves), the more peaked the distribution is around the mean parameter μ . The Gaussian distribution over the real-valued scalar quantity x can be extended and defined over a $T \times 1$ vector \mathbf{x} of continuous random variables. The multivariate extension is given by (Bishop, 2006):

Multivariate Gaussian Distribution

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\} \quad \text{Equation 1.30}$$

Equation 1.30 has two parameters that correspond to the $T \times 1$ vector $\boldsymbol{\mu}$ (called the *mean vector*) and the symmetric, positive-definite $T \times T$ matrix $\boldsymbol{\Sigma}$ (called the *covariance matrix*). Similar to the univariate case, one can summarise the expected values of \mathbf{x} , the mode, and the entropy of the distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ as follows:

$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu} \quad \text{Equation 1.31}$$

$$\text{cov}[\mathbf{x}] = \boldsymbol{\Sigma} \quad \text{Equation 1.32}$$

$$\text{mode}[\mathbf{x}] = \boldsymbol{\mu} \quad \text{Equation 1.33}$$

$$\mathbb{H}[\mathbf{x}] = \frac{1}{2} \ln|\boldsymbol{\Sigma}| + \frac{T}{2} (1 + \ln(2\pi)) \quad \text{Equation 1.34}$$

CHAPTER 2: THEORETICAL BACKGROUND

Figure 1.2 depicts a two-dimensional Gaussian distribution over a vector \mathbf{x} consisting of two continuous random variables such that $\mathbf{x} = [x_1 \ x_2]^T$ (also referred to as a bivariate Normal distribution) with mean vector $\boldsymbol{\mu} = [0 \ 0]^T$ and covariance matrix $\boldsymbol{\Sigma} = \alpha^{-1}\mathbf{I}_{2 \times 2} = 0.5\mathbf{I}_{2 \times 2}$. The term α^{-1} simply scales the distribution variance. This specific form, where the covariance matrix is proportional to the $\mathbf{I}_{2 \times 2}$ identity matrix, is known as an isotropic covariance. Thus, Figure 1.2 shows a surface plot of an isotropic (rotationally invariant) bivariate Gaussian distribution over random variables x_1 and x_2 with the corresponding contour plot depicted in Figure 1.3.

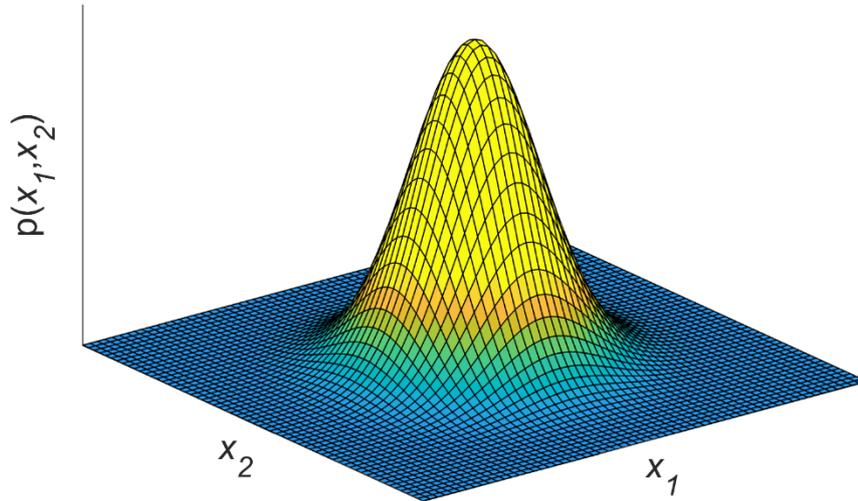


Figure 1.2: Bivariate Gaussian distribution surface plot over two random variables x_1 and x_2 . Figure adapted from Koller and Friedman (Figure 7.1, 2009). The mean vector and covariance matrix correspond to $\boldsymbol{\mu} = [0 \ 0]^T$ and $\boldsymbol{\Sigma} = \alpha^{-1}\mathbf{I}_{2 \times 2} = 0.5\mathbf{I}_{2 \times 2}$, respectively.

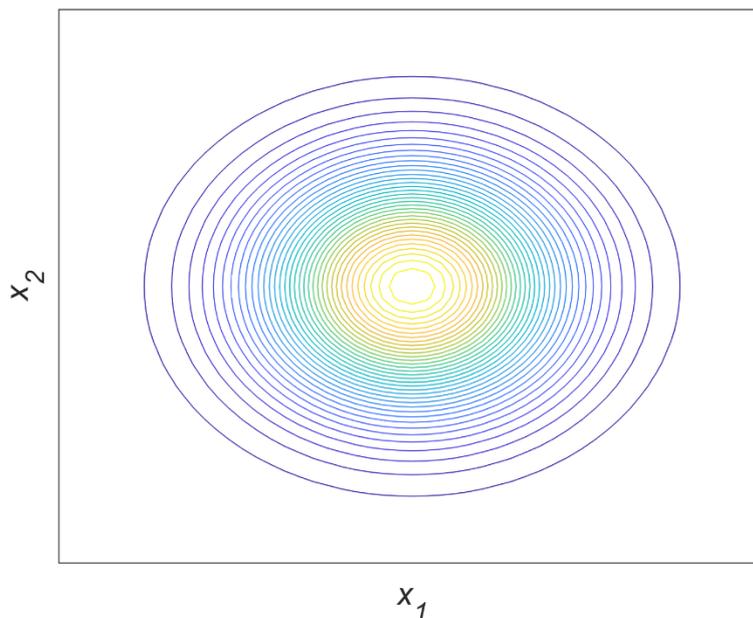


Figure 1.3: Bivariate Gaussian distribution contour plot over two random variables x_1 and x_2 . Figure adapted from Bishop (Figure 2.8, 2006). The mean vector and covariance matrix correspond to $\boldsymbol{\mu} = [0 \ 0]^T$ and $\boldsymbol{\Sigma} = \alpha^{-1}\mathbf{I}_{2 \times 2} = 0.5\mathbf{I}_{2 \times 2}$, respectively.

CHAPTER 2: THEORETICAL BACKGROUND

A surprising characteristic of the multivariate Gaussian distribution is that if two random variables are jointly Gaussian distributed, as in the graphical depiction in Figure 1.2, then the conditional distribution of one random variable, say x_1 , conditioned on the other random variable, say x_2 , is also Gaussian distributed. Similarly, the marginal distribution of both random variables are Gaussian distributions. If one considers a $T \times 1$ vector \mathbf{x} that is Gaussian distributed, i.e. $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ – Equation 1.30, with \mathbf{x} partitioned into two disjoint subsets \mathbf{x}_a and \mathbf{x}_b , then the standard results for conditioning on and marginalisation a Gaussian distribution are given by Bishop (2006) as follows:

Gaussian Conditioning

If one takes \mathbf{x}_a as the first R components of \mathbf{x} and \mathbf{x}_b as the remaining $T - R$ entries, one can define the following partitions for the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ associated with the distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$:

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{bmatrix} \quad \text{Equation 1.35}$$

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix} \quad \text{Equation 1.36}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{bmatrix} \quad \text{Equation 1.37}$$

Following the derivation in Bishop (2006), the mean and covariance for the conditional distribution $p(\mathbf{x}_a|\mathbf{x}_b)$ are expressed as follows:

$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b}), \quad \text{Equation 1.38}$$

where

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b) \quad \text{Equation 1.39}$$

$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba} \quad \text{Equation 1.40}$$

Gaussian Marginalisation

The marginal distribution $p(\mathbf{x}_a)$ has a mean and covariance that is given by Bishop (2006) as:

$$\mathbb{E}[\mathbf{x}_a] = \boldsymbol{\mu}_a \quad \text{Equation 1.41}$$

$$\text{cov}[\mathbf{x}_a] = \boldsymbol{\Sigma}_{aa} \quad \text{Equation 1.42}$$

Thus,

$$p(\mathbf{x}_a) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa}) \quad \text{Equation 1.43}$$

One observes that marginalising over random variables that are jointly Gaussian distributed corresponds to the operation of picking the entries in the partitioned mean vector and covariance matrix for the subset of variables of interest.

CHAPTER 2: THEORETICAL BACKGROUND

2.3.2. The Gamma Distribution

A second important distribution used extensively throughout the current work is the Gamma distribution. Before introducing the Gamma distribution, it is worth noting the standard results given by Equations 2.44 and 2.45. The notation $\Gamma(a)$, used in Equation 1.44, refers to the Gamma function and ensures that the Gamma distribution is normalised correctly, i.e. the distribution integrates to 1 over its domain.

Another important function is the Digamma function, given by Equation 2.45, which is required when evaluating the entropy of a Gamma distributed random variable x and the expected values of x and $\ln(x)$. Evaluating the entropy, $\mathbb{E}[x]$ and the $\mathbb{E}[\ln(x)]$ will be especially important in a subsequent chapter which focus on estimating the parameters of lumped system algebraic dynamic models from sensor measurement data (Bishop, 2006; Murphy, 2012)

| Gamma Distribution Notes | | |
|--------------------------|--|---------------|
| Gamma function | $\Gamma(a) \equiv \int_0^{\infty} u^{a-1} \exp\{-u\} du$ | Equation 1.44 |
| Digamma function | $\psi(a) = \frac{d}{da} [\ln(\Gamma(a))]$ | Equation 1.45 |

| Gamma Distribution | |
|--|---------------|
| $Gam(x a, b) = \frac{1}{\Gamma(a)} b^a x^{a-1} \exp\{-bx\}$ | Equation 1.46 |
| <p>The Gamma distribution is defined over the positive real numbers, and is governed by two parameters a and b. The parameters a and b are constrained such that $a > 0$ and $b > 0$ which ensures the distribution can be normalised. One can summarise the expected values of x and $\ln(x)$, the mode, and the entropy of the distribution $Gam(x a, b)$ as follows:</p> | |
| $\mathbb{E}[x] = \frac{a}{b}$ | Equation 1.47 |
| $\mathbb{E}[\ln(x)] = \psi(a) - \ln(b)$ | Equation 1.48 |
| $Var[x] = \frac{a}{b^2}$ | Equation 1.49 |
| $mode[x] = \frac{a-1}{b} \text{ for } a \geq 1$ | Equation 1.50 |
| $\mathbb{H}[x] = \ln(\Gamma(a)) - (a-1)\psi(a) - \ln(b) + a,$ | Equation 1.51 |
| <p>where $\Gamma(a)$ and $\psi(a)$ are the Gamma and Digamma functions, respectively.</p> | |

Figure 1.4 graphically depicts three Gamma distributions over the random variable x for different settings of the parameters a and b .

CHAPTER 2: THEORETICAL BACKGROUND

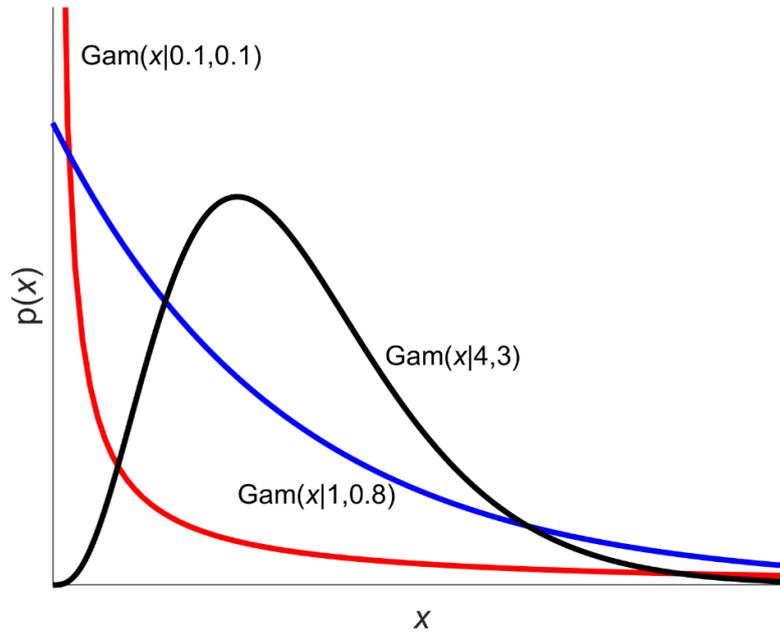


Figure 1.4: Example Gamma distributions with different parameters a and b . Figure adapted from Bishop (Figure 2.13, 2006).

2.3.3. Conjugacy

Note from Equation 1.13 that in order to obtain a normalised posterior distribution over the lumped system dynamic model parameters θ , the engineer has to evaluate the *evidence* quantity, i.e. the integral given by Equation 1.12. To circumvent evaluating the evidence quantity, the engineer can exploit a property referred to as *conjugacy* in the statistics and machine learning literature. The idea is as follows: given a specific form of the likelihood function, one seeks a prior distribution over θ such that the resulting posterior distribution has the same functional form as the prior distribution. If the prior and posterior distribution are of the same functional form, then the prior distribution is said to be conjugate to the likelihood function. Conjugacy allows the engineer to obtain closed-form update solutions for the posterior distribution. Furthermore, the posterior distribution can simply be normalised by looking up the appropriate normalisation constant which will stem from the choice of prior distribution. While this thesis only considers Gaussian likelihood functions for inference purposes, it is important to note that not all likelihood functions are Gaussian. Rather, the likelihood function depends on the problem at hand. Table 1.1 summarises the conjugate prior distributions used in this thesis.

Table 1.1: Conjugate priors for the likelihood functions used in the thesis

| Likelihood Function | Conjugate Prior for Mean | Conjugate Prior for Precision |
|---|--------------------------|-------------------------------|
| Gaussian – [μ unknown, β^{-1} known] | Univariate Gaussian | Not Applicable |
| Gaussian – [μ unknown, β^{-1} known] | Multivariate Gaussian | Not Applicable |
| Gaussian – [μ unknown, β^{-1} unknown] | Multivariate Gaussian | Gamma (Section 2.3.2) |

2.4. Bayesian Linear Regression

The foundations of this thesis rely on the idea of linear regression, how it can be extended to models with explicit closed-form solutions that are nonlinear in the model parameters and, ultimately, to the application of parameter inference for lumped system dynamic models that take the form of ordinary differential equations which, for a limited set of physical systems, are linear in the model parameters.

This section presents two different probabilistic approaches for inferring the parameters of models that can be written as a linear combination of the model parameters. In the first approach, it is assumed that the sensor measurement precision (inverse of the sensor variance parameter) is known *a priori* and that the inference task pertains to making conclusions about the values of the model parameters. The next probabilistic approach relaxes the assumption that the sensor precision is known exactly and infers the sensor precision from data.

Note that from here on this work draws a distinction between models that can be written as a linear combination of the model parameters, for which the parameter vector \mathbf{w} is used, and models that are nonlinear in the model parameters, for which the parameter vector $\mathbf{\Omega}$ is used. Thus, for models that are linear in the model parameters, Bayes' rule for parameter estimation (Equation 1.10) would be:

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})} \quad \text{Equation 1.52}$$

For models that are nonlinear in the model parameters, Bayes' rule for parameter estimation (Equation 1.10) would be:

$$p(\mathbf{\Omega}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{\Omega})p(\mathbf{\Omega})}{p(\mathcal{D})} \quad \text{Equation 1.53}$$

Although the term 'model' is used in the current work, note that the 'model' is simply a *deterministic function*, as derived from fundamental principles, with inputs and parameters that can be used to describe a physical system.

2.4.1. Bayesian Linear Regression - Probabilistic Model I

The first Bayesian linear regression probabilistic model, as outlined in Bishop (2006), considers finding the parameters of a model that can be written as a linear combination of the unknown model parameters with the sensor precision known exactly.

Suppose one is given a deterministic function $y(x, \mathbf{w})$ where the notation \mathbf{w} refers to the unknown model parameters and x is the function input. Furthermore, suppose that the deterministic function $y(x, \mathbf{w})$ models some physical system and that measurements of the response of $y(x, \mathbf{w})$ are taken with a sensor with precision β . Due to the imperfect nature of measurement equipment, the engineer will only observe noise-corrupted versions of the true response of the deterministic function $y(x, \mathbf{w})$. Thus, the engineer observes the following:

$$d = y(x, \mathbf{w}) + \epsilon \quad \text{Equation 1.54}$$

CHAPTER 2: THEORETICAL BACKGROUND

The notation d refers to the target variable that the engineer observes as a result of taking a measurement from the physical system and this measurement is generated from the true deterministic function response $y(x, \mathbf{w})$, corrupted by additive sensor measurement noise, ϵ . The sensor noise throughout this entire thesis is assumed to be *independent and identically distributed* (i.i.d.) following a Gaussian distribution with a zero-mean and precision β (inverse of the sensor variance parameter). The probability of observing the target variable d can therefore be expressed as follows:

$$p(d|x, \mathbf{w}, \beta) = \mathcal{N}(d|y(x, \mathbf{w}), \beta^{-1}) \quad \text{Equation 1.55}$$

Note from Equation 1.55 that the mean of the distribution over the target variable d shifts with $y(x, \mathbf{w})$ as the input x is allowed to change. Assuming that sensor measurements are generated independently (Section 2.1.1, Equation 1.4) from the distribution given by Equation 1.55, the likelihood function, in other words the probability of the data set given \mathbf{w} , β and the input \mathbf{x} , can be written in the form of:

$$p(\mathbf{d}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{i=1}^N \mathcal{N}(d_i|\mathbf{w}^T \boldsymbol{\phi}(x_i), \beta^{-1}) \quad \text{Equation 1.56}$$

Equation 1.56 makes use of the fact that the deterministic function $y(x, \mathbf{w})$, which corresponds to the mean of the distribution, can be written as a linear combination of the function parameters for an input x as follows:

$$y(x, \mathbf{w}) = \mathbf{w}^T \boldsymbol{\phi}(x) \quad \text{Equation 1.57}$$

The vector \mathbf{w} contains the deterministic function parameters such that $\mathbf{w} = [w_0, \dots, w_{M-1}]^T$ where M is the total number of parameters. The notation \mathbf{x} makes explicit that the probability of the data set is conditioned on all of the input points that correspond to sensor measurements, i.e. $\mathbf{x} = \{x_1, x_2, \dots, x_N\}$. Note that regression is a supervised learning problem, thus, the input variable x will always appear in the conditioning set. As a result, to keep the notation uncluttered, the dependency on x , and subsequently \mathbf{x} , is dropped from this point onwards. Also, since the sensor precision is assumed to be known exactly, the notational dependence on β is dropped. The parameter w_0 simply allows the engineer to model any fixed offset in the data and is typically referred to as the bias parameter (not to be confused with the *unbiasedness* heuristic or bias-variance trade-off in the statistical sense). Furthermore, the notation \mathbf{d} groups the target variable measurements into a $N \times 1$ column vector, also sometimes denoted with the symbol $\mathcal{D} = \{d_i\}_{i=1}^N$ for convenience of discussion. The symbol $\boldsymbol{\phi}(x)$ represents the basis functions of $y(x, \mathbf{w})$, where it is made explicit that $\boldsymbol{\phi}(x)$ depends on the input x . A dummy basis function of $\phi_0(x) = 1$ is included for the parameter w_0 . For example, when considering a straight line model, the deterministic function $y(x, \mathbf{w})$ can be represented as follows:

$$y(x, \mathbf{w}) = \mathbf{w}^T \boldsymbol{\phi}(x) = [w_0 \quad w_1] \begin{bmatrix} 1 \\ x \end{bmatrix} \quad \text{Equation 1.58}$$

where $\boldsymbol{\phi}(x) = \begin{bmatrix} \phi_0(x) \\ \phi_1(x) \end{bmatrix}$ with $\phi_0(x) = 1, \phi_1(x) = x$ Equation 1.59

The Bayesian treatment of linear regression requires the engineer to specify a prior probability distribution over the unknown function parameters \mathbf{w} . Since the likelihood function is of a Gaussian form and the deterministic model parameters form part of the mean of the distribution given by Equation 1.55, the conjugate prior for \mathbf{w} is the multivariate normal distribution (Table 1.1, Section 2.3.3) given by:

CHAPTER 2: THEORETICAL BACKGROUND

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0) \quad \text{Equation 1.60}$$

From Bayes' rule for parameter estimation problems (Equation 1.10 and 2.52), the posterior distribution is proportional to the product of the likelihood function and the prior distribution (Appendix A). In other words,

$$p(\mathbf{w}|\mathbf{d}) \propto p(\mathbf{d}|\mathbf{w})p(\mathbf{w}) \quad \text{Equation 1.61}$$

Recall that by exploiting conjugacy (Section 2.3.3), the posterior distribution will have the same functional form as the prior distribution over \mathbf{w} , i.e. the posterior distribution will be multivariate Gaussian. One can obtain the posterior distribution over the deterministic function parameters \mathbf{w} , given the data $\mathcal{D} = \{d_i\}_{i=1}^N$, by completing the multivariate square such that:

$$p(\mathbf{w}|\mathbf{d}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \quad \text{Equation 1.62}$$

where

$$\mathbf{S}_N = (\mathbf{S}_0^{-1} + \beta\Phi^T\Phi)^{-1} \quad \text{Equation 1.63}$$

$$\mathbf{m}_N = \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\Phi^T\mathbf{d}) \quad \text{Equation 1.64}$$

The parameters \mathbf{m}_0 and \mathbf{S}_0 are known as hyperparameters and are set by the engineer to express their prior belief about the unknown deterministic function parameters \mathbf{w} . The notation Φ denotes an $N \times M$ matrix, also referred to as the design matrix, with elements corresponding to:

$$\Phi = \begin{bmatrix} \phi_0(x_1) & \cdots & \phi_{M-1}(x_1) \\ \vdots & \ddots & \vdots \\ \phi_0(x_N) & \cdots & \phi_{M-1}(x_N) \end{bmatrix} \quad \text{Equation 1.65}$$

Note that by exploiting conjugacy, the engineer does not have to evaluate the *evidence* term in Bayes' rule for parameter estimation problems (Equation 1.12 and 2.52). The posterior distribution can simply be normalised by looking up the standard normalisation results for a multivariate Gaussian distribution in any statistics or machine learning textbook.

2.4.2. Connection to Simple Least Squares Regression

A natural question that the engineer might ask is “*How does Bayesian linear regression relate to the traditional least squares approach which is something I'm more familiar with?*” In order to establish the connection, one must first review the traditional linear least squares approach. Say the engineer wishes to fit the model given by Equation 1.58 to an observed data set. One popularly used approach is to minimise an error function $E(\mathbf{w})$ that measures the discrepancy between the predictions from $y(x, \mathbf{w})$ and the observed data set. Typically, the engineer would use the *sum-of-squares* error function defined as follows:

$$E(\mathbf{w}) = \sum_{i=1}^N (y(x_i, \mathbf{w}) - d_i)^2 \quad \text{Equation 1.66}$$

This specific choice of error function minimises the error between the straight line predicted value $y(x_i, \mathbf{w})$ and the associated measured target variable d_i over the entire data set, and optimising it is known as *simple least squares regression* (Englezos and Kalogerakis, 2001; Bishop, 2006).

CHAPTER 2: THEORETICAL BACKGROUND

To establish a connection between least squares and Bayesian linear regression, it is convenient to adjust Equation 1.66 by introducing a factor of $1/2$ such that:

$$E(\mathbf{w}) \propto \frac{1}{2} \sum_{i=1}^N (y(x_i, \mathbf{w}) - d_i)^2 \quad \text{Equation 1.67}$$

Recall that the probability of observing the target variable d , given \mathbf{w} , is expressed by Equation 1.55 with the corresponding likelihood function given by Equation 1.56. By taking the natural logarithm of Equation 1.56, one can show that the *log likelihood* corresponds to:

$$\ln p(\mathbf{d}|\mathbf{w}, \beta) = -\frac{\beta}{2} \sum_{i=1}^N (y(x_i, \mathbf{w}) - d_i)^2 + \frac{N}{2} \ln \beta - \frac{N}{2} \ln 2\pi \quad \text{Equation 1.68}$$

Since the engineer is concerned with finding \mathbf{w} , they must eventually decide on how to select the values of \mathbf{w} . One popular way of selecting \mathbf{w} is to use a heuristic from frequentist statistics known as *Maximum Likelihood*. The maximum likelihood heuristic simply states that the engineer should select the parameters \mathbf{w} that maximises the log likelihood function given by Equation 1.68. The setting of the parameter that are selected by the maximum likelihood heuristic are typically denoted in the statistics and machine learning literature by $\hat{\mathbf{w}}$, however, the author will use the subscript notation \mathbf{w}_{ML} to make explicit that the parameters are obtained from the maximum likelihood (ML) heuristic. Note that \mathbf{w}_{ML} , as a result of using maximum likelihood, is a point estimate for the deterministic function parameters. Contrast this with the Bayesian methodology which produces a posterior distribution $p(\mathbf{w}|\mathbf{d})$ over the deterministic function parameters (Equation 1.62).

For discussion purposes, the author drops the last two terms in Equation 1.68 since these terms do not explicitly depend on \mathbf{w} . Also, note that scaling the log likelihood function by a positive constant does not change the setting of the parameters that correspond to using the maximum likelihood heuristic. As a result, the coefficient $\beta/2$ can be replaced with $1/2$ such that Equation 1.68 can be rewritten as:

$$\ln p(\mathbf{d}|\mathbf{w}, \beta) \propto -\frac{1}{2} \sum_{i=1}^N (y(x_i, \mathbf{w}) - d_i)^2 \quad \text{Equation 1.69}$$

Notice the similarity between Equation 1.67 and 2.69. The last step in connecting the probabilistic perspective taken thus far to the traditionally practised simple least squares methodology is to realise that instead of maximising Equation 1.69 with respect to \mathbf{w} , the engineer can equivalently minimise the negative log likelihood. Therefore, maximising the likelihood function is equivalent to minimising the sum-of-squares error function (Equation 1.67) given the assumption of the i.i.d. Gaussian noise model. In other words, the engineer can motivate the use of the sum-of-squares error function from a maximum likelihood solution perspective. The resulting estimate for \mathbf{w} is given by Bishop (2006) as:

$$\mathbf{w}_{ML} = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{d} \quad \text{Equation 1.70}$$

In setting up the discussion for the Bayesian linear regression probabilistic model, the author stated that the sensor precision parameter is assumed to be known exactly. This is not a limitation of the Bayesian nor the traditional frequentist maximum likelihood viewpoint. In fact,

CHAPTER 2: THEORETICAL BACKGROUND

the maximum likelihood heuristic allows the engineer to obtain a point estimate for the sensor precision parameter by evaluating:

$$\frac{1}{\beta_{ML}} = \frac{1}{N} \sum_{i=1}^N (y(x_i, \mathbf{w}_{ML}) - d_i)^2 \quad \text{Equation 1.71}$$

From the Bayesian perspective, the engineer would want a posterior distribution instead of a point estimate over the noise precision parameter β . This is addressed in a second probabilistic model for Bayesian linear regression discussed in Section 2.4.3.

So far, a connection has been made between the simple least squares error function and the likelihood function via the frequentist maximum likelihood heuristic. This gives the engineer a probabilistic interpretation of the simple least squares methodology. However, the question remains: “*How does this connect to Bayesian linear regression?*”

To establish deeper connection between the simple least squares approach and Bayesian linear regression, recall the sum-of-squares error function given by Equation 1.67 which was discussed in the context of a straight line model. As presented, the sum-of-squares error function is based on measuring the discrepancy between the straight line $y(x, \mathbf{w})$ and the observed data set. However, Equation 1.67 is not restricted to just straight lines and can be applied to any deterministic function that can be written as a linear combination of the function parameters.

Suppose the engineer received only a data set without any knowledge of the underlying deterministic function generating the data. The engineer might have to hypothesise a model and fit it to the data. One might start off with a straight line model, move on to a second order polynomial and continue to higher order polynomials. One thing the engineer will observe is that as the polynomial order increases, the error function $E(\mathbf{w})$ value decreases, implying a better fit. In fact, the engineer can continue increasing the polynomial order until the polynomial passes exactly through every data point with $E(\mathbf{w}) = 0$. However, this will result in a fitted polynomial that will wildly oscillate and poorly represent the true underlying deterministic function. This phenomenon is known as *overfitting*. One way to mitigate the overfitting situation is to use *regularisation*. Regularisation involves the addition of a penalty term to the sum-of-squares error function to inhibit the deterministic function parameters from reaching large values. A popular regularisation term used for simple least squares regression is the *sum-of-squares of parameters* penalty term. The sum-of-squares of parameters penalty term is also referred to as ℓ_2 -regularisation, the squared two-norm or simply the weight decay penalty term. In statistics literature, use of the sum-of-squares of parameters penalty term serves as an example of a parameter shrinkage method (Bishop, 2006; Murphy, 2012). The addition of this penalty term leads to a modified error function given by:

$$\tilde{E}(\mathbf{w}) \propto \frac{1}{2} \sum_{i=1}^N (y(x_i, \mathbf{w}) - d_i)^2 + \frac{\lambda}{2} \mathbf{w}^T \mathbf{w} \quad \text{Equation 1.72}$$

The notation $\mathbf{w}^T \mathbf{w}$ may also be written as $\mathbf{w}^T \mathbf{w} = \|\mathbf{w}\|^2 = w_0^2 + w_1^2 + \dots + w_M^2$ which gives rise to the various penalty term names such as sum-of-squares of parameters or squared two-norm. Here λ governs the importance of the regularisation term. To see the connection to Bayesian linear regression, recall from Equation 1.61 that the posterior distribution is proportional to the product of the Gaussian likelihood function and Gaussian prior over \mathbf{w} .

CHAPTER 2: THEORETICAL BACKGROUND

Given that the engineer is exploiting conjugacy, the posterior distribution will also be Gaussian. Start by setting the hyperparameters for Equation 1.60 to $\mathbf{m}_0 = [0 \ 0]^T$ and $\mathbf{S}_0 = \alpha^{-1} \mathbf{I}_{M \times M}$.

The notation $\mathbf{I}_{M \times M}$ represents the identity matrix and M is the number of deterministic function parameters. The term α^{-1} simply scales the prior distribution variance and can be interpreted as controlling the strength of the prior. Note that the covariance matrix \mathbf{S}_0 is proportional to the identity matrix and is typically referred to as an isotropic covariance (Figure 1.2 and 2.3). By taking the natural logarithm of Equation 1.61, given the selected hyperparameter values, one obtains the following:

$$p(\mathbf{w}|\mathbf{d}) \propto -\frac{\beta}{2} \sum_{i=1}^N (y(x_i, \mathbf{w}) - d_i)^2 + \frac{N}{2} \ln \beta - \frac{N}{2} \ln 2\pi - \ln \frac{\alpha}{2\pi} - \frac{\alpha}{2} \mathbf{w}^T \mathbf{w} \quad \text{Equation 1.73}$$

Since the goal of the engineer is to find a setting of the parameter \mathbf{w} , similar to the previous discussion, one can omit the terms that do not depend on \mathbf{w} . Also, scaling the posterior distribution by a positive constant does not change the posterior maximum location with respect to the deterministic function parameters \mathbf{w} . As a result, one can replace the term $\beta/2$ with $1/2$ and the term $\alpha/2$ with $\alpha/2\beta$. By setting $\alpha/\beta = \lambda$, Equation 1.73 can be rewritten as follows:

$$p(\mathbf{w}|\mathbf{d}) \propto -\frac{1}{2} \sum_{i=1}^N (y(x_i, \mathbf{w}) - d_i)^2 - \frac{\lambda}{2} \mathbf{w}^T \mathbf{w} \quad \text{Equation 1.74}$$

Notice the similarity between Equation 1.72 and 2.74. Following a similar argument, instead of maximising the log posterior with respect to \mathbf{w} , one can equivalently minimise the negative log posterior. Therefore, maximising the log posterior is equivalent to minimising the sum-of-squares error function with the addition of the sum-of-squares of parameters penalty term (Equation 1.72), given that the prior over \mathbf{w} is taken as a zero mean isotropic Gaussian and the noise model is assumed to follow an i.i.d. Gaussian noise model. Maximising the log posterior, or simply the posterior, is known as the *maximum a posteriori* estimate, abbreviated as MAP, with the corresponding parameter estimates denoted by \mathbf{w}_{MAP} . Note that although the MAP estimate for \mathbf{w} is obtained from the posterior distribution $p(\mathbf{w}|\mathbf{d})$, it is still only a point estimate. In other words, by selecting the MAP estimate, the engineer has ‘thrown away’ the distributional information captured by the posterior distribution over \mathbf{w} (Bishop, 2006; Murphy, 2012).

In summary, given the probabilistic model in Equation 1.61, it is possible to recover the traditional simple least squares regression methodology engineers are typically familiar with by simply ignoring the prior over \mathbf{w} , i.e. ignore $p(\mathbf{w})$ in Equation 1.61, and using maximum likelihood to select the unknown function parameters directly from the likelihood function $p(\mathbf{d}|\mathbf{w})$. To avoid overfitting, the engineer might implement regularised least squares estimation; this corresponds to a specific choice of Gaussian prior distribution $p(\mathbf{w})$. The setting of the parameters \mathbf{w} obtained by the engineer through regularised least squares estimation then correspond to the MAP estimates of the posterior distribution $p(\mathbf{w}|\mathbf{d})$. As a result, one observes that every step the engineer took in the discussion above, given certain assumptions, is one step closer to a fully Bayesian treatment of linear regression where the ideal ‘destination’ would be a posterior distribution over the unknown function parameters \mathbf{w} .

2.4.3. Bayesian Linear Regression - Probabilistic Model II

The engineer is more likely to find themselves in a practical setting where the sensor precision parameter is not known *a priori*. The second Bayesian linear regression probabilistic model, as outlined in Bishop (2006), considers inferring the parameters of a model that can be written as a linear combination of the unknown model parameter with an unknown sensor precision. Note that by treating the sensor precision as an unknown, one is implicitly introducing another random variable into the inference procedure. As a result, Bayes' rule for parameter estimation problems, as presented in Equation 1.10, is adjusted to account for the additional random variable such that:

$$p(\mathbf{w}, \beta | \mathcal{D}) = \frac{p(\mathcal{D} | \mathbf{w}, \beta) p(\mathbf{w}, \beta)}{p(\mathcal{D})} \quad \text{Equation 1.75}$$

Note that the engineer is now dealing with two 'types' of parameters, namely, \mathbf{w} which corresponds to the deterministic function parameters of $y(x, \mathbf{w})$ that is used to model some physical system, and the sensor precision parameter β . In order to perform inference, Bayes' rule requires specifying a prior distribution $p(\mathbf{w}, \beta)$ over both random variables \mathbf{w} and β . Using the product rule of probability theory (Equation 1.3), the joint prior probability $p(\mathbf{w}, \beta)$ can be factorised such that $p(\mathbf{w}, \beta) = p(\mathbf{w} | \beta) p(\beta)$. Equation 1.75 can now be rewritten as:

$$p(\mathbf{w}, \beta | \mathcal{D}) = \frac{p(\mathcal{D} | \mathbf{w}, \beta) p(\mathbf{w} | \beta) p(\beta)}{p(\mathcal{D})} \quad \text{Equation 1.76}$$

The quantity $p(\mathcal{D} | \mathbf{w}, \beta)$ corresponds to Equation 1.56. The prior distribution over \mathbf{w} is similar to that of Equation 1.60, however, the dependency on β is now made explicit such that:

$$p(\mathbf{w} | \beta) = \mathcal{N}(\mathbf{w} | \mathbf{m}_0, \beta^{-1} \mathbf{S}_0) \quad \text{Equation 1.77}$$

The conjugate prior for the random variable β , given the target variable distribution is represented by Equation 1.55, is a Gamma distribution (Section 2.3.3, Table 1.1) such that:

$$p(\beta) = \text{Gam}(\beta | a_0, b_0) \quad \text{Equation 1.78}$$

As a result, the joint prior distribution $p(\mathbf{w}, \beta)$ takes the form:

$$p(\mathbf{w}, \beta) = p(\mathbf{w} | \beta) p(\beta) = \mathcal{N}(\mathbf{w} | \mathbf{m}_0, \beta^{-1} \mathbf{S}_0) \text{Gam}(\beta | a_0, b_0) \quad \text{Equation 1.79}$$

The parameters \mathbf{m}_0 , \mathbf{S}_0 , a_0 and b_0 are known as hyperparameters and are set by the engineer to express their prior belief about the unknown deterministic function parameters \mathbf{w} and the sensor precision parameter β . As with the first Bayesian linear regression probabilistic model, the posterior distribution is proportional to the product of the likelihood function and the prior, i.e.

$$p(\mathbf{w}, \beta | \mathcal{D}) \propto p(\mathcal{D} | \mathbf{w}, \beta) p(\mathbf{w} | \beta) p(\beta) \quad \text{Equation 1.80}$$

Since the engineer is exploiting conjugacy, the posterior distribution $p(\mathbf{w}, \beta | \mathcal{D})$ takes the same functional form as Equation 1.79. It can be shown that:

$$p(\mathbf{w}, \beta | \mathcal{D}) = \mathcal{N}(\mathbf{w} | \mathbf{m}_N, \beta^{-1} \mathbf{S}_N) \text{Gam}(\beta | a_N, b_N), \quad \text{Equation 1.81}$$

CHAPTER 2: THEORETICAL BACKGROUND

$$\text{where} \quad \mathbf{S}_N = (\mathbf{S}_0^{-1} + \Phi^T \Phi)^{-1} \quad \text{Equation 1.82}$$

$$\mathbf{m}_N = \mathbf{S}_N (\mathbf{S}_0^{-1} \mathbf{m}_0 + \Phi^T \mathbf{d}) \quad \text{Equation 1.83}$$

$$a_N = a_0 + \frac{N}{2} \quad \text{Equation 1.84}$$

$$b_N = b_0 + \frac{1}{2} (\mathbf{m}_0^T \mathbf{S}_0^{-1} \mathbf{m}_0 - \mathbf{m}_N^T \mathbf{S}_N^{-1} \mathbf{m}_N + \mathbf{d}^T \mathbf{d}) \quad \text{Equation 1.85}$$

Equations 2.82 through 2.85 allow the engineer to obtain a joint posterior distribution over the deterministic function parameters \mathbf{w} and the noise precision parameter β . Contrast this with the frequentist results given by Equations 2.70 and 2.71 which produce point estimates for \mathbf{w} and β .

2.5. Gaussian Processes

In order to motivate the use of Gaussian processes, which find its origins in geostatistics (Cressie, 1993), let us return to the first Bayesian linear regression model in Section 2.4.1. Recall from Equation 1.57 that deterministic functions that are linear in the function parameters can be expressed as $y(x, \mathbf{w}) = \mathbf{w}^T \boldsymbol{\phi}(x)$. The vector \mathbf{w} collects all the function parameters, and $\boldsymbol{\phi}(x)$ represents the basis functions which depend on the input x . Following Bayes' rule for parameter estimation, the engineer constructed a likelihood function, defined a prior distribution over \mathbf{w} by exploiting conjugacy, and obtained a posterior distribution $p(\mathbf{w}|\mathbf{d})$ over the function parameters. If the engineer puts some thought into the process of Bayes' rule for parameter estimation, they should realise that by defining a prior distribution over deterministic functions parameters \mathbf{w} , the engineer is implicitly inducing a prior distribution over functions $y(x, \mathbf{w})$.

If this reasoning seems somewhat obscure, the argument is as follows: say the engineer wishes to fit the straight line given by Equation 1.58 to an observed data set, given that the straight line sufficiently describes the observed data. By defining a prior distribution over \mathbf{w} , the engineer is acknowledging that multiple reasonable straight lines exist. In other words, when the engineer draws samples from the prior distribution $p(\mathbf{w})$, every sample corresponds to a possible straight line that might describe the data set. As a result, if the engineer goes from a prior distribution over \mathbf{w} to a posterior distribution $p(\mathbf{w}|\mathbf{d})$ using Equation 1.61, the engineer is simply updating their belief about how likely various straight lines, as mapped through the posterior distribution over \mathbf{w} , describe the observed data set. As a result, by defining a prior distribution over \mathbf{w} , the engineer has implicitly defined a prior distribution over all straight line functions $y(x, \mathbf{w})$. From a Gaussian process viewpoint, the engineer dispenses with the parameterised deterministic function $y(x, \mathbf{w})$, i.e. the prior distribution $p(\mathbf{w})$, and directly defines a prior distribution over functions.

2.5.1. Bayesian Linear Regression – Gaussian Process Motivation

This section will illustrate how Bayesian linear regression (Probabilistic Model I) can be interpreted as an example of a particular Gaussian process. Consider again the straight line deterministic function given by Equation 1.58. Next, consider the prior distribution over \mathbf{w} , as defined by Equation 1.60. Similar to the regularised least squares estimation discussed in Section 2.4.2, set the prior distribution hyperparameters to $\mathbf{m}_0 = [0 \ 0]^T$ and $\mathbf{S}_0 = \alpha^{-1} \mathbf{I}_{M \times M}$, such that the isotropic Gaussian takes the following form:

CHAPTER 2: THEORETICAL BACKGROUND

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}_{M \times M}) \quad \text{Equation 1.86}$$

For any particular value of \mathbf{w} drawn from Equation 1.86, the engineer will end up with a specific function for $y(x, \mathbf{w}) = \mathbf{w}^T \boldsymbol{\phi}(x)$. Therefore, the prior distribution $p(\mathbf{w})$ induces a distribution over functions of the form $y(x, \mathbf{w}) = \mathbf{w}^T \boldsymbol{\phi}(x)$. In a practical setting, however, the engineer might wish to determine the function value of $y(x)$ at a specific value of x without mapping through the distribution over \mathbf{w} . In this setting, the engineer is interested in directly modeling the joint distribution over function values $y(x_1^*), y(x_2^*), \dots, y(x_N^*)$. Denote the collection of function values as an $N^* \times 1$ vector \mathbf{y} such that:

$$\mathbf{y} = [y(x_1^*) \ y(x_2^*) \ \dots \ y(x_N^*)]^T \quad \text{Equation 1.87}$$

Here the notation N^* simply refers to the number of input points at which the engineer wishes to evaluate the function $y(x)$ values. The collection of function values \mathbf{y} can be written in terms of the deterministic function parameters \mathbf{w} and the design matrix such that:

$$\mathbf{y} = \boldsymbol{\Phi}^* \mathbf{w} \quad \text{Equation 1.88}$$

The design matrix corresponding to the N^* input points is given by:

$$\boldsymbol{\Phi}^* = \begin{bmatrix} \phi_0(x_1^*) & \dots & \phi_{M-1}(x_1^*) \\ \vdots & \ddots & \vdots \\ \phi_0(x_{N^*}) & \dots & \phi_{M-1}(x_{N^*}) \end{bmatrix} \quad \text{Equation 1.89}$$

The question that arises is “How does the engineer find the joint probability distribution of the function values \mathbf{y} ?” From Equation 1.88, observe that \mathbf{y} is written as a linear combination of the Gaussian distributed random variable \mathbf{w} (Equation 1.86), hence \mathbf{y} is Gaussian distributed itself (Bishop, 2006). Thus, the engineer only requires the mean vector and covariance matrix of \mathbf{y} to define the joint distribution of function values. This is readily obtained by realising that (from Equations 2.31, 2.86 and 2.88):

$$\mathbb{E}[\mathbf{y}] = \mathbb{E}[\boldsymbol{\Phi}^* \mathbf{w}] = \boldsymbol{\Phi}^* \mathbb{E}[\mathbf{w}] = \mathbf{0} \quad \text{Equation 1.90}$$

$$\text{cov}[\mathbf{y}] = \mathbb{E}[\mathbf{y}\mathbf{y}^T] = \boldsymbol{\Phi}^* \mathbb{E}[\mathbf{w}\mathbf{w}^T] (\boldsymbol{\Phi}^*)^T = \frac{1}{\alpha} \boldsymbol{\Phi}^* (\boldsymbol{\Phi}^*)^T = \mathbf{K} \quad \text{Equation 1.91}$$

Note that for the Gaussian distributed random variable \mathbf{w} with prior distribution given by Equation 1.86, the quantity $\mathbb{E}[\mathbf{w}\mathbf{w}^T]$ evaluates to $\alpha^{-1}\mathbf{I}_{M \times M}$. The symmetric $N^* \times N^*$ matrix \mathbf{K} is commonly referred to as the Gram matrix (Bishop, 2006). One can expand Equation 1.91 such that:

$$\mathbf{K} = \frac{1}{\alpha} \begin{bmatrix} \phi_0(x_1^*) & \dots & \phi_{M-1}(x_1^*) \\ \vdots & \ddots & \vdots \\ \phi_0(x_{N^*}) & \dots & \phi_{M-1}(x_{N^*}) \end{bmatrix} \begin{bmatrix} \phi_0(x_1^*) & \dots & \phi_{M-1}(x_1^*) \\ \vdots & \ddots & \vdots \\ \phi_0(x_{N^*}) & \dots & \phi_{M-1}(x_{N^*}) \end{bmatrix}^T \quad \text{Equation 1.92}$$

From Equation 1.92, the engineer can directly observe that the element-wise entries of the Gram matrix is given by:

$$K_{im} = \frac{1}{\alpha} \boldsymbol{\phi}^T(x_i^*) \boldsymbol{\phi}(x_m^*) = k(x_i^*, x_m^*) \quad \text{Equation 1.93}$$

In Equation 1.93, $k(x_i^*, x_m^*)$ is called the *kernel function*, also referred to as the covariance function, and the subscripts i and m index the elements of the Gram matrix for $i = 1, \dots, N^*$ and $m = 1, \dots, N^*$, respectively.

CHAPTER 2: THEORETICAL BACKGROUND

The entire procedure discussed in Section 2.5.1 is a particular example of a Gaussian process where the engineer, instead of mapping via the prior distribution $p(\mathbf{w})$ to the space of straight line functions, rather defined a joint distribution directly over the space of function values \mathbf{y} with the mapping from the input space to the output space given by the Gram matrix. Since the Gram matrix captures the covariance between function outputs via the kernel function, it is also the covariance matrix. In other words, the engineer maps directly from the input space, through the basis functions, directly to the function output space via the kernel function, avoiding the need for the prior distribution $p(\mathbf{w})$. Thus, the kernel function specifies a distribution over straight line functions and inference is performed directly in the space of straight line functions. In general, the engineer can specify a kernel function directly (depending on the problem at hand), rather than indirectly obtaining it through the choice of basis functions.

2.5.2. Gaussian Processes for Non-Parametric Regression

A Gaussian process is defined as follows by Bishop (2006):

Gaussian Process

In general, a *Gaussian process* (abbreviated \mathcal{GP}) is a probability distribution over functions $y(x)$ such that the set of values of the function $y(x)$ evaluated at arbitrarily selected input points $x_1^*, x_2^*, \dots, x_N^*$ jointly have a Gaussian distribution.

Note that the definition of a Gaussian process implies a consistency requirement; if the \mathcal{GP} specifies that $p(y(x_1^*), y(x_2^*)) \sim \mathcal{N}(y(x_1^*), y(x_2^*) \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $p(y(x_1^*)) \sim \mathcal{N}(y(x_1^*) \mid \mu_1, \sigma_{11}^2)$ where σ_{11}^2 is the entry of covariance matrix $\boldsymbol{\Sigma}$. This consistency requirement is also known as the marginalisation property (Equation 1.35 through 2.43). In other words, if the engineer examines a larger set of function values, this does not change the distribution over a smaller set of function values, and vice versa. Furthermore, observe that the marginalisation property is fulfilled if the kernel function (otherwise referred to as the covariance function) specifies the entries of the Gaussian process covariance matrix (Gram matrix) which is required to be symmetric and positive-definite (Rasmussen and Williams, 2006).

Rasmussen and Williams (2006) state that a Gaussian process is completely specified by its mean function and the covariance function. However, in most practical applications, the engineer will typically have no prior knowledge about the mean function and it is taken to be zero. This is analogous to the discussion in Section 2.5.1 where the mean for the prior distribution over the deterministic function parameters \mathbf{w} was set to $\mathbf{m}_0 = [0 \ 0]^T$. Thus, the engineer can completely specify the Gaussian process by providing the covariance of the function $y(x)$ at any two values of the input x , say x_i^* and x_m^* , which is given by the kernel function $k(x_i^*, x_m^*)$. As a result, the covariance between two function output values corresponds to (Section 2.1.4):

$$\mathbb{E}[y(x_i^*)y(x_m^*)] = k(x_i^*, x_m^*) \quad \text{Equation 1.94}$$

Note that $k(x_i^*, x_m^*)$ does not necessarily have to correspond to the formulation given by Equation 1.93. As stated previously, the engineer can specify a kernel function directly (depending on the problem at hand), rather than indirectly obtaining it through the choice of basis functions.

CHAPTER 2: THEORETICAL BACKGROUND

The current work will use the following notation for a Gaussian process:

$$p(y(x_i^*)) \sim \mathcal{GP}(0, k(x_i^*, x_i^*)) \quad \text{Equation 1.95}$$

Since the Gaussian process has a consistency requirement that is automatically fulfilled if the kernel function specifies the entries of the covariance matrix, the joint distribution over $y(x_i^*)$ and $y(x_m^*)$ corresponds to:

$$p(y(x_i^*), y(x_m^*)) \sim \mathcal{GP}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} k(x_i^*, x_i^*) & k(x_i^*, x_m^*) \\ k(x_m^*, x_i^*) & k(x_m^*, x_m^*) \end{bmatrix}\right) \quad \text{Equation 1.96}$$

Equation 1.96 can be extended to multiple values of $y(x)$ that correspond to arbitrary input points $x_1^*, x_2^*, \dots, x_N^*$, such that the joint distribution over \mathbf{y} is given by:

$$p(\mathbf{y}) \sim \mathcal{GP}(\mathbf{0}, \mathbf{K}) \quad \text{Equation 1.97}$$

Equation 1.97 is referred to as the Gaussian process prior with covariance matrix \mathbf{K} where the covariance matrix element-wise entries are given by the appropriate kernel function, as defined by the engineer, for $i = 1, \dots, N^*$ and $m = 1, \dots, N^*$, respectively, such that:

$$K_{im} = k(x_i^*, x_m^*) \quad \text{Equation 1.98}$$

A popular kernel function typically encountered in the Gaussian process machine learning literature is the exponentiated quadratic kernel (MacKay, 2004; Bishop, 2006; Rasmussen and Williams, 2006; Barber, 2012).

Exponentiated Quadratic Kernel

The exponentiated quadratic kernel is given by:

$$k(x_i^*, x_m^*) = \sigma_f^2 \exp\left\{-\frac{1}{2\ell^2} (x_i^* - x_m^*)^2\right\} \quad \text{Equation 1.99}$$

Note that the exponentiated quadratic kernel is given in its one-dimensional form since this is the specific form that is used throughout the current work. The *length-scale* ℓ and the *signal variance* σ_f^2 can be varied and are referred to as *hyperparameters*. A justification for why these parameters are called hyperparameters is provided later in the text. In a typical application setting, these hyperparameters are unknown and must be learned from the sensor data.

The specification of the exponentiated quadratic kernel function in Equation 1.99 implies a distribution over functions. To see this, one can draw functions from Equation 1.97 with the corresponding covariance matrix \mathbf{K} whose element-wise entries are calculated from Equation 1.99. Refer to Figure 1.5.

CHAPTER 2: THEORETICAL BACKGROUND

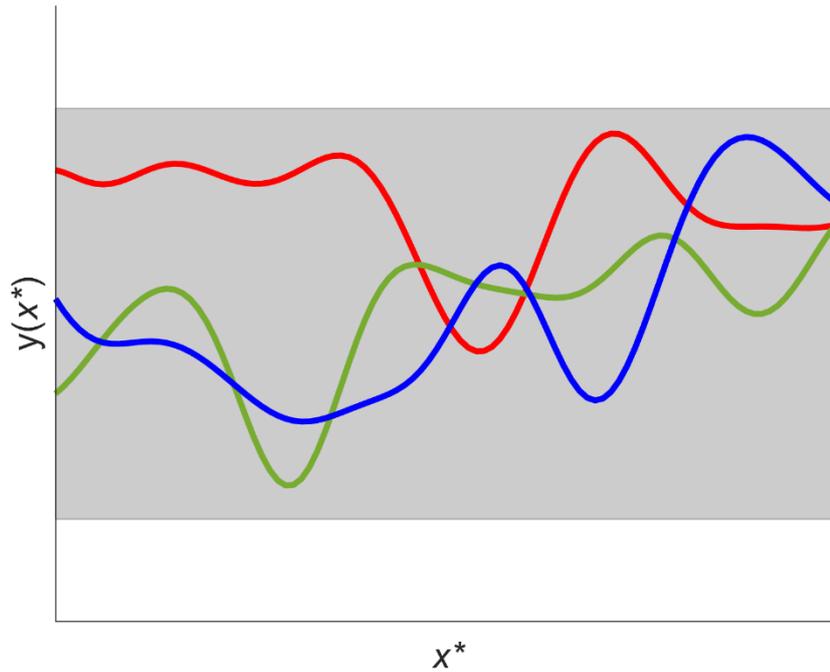


Figure 1.5: Three randomly drawn functions from the \mathcal{GP} prior distribution (Equation 1.97) with all hyperparameter set to unity. The shaded region represents the 99% credibility interval (Section 2.8).

By inspection of Equation 1.99, with the hyperparameters set to unity for discussion purposes, one observes that the covariance for input points close to each other is near unity implying the corresponding function output values are highly correlated. The covariance decreases as input points move further away from each other indicating less correlation between the function output values. Overall, the exponentiated quadratic kernel encodes what are typically referred to in the Gaussian process literature as ‘smoothness assumptions’, as reflected by the random smooth function samples drawn from the Gaussian process prior (Equation 1.97). In other words, the engineer can use the exponentiated quadratic kernel to model underlying deterministic functions of some physical system that are *a priori* believed to be smooth and continuous (Section 4.5.2).

The engineer is typically not interested in drawing functions from the prior distribution given by Equation 1.97. Rather they want to incorporate the knowledge that sensor measurements provide about the underlying deterministic function. In order to apply Gaussian processes to regression, the engineer has to take into account the sensor noise associated with the data due to taking a measurement from the physical system. Suppose that,

$$d_i = y_i(x_i) + \epsilon_i \quad \text{Equation 1.100}$$

Equation 1.100 states that each measured target variable d_i is generated from the true underlying function y_i (evaluated at input x_i) but is independently corrupted with sensor noise ϵ_i that is assumed to be from a zero mean Gaussian with precision parameter β . The probability of observing the target variable d_i can therefore be expressed as follows:

$$p(d_i|x_i, \beta) = \mathcal{N}(d_i|y_i(x_i), \beta^{-1}) \quad \text{Equation 1.101}$$

CHAPTER 2: THEORETICAL BACKGROUND

As per definition, the Gaussian process defines a joint distribution over the target variable measurements \mathbf{d} and the unknown function values \mathbf{y} such that:

$$p(\mathbf{y}, \mathbf{d}) = \mathcal{N}\left(\mathbf{y}, \mathbf{d} \mid \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{C}_{yy} & \mathbf{C}_{yd} \\ \mathbf{C}_{dy} & \mathbf{C}_{dd} \end{bmatrix}\right) \quad \text{Equation 1.102}$$

Here the matrix \mathbf{C}_{yy} contains the covariance entries of the N^* input points at which the engineer wishes to evaluate the function $y(x)$ output values with the element-wise covariance matrix entries given by $k(x_i^*, x_m^*)$. The matrix \mathbf{C}_{dd} contains the covariance entries of the N input points at which sensor measurements are observed from the physical system.

The matrix \mathbf{C}_{dd} is evaluated from $k(x_i, x_m)$ with the addition of independent Gaussian distributed noise to the main diagonal entries corresponding to $\beta^{-1}\delta_{im}$ for $i = 1, \dots, N$ and $m = 1, \dots, N$, respectively. The notation δ_{im} refers to the Kronecker delta which takes on a value of one if and only if $i = m$ and zero otherwise. Notice that the Kronecker delta is on the index of the sensor measurements only. Furthermore, $\mathbf{C}_{dy} = \mathbf{C}_{yd}^T$ where \mathbf{C}_{yd} is referred to as the cross-covariance matrix between \mathbf{y} and \mathbf{d} and is obtained from $k(x_i^*, x_m)$ for $i = 1, \dots, N^*$ and $m = 1, \dots, N$, respectively.

To obtain a posterior distribution over function values, i.e. the Gaussian process $p(\mathbf{y}|\mathbf{d})$, one can use the standard results for conditioning on a Gaussian distribution (Equation 1.35 through 2.40) to obtain from the joint distribution $p(\mathbf{y}, \mathbf{d})$ that:

$$p(\mathbf{y}|\mathbf{d}) \sim \mathcal{GP}(\boldsymbol{\mu}_{y|\mathbf{d}}, \boldsymbol{\Sigma}_{y|\mathbf{d}}) \quad \text{Equation 1.103}$$

where,

$$\boldsymbol{\mu}_{y|\mathbf{d}} = \mathbf{C}_{yd}\mathbf{C}_{dd}^{-1}\mathbf{d} \quad \text{Equation 1.104}$$

$$\boldsymbol{\Sigma}_{y|\mathbf{d}} = \mathbf{C}_{yy} - \mathbf{C}_{yd}\mathbf{C}_{dd}^{-1}\mathbf{C}_{dy} \quad \text{Equation 1.105}$$

The notation $\boldsymbol{\mu}_{y|\mathbf{d}}$ represents the posterior distribution (over function output values) $N^* \times 1$ mean vector with the associated $N^* \times N^*$ posterior covariance matrix $\boldsymbol{\Sigma}_{y|\mathbf{d}}$. The central problem that arises with Gaussian process regression is the inversion of the $N \times N$ matrix \mathbf{C}_{dd} which requires $O(N^3)$ computations. This computational cost prohibits the use of Gaussian process regression on large data sets. Equation 1.97 through 2.105 can visually be interpreted as drawing samples from the Gaussian process prior and only keeping those functions that agree with the sensor data sampled from the physical system. Refer to Figure 1.6 for a visual interpretation. The black dots correspond to sensor measurements taken from the physical system and the three functions, as drawn from the Gaussian process given by Equation 1.103, indicate possible functions from the prior that agree with the sensor measurements. The natural logarithm of the marginal distribution $p(\mathbf{d})$ for Equation 1.102 is given by Bishop (2006) as:

$$\ln p(\mathbf{d}|\boldsymbol{\psi}_{GP}) = -\frac{1}{2}\mathbf{d}^T\mathbf{C}_{dd}^{-1}\mathbf{d} - \frac{1}{2}\ln|\mathbf{C}_{dd}| - \frac{N}{2}\ln 2\pi \quad \text{Equation 1.106}$$

The notation $p(\mathbf{d}|\boldsymbol{\psi}_{GP})$ makes explicit that the evidence term, i.e. $p(\mathbf{d})$ (otherwise referred to as the marginal likelihood), depends on the kernel function parameters $\boldsymbol{\psi}_{GP}$. This mapping occurs through the covariance matrix \mathbf{C}_{dd} whose element-wise entries depend on the kernel function $k(x_i, x_m)$.

CHAPTER 2: THEORETICAL BACKGROUND

Since choosing the kernel function implicitly defines a prior distribution over functions, the kernel function parameters $\boldsymbol{\psi}_{GP}$ are referred to as the Gaussian process hyperparameters. Within a typical setting, techniques for learning $\boldsymbol{\psi}_{GP}$ are based on $p(\mathbf{d}|\boldsymbol{\psi}_{GP})$. The simplest and most widely practised technique is to estimate $\boldsymbol{\psi}_{GP}$ by maximising the evidence term. This technique is also known as a type 2 maximum likelihood procedure (Bishop, 2006). Maximising the evidence can be performed efficiently using any form of gradient-based optimisation since analytical derivatives for Equation 1.106, with respect to the Gaussian process hyperparameters, are available (Bishop, 2006). Taking the derivative of Equation 1.106, which is not a trivial task, with respect to each of the Gaussian process hyperparameters results in:

$$\frac{\partial \ln p(\mathbf{d}|\boldsymbol{\psi}_{GP})}{\partial \psi_i} = \frac{1}{2} \mathbf{d}^T \mathbf{C}_{dd}^{-1} \frac{\partial \mathbf{C}_{dd}}{\partial \psi_i} \mathbf{C}_{dd}^{-1} \mathbf{d} - \frac{1}{2} \text{Tr} \left(\mathbf{C}_{dd}^{-1} \frac{\partial \mathbf{C}_{dd}}{\partial \psi_i} \right) \quad \text{Equation 1.107}$$

The trace operator Tr in Equation 1.107 simply sums over the main diagonal entries of the resulting square matrix $\mathbf{C}_{dd}^{-1}(\partial \mathbf{C}_{dd}/\partial \psi_i)$. Bishop (2006) reports that Equation 1.106 will in general be a nonconvex function, thus it can have multiple maxima. As a result, gradient-based optimisation routines can converge to local maxima. Multiple restarts of the optimisation routine can be performed to find a good local maximum.

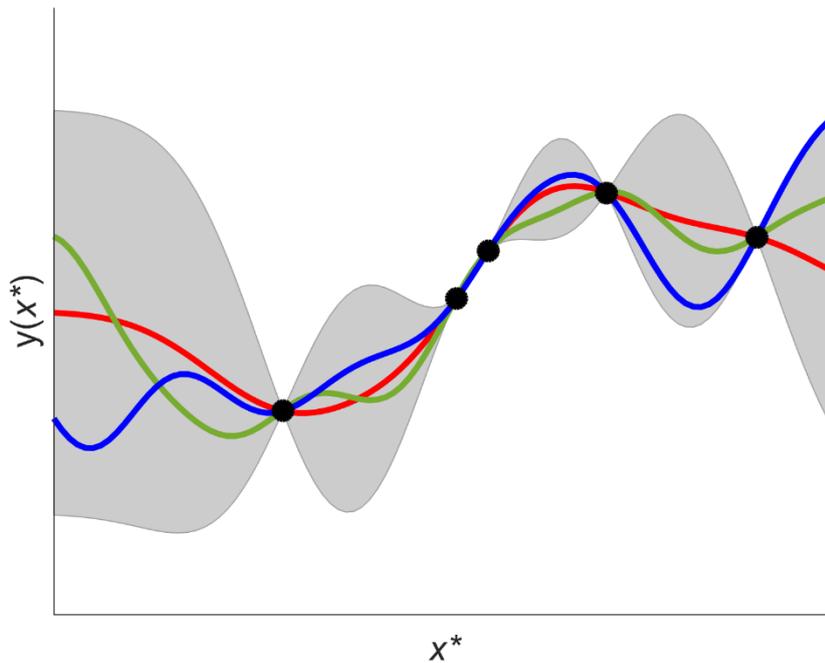


Figure 1.6: Three randomly drawn functions from the \mathcal{GP} posterior distribution (Equation 1.103) with all hyperparameter set to unity. The shaded region represents the 99% credibility interval (Section 2.8).

2.6. Variational Inference

An important task in the application of probabilistic models is the evaluation of the posterior distribution $p(\mathbf{z}|\mathbf{d})$ such that the engineer can make conclusions about unknown quantities of interest, based on observed data.

For most models of interest, it is not feasible to evaluate the posterior distribution $p(\mathbf{z}|\mathbf{d})$ directly, or to calculate exact expectations of this distribution. This is because the dimensionality of the latent space \mathbf{z} is too high to work with or the posterior distribution $p(\mathbf{z}|\mathbf{d})$ has a complex form which results in analytically intractable expectations. This problem typically manifests itself as not being able to evaluate the evidence term $p(\mathbf{d})$ in Bayes' rule which is required to ensure that the posterior distribution $p(\mathbf{z}|\mathbf{d})$ is correctly normalised (Attias, 1999; Jordan et al., 1999; Ghahramani and Beal, 2000; MacKay, 2004; Bishop, 2006; Fox and Roberts, 2012; Murphy, 2012; Blei, Kucukelbir and McAuliffe, 2018).

As a result, one of the central problems that arises in Bayesian inference is the idea of approximating difficult-to-compute probability densities. In situations where one has to resort to such techniques, two broad classes exist depending on whether they rely on stochastic or deterministic approximations. Stochastic techniques, such as the classic Markov chain Monte Carlo sampling methods, have allowed the use of Bayesian methods across various domains. Markov chain Monte Carlo (MCMC) methods generally have the property that given infinite computational resources, the techniques can generate exact results. However, the approximation arises due to finite processor time. In a practical setting, these sampling methods can be computationally demanding and often do not scale well to high dimensional problems (Bishop, 2006; Barber, 2012; Murphy, 2012; Blei, Kucukelbir and McAuliffe, 2018).

An alternative to Markov chain Monte Carlo techniques is deterministic approximation schemes. Deterministic techniques are based on analytic approximations to the posterior distribution $p(\mathbf{z}|\mathbf{d})$ and these methods often scale well to high dimensional problems. The current thesis will specifically focus on *variational inference*, otherwise referred to as *variational Bayes*, which is a method that approximates difficult-to-compute probability densities through optimisation (Bishop, 2006; Blei, Kucukelbir and McAuliffe, 2018).

In this thesis, the vector \mathbf{z} refers to latent variables, while \mathbf{d} is the sensor data, as measured from the physical system, and is collected into an $N \times 1$ vector. For the Bayesian linear regression (Probabilistic Model I) approach outlined in Section 2.4.1, the latent variable corresponds to $\mathbf{z} = \{\mathbf{w}\}$, i.e. the unknown deterministic function parameters. For the second Bayesian linear regression probabilistic model in Section 2.4.3, the author introduced the sensor precision as another unknown variable, thus, the latent variables correspond to $\mathbf{z} = \{\mathbf{w}, \beta\}$.

Bishop (2006) and Charles and Roberts (2012) provide introductions to variational inference with various illustrative examples to emphasise that the idea behind variational inference is to posit a family of probability densities followed by finding the member of the family which is *closest* to the target posterior distribution $p(\mathbf{z}|\mathbf{d})$. *Closeness* is measured by the Kullback-Leibler divergence $\mathcal{KL}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{d}))$ (Equation 1.23), where $q(\mathbf{z})$ is from the approximating family of distributions. A common choice of approximating distributions is the mean-field variational family, where groups of the latent variables comprising vector \mathbf{z} are assumed to be mutually independent.

CHAPTER 2: THEORETICAL BACKGROUND

As a result, a generic member of the mean-field variational family is given by $q(\mathbf{z}) = \prod_{i=1}^V q_i(\mathbf{z}_i)$. Notice that the vector \mathbf{z} has been partitioned into disjoint groups that are denoted by \mathbf{z}_i where $i = 1, \dots, V$. For the second Bayesian linear regression probabilistic model, the variational family comprises $q(\mathbf{w}, \beta) = q(\mathbf{w})q(\beta)$. Following the derivation in Bishop (2006) and Charles and Roberts (2012), one can obtain a general expression for the optimal variational solution for the mean-field family that is given by:

$$\ln q_j^*(\mathbf{z}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{d}, \mathbf{z})] + \text{constant} \quad \text{Equation 1.108}$$

The form of Equation 1.108 provides the basis for the application of the variational methods used throughout this thesis. Equation 1.108 states that the natural logarithm of the optimal solution for each variational factor $q_j(\mathbf{z}_j)$ is obtained by taking the expectation of the natural logarithm of the joint probability distribution over the latent variables and sensor measurement data with respect to all other factors $q_i(\mathbf{z}_i)$ for $i \neq j$. (Bishop, 2006; Blei, Kucukelbir and McAuliffe, 2018).

The solution for each optimal variational factor $q_j^*(\mathbf{z}_j)$ is obtained by first initialising all the relevant variational factors $q_j(\mathbf{z}_j)$ followed by cycling through each variational factor, in turn, replacing each factor with the revised estimate using the current estimates for all other variational factors. This gives rise to the coordinate ascent variational inference algorithm, abbreviated CAVI, whereby each factor is iteratively optimised while keeping the other factors fixed. An important quantity that arises in the derivation of the mean-field family variational approximation is the evidence lower bound, also referred to as the ELBO in machine learning literature.

The ELBO monotonically increases after each iteration allowing the engineer to use it to establish whether convergence has been achieved. However, the ELBO is generally a nonconvex function, thus, the CAVI algorithm only guarantees convergence to a local optimum. As a result, the engineer should perform multiple optimisation runs by initialising the relevant variational factors $q_j(\mathbf{z}_j)$ at random. The optimal mean-field variational posterior approximation to the target posterior distribution $p(\mathbf{z}|\mathbf{d})$ is then the product of each optimal variational solution $q_j^*(\mathbf{z}_j)$. Furthermore, it is worth pointing out that mean-field variational inference generally underestimates the variance of the true posterior distribution (Jordan et al., 1999; Bishop, 2006; Blei, Kucukelbir and McAuliffe, 2018).

2.7. Nonlinear Deterministic Functions

It is important to emphasise that the discussion of deterministic function parameter estimation, whether it be from the frequentist or Bayesian viewpoint, has been restricted to models that can be written as a linear combination of the unknown model parameters \mathbf{w} (Equation 1.57). However, the application of the parameter estimation techniques, as outlined in Sections 2.4.1 through 2.4.3, can be extended to estimate the parameters $\mathbf{\Omega}$ of nonlinear deterministic functions. In other words, functions that cannot be written as a linear combination of the function parameters. A somewhat standard technique for extending the parameter estimation approaches to nonlinear deterministic functions is to linearise the function. In other words, the engineer can use a Taylor expansion to approximate the nonlinear function as a linear deterministic function at the point of linearisation (Marlin, 2000; Englezos and Kalogerakis, 2001; Bishop, 2006; Chappell et al., 2009)

CHAPTER 2: THEORETICAL BACKGROUND

In a typical application setting the nonlinear function, which is denoted by $y(\mathbf{x}, \boldsymbol{\Omega})$, is approximated by a first-order Taylor expansion about some parameter point $\boldsymbol{\Omega}_*$ such that the linear approximation to the nonlinear function is given by:

$$y(\mathbf{x}, \boldsymbol{\Omega}) \approx y(\mathbf{x}, \boldsymbol{\Omega}_*) + \mathbf{J}(\boldsymbol{\Omega} - \boldsymbol{\Omega}_*) + \text{HOT} \quad \text{Equation 1.109}$$

This work assumes that all higher order terms (HOT) are negligible, although this need not be the case. For example, Woolrich and Behrens (2006) considered the application of a second-order Taylor expansion in the problem of estimating the parameters of spatial mixture models. The notation \mathbf{J} refers to the Jacobian matrix whose elements are given by the partial derivatives of the nonlinear function with respect to the unknown function parameters such that:

$$\mathbf{J} = \left[\begin{array}{ccc} \frac{\partial(y(\mathbf{x}, \boldsymbol{\Omega}))}{\partial\Omega_0} & \frac{\partial(y(\mathbf{x}, \boldsymbol{\Omega}))}{\partial\Omega_1} & \dots & \frac{\partial(y(\mathbf{x}, \boldsymbol{\Omega}))}{\partial\Omega_{M-1}} \end{array} \right]_{\boldsymbol{\Omega}_*} \quad \text{Equation 1.110}$$

By using the approximation given by Equation 1.109, all the parameter estimation techniques outlined in Sections 2.4.1 through 2.4.3 are now accessible. Approximating the nonlinear function with a linear counterpart and then solving the simple least squares problem may not be sufficient, especially if the linearisation point $\boldsymbol{\Omega}_*$ is far from the true model parameters. A possible solution to this problem is to use an initial guess $\boldsymbol{\Omega}_{guess}$, linearising the nonlinear function about $\boldsymbol{\Omega}_{guess}$, and then solving the simple least squares estimation problem to obtain a new estimate for the nonlinear function parameters. This procedure can be repeated until the obtained parameter point estimate does not change significantly within some specified tolerance between iterations. This converts the problem into a sequence of linear regression problems.

Within the machine learning, statistics and engineering literature, linearising the nonlinear deterministic function and iteratively solving the least squares problem is known as *nonlinear least squares*, and multiple extensions exist depending on how the problem is implemented (Englezos and Kalogerakis, 2001; Bishop, 2006; Woolrich and Behrens, 2006; Chappell et al., 2009; Murphy, 2012).

2.8. Confidence vs Credibility Intervals

From the frequentist perspective, the most popular and widely understood historical interpretation of a confidence interval follows the Neymanian understanding in which a confidence interval provides a measure of uncertainty by considering the long term frequency of repeated experiments (Neyman, 1937). In other words, if the engineer collects 100 sensor measurement data sets from independent experiments to estimate the parameters of an algebraic (Section 4.4 and 5.2.1) or ODE (Section 4.4, 5.2.2 and 5.2.3) model and constructs a, say 99%, confidence interval for the parameter estimates from each data set, at least 99 of these confidence intervals are expected to contain the true (but fixed) unknown model parameter (Hastie, Tibshirani and Friedman, 2009; Murphy, 2012; Hazra, 2017).

However, in a typical setting, the engineer does not have access to multiple data sets and more often only has a *single* data set since it might be too expensive to perform multiple experiments or repeated experimentation is simply not practical. Whether the confidence interval constructed from this single data set contains the true (but fixed) unknown model parameter is typically unknown.

CHAPTER 2: THEORETICAL BACKGROUND

The constructed confidence interval might or might not contain the true (but fixed) unknown model parameter. All the engineer can claim is the long term proportion of confidence intervals that would contain the true (but fixed) unknown model parameter (Neyman, 1937).

Contrasting the confidence interval is the Bayesian credibility interval which is sometimes put forward as a more practical concept (Edwards, Lindman and Savage, 1963; Bishop, 2006; Box and Tiao, 2011; Murphy, 2012). From the Bayesian perspective, a credibility interval is constructed such that there is a certain probability associated with finding the true (but random) unknown model parameter in that interval.

For example, if the engineer estimates the algebraic or ODE dynamic model parameters from a single sensor measurement data set and constructs a 99% credibility interval, then there is a 99% probability that the true (but random) unknown model parameter is within that interval. The same argument follows when extending either the frequentist confidence interval or the Bayesian credibility interval to a joint confidence or credibility region, respectively.

2.9. Summary

Chapter 2 started with an introduction to the necessary concepts of probability and information theory that will be used throughout this thesis. Following this, the author introduced the Gaussian and Gamma distributions, as well as the concept of conjugacy, which was then used to develop two Bayesian linear regression probabilistic models. A connection between simple least squares regression, which most engineers are familiar with, and Bayesian linear regression was established by showing how the simple least squares regression approach can be motivated from the Bayesian framework under certain assumptions about the noise model and prior distribution over parameters.

The Bayesian linear regression probabilistic model was then itself motivated from a Gaussian process viewpoint which recasts the regression problem into a *function-space viewpoint* instead of a *parameter-space viewpoint*. This was achieved by showing that the engineer can define a prior distribution directly over functions via the kernel function, avoiding the need for a parameter prior distribution. The function-space viewpoint allows the engineer to map directly from the function input space via the kernel function to the function output space.

Variational inference, which is a deterministic approximation scheme for analytically intractable posterior distributions, was then introduced. This approximation scheme will play an important role in Chapter 4. It was also noted that by linearising nonlinear deterministic functions, the Bayesian linear regression methods discussed in this chapter can be used for parameter inference. This will also become apparent in Chapter 4 when the author develops the proposed Bayesian approaches for parameter inference used in this thesis. Furthermore, the conceptual difference between the frequentist confidence and Bayesian credibility interval was discussed.

Chapter 3

Literature Review

“When we try to pick out anything by itself, we find that it is bound fast by a thousand invisible cords that cannot be broken, to everything in the universe.”

- John Muir, 1869

3.1. Overview

In order to maintain coherency between the different chapters in the current thesis, the author will first briefly restate what has been accomplished in the preceding chapters, followed by the outline of the literature review chapter. The parameter estimation literature is vast, however, it can be narrowed down to only pertain to the types of dynamic models (Section 1.5.3) considered in the current thesis, which gives rise to the development of the proposed approaches and methodology section discussed in Chapter 4.

3.1.1. Previous Chapters

Chapter 1 introduced the modeling problem and physical systems (Section 1.1), followed by the idea of using parameters to monitor physical systems (Section 1.3) in applications such as Fault Detection and Isolation (Section 1.3.1) and Condition-based Maintenance (Section 1.3.2). This gave rise to the practical importance of parameters in day-to-day engineering tasks and the need to reliably estimate these parameters from noise-corrupted time series data (Section 1.4).

The underlying mechanism of many physical systems in engineering can be described by algebraic, ordinary differential, and auxiliary equations. However, the principles underpinning the model development phase do not always provide insight into selecting suitable values for all the model parameters (Marlin, 2000; Calderhead, Girolami and Lawrence, 2009). As a result, the engineer has to resort to the inverse modeling approach (Section 1.1) to make conclusions about the values of the model parameters from data (Tarantola, 2005). Ideally, these parameter estimates should be good in quality/accuracy with a pragmatic interpretation of the associated parameter uncertainty (Sections 1.5 and 2.8). Most dynamic models of practical interest to chemical engineers are nonlinear in the model parameters, i.e. it is not possible to write the model in the form of Equation 1.57. This complicates the parameter estimation procedure since traditional parameter estimation techniques are developed for models that are linear in the model parameters.

To circumvent the problem, the engineer can linearise the nonlinear model, with respect to the unknown parameters, using a Taylor expansion (Section 2.7). By doing so, all the parameter estimation techniques outlined in Sections 2.4.1 through 2.4.3 again become accessible for estimating the parameters of nonlinear models. This connection will resurface again when presenting the relevant literature in the context of estimating parameters for models that cannot be written as a linear combination of the model parameters.

3.1.2. Present Chapter

Recall that this thesis exclusively deals with lumped system dynamic models (Section 1.5.3), as derived from fundamental principles such as the conservation of mass and energy, and assumes from the outset that the lumped system dynamic model is available to the chemical engineer. For the reader who might not necessarily be familiar with lumped systems, the term ‘lumped system’ refers to a system in which the system properties do not depend on the position within the system. For a lumped system, the resulting dynamic model reduces to an ordinary differential equation that describes the state variable. Under certain assumptions about the lumped system model inputs, the chemical engineer can explicitly solve the ordinary differential equation to obtain a closed-form algebraic model for the state variable dynamic response. However, in a typical setting, the explicit algebraic model is nonlinear in the model parameters. Algebraic and ordinary differential equation models are of particular interest in the current work due to their practical importance in modeling and monitoring physical systems in a chemical engineering setting (Marlin, 2000).

Note further that the ordinary differential and algebraic equation models may depend on several other physicochemical variables, such as density ρ or heat capacity c_p , which in itself can depend on variables such as temperature, concentration, etc. This thesis assumes that physicochemical variables such as density ρ or heat capacity c_p are time-invariant (Section 1.5.3) as well as independent of temperature, concentration, etc. such that the quantities ρ and c_p can be treated as model parameters. In other words, ρ and c_p take on constant values. The engineer is then faced with the problem of reliably estimating these parameters from noise-corrupted time series data. Depending on the physical system under consideration, the author will explicitly state which physicochemical variables are assumed to be model parameters.

The remainder of Chapter 3 provides a discussion of the relevant literature that pertains to estimating the parameters of algebraic or ordinary differential equation models, with the above-mentioned restrictions, from noise-corrupted time series data.

3.2. Parameter Estimation Techniques

Recall from Section 2.4.2 that in order to estimate the unknown model parameters using simple least squares regression, the engineer had to define an error function (Equation 1.66), sometimes also referred to as the objective function, to measure the discrepancy between the target values d_i and the model’s predicted values. Englezos and Kalogerakis (2001) report that the choice of objective function is very important, as it dictates both the values of the estimated model parameters and their statistical properties. Parameter estimation techniques can broadly be classified into two categories, namely *explicit* and *implicit estimation* techniques.

Implicit estimation techniques are used in scenarios where the model output and input are related via an implicit function. In contrast, in explicit estimation techniques the model output is an explicit function of the inputs to the model. Consequently, the simple least squares objective function (Equation 1.66), which was motivated from a frequentist maximum likelihood perspective (Section 2.4.2), is an explicit estimation technique. This thesis only considers explicit estimation techniques for estimating the unknown parameters of lumped system dynamic models.

3.3. Parameter Estimation Methods for Algebraic Dynamic Models

For a small subset of ordinary differential equation models describing the state variables of a physical system, it is possible to analytically integrate the differential equations to obtain closed-form algebraic solutions (Marlin, 2000).

As a result, the engineer requires a methodology to estimate the parameters of algebraic dynamic models directly from noise-corrupted time series data.

Englezos and Kalogerakis (2001) address the problem of estimating the parameters of models that cannot be written in the form required by Equation 1.57, i.e. models that are nonlinear in the model parameters Ω . The authors propose approximating the nonlinear model with a first-order Taylor expansion, about some initial guess Ω_{guess} for the unknown model parameters (similar to that of Equation 1.109), such that one can construct the sum-of-squares error function associated with simple least squares regression (similar to Equation 1.66).

The authors then proceed by analytically minimising the resulting sum-of-squares error function, with respect to the unknown parameters, to obtain an explicit update equation for $\Delta\Omega = (\Omega - \Omega_{guess})$ which stems from using the first-order Taylor expansion. Note that the update equation is relative to the initial parameter guess Ω_{guess} and provides the step direction for the new parameter guess. The update equation can be used in an iterative manner to estimate the unknown model parameters until some convergence criteria are met. This methodology is referred to as the Gauss-Newton method in optimisation theory and allows the engineer to estimate the parameters of nonlinear models from time series data. A problem that arises is that if Ω_{guess} is not close to the true model parameter values, the resulting update $\Delta\Omega$ might overstep. This can result in the parameter update for the next iteration diverging into a region of parameter space that is not supported by the model causing the parameter estimation procedure to fail.

The over-stepping problem is addressed by introducing a stepping parameter that decreases the size of the update $\Delta\Omega$ at each iteration and makes the parameter estimation procedure more robust. Englezos and Kalogerakis (2001) state that multiple methods exist for selecting the value of the stepping parameter, however, the simplest and most widely used approach is to establish the stepping parameter value from the bisection rule. Typically, one starts with a stepping parameter value of one and continues halving the value of the stepping parameter until the sum-of-squares objective function takes on a value less than the previous iteration.

When the sum-of-squares objective function value is less than in the previous iteration, the stepping parameter value is accepted and used to update the parameter guess for the next iteration. One is essentially performing a line search in the direction corresponding to $\Delta\Omega$ by adjusting the stepping parameter value such that the sum-of-squares error function is reduced in that direction. The introduction of the stepping parameter makes convergence to locally optimal parameter estimates monotonic, since the sum-of-squares objective function decreases after each iteration. The procedure outlined in Englezos and Kalogerakis (2001) can be interpreted as converting the nonlinear regression problem into a series of simple least squares regression problems allowing the engineer to estimate the parameter of algebraic dynamic models that cannot be written as a linear combination of the model parameters. Recall from

CHAPTER 3: LITERATURE REVIEW

Section 2.7 that this type of parameter estimation procedure is typically referred to as nonlinear least squares.

Furthermore, Englezos and Kalogerakis (2001) state that the Gauss-Newton parameter estimates are equivalent to the frequentist maximum likelihood parameter estimates. This can be verified if one assumes that the target variable d is generated independently from a distribution similar to Equation 1.55 (with the distribution mean adjusted to account for the linear approximation to the nonlinear model), corrupted by i.i.d. sensor noise following a zero-mean Gaussian with precision β . Refer back to Section 2.4.2 for a discussion of the connection between the likelihood function and the sum-of-squares error function used in simple least squares regression.

Croaze, Pittman and Reynolds (2012) and **Lai, Kek and Tay (2017)** present a slightly different way of implementing the Gauss-Newton method to estimate the parameters of models that cannot be written as a linear combination of the parameters. Instead of linearising the nonlinear model and constructing the simple least squares objective function (similar to Equation 1.66), the above-mentioned authors construct the objective function such that it directly measures the discrepancy between the data and the nonlinear model. Thus, the resulting objective function is a nonlinear function of the unknown parameters. In order to obtain estimates for the unknown parameters, it is necessary to minimise the nonlinear objective function such that the discrepancy between the data and the model predicted values are minimised.

To achieve the desired goal, Croaze, Pittman and Reynolds (2012) and Lai, Kek and Tay (2017) approximate the nonlinear objective function with a second-order Taylor expansion about some initial guess Ω_{guess} . The second-order Taylor expansion requires evaluating the Hessian matrix which contains the second-order partial derivatives with respect to the unknown model parameters. However, the Hessian matrix is approximated by $J^T J$ where the notation J refers to the Jacobian matrix (Equation 1.110). Using this approximation saves a considerable amount of computational time, since there is no need to evaluate the second-order partial derivatives of the Hessian matrix. The second-order Taylor approximation to the nonlinear objective function, in conjunction with the approximated Hessian matrix, is then analytically minimised to obtain an explicit update equation for $\Delta\Omega = (\Omega - \Omega_{guess})$.

The update equation provides the step direction for the new parameter guess used in the next iteration, relative to the initial guess Ω_{guess} , and can be used in an iterative manner to estimate the unknown model parameters until some convergence criteria are met. Similar to the approach in Englezos and Kalogerakis (2001), over-stepping can occur if Ω_{guess} is far away from the true model parameters. Croaze, Pittman and Reynolds (2012) also remedy the over-stepping problem by introducing a stepping parameter. However, Lai, Kek and Tay (2017) do not address the problem of over-stepping at all. It appears that, by default, the Gauss-Newton methodology presented in the work of Lai, Kek and Tay (2017) assumes a stepping parameter value of one.

Croaze, Pittman and Reynolds (2012) and Lai, Kek and Tay (2017) provide multiple illustrative examples for implementing the Gauss-Newton method. Furthermore, these authors also point out that the Gauss-Newton methodology is not the only approach to minimising the nonlinear objective function. Other optimisation methods include Newton's method, the Quasi-Newton

CHAPTER 3: LITERATURE REVIEW

method and the Levenberg-Marquardt method. Croaze, Pittman and Reynolds (2012) report the Levenberg-Marquardt method to be generally superior to the Gauss-Newton method, because the Hessian matrix is better approximated.

Chappell et al. (2009) approaches the nonlinear least squares regression problem from a completely different perspective by providing a variational inference based (Section 2.6) Bayesian approach to inferring the model parameters $\mathbf{\Omega}$ (cf. Englezos and Kalogerakis (2001) which provided a frequentist maximum likelihood interpretation). Similar to Englezos and Kalogerakis (2001), Chappell et al. (2009) approximates the nonlinear model by a first-order Taylor expansion (Equation 1.109), which allows the authors to construct the sum-of-squares error function (similar to Equation 1.66) associated with simple least squares regression for the linearised model. However, instead of following the sum-of-squares error function route, Chappell et al. (2009) instead follows the likelihood function route by considering the probability of the observed data set, given the unknown model parameters $\mathbf{\Omega}$ and the sensor precision β . Refer to Section 2.4.2 where a connection was established between simple least squares regression and the frequentist maximum likelihood perspective.

By following the probabilistic route via the likelihood function, Chappell et al. (2009) can exploit Bayes' rule for parameter estimation (similar to Equation 1.53) to obtain a posterior distribution over the unknown quantities of interest. The specific quantities of interest in the work of Chappell et al. (2009) correspond to the model parameters $\mathbf{\Omega}$ and the sensor precision parameter β . Observe that the unknown quantities of interest correspond to the quantities inferred by the second probabilistic model for Bayesian linear regression outlined in Section 2.4.3, expect that, here $\mathbf{\Omega}$ corresponds to the unknown parameters of the nonlinear model.

Instead of determining an update for $\Delta\mathbf{\Omega} = (\mathbf{\Omega} - \mathbf{\Omega}_{guess})$ relative to some initial guess $\mathbf{\Omega}_{guess}$, as was the case for all the preceding authors, Chappell et al. (2009) rather approximates the nonlinear model by a first-order Taylor expansion about the mode of the posterior distribution. This might seem strange at first; however, due to linearising the model, Chappell et al. (2009) can construct a likelihood function similar to that in Equation 1.56 for the linearised model. Recall that the likelihood function given by Equation 1.56 takes a Gaussian form. Based on the second Bayesian linear regression probabilistic model discussed in Section 2.4.3, the conjugate priors (Section 2.3.3) for the unknown model parameters $\mathbf{\Omega}$ (based on the linearised model) and sensor precision parameter β correspond to a multivariate Gaussian and Gamma distribution (Table 1.1), respectively. Chappell et al. (2009) exploit this conjugate prior but makes a further independence assumption (Equation 1.4) between the model parameters $\mathbf{\Omega}$ and sensor precision parameter β . In other words, it is assumed that the prior distribution over $\mathbf{\Omega}$ and β factorises as $p(\mathbf{\Omega}, \beta) = p(\mathbf{\Omega})p(\beta)$.

By combining the likelihood function, based on the first-order Taylor approximation to the nonlinear model, with the prior $p(\mathbf{\Omega}, \beta)$, the authors obtain a posterior distribution $p(\mathbf{\Omega}, \beta | \mathcal{D})$. Due to conjugacy as well as the additional independence assumption, the marginal posterior distribution over $\mathbf{\Omega}$ takes the form of a multivariate Gaussian distribution. Thus, the nonlinear model is approximated by a first-order Taylor expansion about the mode of the Gaussian posterior distribution. The problem that arises is that the mode of the Gaussian posterior distribution is not known. A possible approach is to provide an initial guess for the posterior mode and iteratively updating it until some convergence criteria are met. This is similar to the

CHAPTER 3: LITERATURE REVIEW

Gauss-Newton method outlined by the preceding authors where the parameter guess for the next iteration was determined from $\Delta\boldsymbol{\Omega} = (\boldsymbol{\Omega} - \boldsymbol{\Omega}_{guess})$ relative to some initial guess $\boldsymbol{\Omega}_{guess}$.

Chappell et al. (2009) addresses the problem of iteratively updating the Gaussian posterior distribution mode by exploiting variational inference (Section 2.6) where the latent variables correspond to $\mathbf{z} = \{\boldsymbol{\Omega}, \boldsymbol{\beta}\}$. Variational inference is traditionally used to approximate difficult-to-compute probability densities that are infeasible to evaluate directly. Recall that variational inference is inherently an iterative optimisation scheme. The solution for each optimal variational factor is obtained by first initialising all the relevant variational factors $q_j(\mathbf{z}_j)$ followed by cycling through each variational factor and, in turn, replacing each factor with the revised estimate using the current estimates for all other variational factors. After each cycle of updating the variational factors of the CAVI algorithm, Chappell et al. (2009) updates the Gaussian posterior distribution mode. One observes that the variational inference based solution outlined by the Chappell et al. (2009) is not used to approximate any difficult-to-compute probability density, but is rather exploited for its iterative nature to address the nonlinear least squares regression problem.

An additional benefit of following the Bayesian interpretation of nonlinear least squares is that one can use the evidence (otherwise referred to as the marginal likelihood) as a convergence criterion. For any given model and prior, the evidence takes on a constant value. However, in the nonlinear least squares setting, the evidence continuously changes as one updates the (approximate) linearised model between iterations. The evidence is a theoretically justified quantity for model comparison in Bayesian inference (Bishop, 2006). If one views the various linearisations as alternative models, optimising the evidence corresponds to searching for an optimal linearised model. However, the evidence is often intractable for most models of practical interest.

In particular, when using variational inference, where the posterior over the parameters is assumed to be intractable, the evidence will generally be intractable. Variational inference tackles this by optimising the ELBO (Section 2.6), which is a lower bound on the model evidence. Thus, by using maximisation of the ELBO as a convergence criterion, one can naturally find a good variational fit with CAVI, as well as implicitly optimise the model linearisation at each iteration. (Note that, as Chappell et al. (2009) point out, the ELBO in this setting may not follow the typical monotonic increasing behaviour of standard CAVI with a fixed model due to updating of the linear model approximation altering the true evidence value). This can be contrasted with the Gauss-Newton optimisation methodology for which multiple convergence criteria exist with no reason for choosing one convergence criteria above the other. Chappell et al. (2009) then proceed to benchmark their variational Bayesian nonlinear regression approach against the Gauss-Newton nonlinear least squares methodology.

Each of the aforementioned authors demonstrated their proposed parameter estimation methodology on different illustrative examples, reporting different performance criteria results. Table 3.1 presents a comparison of the parameter estimation methodology features based on the illustrative examples.

CHAPTER 3: LITERATURE REVIEW

Table 3.1: Features of the illustrative examples used to evaluate the parameter estimation methodology performance presented by Englezos and Kalogerakis (2001) [EK], Chappell et al. (2009) [CEL], Croaze, Pittman and Reynolds (2012) [CPR] and Lai, Kek and Tay (2017) [LKT].

| Feature | EK | CPR | LKT | CEL |
|--|----------------------|-----------------|-----------------|------------------------------------|
| Parameter estimate accuracy reported/compared to other literature sources/compared to ground truth simulation values | Yes | No | No | Yes |
| Nature of parameter estimates | Point estimates | Point estimates | Point estimates | Distribution over parameter values |
| Real or simulated data sets used | Both | Real | Both | Both |
| Parameter uncertainty representation (confidence/credibility intervals or other) | Confidence intervals | N/A* | N/A* | Other (box plots) |
| Computational runtime reported | No | No | No | No |
| Single/multiple algebraic equation parameter estimation | Both | Single | Single | Single |
| Statistical methodology followed | frequentist | N/A* | N/A* | Bayesian |

*N/A - not applicable

It is worth mentioning that **Au (2012)** attempted to develop a mathematical theory connecting frequentist and Bayesian quantification of parameter uncertainty (Section 2.8) based on what the author refers to as *second order theory*. This requires approximating the posterior distribution with a second order Taylor expansion about the ‘most probable value’ of the unknown parameters, i.e. the posterior distribution mode, followed by fitting a Gaussian distribution centered at that mode. Note that this procedure implicitly requires some form of optimisation routine to find a suitable value for the mode. Within the context of the Machine Learning field, this type of approach is referred to as the *Laplace approximation* and is an alternative to that of variational inference or Markov chain Monte Carlo methods discussed in Section 2.6 (Bishop, 2006). Furthermore, if the distribution under consideration is multimodal, each Laplace approximation will be different depending on which mode is considered.

Au (2012) proceed by explicitly pointing out that there is indeed a connection between the frequentist and Bayesian quantification of parameter uncertainty via the respective covariance matrices (obtained from each statistical methodology) and that the covariance matrices only differ by a weight matrix factor when considering multiple data sets over independent, repeated experiments based on the second order theory approximation. However, these results only hold

CHAPTER 3: LITERATURE REVIEW

true in the absence of any modeling error and can display significant inconsistencies when modeling errors exist. **Au (2012)** explicitly make no claim as to whether Bayesian and frequentist statistics provide similar results in applications with real-world data and state that the results depend on the model, environmental conditions, and the nature of the parameter under consideration.

3.4. Parameter Estimation Methods for ODE Dynamic Models

Recall that lumped system dynamic models take the form of ordinary differential equations, however, is it typically not possible to obtain a closed-form algebraic solution (Section 3.3) for the state variable dynamic response (Marlin, 2000).

As a result, the engineer requires a methodology to estimate the parameters of the ordinary differential equation model directly from noise-corrupted time series data.

Englezos and Kalogerakis (2001) addresses the problem of estimating the parameters of Ordinary differential equation models by positing some function with an unknown functional form that the engineer can fit to the noise-corrupted time series data using simple least squares (Section 2.4.2). Using a first-order Taylor expansion about some initial guess Ω_{guess} (similar to Equation 1.109), the authors obtain a linear approximation to the unknown function. Englezos and Kalogerakis (2001) proceed by assuming that a linear relationship exists between the unknown function and the state variable such that the Taylor approximation can be written as a function of the state variable itself. However, the problem that arises is that the engineer only has a differential equation describing the state variable. Furthermore, the Jacobian matrix can also not be readily obtained without an analytic expression for the state variable. The authors circumvent this by setting up additional differential equations, one for each of the unknown parameters, that, when integrated, produce the appropriate element-wise entries required by the Jacobian matrix.

Setting up additional differential equations might seem strange at first since the engineer's goal is to regress the parameters of the ordinary differential equation describing the state variable. However, Englezos and Kalogerakis (2001) state that one can simply use an external differential equation solver to simultaneously integrate the state variable differential equation as well as the additional differential equations that produce the appropriate element-wise entries of the Jacobian matrix. This gives them access to the numerical values of the state variable and Jacobian matrix, evaluated at the initial guess Ω_{guess} , which can be used to calculate the numerical values of the linear approximation to the initially posited function.

With these numerical values available, Englezos and Kalogerakis (2001) can construct the sum-of-squares error function (similar to Equation 1.66) and analytically minimise the resulting objective function, with respect to the unknown parameters, to obtain an explicit update equation for $\Delta\Omega = (\Omega - \Omega_{guess})$ which stems from using the first-order Taylor expansion (similar to Equation 1.109). This methodology gives rise to the Gauss-Newton method for estimating the parameters of ordinary differential equations from time series data. As with estimating the parameters of algebraic dynamic models discussed in the previous section, $\Delta\Omega$ provides the step direction for the new parameter guess and can be used in an iterative manner to estimate the unknown model parameters until some convergence criteria are met. However,

CHAPTER 3: LITERATURE REVIEW

over-stepping can again occur, and is again addressed by introducing a stepping parameter whose value is determined from the bisection rule. Furthermore, Englezos and Kalogerakis (2001) state that the Gauss-Newton ODE parameter estimates are equal to the frequentist maximum likelihood parameter estimates. This can be verified if one assumes that the target variable d is generated independently from a distribution similar to Equation 1.55 (with the distribution mean adjusted to account for the Taylor approximation), corrupted by i.i.d. sensor measurement noise following a zero-mean Gaussian with precision β .

Calderhead, Girolami and Lawrence (2009) pointed out that existing Bayesian and non-Bayesian methods for estimating the parameters of ODEs all require numerical solutions for the differential equations in order to construct the likelihood function (as in the above procedure outlined by Englezos and Kalogerakis (2001)). Calderhead, Girolami and Lawrence (2009) state that the computational cost associated with obtaining numerical solutions of the ODEs can result in extremely slow computational runtime. To address this, Calderhead, Girolami and Lawrence (2009) propose the use of Gaussian processes (Section 2.5) to directly predict the underlying state variable as well as its derivative from noise-corrupted time series data avoiding the need to numerically solve the ODEs.

Calderhead, Girolami and Lawrence (2009) start by defining a Gaussian process prior (Section 2.5) over the state variables, incorporates noise-corrupted observations of the state variables and obtains a Gaussian process posterior distribution over the state variables. From this Gaussian process posterior distribution, the authors evaluate the marginal likelihood of the observed data and combines this marginal likelihood with two independent prior distributions. The first prior distribution is over the measurement noise associated with each state variable while the second prior distribution is over the Gaussian process hyperparameters. The combination of the prior distributions with the marginal likelihood of the data results in a posterior distribution over the state variables measurement noise and the Gaussian process hyperparameters.

The final step involves constructing two additional statistical models for the state variable derivative. The first model expresses a probability distribution over the state variable derivative, as informed by the Gaussian process interpolating the state variable derivative, while the second model expresses a probability distribution over the state variable derivative informed by the ordinary differential equation describing said state derivative. An overall probability density is obtained by combining the two statistical models using a Product of Experts approach, which was originally introduced by **Hinton (2002)**. Calderhead, Girolami and Lawrence (2009) then proceed to obtain samples of the parameters from the desired marginal posterior distribution over the ODE parameters by sampling from the joint posterior distribution using population-based Markov chain Monte Carlo.

Calderhead, Girolami and Lawrence (2009) provide several examples illustrating their use of Gaussian processes to estimate the parameters of differential equations. However, all illustrative examples involve ODEs that completely describe the system under consideration, with no exogenous inputs to the system. An exogenous input is an independent variable that affects the physical system under consideration whose characteristics and generation process are not known. In other words, the engineer has no means of describing the exogenous input using some mathematical form. One can think of an exogenous input as a way of ‘setting’

CHAPTER 3: LITERATURE REVIEW

arbitrary external system conditions. Hence, the exogenous input affects the system but is not affected by the system itself (Ljung, 1999; Marlin, 2000).

Dondelinger et al. (2013) provides an extension of the Gaussian process based approach presented by Calderhead, Girolami and Lawrence (2009), stating that a disadvantage of their method stems from the fact that the Gaussian process hyperparameters are inferred from data alone, without considering the feedback mechanism associated with the ODEs. In other words, the method proposed by Calderhead, Girolami and Lawrence (2009) does not consider the effect of adapting the ODE parameters on the Gaussian process hyperparameter inference procedure. Furthermore, the authors note that while the approach developed by Calderhead, Girolami and Lawrence (2009) works well for relatively noise-free observations, it provides rather poor parameter estimates when noisier data is encountered.

Dondelinger et al. (2013) propose an improved inference scheme which they refer to as adaptive gradient matching. Their improved inference scheme infers the ODE parameters and the Gaussian process hyperparameters jointly from the posterior distribution. They claim that by doing so, they effectively introduce an information feedback scheme between the Gaussian process hyperparameters and the ODE parameters, improving on the results obtained from the methodology proposed by Calderhead, Girolami and Lawrence (2009). Closely related to the work of Dondelinger et al. (2013) is that of **Campbell and Steele (2012)** which use a method referred to as smooth functional tempering, which is also a population-based MCMC approach, to infer ODE parameters. The approach Campbell and Steele (2012) present is similar to the adaptive gradient matching paradigm outlined in the work of Dondelinger et al. (2013), however, Campbell and Steele (2012) use B-splines to interpolate the state variables instead of Gaussian processes. Similar to Calderhead, Girolami and Lawrence (2009), the examples illustrating the improved inference scheme presented by Dondelinger et al. (2013) are all ODEs that completely describe the system under consideration with no exogenous inputs.

The proposed approaches outlined by Calderhead, Girolami and Lawrence (2009) and Dondelinger et al. (2013) both rely on a Product of Experts heuristic which allows the authors to combine the ODE informed distribution of the state variable derivatives with the Gaussian process posterior distribution interpolating the state variable derivatives. The motivation behind the Product of Experts approach stems from the fact that by multiplying the two distributions, both the Gaussian process data fit and the ODE response have to be satisfied. In other words, the resulting probability distribution will only assign high probability if both the individual probability distributions, referred to as the ‘experts’, assign high probability mass.

The Product of Experts approach was criticised by **Barber and Wang (2014)** who proposed an alternative methodology for inferring the parameter of the ODEs via a probabilistic generative model that can be represented by a directed acyclic graph. However, in turn, **Macdonald, Higham and Husmeier (2015)** criticises the approach outlined by Barber and Wang (2014) and states that their probabilistic generative model leads to an intrinsic identifiability problem.

Gorbach, Bauer and Buhmann (2017) take the work of Calderhead, Girolami and Lawrence (2009) and Dondelinger et al. (2013), retain the Product of Experts heuristic, and propose a variational inference based framework (Section 2.6) which can infer the state variables and ODE model parameters simultaneously. The use of variational inference rather than MCMC has the additional benefit of offering significant computational runtime

CHAPTER 3: LITERATURE REVIEW

improvements. Gorbach, Bauer and Buhmann (2017) state that the Bayesian approaches of the aforementioned authors work well for a fully observed system, but that the proposed approaches cannot simultaneously infer unobserved state variables and ODE parameters and performs rather poorly when only combinations of the state variables are observed.

Gorbach, Bauer and Buhmann (2017) proceed to develop their variational inference based approach for Gaussian process gradient matching by exploiting local linearity of ODEs they consider. This is achieved by restricting attention to models that have reactions based on mass-action kinetics that can be written as a linear combination of the parameters. The mass-action kinetics are allowed to contain an arbitrarily large product of monomials of the state variables. They proceed to obtain a MAP estimate for the ODE parameters, but, the joint posterior distribution over parameters and state variables is analytically intractable. Gorbach, Bauer and Buhmann (2017) proceed by approximating this posterior distribution with mean-field variational inference (Section 2.6) by factorising the joint posterior distribution into a distribution over the ODE parameters and a distribution over state variables. The distribution over state variables is further factorised into distributions for each state variable.

For unobserved states, the authors assume a linear relationship between the observations and the state variables similar to the Gauss-Newton methodology outlined in Englezos and Kalogerakis (2001). The proposed variational inference based approach provides a powerful inference framework that is scalable, outperforms existing approaches in terms of computational runtime and accuracy, and performs reasonably well in scenarios where states are only partially observed. However, the proposed approach is restricted to ODE systems in which the state variables appear as monomials. To the best of the current author's knowledge, the examples illustrating the variational inference based approach for Gaussian process gradient matching presented by Gorbach, Bauer and Buhmann (2017) consists of ODEs that completely describe the system under consideration with no exogenous inputs. Furthermore, Gorbach, Bauer and Buhmann (2017) assume that the Gaussian process hyperparameters are known *a priori* and do not address the problem of estimating the hyperparameters from data.

Wenk et al. (2018) address the controversial debate about the Product of Experts approach used by Calderhead, Girolami and Lawrence (2009), Dondelinger et al. (2013) and Gorbach, Bauer and Buhmann (2017), and in the process discover and explain some theoretical inconsistencies present in the previous works. Based on these insights, they proceed to develop a new graphical model to replace the Product of Experts approach as well as develop a novel algorithm, referred to as Fast Gaussian Process Gradient Matching (FGPGM), which improves the current state-of-the-art ODE parameter inference approaches. Unlike the work of Calderhead, Girolami and Lawrence (2009) and Dondelinger et al. (2013), which requires a population-based MCMC sampling scheme, the approach proposed by Wenk et al. (2018) can use a single-chain Metropolis-Hastings scheme. Furthermore, Wenk et al. (2018) explicitly states that the main results of the preceding authors still hold as their proposed approaches are consistent with the newly developed graphical model.

Each of the aforementioned authors demonstrated their unique proposed parameter estimation methodology on different illustrative examples (with some overlap between the illustrative examples), reporting different performance criteria results. Table 3.2 presents a comparison of the parameter estimation methodology features based on the illustrative examples.

CHAPTER 3: LITERATURE REVIEW

Table 3.2: Features of the illustrative examples used to evaluate the parameter estimation methodology performance presented by Englezos and Kalogerakis (2001) [EK], Calderhead, Girolami and Lawrence (2009) [CGL], Dondelinger et al. (2013) [DEL], Gorbach, Bauer and Buhmann (2017) [GBB] and Wenk et al. (2018) [WEL].

| Feature | EK | CGL | DEL | GBB | WEL |
|--|----------------------|--|------------------------------------|------------------------------------|------------------------------------|
| Parameter estimate accuracy reported/compared to other literature sources/compared to ground truth simulation values | Yes | Yes | Yes | Yes | Yes |
| Nature of parameter estimates | Point Estimates | Distribution over parameter values | Distribution over parameter values | Distribution over parameter values | Distribution over parameter values |
| Real or simulated data sets used | Both | Both | Simulated | Simulated | Simulated |
| Parameter uncertainty representation (confidence/credibility intervals or other) | Confidence intervals | Other (Mean and variance summary statistics) | No | Other (box plots) | Other (box plots) |
| Computational runtime reported | No | Yes | Yes | Yes | Yes |
| Single/multiple ODE parameter estimation | Both | Multiple | Multiple | Multiple | Multiple |
| Statistical methodology followed | frequentist | Bayesian | Bayesian | Bayesian | Bayesian |
| Gaussian process used | No | Yes | Yes | Yes | Yes |
| Approximate inference technique used | N/A* | MCMC | MCMC | Variational Inference | MCMC |
| Exogenous input disturbance considered | No | No | No | No | No |

*N/A - not applicable

3.5. Remark: Suitability of the Gaussian Process Regression Approach

From Section 3.4 and Table 3.2, observe that Gaussian processes regression has established itself as a successful tool in the field of parameter estimation for ODE models. Furthermore, note from Table 3.2 that none of the parameter estimation methodologies developed by Calderhead, Girolami and Lawrence (2009), Dondelinger et al. (2013), Gorbach, Bauer and Buhmann (2017) and Wenk et al. (2018) address the parameter estimation procedure for ODE models with exogenous input disturbances.

Given the definition of the Gaussian process outlined in Section 2.5, it is possible to extend the ODE parameter estimation procedure to include ODEs with an exogenous input disturbance structure. This can be achieved by augmenting the joint distribution given by Equation 1.102 with the additional unknown exogenous input disturbance function values and the input disturbance target variable measurements. Note that here it is assumed that the input disturbance target variable measurements are available to the engineer. One can then use the log marginal likelihood of the augmented model (Section 2.5.2), in conjunction with gradient-based optimisation, to optimise the relevant Gaussian process hyperparameter. This train of thought is further explored in Section 4.5.2.

3.6. Summary

Chapter 3 presented the relevant literature pertaining to parameter estimation methods for lumped system algebraic and ODE models. Arguably, including a discussion on parameter estimation methods for algebraic dynamic models might not seem beneficial since (1) only a small subset of ODE models can be explicitly solved to obtain a closed-form algebraic solution for the state variable dynamic response, and (2) several estimation methodologies exist for inferring the parameters of ODE models from noise-corrupted time series data.

However, in several other applications of chemical engineering, such as estimating the parameters of thermodynamic equations of state, the idea of nonlinear regression is important. Thus, being able to estimate the parameters of algebraic models is generally applicable to a wider scientific audience and other scientific fields of study. Furthermore, the idea of nonlinear regression provides the introduction to the proposed approaches and methodology section discussed in Chapter 4 where the author illustrates, from both the frequentist and Bayesian viewpoints, how it can be used to estimate the parameters of an algebraic model that describes the dynamic response of a CSTR with an assumed exogenous input disturbance structure. The Gauss-Newton methodology presented by Englezos and Kalogerakis (2001) and the Bayesian nonlinear regression approach of Chappell et al. (2009), outlined in Section 3.3, will resurface again in the context of estimating these parameters.

The problems that arise, however, is that the structure of the exogenous input disturbance is typically not known in a practical setting nor is it generally possible to obtain an algebraic solution for the dynamic response of the system state variables of practical interest to chemical engineers. To address these problems, the author uses the algebraic dynamic model parameter inference framework as a guide to develop a regression methodology, based on the first Bayesian linear regression probabilistic model (Section 2.4.1) and Gaussian process regression (Section 2.5), which can infer the parameters of an ODE subject to an exogenous input

CHAPTER 3: LITERATURE REVIEW

disturbance with arbitrary structure. Here the ideas discussed and illustrated in the work of Calderhead, Girolami and Lawrence, (2009), Dondelinger et al. (2013), Gorbach, Bauer and Buhmann (2017) and Wenk et al. (2018) will resurface again as inspiration for the proposed methodology outlined in Chapter 4 (Section 4.5.2).

Chapter 4

Proposed Approaches and Methodology

“The theory of probabilities is at bottom nothing but common sense reduced to calculus; it enables us to appreciate with exactness that which accurate minds feel with a sort of instinct for which oftentimes they are unable to account.”

- Pierre Simon Laplace, 1819

4.1. Overview

Chapters 1 through 3 have provided the necessary foundations of the parameter estimation procedure. Unfortunately, as is the nature of these types of problems, a high level of mathematical engagement is required and it is nowhere in this work as noticeable as in the current chapter. To make the abstract notions used in the subsequent sections somewhat more tangible, the current chapter will ground the presentation of the proposed approaches for estimating the parameters of lumped system algebraic or ODE models by keeping the physical system under consideration in mind. To allow the reader to get the most out of the present chapter and the subsequent discussions, it is worth introducing the two case studies considered within the context of the current work as a means of relating the abstract ideas that follows back to physical, chemical engineering process units that are somewhat more easily interpretable.

4.2. Simulation Case Studies

Two case studies, both originating from Marlin (2000) (which is freely available online), are considered throughout the current work as a vehicle for illustrating the concepts and ideas presented in this chapter. Recall that attention is restricted to a single lumped system dynamic model with physicochemical variables that are spatially and time invariant. This allows the engineer to treat the physicochemical variables as model parameters. Assuming that these physicochemical variables are time invariant is technically not correct but is a close approximation when the model parameters change slowly enough with a small magnitude during the data collection period (Marlin, 2000).

4.2.1. Case Study 1 & 2: Continuous Stirred Tank Reactor

The first case study considers the dynamic response of an isothermal, constant volume CSTR with first-order reaction kinetics for which the mathematical model, as derived from first-principles by exploiting the conservation of mass, is given by Marlin (2000, Second Edition, Example 3.2) as:

$$V \frac{dC_A}{dt} = FC_{A0} - FC_A - kVC_A \quad \text{Equation 3.1}$$

Here dC_A/dt represents the state variable time derivative with the corresponding state variable being C_A (reactor outlet concentration). The flow rate F and reactor inlet concentration C_{A0} are external input variables to the system. The inlet concentration C_{A0} is taken as an exogenous input disturbance to the system while the inlet flow rate is assumed to remain (essentially) constant. The symbol k denotes the reaction rate constant and dictates the chemistry of the

CHAPTER 4: PROPOSED APPROACHES AND METHODOLOGY

reaction. Refer to Figure 4.1 for a graphical depiction of the isothermal, constant volume CSTR, as adapted from Marlin (2000, Second Edition, Example 3.2).

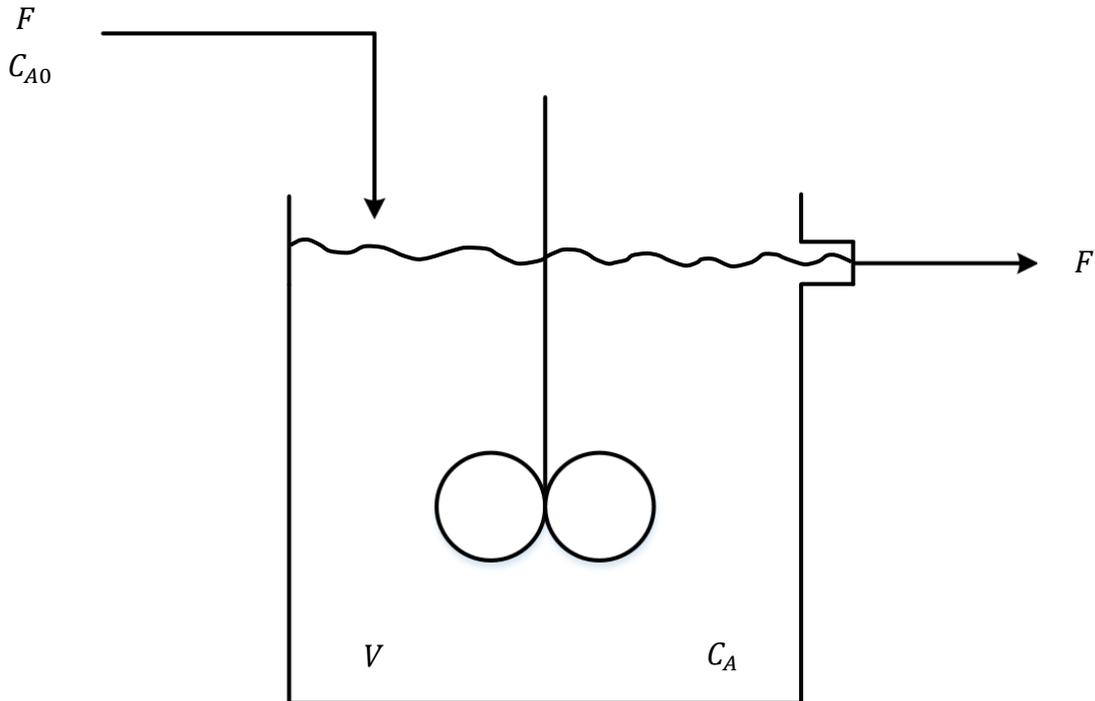


Figure 4.1: Graphical depiction of the isothermal, constant volume CSTR used for Case Studies 1 and 2 [figure adapted from Marlin (2000, Second Edition, Example 3.2)].

Furthermore, it is assumed that no form of process control is applied to the isothermal, constant volume CSTR (Section 1.5.3). For the CSTR case study, the volume V , flow rate F and reaction rate constant k are assumed to be constant quantities and form part of the model parameters that must be inferred from noise-corrupted time series data. Equation 3.1 may be rewritten such that:

$$\frac{dC_A}{dt} = \frac{F}{V}(C_{A0} - C_A) - kC_A \quad \text{Equation 3.2}$$

This particular choice of physical system takes the form of a linear (in the state) nonseparable ODE that can be solved for explicitly using the integrating factor given certain assumptions about the exogenous input disturbance to the system. If one restricts the input disturbance change in C_{A0} to a perfect step change, the analytical expression for the dynamic response of the outlet concentration is given by Marlin (2000) as:

$$C_A(t) = K_p \Delta C_{A0} (1 - \exp\{-t/\tau\}) + (C_A)_{initial} \quad \text{Equation 3.3}$$

The symbol denoting ΔC_{A0} represents the magnitude of the input step disturbance in C_{A0} . Note that in this case, the engineer is explicitly assuming that the characteristics and generation process of the exogenous input disturbance in C_{A0} is known exactly. The process gain K_p , process time constant τ , and the initial reactor steady-state outlet concentration $(C_A)_{initial}$ are calculated as follows:

CHAPTER 4: PROPOSED APPROACHES AND METHODOLOGY

$$K_p = \frac{F}{F + kV} \quad \text{Equation 3.4}$$

$$\tau = \frac{V}{F + kV} \quad \text{Equation 3.5}$$

$$(C_A)_{initial} = \frac{F}{F + kV} (C_{A0})_{initial} = K_p (C_{A0})_{initial} \quad \text{Equation 3.6}$$

Note that the process gain and reaction rate constant are expressed in terms of the model parameters and are, therefore, constant quantities as well. Table 3.1, as outlined in Marlin (2000), summarises the necessary information about the CSTR case study that can be used in the forward modeling approach (Section 1.1) to generate noise-free observations for the physical system in a deterministic manner.

Table 3.1: Isothermal, constant volume CSTR information summary [adapted from Marlin (2000, Second Edition, Example 3.2)].

| Process Variable/Parameter | Description | Variable/Parameter Value |
|----------------------------|--|--------------------------|
| F | Reactor flow rate | $0.085 \frac{m^3}{min}$ |
| V | Reactor volume | $2.1 m^3$ |
| ΔC_{A0} | Step change magnitude in exogenous input disturbance | $0.925 \frac{mole}{m^3}$ |
| k | Reaction rate constant | $0.040 min^{-1}$ |
| $(C_{A0})_{initial}$ | Initial steady state inlet concentration | $0.925 \frac{mole}{m^3}$ |

The explicit difference between Case Study 1 and 2 stem from the assumption made about the exogenous input disturbance. For Case Study 1 (Equation 4.3), one is assuming the generation process and structure of the exogenous input disturbance is exactly known. For Case study 2 (Equation 4.2), the exogenous input disturbance in C_{A0} is allowed to have any arbitrary structure.

The motivation for the CSTR case study is two-fold. Firstly, Equation 3.3 takes the form of an algebraic dynamic model that is linear in the unknown model parameter K_p , but is nonlinear in the model parameter τ . Secondly, the ordinary differential equation (Equation 3.2), which makes no assumption about the characteristics and generation process of the exogenous input disturbance C_{A0} , is linear in the unknown model parameters F/V and k (Table 3.3).

4.2.2. Case Study 3: Liquid Draining Tank

The second case study the current work considers is the dynamic response of a liquid draining tank with a partially opened flow restriction. The mathematical model describing the tank liquid level dynamic response, as derived from first-principles by exploiting the conservation of mass, is given by Marlin (2000, Second Edition, Example 3.6) as:

$$A \frac{dL}{dt} = F_0 - k_v \sqrt{L} \quad \text{Equation 3.7}$$

Here dL/dt represents the state variable time derivative with the corresponding state variable being L (draining tank liquid level). The flow rate F_0 is an exogenous input disturbance variable to the system. For the draining tank case study, the cross-sectional area A and the flow restriction coefficient k_v are assumed to be constant quantities and form part of the model parameters that must be inferred from noise-corrupted time series data. Furthermore, it is assumed that no form of process control is applied to the liquid draining tank (Section 1.5.3). The flow restriction coefficient k_v can be calculated with the following relation given by Marlin (2000, Second Edition, Example 3.6):

$$k_v = \frac{(F_0)_{initial}}{\sqrt{(L)_{initial}}} \quad \text{Equation 3.8}$$

Note that Equation 3.7 is nonlinear in the state variable L due to the presence of the square root level term, i.e. \sqrt{L} . Refer to Figure 4.2 for a graphical depiction of the liquid draining tank, as adapted from Marlin (2000, Second Edition, Example 3.6).

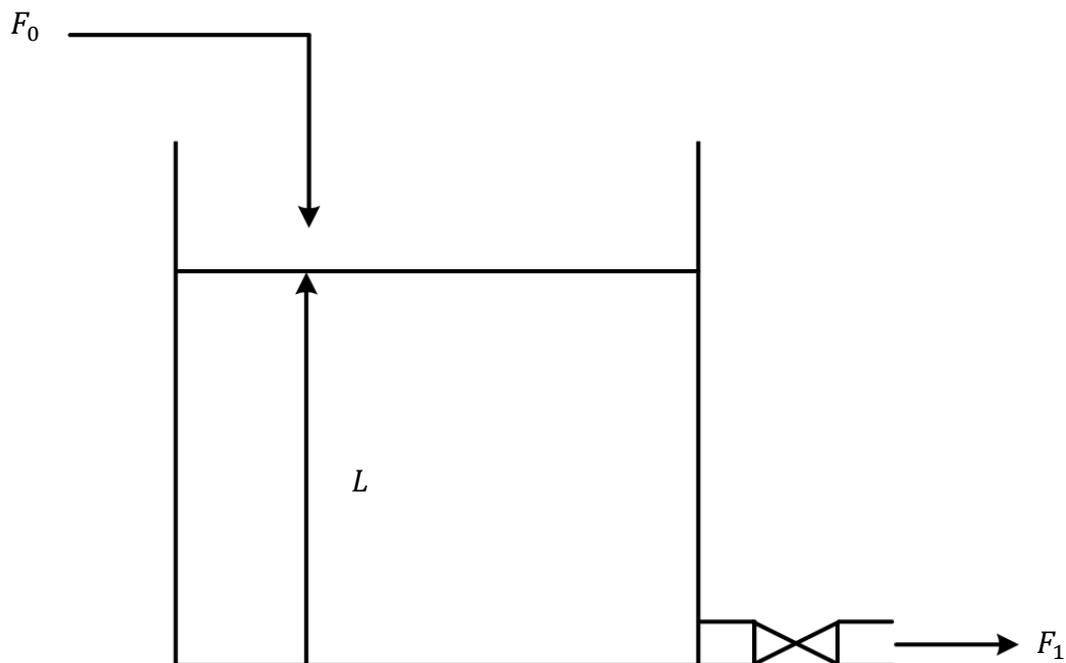


Figure 4.2: Graphical depiction of the liquid draining tank used for Case Study 3 [figure adapted from Marlin (2000, Second Edition, Example 3.6)].

CHAPTER 4: PROPOSED APPROACHES AND METHODOLOGY

General methods for determining a closed-form analytical solution for ODEs that are nonlinear in the state variables are typically not available and, if possible, obtaining an analytical solution is not a trivial task (Marlin, 2000). Thus, the standard methods for estimating the model parameters associated with analytical expressions (algebraic dynamic models) are not accessible. Typically, a linearised dynamic model is developed by approximating each nonlinear state variable term using a first-order Taylor expansion (Marlin, 2000). Equation 3.7 can be rewritten such that:

$$\frac{dL}{dt} = \frac{1}{A}F_0 - \frac{k_v}{A}\sqrt{L} \quad \text{Equation 3.9}$$

The motivation for this particular choice of case study is two-fold. Firstly, observe that Equation 3.9 is nonlinear in the state variable L but is linear in the model parameters $1/A$ and k_v/A (Table 3.3). Secondly, due to the high level of mathematical detail in this chapter, the reader might easily lose track of the overarching theme of the thesis which inevitably reduces to reliably estimating model parameters from noise-corrupted time series data. Thus, by selecting the draining tank case study, the author hopes to establish a connection between the abstract mathematical nature of this and subsequent chapters and the physical engineering process which the reader may be more familiar with.

Draining tanks are common process units encountered in everyday life, e.g. tanks used to collect rainwater or potable water systems installed for recycling and household usage. Thus, familiarity with single draining tank systems can aid by providing intuition for some of the advanced concepts used throughout this chapter (and subsequent chapters) as the goal is to avoid unnecessary complexity. Table 3.2, as outlined in Marlin (2000), summarises the necessary information about the liquid draining tank case study that can be used in the forward modeling approach (Section 1.1) to generate noise-free observations for the physical system in a deterministic manner:

Table 3.2: Liquid draining tank information summary [adapted from Marlin (2000, Second Edition, Example 3.6)].

| Process Variable/Parameter | Description | Variable/Parameter Value |
|----------------------------|------------------------------------|--------------------------|
| $(F_0)_{initial}$ | Tank steady state inlet flow rate | $100 \frac{m^3}{h}$ |
| $(F_1)_{initial}$ | Tank steady state outlet flow rate | $100 \frac{m^3}{h}$ |
| $(L)_{initial}$ | Tank steady state liquid level | $7 m$ |
| k_v | Flow restriction coefficient | $37.8 \frac{m^{2.5}}{h}$ |
| A | Tank cross-sectional area | $7 m^2$ |

4.3. Parameter Estimates Required

Table 3.3 summarises the model parameters that the engineer desires to infer from noise-corrupted time series data for the three case studies considered in the current work. Recall that the parameter vector for dynamic models that can not be written as a linear combination of the model parameters are denoted by the symbol $\boldsymbol{\Omega}$, while models that can be written as a linear combination of the model parameters are denoted by the parameter vector \boldsymbol{w} (Section 2.4).

Table 3.3: Dynamic model parameters to be inferred from noise-corrupted time series data.

| Dynamic Model | Case Study | Parameters to Infer |
|--|------------|---|
| Equation 3.3 (algebraic model) | 1 | $\boldsymbol{\Omega} = [K_p \ \tau]^T = [\Omega_1 \ \Omega_2]^T$ |
| Equation 3.2 (ODE model) – linear in state variable C_A | 2 | $\boldsymbol{w} = \left[\frac{F}{V} \ k\right]^T = [w_1 \ w_2]^T$ |
| Equation 3.9 (ODE model) – nonlinear in state variable L | 3 | $\boldsymbol{w} = \left[\frac{1}{A} \ \frac{k_v}{A}\right]^T = [w_1 \ w_2]^T$ |

4.4. Benchmark Method

The benchmark method used throughout the current thesis for estimating the parameters of lumped system algebraic and ODE models is based on the work of Englezos and Kalogerakis (2001). In other words, the current work uses the Gauss-Newton methodology to minimise the simple least squares objective function in an iterative manner by adjusting the unknown model parameters (Section 3.3 and Section 3.4). *Algorithms 1* and *2* provide the implementation steps, as outlined in Englezos and Kalogerakis (2001), that the engineer can implement to estimate the parameters of algebraic and ODE dynamic models, respectively.

Algorithm 1: Gauss-Newton method for estimating the parameters of lumped system algebraic dynamic models from noise-corrupted time series data (Chapters 4 and 8 - Englezos and Kalogerakis (2001)).

Algorithm 1: Inputs

- A lumped system algebraic model such as Equation 3.3.
- Sensor measurements of the state variable C_A dynamic response.
- An initial guess for the unknown algebraic model parameters τ and K_p collected in the 2×1 vector $\boldsymbol{\Omega}_{guess}$.

1. Initialise the guess ($\boldsymbol{\Omega}_{guess}$) for the unknown algebraic dynamic model parameters (Equation 3.3, Table 3.3).
2. Initialise the NSIG (number of significant digits desired for the unknown model parameters) to establish when convergence is reached.
3. For a pre-specified number of iterations $p = 1, 2, 3, \dots$ repeat the following procedure (the code implemented in this thesis uses 2000 iterations):

CHAPTER 4: PROPOSED APPROACHES AND METHODOLOGY

- 3.1 Compute the dynamic model output (Equation 3.3) and Jacobian matrix \mathbf{J} entries (Equation 1.110) for each of the sensor measurements $i = 1, \dots, N$. Denote the Jacobian matrix for each sensor measurement by \mathbf{J}_i .
- 3.2 Compute the simple least squares objective function value:

$$E(\boldsymbol{\Omega}) = \sum_{i=1}^N \left(d_{C_{A,i}} - \left\{ K_p \Delta C_{A0} \left(1 - \exp \left\{ -\frac{t_i}{\tau} \right\} \right) + (C_A)_{initial} \right\} \right)^2 \quad \text{Equation 3.10}$$

- 3.3 Set up matrix \mathbf{A} by calculating:

$$\mathbf{A} = \sum_{i=1}^N \mathbf{J}_i^T \mathbf{J}_i \quad \text{Equation 3.11}$$

- 3.4 Set up vector \mathbf{b} by calculating:

$$\mathbf{b} = \sum_{i=1}^N \mathbf{J}_i^T \left[d_{C_{A,i}} - \left\{ K_p \Delta C_{A0} \left(1 - \exp \left\{ -\frac{t_i}{\tau} \right\} \right) + (C_A)_{initial} \right\} \right] \quad \text{Equation 3.12}$$

The notation $d_{C_{A,i}}$ refers to the i^{th} sensor measurement of the state variable C_A . Here it is assumed that $(C_A)_{initial}$ is known, however, this need not be the case. The initial concentration can be regarded as a bias parameter that models any fixed offset in the data which can be estimated from the sensor measurement data (Section 2.4.2)

- 3.5 Solve the linear equation $\mathbf{A} \Delta \boldsymbol{\Omega}_{p+1} = \mathbf{b}$ and obtain the new step direction $\Delta \boldsymbol{\Omega}_{p+1}$ for the next iteration – this can be achieved by any *standard external linear equation solver* such as the *backslash* operator in MATLAB[®].
- 3.6 Following the discussion in Englezos and Kalogerakis (2001), use the bisection rule to determine a suitable size for the stepping parameter ρ_{step} and calculate:

$$\boldsymbol{\Omega}_{p+1} = \boldsymbol{\Omega}_p + \rho_{step} \Delta \boldsymbol{\Omega}_{p+1} \quad \text{Equation 3.13}$$

Englezos and Kalogerakis (2001) state that the bisection rule is the most robust and simplest way of determining the stepping parameter ρ_{step} . Here the parameter ρ_{step} is introduced to avoid overstepping which can result due to the first-order Taylor approximation (neglecting all higher order terms).

- 3.7 Continue the iterative procedure until the pre-specified number of iterations are reached or until convergence is achieved by evaluating the convergence criteria:

$$\frac{1}{M} \sum_{i=1}^M \left| \frac{\Omega_{p+1}^i}{\Omega_p^i} \right| \leq 10^{-NSIG} \quad \text{Equation 3.14}$$

Here the notation M refers to the number of unknown model parameters. Englezos and Kalogerakis (2001) report that Equation 3.14 is a general convergence criteria and provides consistent parameter estimate results (although this is not guaranteed) provided none of the dynamic model parameters converge to zero.

4. Compute the statistical properties (Section 4.8) of the estimated algebraic dynamic model parameters $\boldsymbol{\Omega}_{ML} = [K_{p,ML} \ \tau_{ML}]^T$. Recall that these parameter estimates coincide with the (frequentist) maximum likelihood parameter estimates (Section 2.4.2).

CHAPTER 4: PROPOSED APPROACHES AND METHODOLOGY

Algorithm 1: Outputs

- Point estimates for the algebraic dynamic model parameters τ and K_p .
- The simple least squares objective function value (at convergence) which is used to obtain a point estimate for the sensor variance parameter $\sigma_{noise,CA}^2$.
- Matrix \mathbf{A} which is used in conjunction with $\sigma_{noise,CA}^2$ to estimate the model parameter covariance matrix.

Algorithm 2: Gauss-Newton method for estimating the parameters of lumped system ODE dynamic models from noise-corrupted time series data (Chapters 6 and 8 - Englezos and Kalogerakis (2001)).

Algorithm 2: Inputs

- A lumped system ordinary differential equation model such as Equation 3.2 (or Equation 3.9).
- Sensor measurements of the state variable C_A dynamic response and the exogenous input disturbance C_{A0} .
- An initial guess for the unknown ODE model parameters F/V and k collected in the 2×1 vector $\mathbf{\Omega}_{guess}$.

1. Initialise the guess ($\mathbf{\Omega}_{guess}$) for the unknown ODE model parameters (Equation 3.2 or 4.9, Table 3.3).
2. Initialise the NSIG (number of significant digits desired in the unknown model parameters) to establish when convergence is reached.
3. For a pre-specified number of iterations $p = 1, 2, 3, \dots$ repeat the following procedure (the code implemented in this thesis uses 2000 iterations):
 - 3.1. Using an *external numerical differential equation solver*, integrate the state variable ODE (Equations 4.2 or 4.9) and $g_i(t)$ (refer to Algorithm 2: Additional Notes) for each of the ODE model parameters. Any standard numerical differential equation solver can be used such as ode45 in MATLAB[®]/Simulink. From the numerical differential equation solver results, at each of the sensor sampling time points, extract the ODE model predicted value. For each of the additional M unknown ordinary differential equations describing the appropriate Jacobian matrix element-wise entries, extract the corresponding Jacobian matrix element-wise entries $\mathbf{J}_i = [g_1(t_i) \ g_2(t_i) \ \dots \ g_M(t_i)]$ for each of the $i = 1, 2, \dots, N$ sensor measurements.
 - 3.2. Set up matrix \mathbf{A}_R by calculating:

$$\mathbf{A}_R = \mathbf{K} \left[\sum_{i=1}^N \mathbf{J}_i^T \mathbf{J}_i \right] \mathbf{K} = \mathbf{KAK} \quad \text{Equation 3.15}$$

The notation \mathbf{A}_R (similar for \mathbf{b}_R in step 3.3) refers to using the reduced Jacobian matrix \mathbf{J}_R (refer to Algorithm 2: Additional Notes) which is defined as:

$$\mathbf{J}_R = \mathbf{JK} \text{ with } \mathbf{K} = \text{diag}(w_0, w_1, \dots, w_{M-1}) \quad \text{Equation 3.16}$$

CHAPTER 4: PROPOSED APPROACHES AND METHODOLOGY

3.3. Set up vector \mathbf{b}_R by calculating:

$$\mathbf{b}_R = \mathbf{K} \left[\sum_{i=1}^N \mathbf{J}_i^T [d_{C_{A,i}} - C_A(t_i, \mathbf{w}_p)] \right] = \mathbf{Kb} \quad \text{Equation 3.17}$$

The notation $d_{C_{A,i}}$ refers to the i^{th} sensor measurement of the state variable C_A while $C_A(t_i, \mathbf{w}_p)$ refers to the model predicted value obtained from the numerical differential equation solver results at the sensor sampling time point t_i , based on the p^{th} update for the unknown ODE parameters. For the draining tank case study, simply replace $d_{C_{A,i}}$ with $d_{L,i}$ and use the corresponding differential equation describing the draining tank liquid level (Equation 3.9).

3.4. Compute the simple least squares objective function value:

$$E(\mathbf{w}) = \sum_{i=1}^N (d_{C_{A,i}} - C_A(t_i, \mathbf{w}_p))^2 \quad \text{Equation 3.18}$$

3.5. Solve the linear equation $\mathbf{A}_R \Delta \mathbf{w}_{R(p+1)} = \mathbf{b}_R$ and obtain the new step direction $\Delta \mathbf{w}_{R(p+1)}$ – this can be achieved by any *standard external linear equation solver* such as the *backslash* operator in MATLAB[®]

3.6. Using the bisection rule, determine a suitable size for the stepping parameter ρ_{step} and calculate:

$$\mathbf{w}_{p+1} = \mathbf{w}_p + \rho_{step} \mathbf{K} \Delta \mathbf{w}_{R(p+1)} \quad \text{Equation 3.19}$$

3.7. Continue the iterative procedure until the pre-specified number of iterations are reached or until convergence is achieved by evaluating the convergence criteria:

$$\frac{1}{M} \sum_{i=1}^M \left| \frac{w_{p+1}^i}{w_p^i} \right| \leq 10^{-NSIG} \quad \text{Equation 3.20}$$

Here the notation M refers to the number of unknown ODE model parameters.

4. Compute the statistical properties (Section 4.8) of the estimated differential equation model parameters, i.e. $\mathbf{w}_{ML} = [(F/V)_{ML} \ k_{ML}]^T$ or $\mathbf{w}_{ML} = [(1/A)_{ML} \ (k_v/A)_{ML}]^T$. Recall that these parameter estimates coincide with the (frequentist) maximum likelihood parameter estimates (Section 2.4.2).

Algorithm 2: Outputs

- Point estimates for the ordinary differential equation model parameters F/V and k .
- The simple least squares objective function value (at convergence) which is used to obtain a point estimate for the sensor variance parameter σ_{noise,C_A}^2 .
- Matrix \mathbf{A} which is used in conjunction with σ_{noise,C_A}^2 to estimate the model parameter covariance matrix.

Algorithm 2: Additional Notes***Jacobian Matrix***

The corresponding Jacobian matrix column-wise entries are obtained by solving the additional differential equations - in conjunction with the differential equation describing the state variable derivative - for each of the $i = 1, 2, \dots, M$ model parameters:

$$\frac{dg_i(t)}{dt} = \left(\frac{\partial f(t, \mathbf{w})}{\partial C_A} \right) g_i(t) + \frac{\partial f(t, \mathbf{w})}{\partial w_i} ; g_i(t_0) = 0 \quad \text{Equation 3.21}$$

The notation $f(t, \mathbf{w})$ refers to the differential equation describing the state variable derivative, i.e. $f(t, \mathbf{w}) = dC_A/dt$ (Equation 3.2). For the liquid draining tank case study, simply replace $f(t, \mathbf{w}) = dC_A/dt$ with $f(t, \mathbf{w}) = dL/dt$ (Equation 3.9).

Reduced Jacobian Matrix

If the parameters of the ODE model differ by more than one order of magnitude, the matrix \mathbf{A} can appear to be ill-conditioned despite having a well-defined parameter estimation problem. Englezos and Kalogerakis (2001) report that the best way to overcome this is by using the reduced Jacobian matrix \mathbf{J}_R (Steps 3.2 and 3.3). Note that the reduced Jacobian matrix adjustment is equally valid for *Algorithm 1* if one suspects that the algebraic model parameters differ by more than one order of magnitude. This technique is known as “Scaling of Matrix \mathbf{A} ” (Englezos and Kalogerakis, 2001).

4.5. Bayesian Parameter Estimation Methods

The following section outlines the proposed Bayesian approaches (Section 1.5.2) contrasting the Gauss-Newton methodology for estimating the parameters of algebraic (*Algorithm 1*) and ordinary differential equation (*Algorithm 2*) models.

4.5.1. Variational Bayesian Nonlinear Regression

The current work adapts the proposed variational Bayesian inference approach outlined in Chappell et al. (2009) (Section 3.3) with a minor modification. Chappell et al. (2009) infers a posterior distribution over the unknown algebraic model parameters $\mathbf{\Omega}$ and the sensor precision parameter $\beta = 1/\sigma_{noise, C_A}^2$ by assuming that the joint prior distribution $p(\mathbf{\Omega}, \beta)$ factorises into a product of marginal distributions $p(\mathbf{\Omega}, \beta) = p(\mathbf{\Omega})p(\beta)$, i.e. Chappell et al. (2009) makes an independence assumption (Equation 1.4). Note that Chappell et al. (2009) denotes the sensor precision parameter by ϕ whereas the current work uses the symbol β to keep the notation consistent with the second Bayesian linear regression probabilistic model discussed in Section 2.4.3. The current work makes use of a fully conjugate prior by using a prior where $\mathbf{\Omega}$ and β are dependent on each other, i.e. $p(\mathbf{\Omega}, \beta) = p(\mathbf{\Omega}|\beta)p(\beta)$.

The main reason the current work uses the fully conjugate prior arises from approximating the nonlinear model by a first-order Taylor approximation. As one is approximating the nonlinear model by a linear counterpart, the convergence of the variational Bayesian algorithm is no longer guaranteed, since the proposed algorithm may converge within a region in parameter space not supported by the algebraic model.

CHAPTER 4: PROPOSED APPROACHES AND METHODOLOGY

However, by using the conditional probability distribution $p(\boldsymbol{\Omega}|\beta)$ instead of the marginal distribution $p(\boldsymbol{\Omega})$, the coefficient β acts as a regulariser similar to that of regularised least squares (Section 2.4.2), with the aim to encourage convergence within the region of parameter space supported by the algebraic model (MacKay, 1996; Murphy, 2012). From Bayes' rule (similar to that of Equation 1.10) the joint parameter posterior distribution over the unknown parameter vector $\boldsymbol{\Omega}$ and the sensor precision parameter β is given by:

$$p(\boldsymbol{\Omega}, \beta | \mathbf{d}_{C_A}) = \frac{p(\mathbf{d}_{C_A} | \boldsymbol{\Omega}, \beta) p(\boldsymbol{\Omega} | \beta) p(\beta)}{p(\mathbf{d}_{C_A})} \quad \text{Equation 3.22}$$

Note that Equation 3.22 corresponds to the second probabilistic model for Bayesian linear regression (Section 2.4.3) with $\mathcal{D}_{C_A} = \{d_{C_A,i}\}_{i=1}^N$ representing the noise-corrupted state variable sensor measurements. The symbol \mathbf{d}_{C_A} denotes the sensor measurements collected in an $N \times 1$ column vector. The problem that arises with the regression procedure is that it is no longer possible to write the algebraic model as a linear combination of the unknown model parameters. However, similar to the Gauss-Newton methodology for algebraic dynamic models (Section 3.3 and 4.4), one can build a proxy for the nonlinear algebraic dynamic model by using a first-order Taylor approximation (similar to Equation 1.109) (Englezos and Kalogerakis, 2001; Chappell et al., 2009). In general, any algebraic dynamic model can be approximated about the mode of the parameter posterior probability distribution \mathbf{m}_N (which takes the form of a multivariate Gaussian distribution in this case (details discussed below) due to exploiting conjugacy (Table 1.1, Section 2.3.3)) by a first-order Taylor approximation such that for each sensor measurement at time t_i :

$$C_A(t_i, \boldsymbol{\Omega}) \approx C_A(t_i, \mathbf{m}_N) + \mathbf{J}_i(\boldsymbol{\Omega} - \mathbf{m}_N) + \text{H. O. T} \quad \text{Equation 3.23}$$

It is possible to linearise the algebraic dynamic model about the mode of the prior distribution. However, if the prior distribution mode is far away from the true underlying parameter values, the algorithm is at greater risk of diverging. All higher order terms (HOT) are assumed to be negligible, although this need not be the case. This procedure can be repeated for all N sensor measurements such that the notation \mathbf{J} refers to the stacked $N \times M$ Jacobian matrix whose elements are given by the partial derivatives of the nonlinear model with respect to the unknown model parameters – evaluated at the posterior mode \mathbf{m}_N – for the $i = 1, \dots, M$ unknown model parameters. With this in mind, the likelihood function $p(\mathbf{d}_{C_A} | \boldsymbol{\Omega}, \beta)$ can be written as:

$$p(\mathbf{d}_{C_A} | \boldsymbol{\Omega}, \beta) = \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \exp\left\{-\frac{\beta}{2}(\mathbf{k} - \mathbf{J}(\boldsymbol{\Omega} - \mathbf{m}_N))^T(\mathbf{k} - \mathbf{J}(\boldsymbol{\Omega} - \mathbf{m}_N))\right\} \quad \text{Equation 3.24}$$

The $N \times 1$ vector \mathbf{k} is defined as:

$$\mathbf{k} = \mathbf{d}_{C_A} - [C_A(t_1, \mathbf{m}_N), C_A(t_2, \mathbf{m}_N), \dots, C_A(t_N, \mathbf{m}_N)]^T \quad \text{Equation 3.25}$$

The current work exploits the conjugacy property (Table 1.1, Section 2.3.3) and defines the parameter prior distribution over the unknown model parameter vector $\boldsymbol{\Omega}$ as a multivariate Gaussian distribution such that:

$$p(\boldsymbol{\Omega} | \beta) = \left(\frac{\beta}{2\pi}\right)^{\frac{M}{2}} |\Lambda_0|^{-\frac{1}{2}} \exp\left\{-\frac{\beta}{2}(\boldsymbol{\Omega} - \mathbf{m}_0)^T \Lambda_0 (\boldsymbol{\Omega} - \mathbf{m}_0)\right\} \quad \text{Equation 3.26}$$

CHAPTER 4: PROPOSED APPROACHES AND METHODOLOGY

Recall that the notation M refers to the number of unknown algebraic model parameters. The symbol $|\cdot|$ refers to evaluating the determinant of the $M \times M$ precision matrix Λ_0 (inverse of covariance matrix). Furthermore, the symbols \mathbf{m}_0 and Λ_0 denote the prior hyperparameters. The conjugate prior for the sensor precision parameter β is the Gamma distribution (Table 1.1, Section 2.3.3). When referring to the work of Chappell et al. (2009), it is worth noting that the authors use an alternative parameterisation for the Gamma distribution given by:

$$p(\beta) = \frac{1}{\Gamma(c_0)} \frac{\beta^{c_0-1}}{s_0^{c_0}} \exp\left\{\frac{-\beta}{s_0}\right\} \quad \text{Equation 3.27}$$

The parameterisation given by Equation 3.27 is related to the parameterisation outlined in Section 2.3.2 (Equation 1.46) by observing that $b = 1/s_0$. The current work adopts the parameterisation outlined in Chappell et al. (2009). The symbols c_0 and s_0 denote the Gamma prior distribution hyperparameters. Given that the likelihood function (Equation 3.24) and the fully conjugate prior (Equations 4.26 and 4.27) are defined, the question that might arise is: “*Why use a variational Bayesian approach to infer the unknown algebraic model parameters?*” Recall from the discussion on variational inference (Section 2.6) that the optimal variational factors $q_j^*(\mathbf{z}_j)$ are obtained by first initialising all the relevant variational factors $q_j(\mathbf{z}_j)$ followed by cycling through each variational factor and, in turn, replacing each factor with the revised estimate using the current estimates for all other variational factors. For each cycle of the variational inference algorithm, the Gaussian posterior distribution mode \mathbf{m}_N (about which the algebraic dynamic model is linearised) can be updated until the ELBO converges. This allows the engineer to recast the nonlinear regression problem within a Bayesian inference framework. In order to derive the CAVI posterior distribution results, recall that the optimal variational solution (assuming that the factorising family of distributions is the mean-field variational family) for each variational factor is given by Equation 1.108. Furthermore, recall that the latent variables correspond to $\mathbf{z} = \{\boldsymbol{\Omega}, \beta\}$.

The optimal variational solution for the approximating posterior distribution over the unknown algebraic model parameters $\boldsymbol{\Omega}$ is given by:

$$\ln q_{\boldsymbol{\Omega}}^*(\boldsymbol{\Omega}) = \mathbb{E}_{q(\beta)}[\ln p(\mathbf{d}_{C_A}, \boldsymbol{\Omega}, \beta)] + \text{constant} \quad \text{Equation 3.28}$$

Note that from here on the subscript notation on $q_{\boldsymbol{\Omega}}^*$ is dropped since no ambiguity arises to which distribution is being used. The joint probability distribution over $\boldsymbol{\Omega}$ and β takes the form $p(\mathbf{d}_{C_A}, \boldsymbol{\Omega}, \beta) = p(\mathbf{d}_{C_A}|\boldsymbol{\Omega}, \beta)p(\boldsymbol{\Omega}|\beta)p(\beta)$ such that the optimal variational solution for the distribution over $\boldsymbol{\Omega}$ is given by:

$$\ln q^*(\boldsymbol{\Omega}) = \mathbb{E}_{\beta}[\ln p(\mathbf{d}_{C_A}, \boldsymbol{\Omega}, \beta)] + \text{constant} \quad \text{Equation 3.29}$$

$$\ln q^*(\boldsymbol{\Omega}) = \mathbb{E}_{\beta}[\ln [p(\mathbf{d}_{C_A}|\boldsymbol{\Omega}, \beta)p(\boldsymbol{\Omega}|\beta)p(\beta)]] + \text{constant} \quad \text{Equation 3.30}$$

With some algebraic manipulation, it can be shown that:

$$\ln q^*(\boldsymbol{\Omega}) = \frac{-\mathbb{E}_{\beta}[\beta]}{2} \left[(\mathbf{k} - \mathbf{J}(\boldsymbol{\Omega} - \mathbf{m}_N))^T (\mathbf{k} - \mathbf{J}(\boldsymbol{\Omega} - \mathbf{m}_N)) + (\boldsymbol{\Omega} - \mathbf{m}_0)^T \Lambda_0 (\boldsymbol{\Omega} - \mathbf{m}_0) \right] + \text{constant} \quad \text{Equation 3.31}$$

CHAPTER 4: PROPOSED APPROACHES AND METHODOLOGY

After completing the multivariate square, one can show that the optimal solution for the factor $q^*(\mathbf{\Omega})$ is a multivariate Gaussian distribution parameterised by:

$$q^*(\mathbf{\Omega}) = \mathcal{N}(\mathbf{\Omega} | \mathbf{m}_N, \mathbf{\Lambda}_N^{-1}) \quad \text{Equation 3.32}$$

$$\mathbf{m}_N = \mathbf{\Lambda}_N^{-1} (s_N c_N [\mathbf{J}^T (\mathbf{k} + \mathbf{J} \mathbf{m}_N) + \mathbf{\Lambda}_0 \mathbf{m}_0]) \quad \text{Equation 3.33}$$

$$\mathbf{\Lambda}_N = s_N c_N (\mathbf{J}^T \mathbf{J} + \mathbf{\Lambda}_0) \quad \text{Equation 3.34}$$

Observe that the parameter posterior update equation for \mathbf{m}_N (Equation 3.33) is a function of itself. Chappell et al. (2009) proposes that during the CAVI iterative cycling procedure, one can initialise \mathbf{m}_N on the right-hand side of Equation 3.33 as $\mathbf{m}_{N,old}$, and, in conjunction with the remaining variational updates, update \mathbf{m}_N on the left-hand side of Equation 3.33 as $\mathbf{m}_{N,new}$. For the next iteration of the CAVI algorithm, $\mathbf{m}_{N,old}$ is then assigned the value of $\mathbf{m}_{N,new}$ from the previous iteration, and so on, until convergence is achieved. With this in mind, the parameter posterior update equation for \mathbf{m}_N (Equation 3.33) can be rewritten as follows:

$$\mathbf{m}_{N,new} = \mathbf{\Lambda}_N^{-1} (s_N c_N [\mathbf{J}^T (\mathbf{k} + \mathbf{J} \mathbf{m}_{N,old}) + \mathbf{\Lambda}_0 \mathbf{m}_0]) \quad \text{Equation 3.35}$$

The optimal variational solution for the approximating posterior distribution over the unknown sensor precision parameter β is given by:

$$\ln q_\beta^*(\beta) = \mathbb{E}_{q(\mathbf{\Omega})} [\ln p(\mathbf{d}_{C_A}, \mathbf{\Omega}, \beta)] + constant \quad \text{Equation 3.36}$$

Note that from here on the subscript notation on q_β^* is dropped since no ambiguity arises to which distribution is being used. Recall that the joint probability distribution over $\mathbf{\Omega}$ and β takes the form $p(\mathbf{d}_{C_A}, \mathbf{\Omega}, \beta) = p(\mathbf{d}_{C_A} | \mathbf{\Omega}, \beta) p(\mathbf{\Omega} | \beta) p(\beta)$ such that the optimal variational solution for the parameter β is given by:

$$\ln q^*(\beta) = \mathbb{E}_\Omega [\ln p(\mathbf{d}_{C_A}, \mathbf{\Omega}, \beta)] + constant \quad \text{Equation 3.37}$$

$$\ln q^*(\beta) = \mathbb{E}_\Omega [\ln [p(\mathbf{d}_{C_A} | \mathbf{\Omega}, \beta) p(\mathbf{\Omega} | \beta) p(\beta)]] + constant \quad \text{Equation 3.38}$$

With some algebraic manipulation, it can be shown that:

$$\begin{aligned} \ln q^*(\beta) &= \left(\frac{N}{2} + \frac{M}{2} + c_0 - 1 \right) \ln \beta + constant \\ &- \beta \left(\frac{1}{s_0} + \frac{1}{2} [\mathbf{k}^T \mathbf{k} + \text{Tr}[\mathbf{\Lambda}_N^{-1} (\mathbf{J}^T \mathbf{J} + \mathbf{\Lambda}_0)] + (\mathbf{m}_N - \mathbf{m}_0)^T \mathbf{\Lambda}_0 (\mathbf{m}_N - \mathbf{m}_0)] \right) \end{aligned} \quad \text{Equation 3.39}$$

One can show that the optimal solution for the factor $q^*(\beta)$ is a Gamma distribution parameterised by:

$$q^*(\beta) = \text{Gam}(\beta | c_N, s_N) \quad \text{Equation 3.40}$$

$$c_N = \frac{N}{2} + \frac{M}{2} + c_0 \quad \text{Equation 3.41}$$

$$\frac{1}{s_N} = \frac{1}{s_0} + \frac{1}{2} [\mathbf{k}^T \mathbf{k} + \text{Tr}[\mathbf{\Lambda}_N^{-1} (\mathbf{J}^T \mathbf{J} + \mathbf{\Lambda}_0)] + (\mathbf{m}_N - \mathbf{m}_0)^T \mathbf{\Lambda}_0 (\mathbf{m}_N - \mathbf{m}_0)] \quad \text{Equation 3.42}$$

CHAPTER 4: PROPOSED APPROACHES AND METHODOLOGY

As discussed in Section 3.3, the benefit of using the variational Bayesian nonlinear regression approach stems from the fact that the ELBO arises as a natural choice for a suitable convergence criterion from the Bayesian framework. However, if one wishes to use the ELBO as a convergence criterion, one has to evaluate the ELBO explicitly. Bishop (2006) outlines that the ELBO for the second Bayesian linear regression probabilistic model can be calculated by evaluating the following integral:

$$\mathcal{L}(\boldsymbol{\Omega}, \beta) = \int_{\boldsymbol{\Omega}} \int_{\beta} q^*(\boldsymbol{\Omega})q^*(\beta) \ln \frac{p(\mathbf{d}_{C_A}|\boldsymbol{\Omega}, \beta)p(\boldsymbol{\Omega}|\beta)p(\beta)}{q^*(\boldsymbol{\Omega})q^*(\beta)} d\boldsymbol{\Omega}d\beta \quad \text{Equation 3.43}$$

Evaluating the ELBO results in the following analytical expression (**Equation 3.44**):

$$\begin{aligned} \mathcal{L}(\boldsymbol{\Omega}, \beta) = & \frac{N}{2}(\psi(c_N) + \ln s_N) - \frac{N}{2} \ln 2\pi - \frac{c_N s_N}{2} [\mathbf{k}^T \mathbf{k} + \text{Tr}(\boldsymbol{\Lambda}_N^{-1} \mathbf{J}^T \mathbf{J})] - \frac{M}{2} \ln 2\pi + \frac{1}{2} \ln |\boldsymbol{\Lambda}_0| \\ & + \frac{M}{2}(\psi(c_N) + \ln s_N) - \frac{c_N s_N}{2} [(\mathbf{m}_N - \mathbf{m}_0)^T \boldsymbol{\Lambda}_0 (\mathbf{m}_N - \mathbf{m}_0) + \text{Tr}(\boldsymbol{\Lambda}_N^{-1} \boldsymbol{\Lambda}_0)] - \ln \Gamma(c_0) \\ & + (c_0 - 1)(\psi(c_N) + \ln s_N) - c_0 \ln s_0 - \frac{c_N s_N}{s_0} + \frac{1}{2} \ln |\boldsymbol{\Lambda}_N^{-1}| + \frac{M}{2} (1 + \ln 2\pi) \\ & + c_N + \ln s_N + \ln \Gamma(c_N) - (c_N - 1)\psi(c_N) \end{aligned}$$

The notation $|\cdot|$ refers to calculating the determinant while $\text{Tr}(\cdot)$ requires evaluating the trace, i.e. the sum of the main diagonal entries of the associated matrix. The symbols $\psi(\cdot)$ and $\Gamma(\cdot)$ refer to the Digamma and Gamma functions, respectively as outlined in Section 2.3.2.

Algorithm 3 outlines the implementation of the resulting variational Bayesian nonlinear regression approach that can be used to estimate the parameters of lumped system algebraic models from noise-corrupted time series data.

Algorithm 3: Variational Bayesian nonlinear regression for lumped system algebraic models.

Algorithm 3: Inputs

- A lumped system algebraic model such as Equation 3.3.
- Sensor measurements of the state variable C_A dynamic response.
- The prior distribution hyperparameters c_0, s_0, \mathbf{m}_0 and $\boldsymbol{\Lambda}_0$.
- Initial values (can be randomly initialised) for the variational posterior distribution hyperparameters $c_N, s_N, \mathbf{m}_{N,old}$ and $\boldsymbol{\Lambda}_N$.

1. Initialise the parameter prior distribution hyperparameters c_0, s_0, \mathbf{m}_0 and $\boldsymbol{\Lambda}_0$. Initialise the evidence lower bound value \mathcal{L}_1 and provide a tolerance level \mathcal{L}_{tol} to establish when the ELBO has converged. The current work sets the value of \mathcal{L}_1 to $-\text{Inf}$ in MATLAB[®] which returns the IEEE arithmetic representation of negative infinity.
2. Initialise the variational posterior distribution hyperparameters $c_N, s_N, \mathbf{m}_{N,old}$ and $\boldsymbol{\Lambda}_N$.
3. Provide a maximum pre-specified number of trial iterations $trial_{maximum}$ (the code implemented in this thesis uses 500 trial iterations). Refer to Chappell et al. (2009) and step 4.6 for a discussion as to why trial iterations are used.
4. For a pre-specified number of iterations $p = 1, 2, 3, \dots$ repeat the following procedure (the code implemented in this thesis uses 2000 iterations):

CHAPTER 4: PROPOSED APPROACHES AND METHODOLOGY

- 4.1. Compute the vector \mathbf{k} (Equation 3.25) and the stacked $N \times M$ Jacobian matrix \mathbf{J} evaluated at $\mathbf{m}_{N,old}$ for the $i = 1, 2, 3, \dots, N$ sensor measurements.
- 4.2. Evaluate the evidence lower bound value \mathcal{L}_{p+1} using Equation 4.44.
- 4.3. If $|\mathcal{L}_{p+1} - \mathcal{L}_p| \leq \mathcal{L}_{tol}$, stop the algorithm and use the prior distribution hyperparameters as the variational posterior distribution hyperparameters. Here the notation $|\cdot|$ refers to evaluating the absolute value.
- 4.4. Otherwise, if $\mathcal{L}_{p+1} > \mathcal{L}_p$, update the Gamma distribution variational posterior parameters with Equation 3.41 and 4.42, respectively, for the N sensor measurements. Update the Gaussian distribution variational approximation precision matrix $\mathbf{\Lambda}_N$ with Equation 3.34.
- 4.5. For a pre-specified number of iterations $u = 1, 2, 3, \dots$ (the code in this thesis uses 50 iterations) update the Gaussian distribution variational approximation mean $\mathbf{m}_{N,new}$ using $\mathbf{m}_{N,old}$ for the N sensor measurements using Equation 3.35. Within this iterative structure, for every iteration u , set the initial model parameter guess (prior mean vector) $\mathbf{m}_0 = \mathbf{m}_{N,new}$. From implementation experience, the author finds that this approach works well since this is equivalent to providing a better initial guess for the unknown parameters during every iteration u . This is important since the first-order Taylor approximation can cause the variational parameter posterior distribution to converge in a parameter space region not supported by the algebraic model.
- 4.6. Due to using a first-order Taylor approximation for the nonlinear algebraic model, the ELBO \mathcal{L} may pass through some maximum value and reverse, i.e. the numerical value of the ELBO can decrease. Chappell et al. (2009) report that if the algorithm is allowed to proceed past the point where the ELBO starts decreasing, further improvement in \mathcal{L} may reoccur. Therefore, Chappell et al. (2009) only stop executing the algorithm if no improvement in the ELBO is observed after allowing an additional number of trial iterations after \mathcal{L} last increased. Thus, for the maximum pre-specified number of trials (step 3), initialise a counter for the number of additional trial iterations used.
- 4.7. If $\mathcal{L}_{p+1} < \mathcal{L}_p$ and $trial_{counter} \leq trial_{maximum}$, save the current variational posterior distribution hyperparameter solutions. Update the Gamma distribution variational approximation with Equation 3.41 and 4.42, respectively, for the N sensor measurements. Update the Gaussian variational approximation precision matrix $\mathbf{\Lambda}_N$ with Equation 3.34. Repeat step 4.5 to update the Gaussian distribution variational approximation mean vector $\mathbf{m}_{N,new}$ using $\mathbf{m}_{N,old}$. Increment the trial counter ($trial_{counter}$) such that the maximum number of trials (step 3) are not exceeded.
- 4.8. Otherwise, if no improvement is observed in the ELBO after the pre-specified number of trial iterations, revert back to the saved variational posterior distribution hyperparameter solutions.
5. Calculate the statistical properties of the inferred parameters (Section 4.8).

Algorithm 3: Outputs

- Variational posterior approximations for the distribution over the unknown algebraic model parameters and the sensor precision parameter, respectively, – i.e. $q^*(\boldsymbol{\Omega})$ and $q^*(\beta)$. Recall that $\beta = 1/\sigma_{noise,CA}^2$ for Equation 3.3.
- Values for the evidence lower bound \mathcal{L} at each iteration which can be used to establish whether the CAVI algorithm has converged.

4.5.2. ODE Parameter Estimation via Gaussian Process Gradient Matching

The Gaussian process gradient matching procedure outlined below draws inspiration from the work of Gorbach, Bauer, and Buhmann (2017), Wenk et al. (2018), Barber and Wang (2014), Macdonald, Higham, and Husmeier (2015), Calderhead, Girolami, and Lawrence (2009) and Dondelinger et al. (2013), as discussed in Section 3.4.

The Gaussian process based methodology developed below only applies to lumped system ODE models that can be written as a linear combination of the model parameters \mathbf{w} (Equations 4.2 and 4.9, Table 3.3). However, note that the Gauss-Newton benchmark presented in *Algorithm 2* (Section 4.4) can also be applied to lumped system ODE models that are nonlinear in the model parameters and is not limited to the imposed linearity restriction used in the development of the Gaussian process based methodology.

For Case Study 1 it was possible to obtain a closed-form solution (Equation 3.3) for the CSTR dynamic response when the engineer restricted the input disturbance in C_{A0} to a perfect step change. However, in a typical engineering setting, the engineer does not have any prior knowledge about the structure or generation process of the exogenous input disturbance. Furthermore, general methods for determining closed-form solutions for lumped system ODE models that are nonlinear in the state variables is not a trivial task, if possible at all. The proposed Gaussian process based methodology relaxes the assumption that the structure of the exogenous input disturbance is known exactly and allows the exogenous input disturbance to take any form (Sections 3.4 and 3.5).

The Joint Distribution Gaussian Process Model

One can define a joint distribution (similar to Equation 1.102, Section 3.5), as per the definition of a Gaussian process (Section 2.5.2), over the state variable C_A , state variable derivative C'_A and exogenous input disturbance C_{A0} function values, as well as the state variable and exogenous input disturbance sensor measurement values such that:

$$p \left(\begin{bmatrix} C'_A \\ C_A \\ C_{A0} \\ d_{C_A} \\ d_{C_{A0}} \end{bmatrix} \right) \sim \mathcal{GP} \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \begin{bmatrix} B_{C'_A C'_A} & B_{C'_A C_A} & B_{C'_A C_{A0}} \\ B_{C_A C'_A} & B_{C_A C_A} & B_{C_A C_{A0}} \\ B_{C_{A0} C'_A} & B_{C_{A0} C_A} & B_{C_{A0} C_{A0}} \end{bmatrix} & \begin{bmatrix} B_{C'_A d_{C_A}} & B_{C'_A d_{C_{A0}}} \\ B_{C_A d_{C_A}} & B_{C_A d_{C_{A0}}} \\ B_{C_{A0} d_{C_A}} & B_{C_{A0} d_{C_{A0}}} \end{bmatrix} \\ \begin{bmatrix} B_{d_{C_A} C'_A} & B_{d_{C_A} C_A} & B_{d_{C_A} C_{A0}} \\ B_{d_{C_{A0}} C'_A} & B_{d_{C_{A0}} C_A} & B_{d_{C_{A0}} C_{A0}} \end{bmatrix} & \begin{bmatrix} B_{d_{C_A} d_{C_A}} & B_{d_{C_A} d_{C_{A0}}} \\ B_{d_{C_{A0}} d_{C_A}} & B_{d_{C_{A0}} d_{C_{A0}}} \end{bmatrix} \end{bmatrix} \right) \quad \text{Equation 3.45}$$

Following the discussion outlined in Gorbach, Bauer and Buhmann (2017), one can assume that the state variable C_A , state variable derivative C'_A and the exogenous input disturbance C_{A0} function values are statistically independent – i.e. their covariance matrices reduce to zero (Section 2.1.4) – such that the joint distribution given by Equation 3.45 simplifies to:

$$p \left(\begin{bmatrix} \mathbf{C}'_A \\ \mathbf{C}_A \\ \mathbf{C}_{A0} \\ \mathbf{d}_{C_A} \\ \mathbf{d}_{C_{A0}} \end{bmatrix} \right) \sim \mathcal{GP} \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{B}_{C'_A C'_A} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_{C_A C_A} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{B}_{C_{A0} C_{A0}} \\ \mathbf{B}_{d_{C_A} C'_A} & \mathbf{B}_{d_{C_A} C_A} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{B}_{d_{C_{A0}} C_{A0}} \end{bmatrix} \begin{bmatrix} \mathbf{B}_{C'_A d_{C_A}} & \mathbf{0} \\ \mathbf{B}_{C_A d_{C_A}} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_{C_{A0} d_{C_{A0}}} \\ \mathbf{B}_{d_{C_A} d_{C_A}} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_{d_{C_{A0}} d_{C_{A0}}} \end{bmatrix} \right) \quad \text{Equation 3.46}$$

Kernel Function Selection

The entries of the covariance matrices pertaining to the state variable function values \mathbf{C}_A , as indicated by the subscript notation associated with each matrix \mathbf{B} , is determined in an element-wise manner using the following kernel function:

$$k_{C_A}(t_i, t_m) = \sigma_{f_{1,C_A}}^2 \exp \left\{ \frac{-1}{2\ell_{1,C_A}^2} (t_i - t_m)^2 \right\} + \sigma_{f_{2,C_A}}^2 \exp \left\{ \frac{-1}{2\ell_{2,C_A}^2} (t_i - t_m)^2 \right\} + \sigma_{noise,C_A}^2 \delta_{im} \quad \text{Equation 3.47}$$

The kernel function given by Equation 3.47 is the sum of two exponentiated quadratic kernels (Equation 1.99) plus the addition of noise to reflect the sensor measurements. The primary argument behind using two exponentiated quadratic kernels is to reflect the notion of global and local smooth variations in the underlying state variable. The first exponentiated quadratic kernel models global smooth variations in the state variable, i.e. over long periods of time one expects global smooth behaviour in the state variable. However, since no assumption is made about the input disturbance structure, the input disturbance can cause the state variable to vary on a local scale, i.e. the state variable may vary over short periods of time. Thus, a second kernel function is introduced to model the local smooth variations in the state variable over short periods of time (Duvenaud, 2014). The symbol σ_{noise,C_A}^2 denotes the noise variance parameter associated with the sensor measuring the state variable C_A . Similarly, the covariance matrix entries pertaining to the exogenous input disturbance function values \mathbf{C}_{A0} , as indicated by the subscript notation associated with each matrix \mathbf{B} , is determined in an element-wise manner using the following kernel function:

$$k_{C_{A0}}(t_i, t_m) = \sigma_{f_{1,C_{A0}}}^2 \exp \left\{ \frac{-1}{2\ell_{1,C_{A0}}^2} (t_i - t_m)^2 \right\} + \sigma_{f_{2,C_{A0}}}^2 \exp \left\{ \frac{-1}{2\ell_{2,C_{A0}}^2} (t_i - t_m)^2 \right\} + \sigma_{noise,C_{A0}}^2 \delta_{im} \quad \text{Equation 3.48}$$

The symbol $\sigma_{noise,C_{A0}}^2$ denotes the noise variance parameter associated with the sensor measuring the exogenous input disturbance C_{A0} . Note that the kernel function given by Equation 3.48 again encodes the same behaviour as Equation 3.47. Although the specified kernel function does impose an expected structure on the exogenous input, it is not as restrictive as assuming a perfect step (Case Study 1). Furthermore, if the engineer suspects that the exogenous input will not exhibit global and local smooth behaviour, the expected behaviour can simply be encoded by considering a different kernel function. Table 3.4 summarises the Gaussian process kernel function hyperparameters used to encode the expected underlying behaviour of the state variable and exogenous input disturbance functions.

CHAPTER 4: PROPOSED APPROACHES AND METHODOLOGY

Table 3.4: Gaussian process kernel function hyperparameters required to construct the joint distribution given by Equation 4.46.

| Kernel Function | Global Function Variation | Local Function Variation |
|-------------------------|---------------------------|--------------------------|
| $k_{C_A}(t_i, t_m)$ | σ_{f_1, C_A} | σ_{f_2, C_A} |
| | ℓ_{1, C_A} | ℓ_{2, C_A} |
| $k_{C_{A_0}}(t_i, t_m)$ | $\sigma_{f_1, C_{A_0}}$ | $\sigma_{f_2, C_{A_0}}$ |
| | $\ell_{1, C_{A_0}}$ | $\ell_{2, C_{A_0}}$ |

In addition to the hyperparameters outlined in Table 3.4, the noise variance parameters associated with the sensor measurements for the state variable and the exogenous input disturbance are also unknown.

Optimising the Gaussian Process Kernel Hyperparameters

Within a practical setting, the variance parameters as well as the hyperparameters associated with the kernel functions given by Equations 4.47 and 4.48 must be learned from the state variable and exogenous input disturbance sensor measurement data. Techniques for learning the hyperparameters are based on maximising the marginal likelihood (evidence) using any form of gradient-based optimisation (Section 2.5.2). This is also known as a type 2 maximum likelihood procedure (Bishop, 2006). The log marginal likelihood for the joint distribution given by Equation 3.46 can be expressed as follows:

$$\ln p(\mathbf{d}_{C_A}, \mathbf{d}_{C_{A_0}} | \boldsymbol{\psi}_{GP}) = -\frac{1}{2} \begin{bmatrix} \mathbf{d}_{C_A} \\ \mathbf{d}_{C_{A_0}} \end{bmatrix}^T \mathbf{B}_{GP}^{-1} \begin{bmatrix} \mathbf{d}_{C_A} \\ \mathbf{d}_{C_{A_0}} \end{bmatrix} - \frac{1}{2} \ln |\mathbf{B}_{GP}| - N \ln 2\pi \quad \text{Equation 3.49}$$

The matrix \mathbf{B}_{GP} is defined such that:

$$\mathbf{B}_{GP} = \begin{bmatrix} \mathbf{B}_{d_{C_A} d_{C_A}} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_{d_{C_{A_0}} d_{C_{A_0}}} \end{bmatrix} \quad \text{Equation 3.50}$$

The notation $\boldsymbol{\psi}_{GP}$ (as in Section 2.5.2) collectively refers to the kernel hyperparameters and the sensor variance parameters required for Equations 4.47 and 4.48. The notation $|\cdot|$ refers to evaluating the determinant of \mathbf{B}_{GP} . Gradient-based optimisation is possible for the log marginal likelihood (Equation 3.49) since the engineer can obtain analytical derivatives for $\ln p(\mathbf{d}_{C_A}, \mathbf{d}_{C_{A_0}} | \boldsymbol{\psi}_{GP})$ with respect to each parameter ψ_i such that:

$$\frac{\partial \ln p(\mathbf{d}_{C_A}, \mathbf{d}_{C_{A_0}} | \boldsymbol{\psi}_{GP})}{\partial \psi_i} = \frac{1}{2} \begin{bmatrix} \mathbf{d}_{C_A} \\ \mathbf{d}_{C_{A_0}} \end{bmatrix}^T \mathbf{B}_{GP}^{-1} \frac{\partial \mathbf{B}_{GP}}{\partial \psi_i} \mathbf{B}_{GP}^{-1} \begin{bmatrix} \mathbf{d}_{C_A} \\ \mathbf{d}_{C_{A_0}} \end{bmatrix} - \frac{1}{2} \text{Tr} \left(\mathbf{B}_{GP}^{-1} \frac{\partial \mathbf{B}_{GP}}{\partial \psi_i} \right) \quad \text{Equation 3.51}$$

The form of the derivative $\partial \mathbf{B}_{GP} / \partial \psi_i$ depends on the form of the kernel functions used as well as which parameter the derivative is taken with respect to. For the specific choice of kernel functions given by Equations 4.47 and 4.48, all the kernel hyperparameters as well as the sensor variance parameters are positive, i.e. $\psi_i > 0$. One can constrain the parameter search space to ensure $\psi_i > 0$ by defining $\psi_{c,i} = \ln(\psi_i)$, and simply use the chain rule to obtain the appropriate derivatives required for Equation 3.51.

CHAPTER 4: PROPOSED APPROACHES AND METHODOLOGY

The current work uses gradient ascent in conjunction with Equations 4.49 and 4.51 (with the constraint $\psi_i > 0$) to find point estimates for the kernel hyperparameters and sensor variance parameters required for Equations 4.47 and 4.48.

Conditioning to Obtain the Gaussian Process Posterior Distributions

Since the Gaussian process has a consistency requirement that is automatically fulfilled if the kernel functions specify the entries of the covariance matrix, the marginal distribution over a subset of the variables is another Gaussian process (Sections 2.3.1 and 2.5.2). In other words,

$$p(\mathbf{C}'_A, \mathbf{d}_{C_A}, \mathbf{d}_{C_{A0}}) = \int_{\text{dom}(\mathbf{C}_A)} \int_{\text{dom}(\mathbf{C}_{A0})} p(\mathbf{C}'_A, \mathbf{C}_A, \mathbf{C}_{A0}, \mathbf{d}_{C_A}, \mathbf{d}_{C_{A0}}) d\mathbf{C}_{A0} d\mathbf{C}_A \quad \text{Equation 3.52}$$

Equation 3.52 can equivalently be expressed as:

$$p\left(\begin{bmatrix} \mathbf{C}'_A \\ \mathbf{d}_{C_A} \\ \mathbf{d}_{C_{A0}} \end{bmatrix}\right) \sim \mathcal{GP}\left(\begin{bmatrix} [\mathbf{0}] \\ [\mathbf{0}] \\ [\mathbf{0}] \end{bmatrix}, \begin{bmatrix} [\mathbf{B}_{C'_A C'_A}] & [\mathbf{B}_{C'_A d_{C_A}}] & [\mathbf{B}_{C'_A d_{C_{A0}}}] \\ [\mathbf{B}_{d_{C_A} C'_A}] & [\mathbf{B}_{d_{C_A} d_{C_A}}] & [\mathbf{B}_{d_{C_A} d_{C_{A0}}}] \\ [\mathbf{0}] & [\mathbf{0}] & [\mathbf{B}_{d_{C_{A0}} d_{C_{A0}}}] \end{bmatrix}\right) \quad \text{Equation 3.53}$$

Using the standard results for conditioning on a Gaussian distribution (Section 2.3.1), one can obtain a posterior distribution over the state variable derivative function values \mathbf{C}'_A which is another Gaussian process such that:

$$p(\mathbf{C}'_A | \mathbf{d}_{C_A}, \mathbf{d}_{C_{A0}}) \sim \mathcal{GP}\left(\boldsymbol{\mu}_{C'_A | d_{C_A}, d_{C_{A0}}}, \boldsymbol{\Sigma}_{C'_A | d_{C_A}, d_{C_{A0}}}\right) \quad \text{Equation 3.54}$$

The mean vector $\boldsymbol{\mu}_{C'_A | d_{C_A}, d_{C_{A0}}}$ and covariance matrix $\boldsymbol{\Sigma}_{C'_A | d_{C_A}, d_{C_{A0}}}$ are defined such that:

$$\boldsymbol{\mu}_{C'_A | d_{C_A}, d_{C_{A0}}} = [\mathbf{B}_{C'_A d_{C_A}} \quad \mathbf{0}] \begin{bmatrix} \mathbf{B}_{d_{C_A} d_{C_A}} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_{d_{C_{A0}} d_{C_{A0}}} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{d}_{C_A} \\ \mathbf{d}_{C_{A0}} \end{bmatrix} \quad \text{Equation 3.55}$$

$$\boldsymbol{\Sigma}_{C'_A | d_{C_A}, d_{C_{A0}}} = \mathbf{B}_{C'_A C'_A} - [\mathbf{B}_{C'_A d_{C_A}} \quad \mathbf{0}] \begin{bmatrix} \mathbf{B}_{d_{C_A} d_{C_A}} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_{d_{C_{A0}} d_{C_{A0}}} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{B}_{d_{C_A} C'_A} \\ \mathbf{0} \end{bmatrix} \quad \text{Equation 3.56}$$

The problem that arises is the specification of the kernel function encoding the underlying behaviour of the state variable derivative. Papoulis and Pillai (2002), Gorbach, Bauer, and Buhmann (2017), and Wenk et al. (2018) state that any finite set of function values of the state variable is jointly Gaussian distributed with the finite set of function values of the state variable derivative such that one can define a joint Gaussian process over \mathbf{C}'_A and \mathbf{C}_A as follows:

$$p\left(\begin{bmatrix} \mathbf{C}_A \\ \mathbf{C}'_A \end{bmatrix}\right) \sim \mathcal{GP}\left(\begin{bmatrix} [\mathbf{0}] \\ [\mathbf{0}] \end{bmatrix}, \begin{bmatrix} \mathbf{B}_{C_A C_A} & \mathbf{B}_{C_A C'_A} \\ \mathbf{B}_{C'_A C_A} & \mathbf{B}_{C'_A C'_A} \end{bmatrix}\right) \quad \text{Equation 3.57}$$

Based on the discussion outlined in Gorbach, Bauer, and Buhmann (2017) the matrix $\mathbf{B}_{C'_A C_A}$ is populated in an element-wise manner using the kernel function obtained from evaluating:

$$\frac{d}{da} [k_{C_A}(a, t_m)] \Big|_{a=t_i} \quad \text{Equation 3.58}$$

CHAPTER 4: PROPOSED APPROACHES AND METHODOLOGY

Similarly, the matrix $\mathbf{B}_{C_A C'_A}$ is obtained by noting that $\mathbf{B}_{C_A C'_A} = [\mathbf{B}_{C'_A C_A}]^T$. Lastly, one can populate the matrix $\mathbf{B}_{C'_A C'_A}$ in an element-wise manner by using the kernel function obtained from evaluating:

$$\frac{\partial}{\partial a} \frac{\partial}{\partial b} [k_{C_A}(a, b)] \Big|_{a=t_i, b=t_m} \quad \text{Equation 3.59}$$

One can obtain a Gaussian process posterior distribution describing the exogenous input disturbance by noting that:

$$p \left(\begin{bmatrix} \mathbf{C}_{A0} \\ \mathbf{d}_{C_A} \\ \mathbf{d}_{C_{A0}} \end{bmatrix} \right) \sim \mathcal{GP} \left(\begin{bmatrix} [\mathbf{0}] \\ [\mathbf{0}] \\ [\mathbf{0}] \end{bmatrix}, \begin{bmatrix} [\mathbf{B}_{C_{A0} C_{A0}}] & [\mathbf{0} \ \mathbf{B}_{C_{A0} d_{C_{A0}}}] \\ [\mathbf{0}] & [\mathbf{B}_{d_{C_A} d_{C_A}} \ \mathbf{0}] \\ [\mathbf{B}_{d_{C_{A0}} C_{A0}}] & [\mathbf{0} \ \mathbf{B}_{d_{C_{A0}} d_{C_{A0}}}] \end{bmatrix} \right) \quad \text{Equation 3.60}$$

$$p(\mathbf{C}_{A0} | \mathbf{d}_{C_A}, \mathbf{d}_{C_{A0}}) \sim \mathcal{GP}(\boldsymbol{\mu}_{C_{A0} | d_{C_A}, d_{C_{A0}}}, \boldsymbol{\Sigma}_{C_{A0} | d_{C_A}, d_{C_{A0}}}) \quad \text{Equation 3.61}$$

The mean vector $\boldsymbol{\mu}_{C_{A0} | d_{C_A}, d_{C_{A0}}}$ and covariance matrix $\boldsymbol{\Sigma}_{C_{A0} | d_{C_A}, d_{C_{A0}}}$ are defined such that:

$$\boldsymbol{\mu}_{C_{A0} | d_{C_A}, d_{C_{A0}}} = [\mathbf{0} \ \mathbf{B}_{C_{A0} d_{C_{A0}}}] \begin{bmatrix} \mathbf{B}_{d_{C_A} d_{C_A}} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_{d_{C_{A0}} d_{C_{A0}}} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{d}_{C_A} \\ \mathbf{d}_{C_{A0}} \end{bmatrix} \quad \text{Equation 3.62}$$

$$\boldsymbol{\Sigma}_{C_{A0} | d_{C_A}, d_{C_{A0}}} = \mathbf{B}_{C_{A0} C_{A0}} - [\mathbf{0} \ \mathbf{B}_{C_{A0} d_{C_{A0}}}] \begin{bmatrix} \mathbf{B}_{d_{C_A} d_{C_A}} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_{d_{C_{A0}} d_{C_{A0}}} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{0} \\ \mathbf{B}_{d_{C_{A0}} C_{A0}} \end{bmatrix} \quad \text{Equation 3.63}$$

Lastly, one can obtain a Gaussian process posterior distribution describing the state variable C_A by noting that:

$$p \left(\begin{bmatrix} \mathbf{C}_A \\ \mathbf{d}_{C_A} \\ \mathbf{d}_{C_{A0}} \end{bmatrix} \right) \sim \mathcal{GP} \left(\begin{bmatrix} [\mathbf{0}] \\ [\mathbf{0}] \\ [\mathbf{0}] \end{bmatrix}, \begin{bmatrix} [\mathbf{B}_{C_A C_A}] & [\mathbf{B}_{C_A d_{C_A}} \ \mathbf{0}] \\ [\mathbf{B}_{d_{C_A} C_A}] & [\mathbf{B}_{d_{C_A} d_{C_A}} \ \mathbf{0}] \\ [\mathbf{0}] & [\mathbf{0} \ \mathbf{B}_{d_{C_{A0}} d_{C_{A0}}}] \end{bmatrix} \right) \quad \text{Equation 3.64}$$

$$p(\mathbf{C}_A | \mathbf{d}_{C_A}, \mathbf{d}_{C_{A0}}) \sim \mathcal{GP}(\boldsymbol{\mu}_{C_A | d_{C_A}, d_{C_{A0}}}, \boldsymbol{\Sigma}_{C_A | d_{C_A}, d_{C_{A0}}}) \quad \text{Equation 3.65}$$

The mean vector $\boldsymbol{\mu}_{C_A | d_{C_A}, d_{C_{A0}}}$ and covariance matrix $\boldsymbol{\Sigma}_{C_A | d_{C_A}, d_{C_{A0}}}$ are defined such that:

$$\boldsymbol{\mu}_{C_A | d_{C_A}, d_{C_{A0}}} = [\mathbf{B}_{C_A d_{C_A}} \ \mathbf{0}] \begin{bmatrix} \mathbf{B}_{d_{C_A} d_{C_A}} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_{d_{C_{A0}} d_{C_{A0}}} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{d}_{C_A} \\ \mathbf{d}_{C_{A0}} \end{bmatrix} \quad \text{Equation 3.66}$$

$$\boldsymbol{\Sigma}_{C_A | d_{C_A}, d_{C_{A0}}} = \mathbf{B}_{C_A C_A} - [\mathbf{B}_{C_A d_{C_A}} \ \mathbf{0}] \begin{bmatrix} \mathbf{B}_{d_{C_A} d_{C_A}} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_{d_{C_{A0}} d_{C_{A0}}} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{B}_{d_{C_A} C_A} \\ \mathbf{0} \end{bmatrix} \quad \text{Equation 3.67}$$

Recall that Equations 4.54 through 4.56 provide a Gaussian process posterior distribution over the state variable derivative function values \mathbf{C}'_A . From an engineering perspective, it is not at all obvious how to interpret this Gaussian process posterior distribution. One possibly helpful analogy is to think of placing a hypothetical sensor at the CSTR outlet that is capable of measuring the state variable derivative. Refer to Figure 4.3.

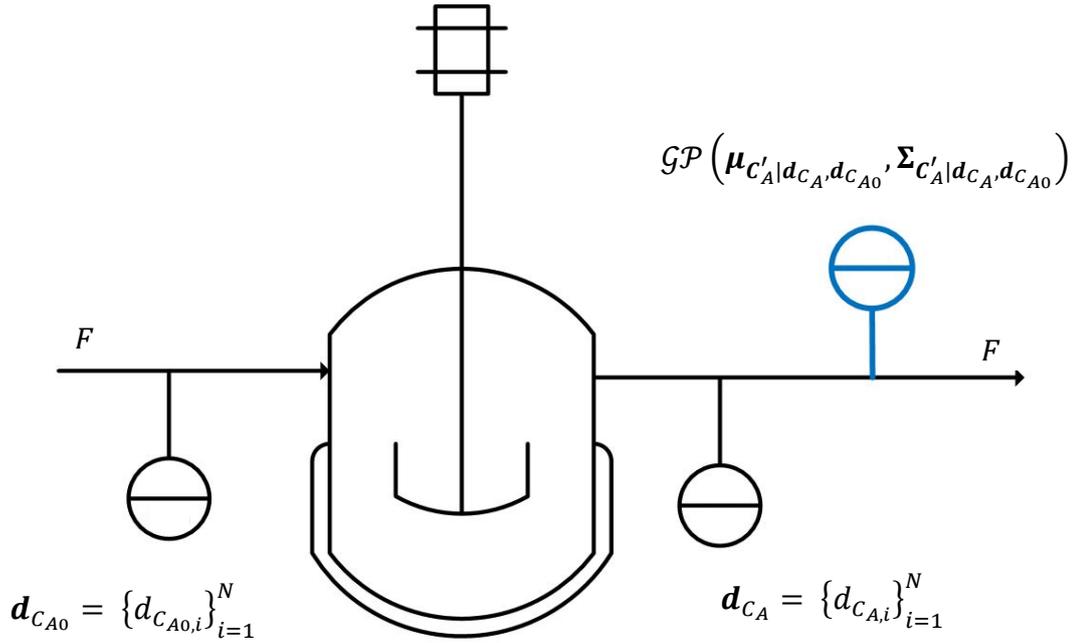


Figure 4.3: Visual interpretation of the hypothetical Gaussian process state variable derivative “sensor” (blue sensor).

Since the Gaussian process posterior is a probability distribution over the function dC_A/dt such that the set of values of the function dC_A/dt evaluated at the selected input time points is jointly Gaussian distributed, samples from this Gaussian process posterior distribution correspond to possible data set manifestations. In other words, if one thinks of the Gaussian process posterior distribution as a hypothetical sensor, then each sample from the Gaussian process corresponds to a possible data set that the hypothetical sensor could have measured.

Implementing the Regression Model

If the engineer uses the data set that maximises the Gaussian process state variable derivative posterior distribution given by Equation 3.54, i.e. the MAP estimate, then:

$$\boldsymbol{\mu}_{c'_A|d_{C_A}, d_{C_{A0}}} = \underset{c'_A}{\operatorname{argmax}} p(c'_A | d_{C_A}, d_{C_{A0}}) \quad \text{Equation 3.68}$$

Since Equation 3.2 is linear in the model parameters \mathbf{w} , it can be rewritten such that (Section 2.4.1):

$$\frac{dC_A}{dt} = \begin{bmatrix} F \\ \frac{F}{V} \quad k \end{bmatrix} \begin{bmatrix} C_{A0} - C_A \\ -C_A \end{bmatrix} \quad \text{Equation 3.69}$$

$$\frac{dC_A}{dt} = \mathbf{w}^T \boldsymbol{\phi}(t) \quad \text{Equation 3.70}$$

However, note that the engineer only has noise-corrupted sensor measurements for the basis functions $\boldsymbol{\phi}(t)$. Smoothed out estimates for the basis functions can be obtained by considering

CHAPTER 4: PROPOSED APPROACHES AND METHODOLOGY

the MAP estimates for the state variable (Equation 3.65) and exogenous input disturbance (Equation 3.61) Gaussian process posterior distributions such that.

$$\boldsymbol{\mu}_{C_A|d_{C_A}, d_{C_{A0}}} = \underset{C_A}{\operatorname{argmax}} p(C_A | d_{C_A}, d_{C_{A0}}) \quad \text{Equation 3.71}$$

$$\boldsymbol{\mu}_{C_{A0}|d_{C_A}, d_{C_{A0}}} = \underset{C_{A0}}{\operatorname{argmax}} p(C_{A0} | d_{C_A}, d_{C_{A0}}) \quad \text{Equation 3.72}$$

Equations 4.68 through 4.72 make it possible for the engineer to exploit a methodology similar to the first Bayesian linear regression probabilistic model discussed in Section 2.4.1 to estimate the parameters \boldsymbol{w} of the ODE model given by Equation 3.2. From Bayes' rule (Section 2.4), the posterior distribution over the ODE parameters \boldsymbol{w} , given the MAP estimate (Equation 3.68) for the hypothetical state variable derivative sensor measurements, can be expressed as:

$$p(\boldsymbol{w} | \boldsymbol{\mu}_{C'_A|d_{C_A}, d_{C_{A0}}}) = \frac{p(\boldsymbol{\mu}_{C'_A|d_{C_A}, d_{C_{A0}}} | \boldsymbol{w}) p(\boldsymbol{w})}{p(\boldsymbol{\mu}_{C'_A|d_{C_A}, d_{C_{A0}}})} \quad \text{Equation 3.73}$$

The problem that arises is constructing the likelihood function $p(\boldsymbol{\mu}_{C'_A|d_{C_A}, d_{C_{A0}}} | \boldsymbol{w})$ required by Equation 3.73. From Section 2.4.1 the probability of observing a target variable d at an arbitrary input x is given by Equation 1.55 which is centered at the function predicted value $y(x, \boldsymbol{w})$ with variance $\sigma_{noise}^2 = 1/\beta$ that arises from assuming zero-mean i.i.d. sensor measurements. The data generation process underlying Equation 1.55 assumes that the target variable d is generated from the true underlying function $y(x, \boldsymbol{w})$, however, due to imperfect measurement equipment, the engineer only observes a noise-corrupted version of the true underlying function value at input x .

However, if one follows the Gaussian process route discussed above whereby data, i.e. measurements from the hypothetical sensor (Figure 4.3), for the state variable derivative are obtained from the Gaussian process posterior distribution (Equation 4.54 and 4.68), then the data generation process for target variable d does not follow that of Equation 1.55. Consider some arbitrary input time point t_i . From the consistency requirement for Gaussian processes (Section 2.5.2), the marginal distribution over the state variable derivative function value at time point t_i is given by:

$$p((C'_A)_i | t_i) = \mathcal{N}\left((C'_A)_i | \mu_{(C'_A|d_{C_A}, d_{C_{A0}})_i}, \sigma_{(C'_A|d_{C_A}, d_{C_{A0}})_{ii}}^2\right) \quad \text{Equation 3.74}$$

The notation $\mu_{(C'_A|d_{C_A}, d_{C_{A0}})_i}$ and $\sigma_{(C'_A|d_{C_A}, d_{C_{A0}})_{ii}}^2$ corresponds to extracting the relevant entries for t_i from the $N \times 1$ column vector given by Equation 3.55 and the $N \times N$ covariance matrix given by Equation 3.56 (Section 2.3.1).

Suppose that, based on the data generating process for Equation 1.55 (Section 2.4.1), the target variable data generation process associated with Equation 3.70 is given by the following:

$$d_{\left(\frac{dC_A}{dt}\right)_i} = \left(\frac{dC_A}{dt}\right)_i + \epsilon_i \quad \text{Equation 3.75}$$

CHAPTER 4: PROPOSED APPROACHES AND METHODOLOGY

$$d\left(\frac{dC_A}{dt}\right)_i = \mathbf{w}^T \boldsymbol{\phi}(t_i) + \epsilon_i \quad \text{Equation 3.76}$$

Here it is assumed that the target variable $d_{(dC_A/dt)_i}$ for time point t_i is generated from the underlying differential equation describing the state variable derivative and that each target variable $d_{(dC_A/dt)_i}$ is independently corrupted by zero-mean Gaussian noise with variance parameter $\sigma_{t_i}^2$. The probability of observing target variable $d_{(dC_A/dt)_i}$ at time point t_i can therefore be expressed as follows:

$$p\left(d_{(dC_A/dt)_i} | t_i\right) = \mathcal{N}\left(d_{(dC_A/dt)_i} | \mathbf{w}^T \boldsymbol{\phi}(t_i), \sigma_{t_i}^2\right) \quad \text{Equation 3.77}$$

However, the problem that arises is that the variance parameter $\sigma_{t_i}^2$ and the target variable $d_{(dC_A/dt)_i}$ are not known and cannot be measured directly from the physical system. One way to circumvent this problem is to exploit the Gaussian process posterior distribution over the state variable derivative function values. Recall that the Gaussian posterior distribution over the state variable derivative function values acts as a hypothetical sensor that provides the engineer with a way to generate possible data sets for the state variable derivative based on sensor measurements of the state variable and exogenous input disturbance (Figure 4.3). If the engineer assumes that the Gaussian posterior distribution interpolating the state variable derivative function value at time point t_i is approximating the target variable distribution generating $d_{(dC_A/dt)_i}$, then Equation 3.77 can be rewritten such that:

$$p_{GP}((C'_A)_i | t_i) = \mathcal{N}((C'_A)_i | \mathbf{w}^T \boldsymbol{\phi}(t_i), \sigma_{t_i}^2) \quad \text{Equation 3.78}$$

In order to determine a suitable value for the variance parameter $\sigma_{t_i}^2$, the current work provides the following information theory based argument. In particular, consider the Kullback-Leibler divergence between the distributions given by Equation 4.74 and 4.78 (Section 2.2). In other words,

$$\mathcal{KL}(p_{GP}((C'_A)_i | t_i) || p((C'_A)_i | t_i)) = - \int_{-\infty}^{\infty} p_{GP}((C'_A)_i | t_i) \ln \frac{p((C'_A)_i | t_i)}{p_{GP}((C'_A)_i | t_i)} dC'_A \quad \text{Equation 3.79}$$

By evaluating Equation 3.79, one obtains the Kullback-Leibler divergence between the two univariate Gaussian distributions such that:

$$\begin{aligned} \mathcal{KL}(p_{GP}((C'_A)_i | t_i) || p((C'_A)_i | t_i)) &= \frac{1}{2} \ln \left(\frac{\sigma_{(C'_A | d_{C_A}, d_{C_{A0}})_{ii}}^2}{\sigma_{t_i}^2} \right) \\ &+ \frac{\sigma_{t_i}^2}{2\sigma_{(C'_A | d_{C_A}, d_{C_{A0}})_{ii}}^2} + \frac{\left(\mathbf{w}^T \boldsymbol{\phi}(t_i) - \mu_{(C'_A | d_{C_A}, d_{C_{A0}})_i} \right)^2}{2\sigma_{(C'_A | d_{C_A}, d_{C_{A0}})_{ii}}^2} - \frac{1}{2} \end{aligned} \quad \text{Equation 3.80}$$

If the goal is to minimise the Kullback-Leibler divergence between the two univariate Gaussian distributions $p_{GP}((C'_A)_i | t_i)$ and $p((C'_A)_i | t_i)$, then this can be achieved by first setting the value of $\sigma_{t_i}^2 = \sigma_{(C'_A | d_{C_A}, d_{C_{A0}})_{ii}}^2$. In other words, the variance parameter associated with ϵ_i that forms

CHAPTER 4: PROPOSED APPROACHES AND METHODOLOGY

part of the assumed generation process of target variable $d_{(dc_A/dt)_i}$ is set equal to the variance of the marginal distribution over the state variable derivative function value (Equation 3.74). Given that $\sigma_{t_i}^2 = \sigma_{(C'_A|d_{C_A}, d_{C_{A0}})_{ii}}^2$, the Kullback-Leibler divergence reduces to:

$$\mathcal{KL}(p((C'_A)_i|t_i)||p_{GP}((C'_A)_i|t_i)) = \frac{\left(\mathbf{w}^T \boldsymbol{\phi}(t_i) - \mu_{(C'_A|d_{C_A}, d_{C_{A0}})_i}\right)^2}{2\sigma_{(C'_A|d_{C_A}, d_{C_{A0}})_{ii}}^2} \quad \text{Equation 3.81}$$

From Equation 3.81, given that the engineer takes $\sigma_{t_i}^2 = \sigma_{(C'_A|d_{C_A}, d_{C_{A0}})_{ii}}^2$, one observes that the Kullback-Leibler divergence between $p_{GP}((C'_A)_i|t_i)$ and $p((C'_A)_i|t_i)$ would be equal to zero, i.e. $p((C'_A)_i|t_i) = p_{GP}((C'_A)_i|t_i)$, if the mean of the marginal distribution interpolating the state variable derivative function value at time point t_i is identical to the true underlying ODE state variable predicted value, i.e. the mean of Equation 3.78. Equation 3.78 can be rewritten such that:

$$p_{GP}((C'_A)_i|t_i) = \mathcal{N}\left((C'_A)_i | \mathbf{w}^T \boldsymbol{\phi}(t_i), \sigma_{(C'_A|d_{C_A}, d_{C_{A0}})_{ii}}^2\right) \quad \text{Equation 3.82}$$

Due to sensor measurement noise associated with the state variable and exogenous input disturbance, there will be a discrepancy between $p_{GP}((C'_A)_i|t_i)$ and $p((C'_A)_i|t_i)$ given by:

$$\left(\mathbf{w}^T \boldsymbol{\phi}(t_i) - \mu_{(C'_A|d_{C_A}, d_{C_{A0}})_i}\right)^2 \quad \text{Equation 3.83}$$

Based on Equation 3.83 one is matching the gradient interpolated by the Gaussian process at time point t_i to the value predicted by the ODE. Note the similarity of the discrepancy to the simple least squares objective function given by Equation 1.66 (Section 2.4.2) for the case where $i = 1$, which can be derived from the first Bayesian linear regression probabilistic model outlined in Section 2.4.1. One observes that if the mean of the Gaussian posterior distribution over the state variable derivative function value at t_i given by Equation 3.74, which is also the MAP estimate, is interpreted as a data point measured by the hypothetical sensor, then the ODE parameters can be estimated using the first Bayesian linear regression probabilistic model. This serves as motivation as to why the author uses the MAP estimate given by Equation 4.68 which is used to formulate Bayes' rule (Equation 3.73) for inferring \mathbf{w} . If the engineer assumes that $(C'_A)_i$ is drawn independently from the probability distribution given by Equation 3.82 for $i = 1, 2, \dots, N$ hypothetical sensor measurements, then one can construct the likelihood function (Equation 4.84) as follows:

$$p\left(\boldsymbol{\mu}_{C'_A|d_{C_A}, d_{C_{A0}}} | \mathbf{w}\right) = \frac{1}{(2\pi)^{\frac{N}{2}}} \frac{1}{|\mathbf{R}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{\mu}_{C'_A|d_{C_A}, d_{C_{A0}}} - \boldsymbol{\Phi}\mathbf{w}\right)^T \mathbf{R}^{-1}\left(\boldsymbol{\mu}_{C'_A|d_{C_A}, d_{C_{A0}}} - \boldsymbol{\Phi}\mathbf{w}\right)\right\}$$

The notation $\boldsymbol{\mu}_{C'_A|d_{C_A}, d_{C_{A0}}}$ refers to the MAP estimate for the Gaussian process posterior distribution over the state variable derivative function values (Equation 3.68) and $\boldsymbol{\Phi}$ is the design matrix (Section 2.4.1) that is constructed with the MAP estimates obtained from Equations 4.71 and 4.72. The covariance matrix \mathbf{R} , due to drawing independent samples from

CHAPTER 4: PROPOSED APPROACHES AND METHODOLOGY

Equation 3.82 for $i = 1, 2, \dots, N$ hypothetical sensor measurements, is obtained from Equation 3.56 as follows:

$$\mathbf{R} = \text{diag} \left(\boldsymbol{\Sigma}_{C'_A | d_{C_A}, d_{C_{A0}}} \right) \quad \text{Equation 3.84}$$

By exploiting conjugacy (Section 2.3.3, Table 1.1), one can define a parameter prior probability distribution such that:

$$p(\mathbf{w}) = \frac{1}{(2\pi)^{\frac{M}{2}}} \frac{1}{|\mathbf{S}_0|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (\mathbf{w} - \mathbf{m}_0)^T \mathbf{S}_0^{-1} (\mathbf{w} - \mathbf{m}_0) \right\} \quad \text{Equation 3.85}$$

One can show that the parameter posterior probability distribution $p(\mathbf{w} | \boldsymbol{\mu}_{C'_A | d_{C_A}, d_{C_{A0}}})$, after completing the multivariate square, takes the form of a multivariate Gaussian distribution parameterised by \mathbf{m}_N and \mathbf{S}_N such that:

$$p(\mathbf{w} | \boldsymbol{\mu}_{C'_A | d_{C_A}, d_{C_{A0}}}) = \mathcal{N}(\mathbf{w} | \mathbf{m}_N, \mathbf{S}_N) \quad \text{Equation 3.86}$$

$$\mathbf{m}_N = \mathbf{S}_N \left(\mathbf{S}_0^{-1} \mathbf{m}_0 + \boldsymbol{\Phi}^T \mathbf{R}^{-1} \boldsymbol{\mu}_{C'_A | d_{C_A}, d_{C_{A0}}} \right) \quad \text{Equation 3.87}$$

$$\mathbf{S}_N = (\mathbf{S}_0^{-1} + \boldsymbol{\Phi}^T \mathbf{R}^{-1} \boldsymbol{\Phi})^{-1} \quad \text{Equation 3.88}$$

The algorithmic implementation of the Gaussian process based approach for estimating the parameters of a lumped system ODE model is outlined in *Algorithm 4*.

Algorithm 4: Gaussian process based gradient matching for inferring the parameters of a lumped system ODE model.

Algorithm 4: Inputs

- A lumped system ODE model such as Equation 3.2 (or Equation 3.9).
- Sensor measurements of the state variable C_A dynamic response.
- Sensor measurements of the exogenous input disturbance C_{A0} .
- The prior probability hyperparameters \mathbf{m}_0 and \mathbf{S}_0 .
- Initial values for the Gaussian process kernel function hyperparameters (Table 3.4).
- Initial values for the state variable and exogenous input disturbance sensor variance parameters.

1. Optimise the Gaussian process kernel hyperparameters outlined in Table 3.4 and the sensor variance parameters using gradient ascent in conjunction with Equations 4.49 and 4.51 (with the constraint $\psi_i > 0$) to obtain point estimates.
2. For the point estimates obtained in step 1 – construct the covariance matrices associated with the state variable, state variable derivative and exogenous input disturbance function values (as indicated by subscript notation for each matrix \mathbf{B} in Equation 3.46) using the kernel functions given by Equations 4.47 and 4.48.
3. Using the standard results for conditioning on a Gaussian distribution, obtain Gaussian process posteriors for the state variable derivative, exogenous input disturbance and state variable function values, respectively, by evaluating:

CHAPTER 4: PROPOSED APPROACHES AND METHODOLOGY

- 3.1. Equation 3.55 $(\boldsymbol{\mu}_{\mathbf{C}'_A|\mathbf{d}_{C_A},\mathbf{d}_{C_{A0}}})$ and 4.56 $(\boldsymbol{\Sigma}_{\mathbf{C}'_A|\mathbf{d}_{C_A},\mathbf{d}_{C_{A0}}})$ to obtain $p(\mathbf{C}'_A|\mathbf{d}_{C_A}, \mathbf{d}_{C_{A0}})$
- 3.2. Equation 3.62 $(\boldsymbol{\mu}_{\mathbf{C}_{A0}|\mathbf{d}_{C_A},\mathbf{d}_{C_{A0}}})$ and 4.63 $(\boldsymbol{\Sigma}_{\mathbf{C}_{A0}|\mathbf{d}_{C_A},\mathbf{d}_{C_{A0}}})$ to obtain $p(\mathbf{C}_{A0}|\mathbf{d}_{C_A}, \mathbf{d}_{C_{A0}})$
- 3.3. Equation 3.66 $(\boldsymbol{\mu}_{\mathbf{C}_A|\mathbf{d}_{C_A},\mathbf{d}_{C_{A0}}})$ and 4.67 $(\boldsymbol{\Sigma}_{\mathbf{C}_A|\mathbf{d}_{C_A},\mathbf{d}_{C_{A0}}})$ to obtain $p(\mathbf{C}_A|\mathbf{d}_{C_A}, \mathbf{d}_{C_{A0}})$
4. Construct the design matrix Φ from the MAP estimates given by Equation 4.71 and 4.72.
5. Construct the covariance matrix \mathbf{R} by evaluating Equation 3.84.
6. Obtain the hypothetical sensor measurements for the state variable derivative function values by evaluating Equation 3.68.
7. Set the prior hyperparameters \mathbf{m}_0 and \mathbf{S}_0 .
8. Obtain the posterior over the lumped system ODE model parameters by evaluating Equations 4.87 and 4.88.
9. Calculate the statistical properties of the inferred parameters (Section 4.8).

Algorithm 4: Outputs

- A posterior distribution over the lumped system ordinary differential equation model parameters.
- Point estimates for the state variable and exogenous input disturbance sensor variance parameters.

Algorithm 2 and 4 can easily be extended to consider parameter inference for the liquid draining tank case study (Equation 3.9, Table 3.3) by simply replacing the state variable, state variable derivative and exogenous input disturbance with the appropriate system variables pertaining to the liquid draining tank physical system.

4.6. Parameter Tracking Applications

In order to illustrate the benefit of the Bayesian methodology, i.e. the advantage of obtaining a posterior distribution instead of point estimates for the lumped system ODE model parameters, Case Studies 2 and 3 are extended to include inference about catalyst decay and valve degradation, respectively. Furthermore, a simplistic cost-benefit analysis is implemented for each case study to demonstrate the effect of the parameter posterior distribution on engineering decision making.

4.6.1. Extended Case Study 2: CSTR with Catalyst Decay

In order to model the effect of catalyst decay on the outlet concentration of the isothermal, constant volume CSTR outlined in Section 4.2.1 (Case Study 2, Table 3.3), the reaction rate constant k is taken as a proxy for the catalyst and the following catalyst decay model is assumed (Fogler, 2006):

$$k = k_0 \exp\{-k_d t\} \quad \text{Equation 3.89}$$

Furthermore, it is assumed that k_0 and k_d are time-invariant and do not depend on any other system variables such as concentration, temperature, etc. (Section 1.5.3, 3.1.2) such that k_0 and k_d can be treated as constant values that the engineer desires to learn from noise-corrupted time series data. Equation 3.1 can be rewritten to incorporate the catalyst decay model such that:

CHAPTER 4: PROPOSED APPROACHES AND METHODOLOGY

$$V \frac{dC_A}{dt} = FC_{A0} - FC_A - k_0 \exp\{-k_d t\} VC_A \quad \text{Equation 3.90}$$

$$\frac{dC_A}{dt} = \frac{F}{V} (C_{A0} - C_A) - k_0 \exp\{-k_d t\} C_A \quad \text{Equation 3.91}$$

Recall that the Bayesian methodology presented in Section 4.5.2 (*Algorithm 4*) can only infer the parameters of a lumped system ODE model that can be written as a linear combination of the model parameters. Consequently, it is not possible to infer the model parameter k_d . Therefore, the parameter k_d is set to its true value (Table 3.5) during the data generation (Section 4.7) and parameter inference procedure. From Equation 3.91 the parameters to infer correspond to F/V and k_0 .

Table 3.5: Isothermal, constant volume CSTR catalyst decay model parameters used in the forward modeling approach.

| Model Parameter | Parameter Value |
|-----------------|---------------------------------------|
| k_0 | 0.040 min^{-1} |
| k_d | $7.5 \times 10^{-6} \text{ min}^{-1}$ |

To illustrate the benefit of the parameter posterior distribution on engineering decision making, the isothermal CSTR with catalyst decay case study is further extended to include a simple cost-benefit analysis. The trade-off is determined between the reduction in profit associated with the catalyst replacement time, which for the purposes of the current work corresponds to the time point at which the catalyst proxy k reaches a recommended value of 0.0320 min^{-1} , and the reduction in profit between the predicted maximum profit value and the ground truth profit value (as given by some profit function).

For the current study, it is assumed that the profit function associated with the isothermal, constant volume CSTR process unit is given by:

$$P_{CSTR}(t) = 9800 - 5000t(k_0^2 - 064k_0 + 0.0218) - 50(t^2 - 28t + 196) \quad \text{Equation 3.92}$$

This profit function, whose output is South African rand, explicitly depends on the lumped system ODE model parameter k_0 such that the effect of estimating the parameter from noise-corrupted time series data can be propagated through to engineering decision-making based on the cost-benefit analysis associated with catalyst replacement. The symbol t denotes time in days. Observe that Equation 4.92 is a quadratic function of time t and can be analytically maximised to obtain the time point of maximum profit for a selected value of the parameter k_0 . The current work only considers evaluating the profit function over a 30 day period. The quadratic dependence on time is selected to reflect that over the 30 day period the profit associated with the CSTR process unit increases to a maximum after which it decreases due to operating costs.

4.6.2. Extended Case Study 3: Liquid Draining Tank with Valve Degradation

In order to model the effect of valve degradation on the liquid draining tank outlined in Section 4.2.2 (Case Study 3, Table 3.3), the flow restriction coefficient is taken as a proxy for the valve flow coefficient and the following valve degradation model is assumed (Van Noortwijk and Pandey, 2003):

$$k_v = k_{v_0} + A_v t^{b_v} \quad \text{Equation 3.93}$$

Furthermore, it is assumed that k_{v_0} , A_v and b_v are time-invariant and do not depend on any other system variables (Section 1.5.3, 3.1.2) such that k_{v_0} , A_v and b_v can be treated as constant values that the engineer desires to learn from noise-corrupted time series data. Equation 3.7 can be rewritten to incorporate the valve degradation model such that:

$$A \frac{dL}{dt} = F_0 - (k_{v_0} + A_v t^{b_v}) \sqrt{L} \quad \text{Equation 3.94}$$

$$\frac{dL}{dt} = \frac{1}{A} F_0 - \frac{k_{v_0}}{A} \sqrt{L} + \frac{A_v}{A} t^{b_v} \sqrt{L} \quad \text{Equation 3.95}$$

The parameter k_{v_0} represents the initial valve flow coefficient, A_v is the rate of degradation and b_v reflects the nonlinear trend of the valve degradation law. Since the Bayesian methodology presented in Section 4.5.2 (*Algorithm 4*) can only infer the parameters of a lumped system ODE model that can be written as a linear combination of the model parameters, the model parameter b_v is set to its true value (Table 3.6) during the data generation (Section 4.7) and parameter inference procedure. From Equation 3.95 the parameters to infer correspond to $1/A$, k_{v_0}/A and A_v/A .

Table 3.6: Liquid draining tank valve degradation model parameters used in the forward modeling approach.

| Model Parameter | Parameter Value |
|-----------------|------------------------------------|
| k_{v_0} | $37.8 \frac{m^{2.5}}{h}$ |
| A_v | $0.00959 \frac{m^{2.5}}{h^{2.07}}$ |
| b_v | 1.07 (dimensionless) |

To illustrate the benefit of the parameter posterior distribution on engineering decision making, the liquid draining tank with valve degradation case study is further extended to include a simplistic cost-benefit analysis.

The trade-off is determined between the reduction in profit associated with the valve replacement time, which for the purposes of the current work corresponds to the time point at which the valve flow coefficient degrades such that it reaches a recommended value of $1.2k_{v_0}$, and the reduction in profit between the predicted maximum profit value and the ground truth

CHAPTER 4: PROPOSED APPROACHES AND METHODOLOGY

profit value (as given by some profit function). For the current study, it is assumed that the profit function associated with the liquid draining tank unit is given by:

$$P_{Tank}(t) = 9800 - 5000t \left(\left(\frac{k_{v_0}}{A} \right)^2 - 1.44 \frac{k_{v_0}}{A} + 0.0648 \right) - 50(t^2 - 28t + 196) \quad \text{Equation 3.96}$$

The draining tank profit function, whose output is 100 times South African rand, explicitly depends on the lumped system ODE model parameter k_{v_0}/A such that the effect of estimating the parameter from noise-corrupted time series data can be propagated through to engineering decision making based on the cost-benefit analysis associated with valve replacement. The symbol t denotes time in days. Equation 4.96 is constructed such that it shares the same characteristics as the CSTR process unit profit function (Equation 4.92).

4.7. Data Generation Process

This thesis makes use of synthetically generated sensor measurement data using the forward modeling approach (Section 1.5.2). In other words, the ground truth lumped system dynamic model (with the corresponding model parameters) is known exactly and is used to generate data for the physical system (Section 1.5.3) under consideration. The synthetic data will then be used in conjunction with the algorithms outlined in Sections 4.4 and 4.5 to infer the model parameters which generated the synthetic sensor measurement data set (Section 1.5.2).

4.7.1. Case Study 1: Isothermal CSTR – Perfect Input Step Disturbance

Synthetic data is generated from Equation 3.3 (using the information summary in Table 3.1) at uniformly spaced intervals between $t \in \{0, 4.21, \dots, 80\}$ minutes, with an interval spacing of 4.21 minutes, by evaluating the CSTR algebraic model (Equation 3.3) at the uniformly spaced intervals to obtain the deterministic function response values. The interval spacing of 4.21 minutes simulate a sensor sampling the CSTR outlet concentration C_A every 4.21 minutes for a total of 20 sensor measurements over the 80 minute data collection period. However, due to imperfect measurement equipment, the engineer will only observe a noise-corrupted version of the true underlying CSTR outlet concentration in a practical setting.

The imperfect nature of the sensor can be modeled by the addition of independent zero-mean Gaussian noise with a variance parameter σ_{noise, C_A}^2 to the deterministic function response values (evaluated at the uniformly spaced intervals $t \in \{0, 4.21, \dots, 80\}$) to generate a noise-corrupted data set that would result from taking outlet concentration measurements with a sensor. The current thesis sets the value of $\sigma_{noise, C_A}^2 = 2.25 \times 10^{-4}$. The additive Gaussian noise is generated using the *normrnd* function in MATLAB® in conjunction with the Mersenne Twister random number generator with a seed value of zero.

Under the exact same conditions described above, one hundred additional independent sensor data sets are generated by changing the seed value from zero to ‘Shuffle’ in MATLAB®. The ‘Shuffle’ command seeds the random number generator based on the current time which ensures that each generated data set is unique.

CHAPTER 4: PROPOSED APPROACHES AND METHODOLOGY

By generating sensor measurement data in this manner, one is capturing a property of many real-world data sets, namely that these data sets have an underlying regularity, which one wishes to learn, but individual target variable measurements are corrupted with noise.

4.7.2. Case Study 2: Isothermal CSTR – Random Exogenous Input Disturbance

The exogenous input disturbance for C_{A0} is generated by drawing a random sample of $N = 121$ function values C_{A0} for $t \in \{0, 1, \dots, 120\}$ from a Gaussian process prior where each entry in the mean vector is set to $(C_{A0})_{initial}$ (Table 3.1). The Gaussian process prior covariance matrix is populated in an element-wise manner by evaluating Equation 1.99. In order to compute Equation 1.99, the kernel function hyperparameters σ_f and ℓ are drawn uniformly using the *rand* function in MATLAB[®] between the ranges $\sigma_f \in (0; 1)$ [seed set to ‘Shuffle’] and $\ell \in (0; 35)$ [seed set to ‘Shuffle’], respectively, using the Mersenne Twister random number generator. Random samples are repeatedly drawn from the Gaussian process prior until all function values C_{A0} for a random sample are within the range $0.95(C_{A0})_{initial} \leq (C_{A0})_{initial} \leq 1.05(C_{A0})_{initial}$. The first sample that falls within the specified range is taken as the noise-free data set for the input disturbance at Normal Operating Conditions (NOC).

The input disturbance function values are then fed into Simulink where Equation 3.2 (using the information summary in Table 3.1) is numerically integrated using ode45 to produce the CSTR outlet concentration C_A deterministic function response values for the random exogenous input disturbance sample. The 121 function values for the exogenous input disturbance C_{A0} produce 121 corresponding outlet concentration deterministic response values. For a sensor measurement sampling time of $t = 1$ minute, this corresponds to an exogenous input data set and outlet concentration data set of 121 sensor measurements.

However, due to imperfect measurement equipment, the engineer will only observe a noise-corrupted version of the true underlying exogenous input disturbance and CSTR outlet concentration in a practical setting. Thus, zero-mean Gaussian noise with variance parameters of $\sigma_{noise, C_{A0}}^2$ and σ_{noise, C_A}^2 are added to the random exogenous input disturbance sample (as obtained from the Gaussian process prior) and the CSTR outlet concentration deterministic response values, respectively. The additive Gaussian noise is implemented in the Simulink model using the *Random Number* generator block. For the current study, the values of the sensor variance parameters are set to $\sigma_{noise, C_{A0}}^2 = 7 \times 10^{-6}$ and $\sigma_{noise, C_A}^2 = 2 \times 10^{-6}$, respectively. The seed values required for the *Random Number* generator blocks are obtained by drawing interger values on the interval $(1; 1 \times 10^9)$ using the *randi* function in MATLAB[®] where the required seeds to initialise the *randi* function (Mersenne Twister random number generator) associated with each of the sensor variance parameters $\sigma_{noise, C_{A0}}^2$ and σ_{noise, C_A}^2 are set to 2 and 3, respectively.

One hundred additional independent sensor measurement data sets are generated by changing the seed values associated with $\sigma_{noise, C_{A0}}^2$ and σ_{noise, C_A}^2 from 2 and 3 to ‘Shuffle’, respectively, to ensure that each generated data set is unique. Note that the sensor variance parameters $\sigma_{noise, C_{A0}}^2$ and σ_{noise, C_A}^2 as well as the input disturbance signal are the same for each of the generated data sets.

CHAPTER 4: PROPOSED APPROACHES AND METHODOLOGY

Data sets for the CSTR case study with catalyst decay is generated in the exact same way as described above. However, the input disturbance function values are fed into Simulink where Equation 3.91 (using the information in Tables 4.1 and 4.5) is numerically integrated using ode45 to produce the CSTR outlet concentration response values.

4.7.3. Case Study 3: Draining Tank– Random Exogenous Input Disturbance

The exogenous input disturbance for F_0 is generated by drawing a random sample of $N = 121$ function values \mathbf{F}_0 for $t \in \{0,1, \dots, 120\}$ from a Gaussian process prior where each entry in the mean vector is set to $(F_0)_{initial}$ (Table 3.2). The Gaussian process prior covariance matrix is constructed in the exact same manner as outlined in Section 4.7.2. Random samples are repeatedly drawn from the Gaussian process prior until all function values \mathbf{F}_0 for a random sample are within the range $0.95(F_0)_{initial} \leq (F_0)_{initial} \leq 1.05(F_0)_{initial}$. The first sample that falls within the specified range is taken as the noise-free data set for the input disturbance at Normal Operating Conditions.

The input disturbance function values are then fed into Simulink where Equation 3.9 (using the information summary in Table 3.2) is numerically integrated using ode45 to produce the tank liquid level deterministic function response values for the random exogenous input disturbance sample. The 121 function values for the exogenous input disturbance F_0 produce 121 corresponding tank liquid level response values. For a sensor measurement sampling time of $t = 1$ minute, this corresponds to an exogenous input data set and tank liquid level data set of 121 sensor measurements.

Zero-mean Gaussian noise with variance parameters of σ_{noise,F_0}^2 and $\sigma_{noise,L}^2$ are added to the random exogenous input disturbance sample and the tank liquid level response values, respectively. For the current study, the values of the sensor variance parameters are set to $\sigma_{noise,F_0}^2 = 2 \times 10^{-5}$ and $\sigma_{noise,L}^2 = 4 \times 10^{-5}$, respectively. The seed values required for the *Random Number* generator blocks are obtained by drawing integer values on the interval $(1; 1 \times 10^9)$ using the *randi* function in MATLAB[®] where the required seeds to initialise the *randi* function (Mersenne Twister random number generator) associated with each of the sensor variance parameters σ_{noise,F_0}^2 and $\sigma_{noise,L}^2$ are set to 1 and 2, respectively.

One hundred additional independent sensor measurement data sets are generated in the exact same way as outlined in Section 4.7.2 by using the ‘Shuffle’ command in MATLAB[®].

Data sets for the liquid draining tank case study with the valve degradation model is generated in the exact same way as described above. However, a sample of $N = 301$ equally spaced function values \mathbf{F}_A for $t \in \{0,1, \dots, 300\}$ is drawn from the Gaussian process prior. The input disturbance function values are then fed into Simulink where Equation 3.95 (using the information in Tables 4.2 and 4.6) is numerically integrated using ode45 to produce the tank liquid level response values for the random exogenous input disturbance sample drawn from the Gaussian process prior.

4.8. Performance Criteria

As mentioned in Section 4.5, the proposed Bayesian methodologies (*Algorithm 3* and *4*) will be benchmarked against the more common Gauss-Newton simple least squares implementation (*Algorithm 1* and *2*) for estimating the parameters of a lumped system algebraic or ODE model. Both Bayesian and frequentist (Sections 4.4 and 4.5) approaches will be evaluated for a *single* data set initialisation by considering the following performance criteria:

1. Algorithm execution time.
2. Inferred parameter accuracy/reliability against the simulation ground truth.
3. Marginal parameter confidence/credibility intervals* (Section 2.8).
4. Joint parameter confidence/credibility interval* (Section 2.8).
5. Expected mean response*.
6. The cost-benefit trade-off associated with the use of the parameter point estimates (*Algorithm 1* and *2*) compared to the parameter posterior distribution (*Algorithm 3* and *4*) and the resulting effect on engineering decision-making.

For Case Study 1, 2 and 3 (Section 4.2.1, 4.2.2, Table 3.3), only the first five performance criteria outlined above are applied. Performance criterion six pertains to the extended case studies, i.e. the isothermal CSTR case study with catalyst decay (Section 4.6.1) and the draining tank case study with valve degradation (Section 4.6.2). All confidence and credibility intervals are evaluated at an $\alpha = 0.01$, i.e. $100(1 - \alpha)\% = 99\%$. Details for calculating the sensor variance parameter σ_{noise}^2 associated with *Algorithm 1* and *2*, the marginal confidence interval and joint confidence region, as well as the expected mean response, for the Gauss-Newton simple least squares methodology can be found in Chapter 11 of Englezos and Kalogerakis (2001).

In order to establish how consistent the different algorithms are, experiments are repeated for each case study (including the extended case studies). One hundred additional independent data sets are generated with the same exogenous input disturbance signal and sensor variance parameters as outlined in Section 4.7, however, the sensor seeds are changed to ensure that each manifestation of the data sets are unique. By keeping the exogenous input disturbance signal and sensor variance parameters the same, one is essentially restarting each experiment under the exact same conditions and ‘measuring’ the state variable and exogenous input disturbance using the same sensors. Both Bayesian and frequentist approaches (Sections 4.4 and 4.5) will be evaluated for the multiple data sets by considering the following performance criteria:

7. Mean of the inferred model parameter estimates.
8. Standard deviation of the inferred model parameter estimates.
9. The proportion (coverage frequency) of the simulation ground truth parameter values that fall within the 99% marginal parameter confidence/credibility intervals.
10. The average marginal parameter confidence/credibility interval width.

* Indicates statistical property

4.9. Summary

Chapter 4 started by introducing the two physical systems considered throughout this thesis, namely, the isothermal, constant volume CSTR and the liquid draining tank (Section 1.5.3). Following the introduction of the physical systems, the author presented the implementation procedure for the Gauss-Newton parameter estimation benchmarks (*Algorithms 1* and *2*) which stem from the work of Englezos and Kalogerakis (2001) discussed in Sections 3.3 and 3.4. The author then continued by proposing two Bayesian methodologies (*Algorithms 3* and *4*), in contrast to the Gauss-Newton simple least squares approach, for estimating the parameters of lumped system algebraic or ODE models. Both Bayesian methodologies find inspiration in the work of Calderhead, Girolami and Lawrence (2009), Chappell et al. (2009), Dondelinger et al. (2013), Gorbach, Bauer and Buhmann (2017), and Wenk et al. (2018).

The CSTR and draining tank case studies were then extended to include catalyst decay and valve degradation, respectively, followed by presenting individual process unit profit functions to propagate the effect of the estimated model parameters through to engineering decision making based on a simple cost-benefit analysis. Lastly, in order to establish whether the proposed Bayesian approaches provide comparable results to the more commonly used Gauss-Newton simple least squares methodology, the author provided several performance criteria to assess the proposed Bayesian methodologies.

Chapter 5 (Results and Discussion) will build on Chapter 4 by presenting the inferred parameter results for Case Studies 1 through 3 (Table 3.3) (as inferred from the synthetically generated data using the forward modeling approach, Section 4.7). Results will be presented for the benchmark approaches (*Algorithms 1* and *2*) and the proposed Bayesian methodologies (*Algorithms 3* and *4*) as well as for the extended case studies that incorporate catalyst decay and valve degradation.

Furthermore, the performance criteria results (Section 4.8, criteria 1 through 6) of the proposed Bayesian methodologies will be compared to the benchmark approaches, with particular emphasis on how the parameter point estimates and parameter posterior distribution influence engineering decision-making given a single data set manifestation. Additionally, the results obtained from performance criteria 7 through 10, as applied to the multiple independently generated data sets, will be presented to determine whether the frequentist benchmark and the proposed Bayesian methodologies provide consistent results, when compared to the simulation ground truth parameter values.

Chapter 5

Results and Discussion

“The actual science of logic is conversant at present only with things either certain, impossible, or entirely doubtful...Therefore the true logic for this world is the calculus of probabilities, which takes account of the magnitude of the probability which is, or ought to be, in a reasonable (person’s) mind”

- James Clerk Maxwell, 1850

5.1. Overview

Chapters 1 through 3 presented the necessary introductory material, theoretical foundations and literature to understand the parameter inference procedure. The content and insights from these chapters were then used to develop the proposed Bayesian approaches outlined in Chapter 4 which can be used by the engineer to make conclusions about the parameters of a lumped system algebraic or ODE dynamic model from noise-corrupted time series data.

Chapter 5 presents and discusses the results obtained from the benchmark approaches (Section 4.4) and proposed Bayesian methodologies (Sections 4.5), as applied to synthetically generated sensor measurement data (Section 4.7), for each of the case studies considered (Sections 4.2 and 4.6). If the proposed Bayesian methodologies provide comparable results to the benchmark approaches, based on the performance criteria outlined in Section 4.8, this can serve as motivation to further explore the benefit and application of Bayesian inference in the chemical engineering setting.

5.2. Case Study Parameter Inference Results

This section presents the parameter inference results obtained from the benchmark approaches (Section 4.4) and the proposed Bayesian methodologies (Section 4.5), as applied to the synthetically generated sensor measurement data (Section 4.7), for each individual case study (Section 4.2, Table 3.3). The MAP estimates are used as representative values from the Bayesian approaches to compare to the maximum likelihood estimates obtained from the benchmark approaches.

5.2.1. Case Study 1: Isothermal CSTR – Perfect Step Input Disturbance

Table 4.1 summarises the algorithm execution time for *Algorithms 1* and *3*, as well as the inferred parameter accuracy/reliability (expressed as the percentage error against the simulation ground truth parameter values used during the synthetic data generation process for Equation 3.3).

CHAPTER 5: RESULTS AND DISCUSSION

Table 4.1: Summary of the algorithm execution time, inferred lumped system algebraic model parameters and sensor standard deviation parameter for each approach. Note that these results are given for a single data set manifestation.

| Probabilistic Interpretation | Algorithm Number | Algorithm Execution Time | Unknown Parameter | Simulation Ground Truth | Inferred Value | % Error |
|------------------------------|------------------|--------------------------|----------------------|---------------------------|---------------------------|---------|
| Frequentist | 1 | 11.8 ms | K_p | 0.503 (-) | 0.520 (-) | 3.48% |
| | | | τ | 12.43 min | 13.17 min | 6.02% |
| | | | σ_{noise,C_A} | $0.0150 \frac{mole}{m^3}$ | $0.0222 \frac{mole}{m^3}$ | 47.78% |
| Bayesian | 3 | 52.7 ms | K_p | 0.503 (-) | 0.521 (-) | 3.50% |
| | | | τ | 12.43 min | 13.19 min | 6.15% |
| | | | σ_{noise,C_A} | $0.0150 \frac{mole}{m^3}$ | $0.0221 \frac{mole}{m^3}$ | 47.06% |

(-) dimensionless quantity

Table 5.2 summarises the results obtained from applying *Algorithms 1* and *3* to the one hundred independently generated sensor measurement data sets (Sections 4.7 and 4.8 - performance criteria 7 through 10).

Table 5.2: Summary of the mean inferred parameter estimates, standard deviation, proportion of confidence/credibility intervals containing the simulation ground truth parameter values and average confidence/credibility interval width for the one hundred independently generated sensor measurement data sets (performance criteria 7 through 10, Section 4.8).

| Algorithm Number | Unknown Parameter | Simulation Ground Truth | Mean Inferred Value | Standard Deviation | Percentage Contained* | Average Interval Width |
|------------------|-------------------|-------------------------|---------------------|--------------------|-----------------------|------------------------|
| 1 | K_p | 0.503 (-) | 0.504 (-) | 0.006 (-) | 99% | 0.031 (-) |
| | τ | 12.43 min | 12.52 min | 0.675 min | 96% | 3.51 min |
| 3 | K_p | 0.503 (-) | 0.504 (-) | 0.006 (-) | 92% | 0.021 (-) |
| | τ | 12.43 min | 12.53 min | 0.678 min | 46% | 0.703 min |

*Percentage Contained refers to the proportion of the simulation ground truth parameter values that fall within the 99% marginal parameter confidence/credibility interval.

Figures 5.1 through 5.9 visually depict performance criteria 3 to 5 (Section 4.8) for *Algorithms 1* and *3* applied to a single sensor measurement data set.

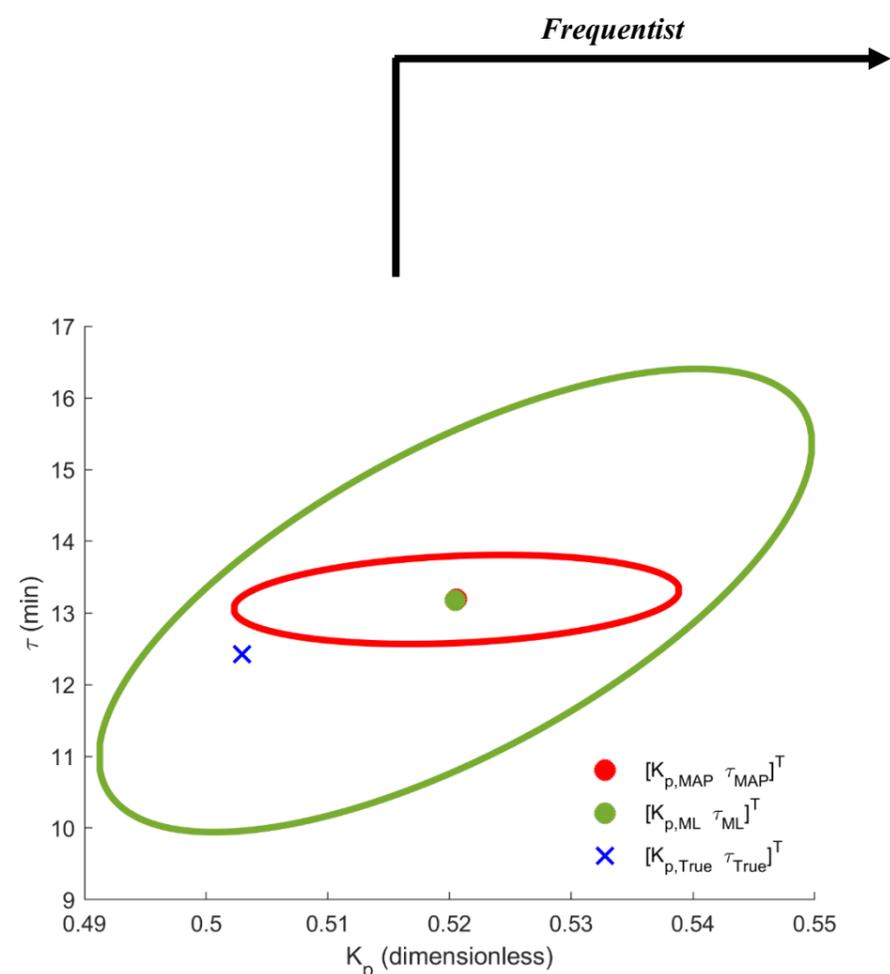
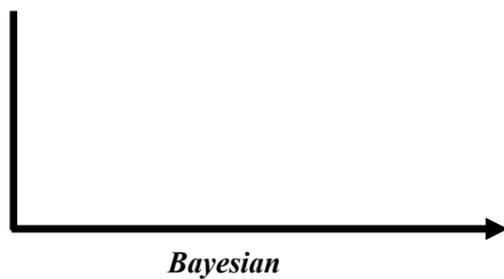


Figure 5.1: 99% joint parameter confidence (green) and credibility (red) regions (Section 2.8). The blue cross denotes the simulation ground truth. The confidence region is centered at the maximum likelihood estimate for the model parameters (Equation 3.3) while the credibility region is centered at the mean (MAP estimate) of the joint Gaussian posterior distribution.



Bayesian

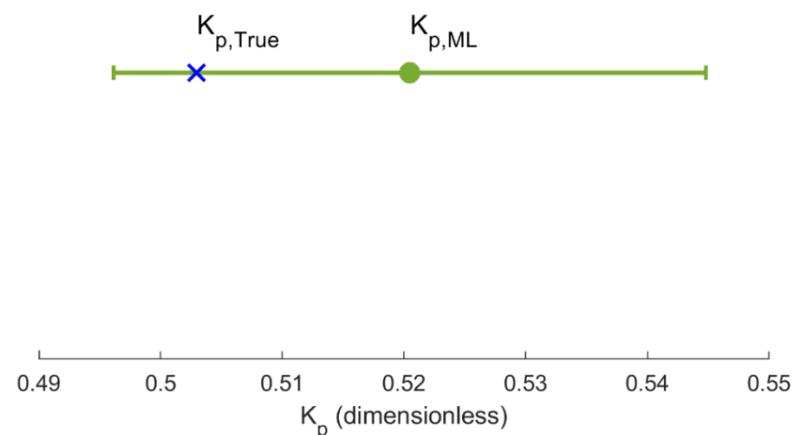


Figure 5.2: 99% confidence interval for the unknown model parameter K_p . The blue cross denotes the simulation ground truth. The confidence interval (Section 2.8) is centered at the maximum likelihood estimate for K_p .

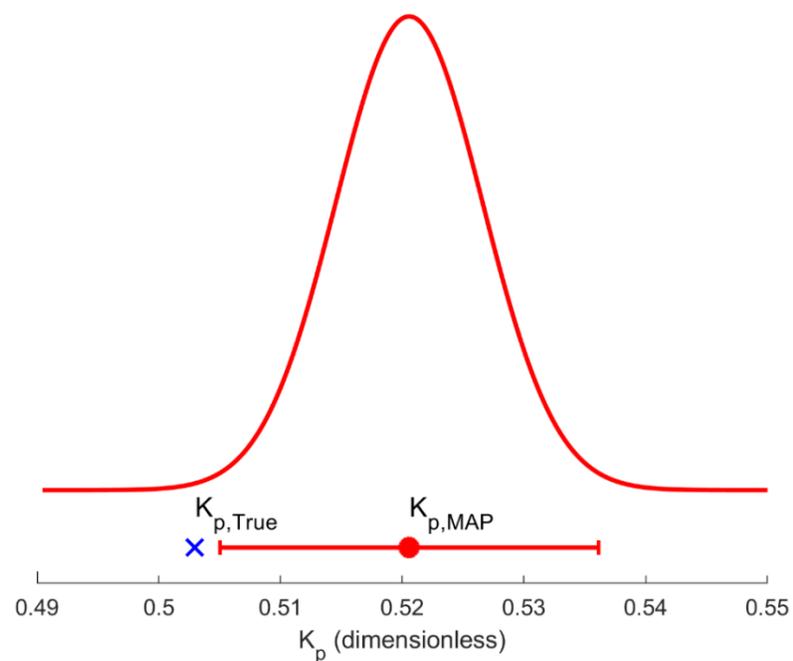


Figure 5.4: Marginal posterior distribution over the unknown model parameter K_p with the corresponding 99% credibility interval (Section 2.8). The blue cross denotes the simulation ground truth.

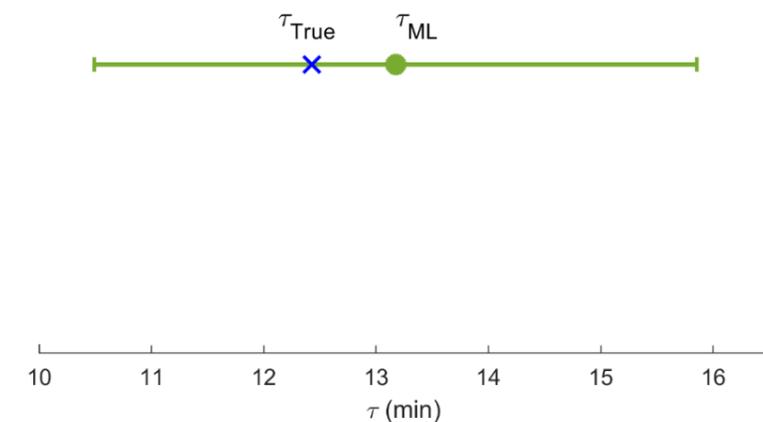


Figure 5.3: 99% confidence interval for the unknown model parameter τ . The blue cross denotes the simulation ground truth. The confidence interval (Section 2.8) is centered at the maximum likelihood estimate for τ .

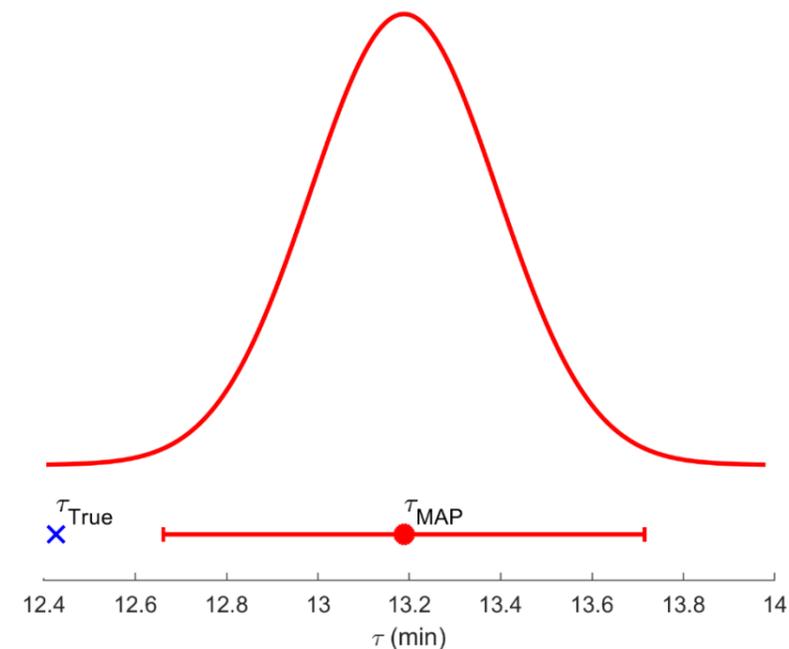


Figure 5.5: Marginal posterior distribution over the unknown model parameter τ with the corresponding 99% credibility interval (Section 2.8). The blue cross denotes the simulation ground truth.

Note that the x-axis scale for Figure 5.5 has been adjusted to aid in the visual interpretation of the results. Furthermore, note that all results are given for a single data set manifestation.

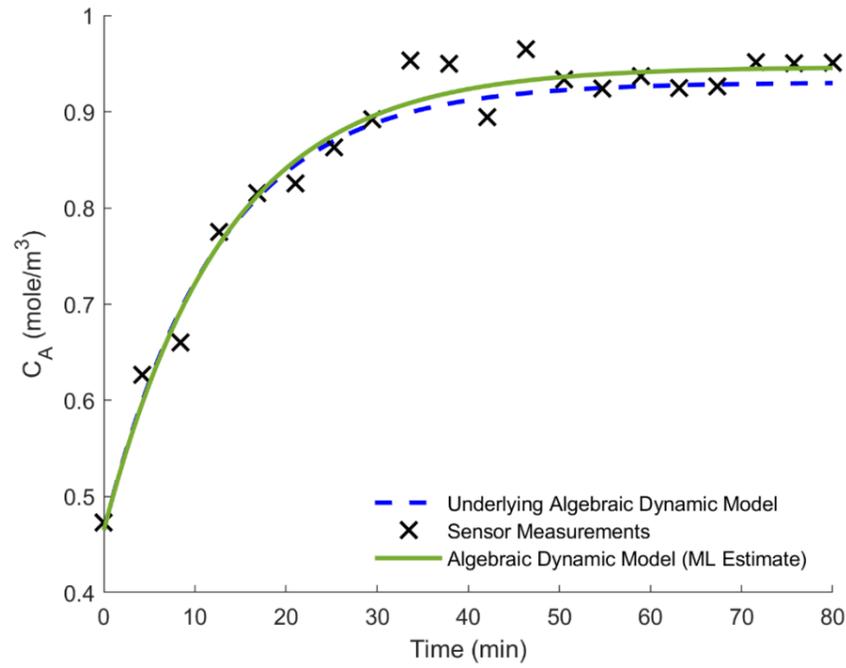


Figure 5.6 Maximum likelihood fit for the expected mean response of Equation 3.3 obtained from applying *Algorithm 1* (Section 4.4) to the synthetically generated sensor measurement data (Section 4.7.1).

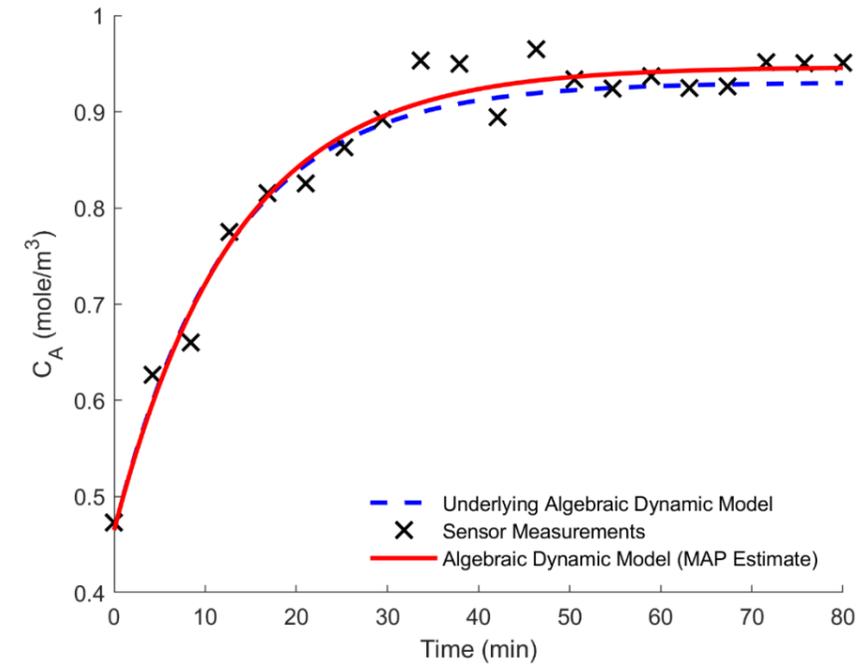


Figure 5.7: MAP estimate fit for Equation 3.3 obtained from applying *Algorithm 3* (Section 4.5.1) to the synthetically generated sensor measurement data (Section 4.7.1).

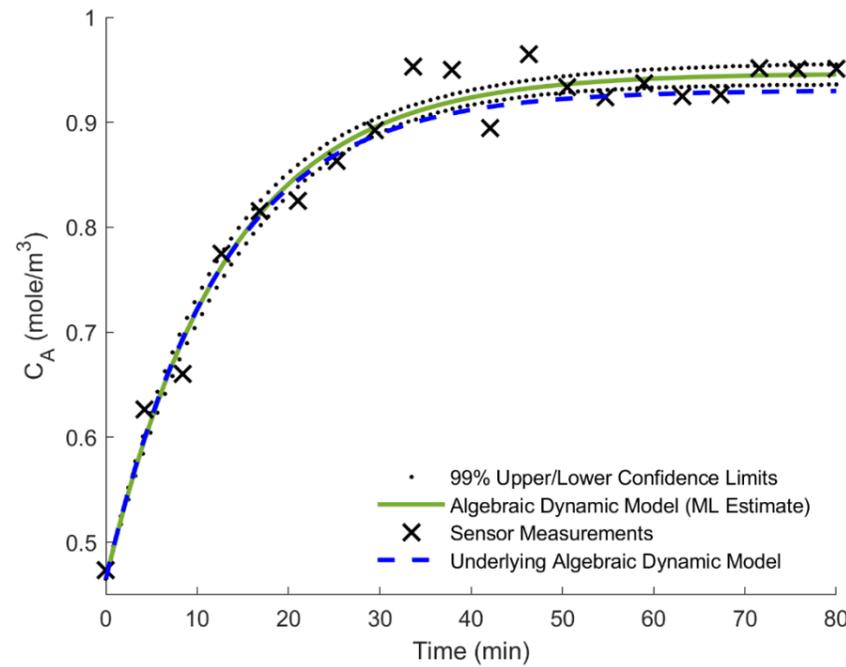


Figure 5.8: 99% confidence interval (Section 2.8) of the expected mean response of Equation 3.3. The dotted lines represent the lower and upper confidence limits.

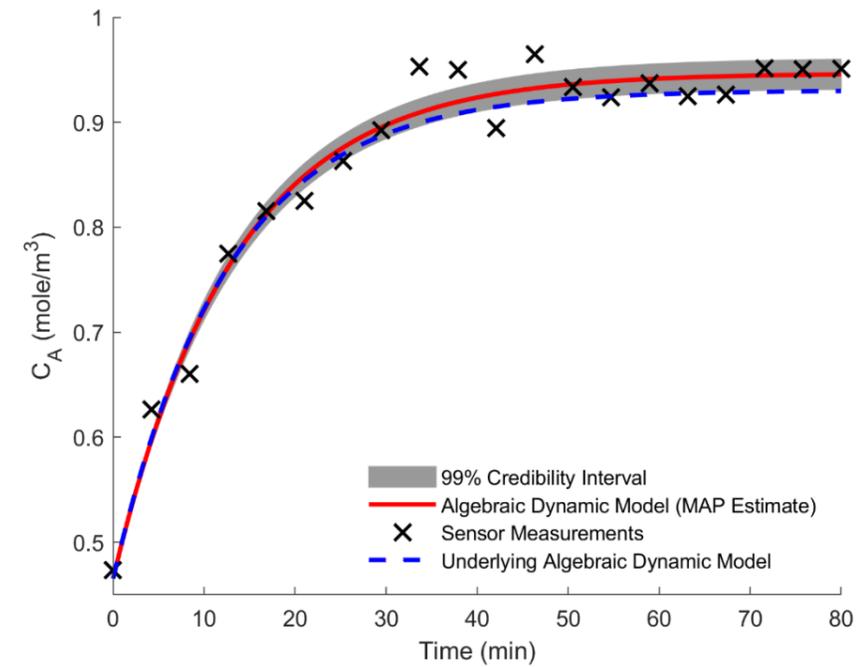


Figure 5.9: 99% credibility interval (Section 2.8) of the expected mean response of Equation 3.3. Calculated by evaluating Equation 4.9. The shaded region represents the 99% credibility interval.

CHAPTER 5: RESULTS AND DISCUSSION

The marginal posterior distributions associated with Figures 5.4 and 5.5 are obtained from Equation 3.32 by simply marginalising out the model parameter that the engineer is not interested in. Since the posterior distribution over the unknown algebraic model parameters take the form of a bivariate Gaussian distribution (Section 2.3.1), the marginal posterior distribution over each algebraic model parameter corresponds to picking the entries of the mean vector (Equation 3.35) and the covariance matrix (inverse of Equation 3.34) associated with each unknown model parameter such that:

$$\mathbf{S}_N = \mathbf{A}_N^{-1} \quad \text{Equation 4.1}$$

Figure 5.4
$$p(K_p) = \mathcal{N}\left(K_p | [\mathbf{m}_{N,new}]_1, [\mathbf{S}_N]_{11}\right) \quad \text{Equation 4.2}$$

Figure 5.5
$$p(\tau) = \mathcal{N}\left(\tau | [\mathbf{m}_{N,new}]_2, [\mathbf{S}_N]_{22}\right) \quad \text{Equation 4.3}$$

The expected mean response and credibility interval illustrated in Figure 5.9 is determined by observing that the first-order Taylor expansion for some arbitrary input time point t_i is linear in the unknown model parameters $\boldsymbol{\Omega}$ (Equation 3.23). Since $\boldsymbol{\Omega}$ is normally distributed according to Equation 3.32, for an arbitrary input time point t_i , $C_A(t_i, \boldsymbol{\Omega})$ will also be normally distributed (Bishop, 2006). Thus, the engineer only requires the mean and covariance of $C_A(t_i, \boldsymbol{\Omega})$ at the input time point t_i . The mean can readily be obtained by evaluating the expected value (Sections 2.1.4 and 2.3.1) of $C_A(t_i, \boldsymbol{\Omega})$ under the variational approximating posterior distribution $q^*(\boldsymbol{\Omega})$ such that:

$$\mathbb{E}_{q^*(\boldsymbol{\Omega})}[C_A|t_i] = \int_{-\infty}^{\infty} q^*(\boldsymbol{\Omega}) (C_A(t_i, \mathbf{m}_N) + \mathbf{J}(\boldsymbol{\Omega} - \mathbf{m}_N)) d\boldsymbol{\Omega} \quad \text{Equation 4.4}$$

$$\mathbb{E}_{q^*(\boldsymbol{\Omega})}[C_A|t_i] = C_A(t_i, \mathbf{m}_N) \quad \text{Equation 4.5}$$

The variance can readily be obtained by evaluating the expected values (Sections 2.1.4 and 2.3.1) given by Equation 4.6 under the variational approximating posterior distribution $q^*(\boldsymbol{\Omega})$ such that:

$$\text{Var}[C_A|t_i] = \mathbb{E}_{q^*(\boldsymbol{\Omega})}[(C_A|t_i)^2] - (\mathbb{E}_{q^*(\boldsymbol{\Omega})}[C_A|t_i])^2 \quad \text{Equation 4.6}$$

$$\text{Var}[C_A|t_i] = \int_{-\infty}^{\infty} q^*(\boldsymbol{\Omega}) (C_A(t_i, \mathbf{m}_N) + \mathbf{J}_i(\boldsymbol{\Omega} - \mathbf{m}_N))^2 d\boldsymbol{\Omega} - (C_A(t_i, \mathbf{m}_N))^2 \quad \text{Equation 4.7}$$

$$\text{Var}[C_A|t_i] = \mathbf{J}_i \mathbf{S}_N \mathbf{J}_i^T \quad \text{Equation 4.8}$$

As a result, the probability of observing C_A at an arbitrary input time point t_i is given by:

Figure 5.9
$$p(C_A|t_i) = \mathcal{N}(C_A | C_A(t_i, \mathbf{m}_N), \mathbf{J}_i \mathbf{S}_N \mathbf{J}_i^T) \quad \text{Equation 4.9}$$

CHAPTER 5: RESULTS AND DISCUSSION

From Table 4.1 and Figures 5.1 through 5.9, one observes that the proposed variational Bayesian nonlinear regression approach provides comparable results to the benchmark methodology for Case Study 1 (Equation 3.3) for a single data set manifestation. However, the 99% credibility region (Figure 5.1), as well as the credibility interval for each model parameter (Figure 5.4 and 5.5), does not contain the simulation ground truth parameter values.

This is a result of using mean-field variational inference which generally underestimates the variance of the true posterior distribution. Furthermore, observe from Table 5.2 that both *Algorithms 1* and *3* provide consistent results when applied to the one hundred independently generated data sets. For the one hundred independent data sets, only 46% of the constructed credibility intervals (Section 2.8) contain the simulation ground truth parameter value for τ . However, the average confidence interval width for parameter τ is approximately 5 times larger than the average credibility interval width which indicates that the confidence interval estimate for τ is conservative. This explains why 98% of the constructed confidence intervals (Section 2.8) contain the simulation ground truth value of parameter τ .

5.2.2. Case Study 2: Isothermal CSTR – Exogenous Input Disturbance

Table 5.3 summarises the algorithm execution time for *Algorithms 2* and *4*, as well as the inferred parameter accuracy/reliability (expressed as the percentage error against the simulation ground truth parameter values used during the synthetic data generation process for Equation 3.2).

Table 5.3: Summary of the algorithm execution time, inferred lumped system ODE model parameters and sensor standard deviation parameters. Note that these results are given for a single data set manifestation.

| Probabilistic Interpretation | Algorithm Number | Algorithm Execution Time | Unknown Parameter | Simulation Ground Truth | Inferred Value | % Error |
|------------------------------|------------------|--------------------------|-------------------------|--|--|---------|
| Frequentist | 2 | 5.42 s | F/V | 0.0405 min^{-1} | 0.0445 min^{-1} | 9.96% |
| | | | k | 0.0400 min^{-1} | 0.0440 min^{-1} | 9.92% |
| | | | σ_{noise,C_A} | $0.00141 \frac{\text{mole}}{\text{m}^3}$ | $0.00153 \frac{\text{mole}}{\text{m}^3}$ | 7.97% |
| | | | $\sigma_{noise,C_{AO}}$ | $0.00265 \frac{\text{mole}}{\text{m}^3}$ | N/A* | N/A* |
| Bayesian | 4 | 6.21 s | F/V | 0.0405 min^{-1} | 0.0404 min^{-1} | 0.20% |
| | | | k | 0.0400 min^{-1} | 0.0399 min^{-1} | 0.27% |
| | | | σ_{noise,C_A} | $0.00141 \frac{\text{mole}}{\text{m}^3}$ | $0.00151 \frac{\text{mole}}{\text{m}^3}$ | 6.70% |
| | | | $\sigma_{noise,C_{AO}}$ | $0.00265 \frac{\text{mole}}{\text{m}^3}$ | $0.00247 \frac{\text{mole}}{\text{m}^3}$ | 7.11% |

*N/A - not applicable

CHAPTER 5: RESULTS AND DISCUSSION

Table 5.4 summarises the results obtained from applying *Algorithm 2* and *4* to the one hundred independently generated sensor measurement data sets (Sections 4.7 and 4.8 - performance criteria 7 through 10). Recall that the results outlined in Table 5.4 pertain to repeated experiments where the exogenous input disturbance signal for C_{A0} , i.e. the random exogenous input disturbance draw from the Gaussian process prior, and the sensor variance parameters are kept the same to reflect repeating the experiments under the exact same starting conditions.

Table 5.4: Summary of the mean inferred parameter estimates, standard deviation, proportion of confidence/credibility intervals containing the simulation ground truth parameter values and average confidence/credibility interval width for the one hundred independently generated sensor measurement data sets. The units for all numerical entries in Table 5.4 are min^{-1} (excluding the Percentage Contained* column) [results are given for performance criteria 7 through 10, Section 4.8].

| Algorithm Number | Unknown Parameter | Simulation Ground Truth | Mean Inferred Value | Standard Deviation | Percentage Contained* | Average Interval Width |
|------------------|-------------------|-------------------------|---------------------|--------------------|-----------------------|------------------------|
| 2 | F/V | 0.0405 | 0.0403 | 0.0020 | 98% | 0.0087 |
| | k | 0.0400 | 0.0398 | 0.0020 | 98% | 0.0086 |
| 4 | F/V | 0.0405 | 0.0380 | 0.0020 | 23% | 0.0022 |
| | k | 0.0400 | 0.0376 | 0.0020 | 24% | 0.0022 |

*Percentage Contained refers to the proportion of the simulation ground truth parameter values that fall within the 99% marginal parameter confidence/credibility interval.

Observe from Table 5.4 that both the frequentist (*Algorithm 2*) and Bayesian (*Algorithm 4*) approaches provide consistent result when compared to the simulation ground parameter values. Note that only 23 of the 100 constructed credibility intervals (one for each of the independently generated sensor measurement data sets) contain the simulation ground truth parameter value for F/V while only 24 of the constructed confidence intervals contain the ground truth parameter value for k .

However, the average credibility interval width for F/V and k is approximately four times smaller than the average confidence interval width (Section 2.8). This indicates that the confidence interval estimates for F/V and k , as obtained from applying *Algorithm 2* to the one hundred independently generated data sets, are very conservative which explains why 98% of the constructed confidence intervals contain the simulation ground truth values of F/V and k , respectively.

CHAPTER 5: RESULTS AND DISCUSSION

Figures 5.10 through 5.20 visually depict performance criteria 3 to 5 (Section 4.8) for *Algorithms 2* and *4* applied to a single sensor measurement data set. Refer to Chapter 11 of Englezos and Kalogerakis (2001) for details on calculating the marginal confidence interval and joint confidence region, as well as the expected mean response, for the Gauss-Newton nonlinear least squares methodology.

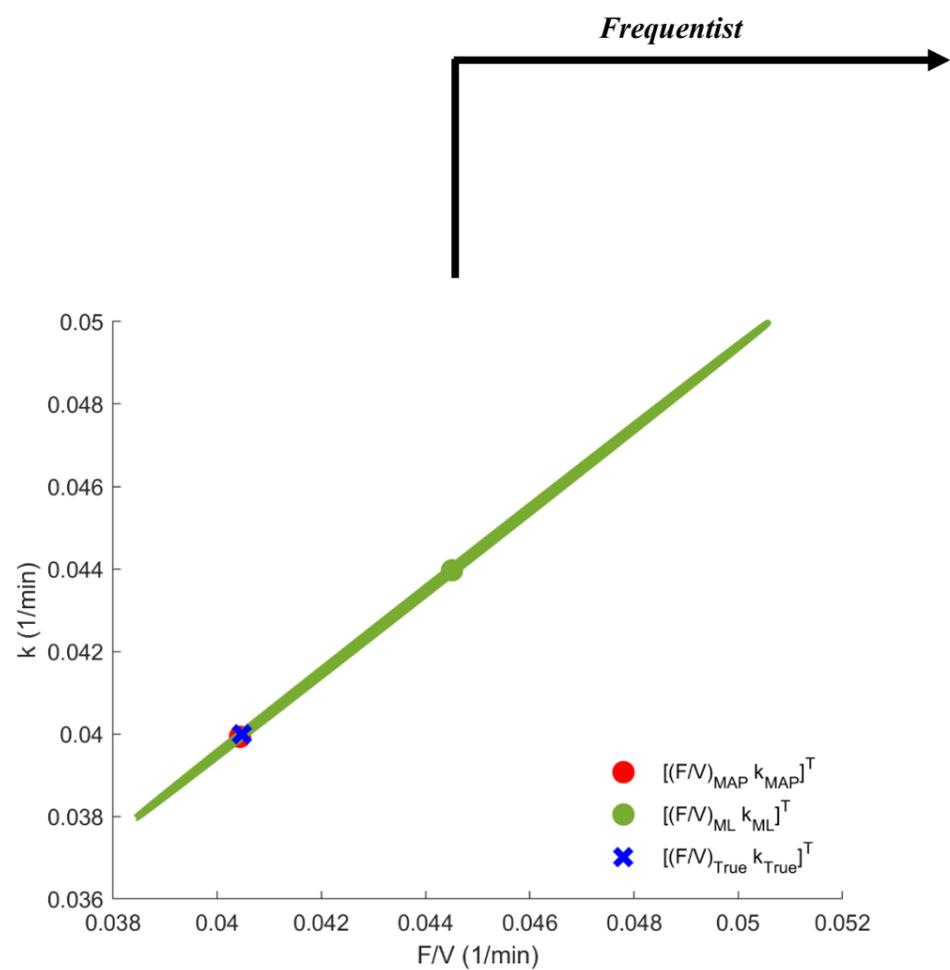


Figure 5.10: 99% joint parameter confidence (green) and credibility (red) regions (Section 2.8). The blue cross denotes the simulation ground truth. The confidence region is centered at the maximum likelihood estimate for the model parameters (Equation 3.2) while the credibility region is centered at the mean of the joint Gaussian posterior distribution (MAP estimate). The confidence and credibility regions lie above each other visually. Note that parameters k and F/V are highly correlated.

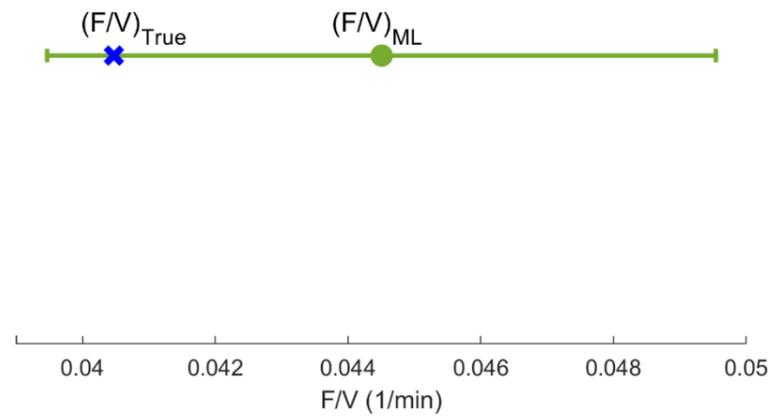
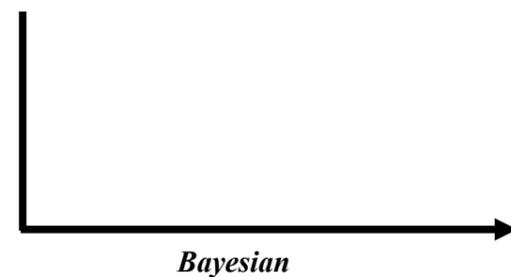


Figure 5.11: 99% confidence interval (Section 2.8) for the unknown model parameter F/V . The blue cross denotes the simulation ground truth. The confidence interval is centered at the maximum likelihood estimate for F/V .

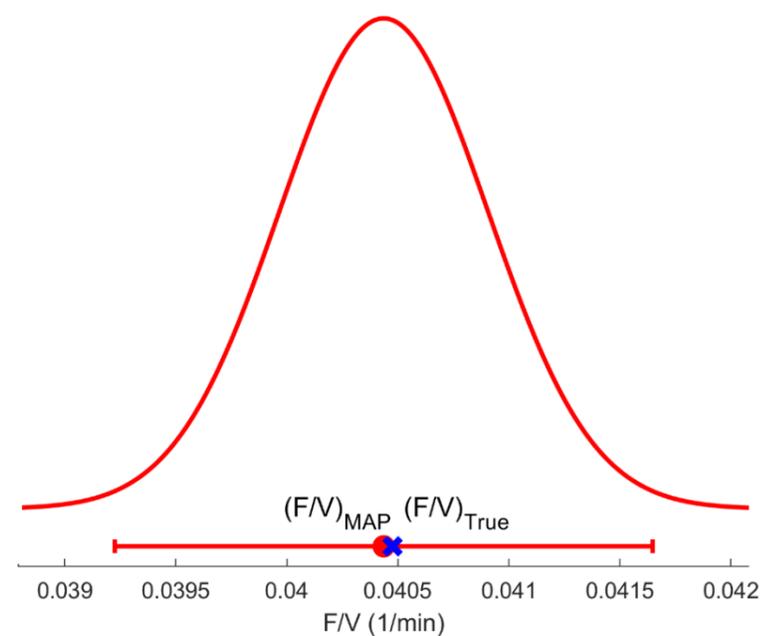


Figure 5.13: Marginal posterior distribution over the unknown model parameter F/V with the corresponding 99% credibility interval (Section 2.8). The blue cross denotes the simulation ground truth.

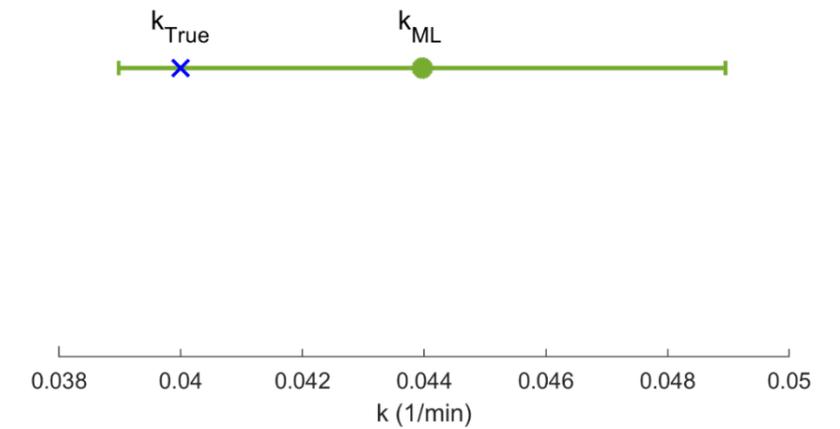


Figure 5.12: 99% confidence interval (Section 2.8) for the unknown model parameter k . The blue cross denotes the simulation ground truth. The confidence region is centered at the maximum likelihood estimate for k .

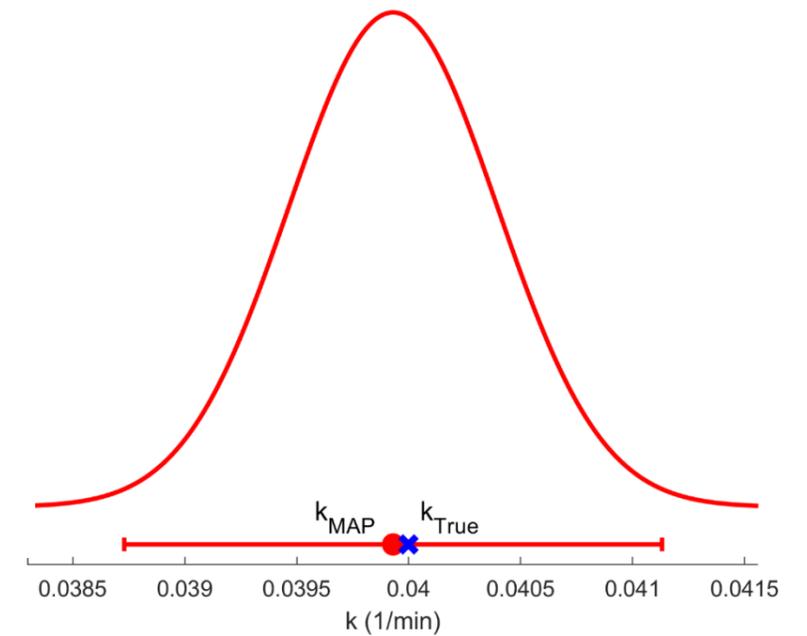


Figure 5.14: Marginal posterior distribution over the unknown model parameter k with the corresponding 99% credibility interval (Section 2.8). The blue cross denotes the simulation ground truth.

Note that the x-axis scale for Figures 5.13 and 5.14 have been adjusted to aid in the visual interpretation of the results. Furthermore, note that all results are given for a single data set manifestation.

CHAPTER 5: RESULTS AND DISCUSSION

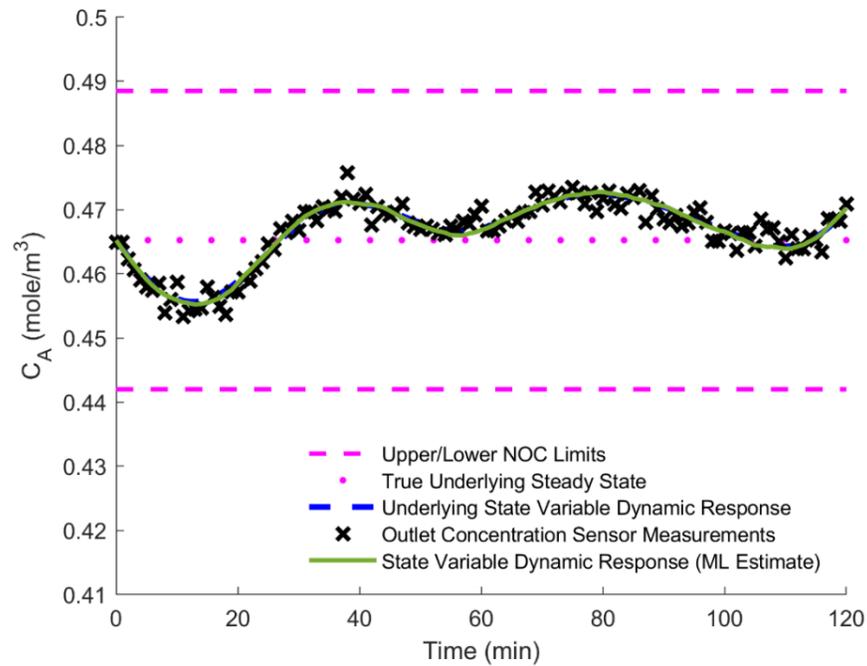


Figure 5.15: Maximum likelihood fit for the expected mean response of Equation 3.2 obtained from applying *Algorithm 2* (Section 4.4) to the synthetically generated sensor measurement data (Section 4.7.2). Note that the maximum likelihood fit track the underlying behaviour so closely that the curves lie above each other visually.

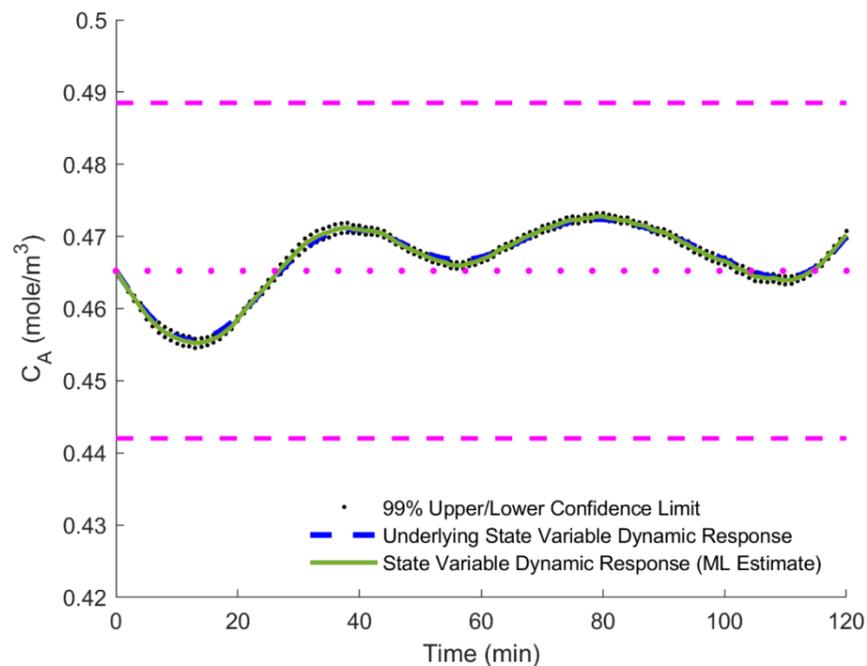


Figure 5.18: 99% confidence interval (Section 2.8) of the expected mean response of Equation 3.2. The dotted lines represent the lower and upper confidence limits.

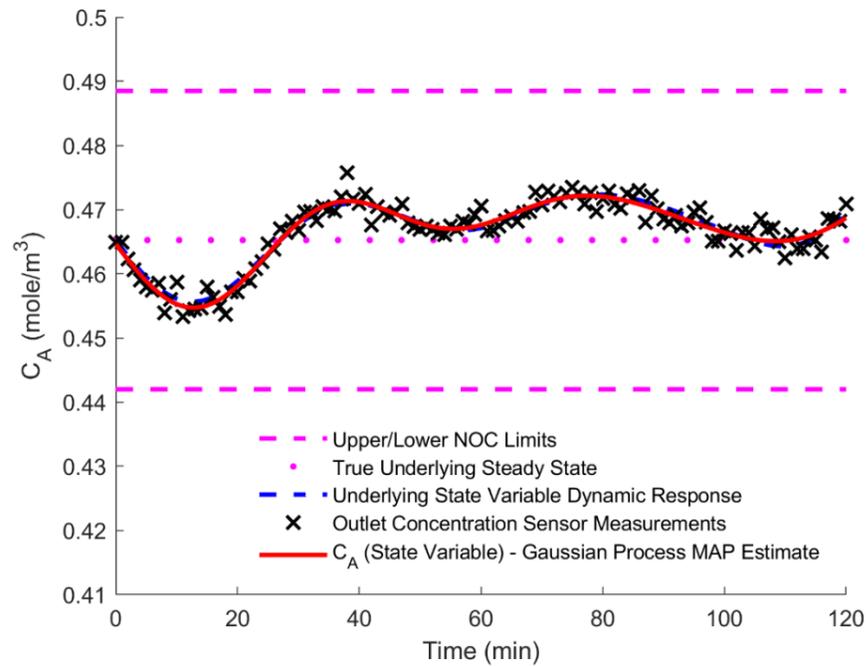


Figure 5.16: Gaussian process MAP estimate fit for the state variable C_A obtained from applying *Algorithm 4* (Section 4.5.2) to the synthetically generated sensor measurement data (Section 4.7.2). Note that the MAP fit track the underlying behaviour so closely that the curves lie above each other visually.

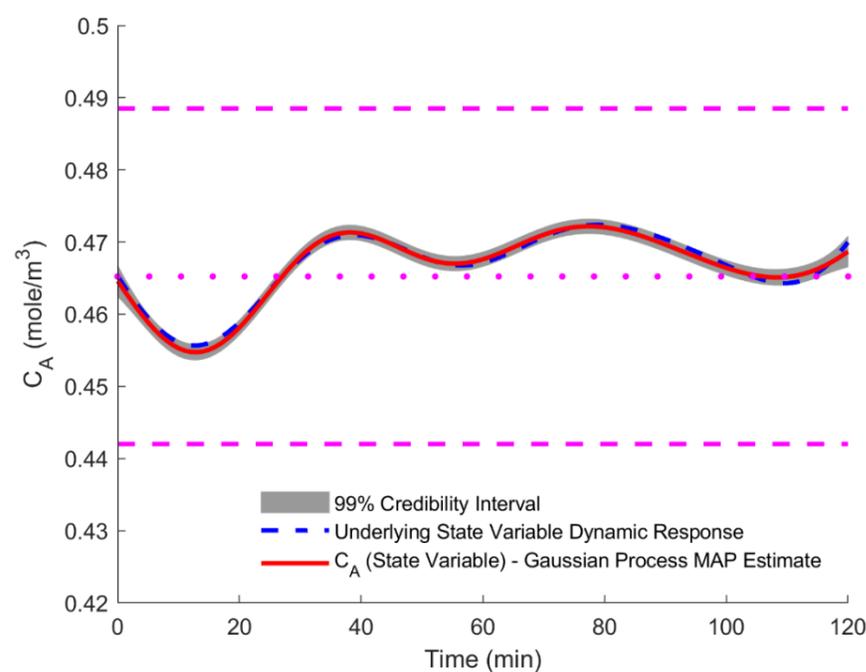


Figure 5.19: Gaussian process mean (MAP) estimate over the state variable C_A function values for Equation 3.2. The shaded region represents the 99% credibility interval (Section 2.8).

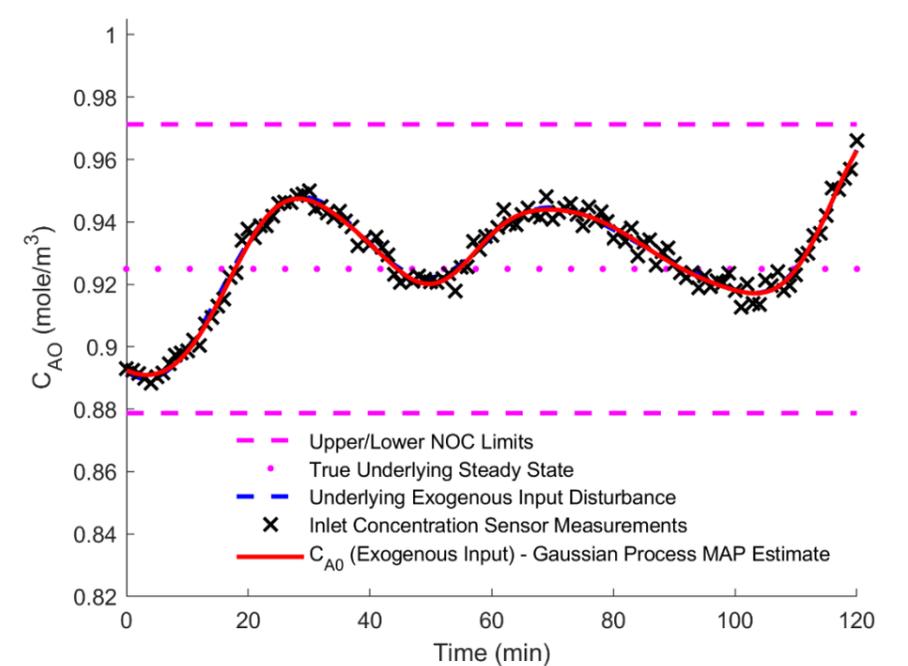


Figure 5.17: Gaussian process MAP estimate fit for the exogenous input C_{A0} obtained from applying *Algorithm 4* (Section 4.5.2) to the synthetically generated sensor measurement data (Section 4.7.2). Note that the MAP fit track the underlying behaviour so closely that the curves lie above each other visually.

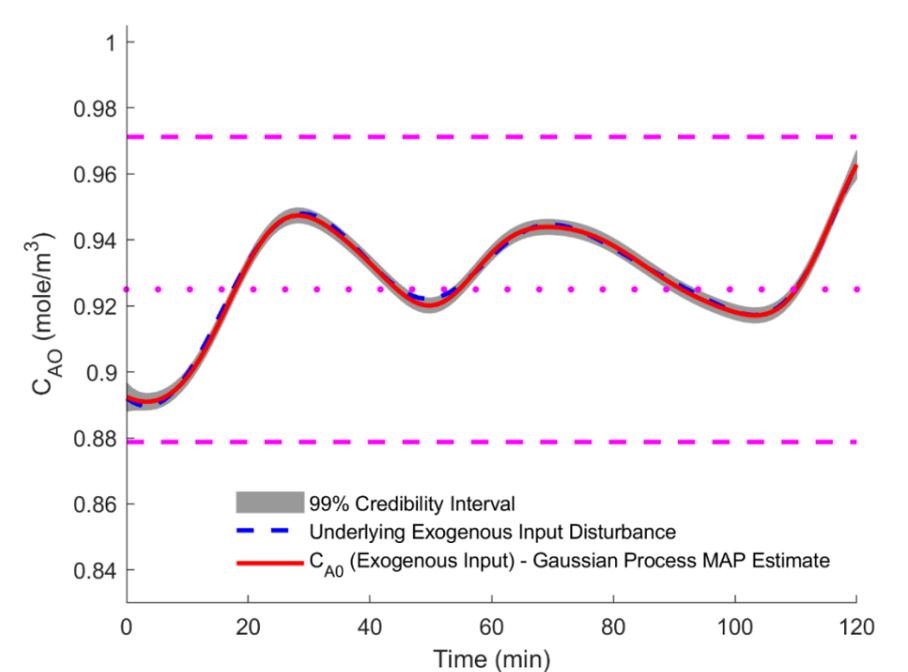


Figure 5.20: Gaussian process mean (MAP) estimate over the exogenous input C_{A0} function values for Equation 3.2. The shaded region represents the 99% credibility interval (Section 2.8).

Note that the sensor measurements have been removed from Figures 5.18, 5.19 and 5.20 to aid in the visual interpretation of the results. Furthermore, note that all results are given for a single data set manifestation.

CHAPTER 5: RESULTS AND DISCUSSION

The marginal posterior distributions associated with Figures 5.13 and 5.14 are calculated from Equation 3.86, i.e. $p(\mathbf{w} | \boldsymbol{\mu}_{C'_A} | d_{C_A}, d_{C_{A0}})$, in the exact same manner as Figures 5.4 and 5.5 by simply marginalising out the model parameter that the engineer is not interested in (Section 2.3.1). Recall that Equation 3.86 takes the form of a bivariate Gaussian distribution such that marginalisation corresponds to picking the relevant entries of the mean vector (Equation 3.87) and covariance matrix (Equation 3.88).

From Table 5.3 and Figures 5.10 through 5.20, one observes that the proposed Gaussian process based approach provides comparable results to the benchmark methodology for Case Study 2 (Equation 3.2) for a single data set manifestation. However, the proposed Bayesian methodology does provide more accurate estimates for the ordinary differential equation model parameters. Furthermore, the proposed methodology also allows the engineer to estimate the exogenous input disturbance variable C_{A0} for which the characteristics and generation process are typically not known in a practical setting. Lastly, from the proposed Bayesian methodology one can obtain a point estimate for the exogenous input disturbance sensor standard deviation parameter which is not possible with the benchmark approach.

5.2.3. Case Study 3: Liquid Draining Tank – Exogenous Input Disturbance

Table 5.5 summarises the algorithm execution time for *Algorithm 2* and *4*, as well as the inferred parameter accuracy/reliability (expressed as the percentage error against the simulation ground truth parameter values used during the synthetic data generation process for Equation 3.9). The results in Table 5.5 are presented for a single data set manifestation.

Table 5.6 summarises the results obtained from applying *Algorithm 2* and *4* to the one hundred independently generated sensor measurement data sets (Sections 4.7 and 4.8 - results are given for performance criteria 7 through 10). Observe from Table 5.6 that both the frequentist (*Algorithm 2*) and Bayesian (*Algorithm 4*) approaches provide consistent results when compared to the simulation ground parameter values. Note that only 50% of the constructed credibility intervals (one for each of the independently generated data sets) contain the ground truth parameter values of $1/A$ and k_v/A , respectively. However, the average confidence interval width of parameters $1/A$ and k_v/A is approximately 1.35 times larger than the average credibility interval width (Section 2.8). This indicates that the confidence interval estimates of $1/A$ and k_v/A , as obtained from applying *Algorithm 2*, are more conservative which explains why 74% of the constructed confidence intervals contain the simulation ground truth parameter values of $1/A$ and k_v/A , respectively.

Recall from Section 2.8 that the most popular and widely understood historical interpretation of a confidence interval follows the Neymanian understanding in which a confidence interval provides a measure of uncertainty by considering the long term frequency of repeated experiments (Neyman, 1937). In other words, if the engineer collects one hundred sensor measurement data sets from independent experiments to estimate the model parameters and constructs a 99% confidence interval for each of the parameter estimates from each data set, at least 99 of these confidence intervals are expected to contain the true (but fixed) unknown model parameter. However, from the results in Table 5.6, only 74% of the constructed confidence intervals contain the ground truth simulation values of $1/A$ and k_v/A , respectively.

CHAPTER 5: RESULTS AND DISCUSSION

Table 5.5: Summary of the algorithm execution time, inferred lumped system ODE model parameters and sensor standard deviation parameters. Note that these results are given for a single data set manifestation [*N/A - not applicable].

| Probabilistic Interpretation | Algorithm Number | Algorithm Execution Time | Unknown Parameter | Simulation Ground Truth | Inferred Value | % Error |
|------------------------------|------------------|--------------------------|----------------------|------------------------------|------------------------------|---------|
| Frequentist | 2 | 7.78 s | $1/A$ | $0.1429 m^{-2}$ | $0.1395 m^{-2}$ | 2.32% |
| | | | k_v/A | $0.0900 \frac{m^{0.5}}{min}$ | $0.0879 \frac{m^{0.5}}{min}$ | 2.30% |
| | | | $\sigma_{noise,L}$ | $0.00632 m$ | $0.00617 m$ | 2.42% |
| | | | σ_{noise,F_0} | $0.00447 \frac{m^3}{min}$ | N/A* | N/A* |
| Bayesian | 4 | 38.3 s | $1/A$ | $0.1429 m^{-2}$ | $0.1387 m^{-2}$ | 2.92% |
| | | | k_v/A | $0.0900 \frac{m^{0.5}}{min}$ | $0.0874 \frac{m^{0.5}}{min}$ | 2.90% |
| | | | $\sigma_{noise,L}$ | $0.00632 m$ | $0.00622 m$ | 1.67% |
| | | | σ_{noise,F_0} | $0.00447 \frac{m^3}{min}$ | $0.00459 \frac{m^3}{min}$ | 2.59% |

Table 5.6: Summary of the mean inferred parameter estimates, standard deviation, proportion of confidence/credibility intervals containing the simulation ground truth parameter values and average confidence/credibility interval width for the one hundred independently generated sensor measurement data sets (results are given for performance criteria 7 through 10, Section 4.8).

| Algorithm Number | Unknown Parameter | Simulation Ground Truth | Mean Inferred Value | Standard Deviation | Percentage Contained* | Average Interval Width |
|------------------|-------------------|------------------------------|------------------------------|------------------------------|-----------------------|------------------------------|
| 2 | $1/A$ | $0.1429 m^{-2}$ | $0.1431 m^{-2}$ | $0.0039 m^{-2}$ | 74 % | $0.0093 m^{-2}$ |
| | k_v/A | $0.0900 \frac{m^{0.5}}{min}$ | $0.0901 \frac{m^{0.5}}{min}$ | $0.0025 \frac{m^{0.5}}{min}$ | 74 % | $0.0058 \frac{m^{0.5}}{min}$ |
| 4 | $1/A$ | $0.1429 m^{-2}$ | $0.1407 m^{-2}$ | $0.0040 m^{-2}$ | 50 % | $0.0069 m^{-2}$ |
| | k_v/A | $0.0900 \frac{m^{0.5}}{min}$ | $0.0886 \frac{m^{0.5}}{min}$ | $0.0025 \frac{m^{0.5}}{min}$ | 50 % | $0.0043 \frac{m^{0.5}}{min}$ |

*Percentage Contained refers to the proportion of the simulation ground truth parameter values that fall within the 99% marginal parameter confidence/credibility interval.

Figures 5.21 through 5.31 visually depict performance criteria 3 to 5 (Section 4.8) for *Algorithms 2* and *4* applied to a single sensor measurement data set.

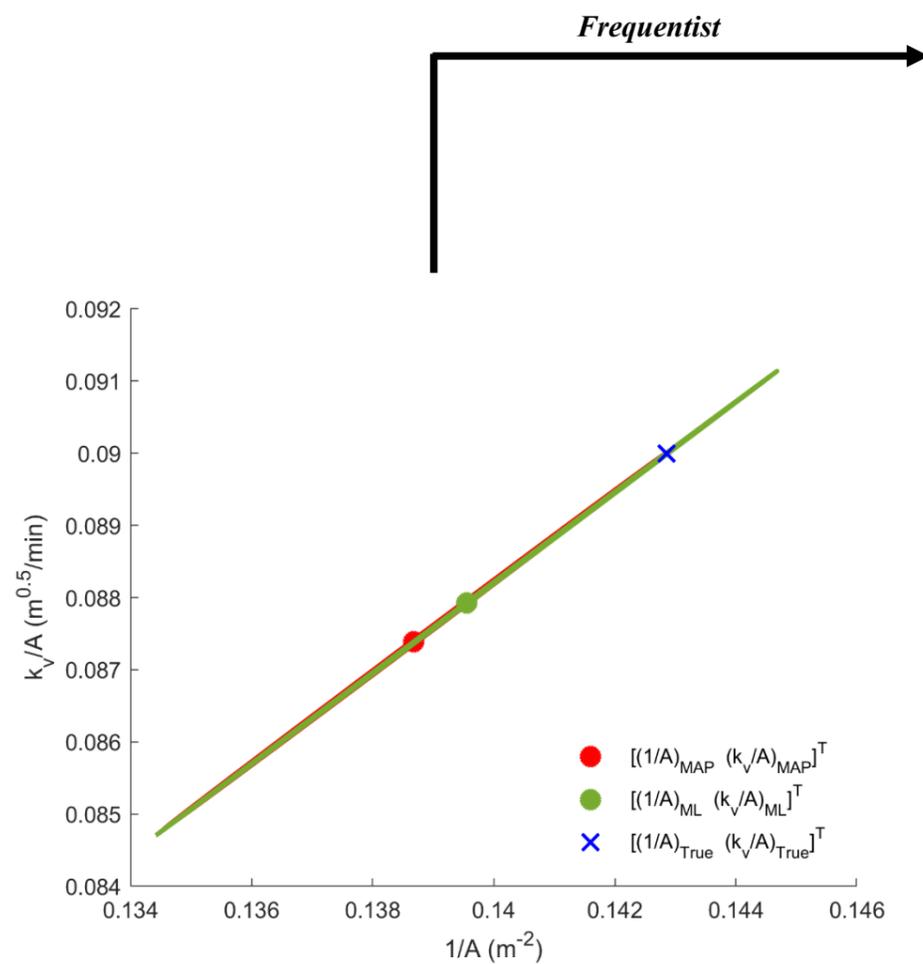


Figure 5.21: 99% joint parameter confidence (green) and credibility (red) regions (Section 2.8). The blue cross denotes the simulation ground truth. The confidence region is centered at the maximum likelihood estimate for the model parameters (Equation 3.9) while the credibility region is centered at the mean of the joint Gaussian posterior distribution (MAP estimate). The confidence and credibility regions lie above each other visually. Note that parameters k_v/A and $1/A$ are highly correlated.

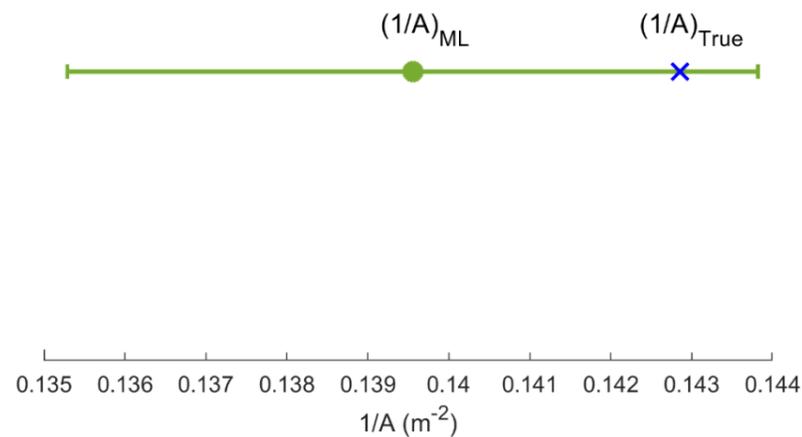


Figure 5.22: 99% confidence interval (Section 2.8) for the unknown model parameter $1/A$. The blue cross denotes the simulation ground truth. The confidence interval is centered at the maximum likelihood estimate for $1/A$.

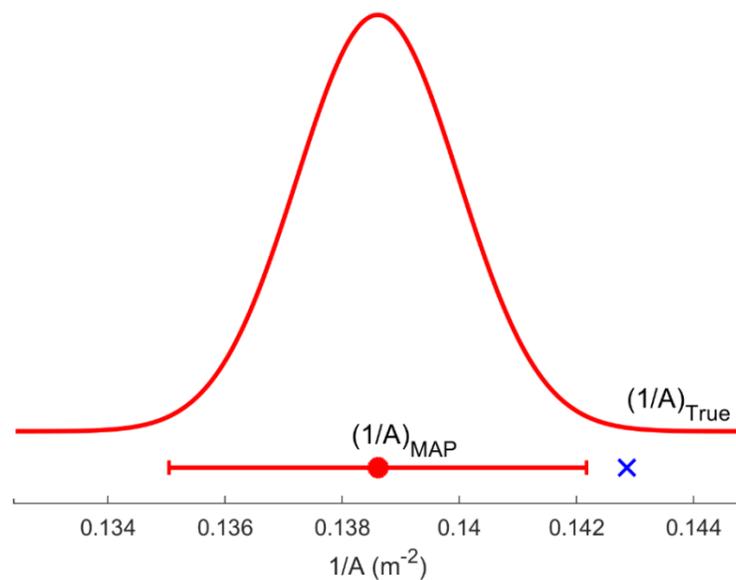


Figure 5.24: Marginal posterior distribution over the unknown model parameter $1/A$ with the corresponding 99% credibility interval (Section 2.8). The blue cross denotes the simulation ground truth.

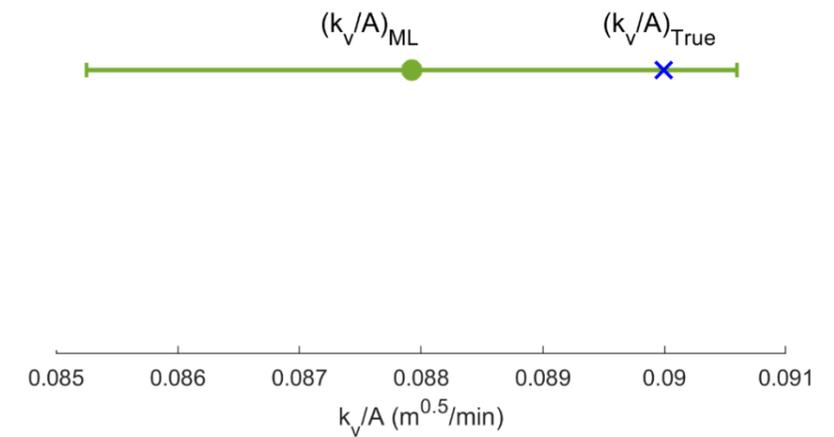


Figure 5.23: 99% confidence interval (Section 2.8) for the unknown model parameter k_v/A . The blue cross denotes the simulation ground truth. The confidence interval is centered at the maximum likelihood estimate for k_v/A .

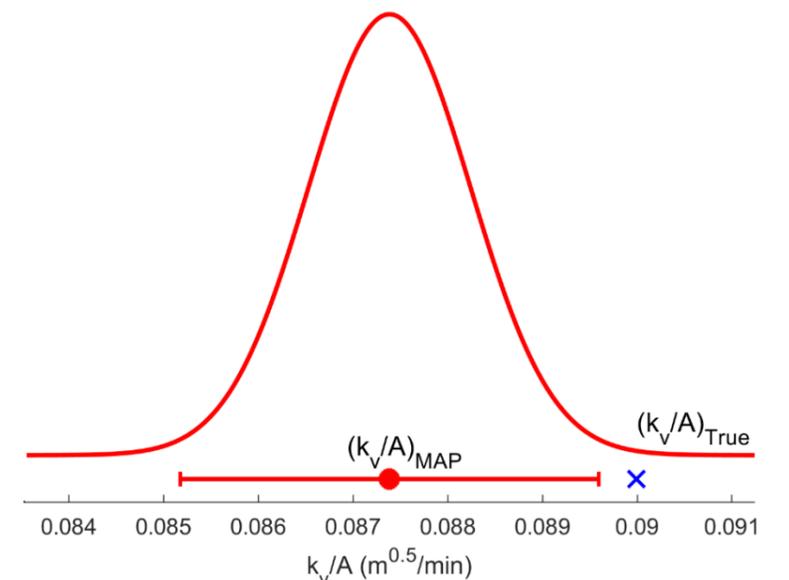


Figure 5.25: Marginal posterior distribution over the unknown model parameter k_v/A with the corresponding 99% credibility interval (Section 2.8). The blue cross denotes the simulation ground truth.

Note that the x-axis scale for Figures 5.24 and 5.25 have been adjusted to aid in the visual interpretation of the results. Furthermore, note that all results are given for a single data set manifestation.

CHAPTER 5: RESULTS AND DISCUSSION

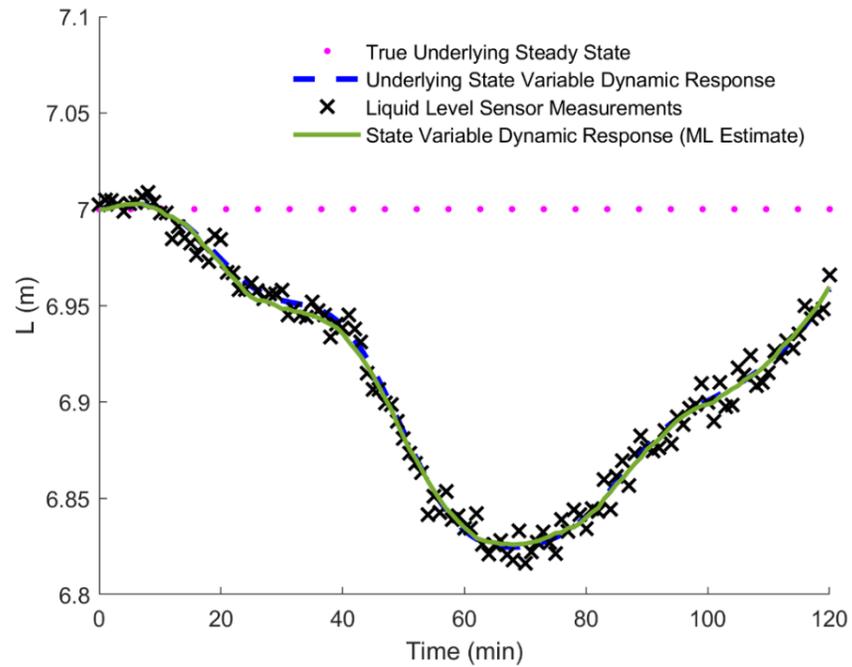


Figure 5.26: Maximum likelihood fit for the expected mean response of Equation 3.9 obtained from applying *Algorithm 2* (Section 4.4) to the synthetically generated sensor measurement data (Section 4.7.3). Note that the maximum likelihood fit track the underlying behaviour so closely that the curves lie above each other visually.

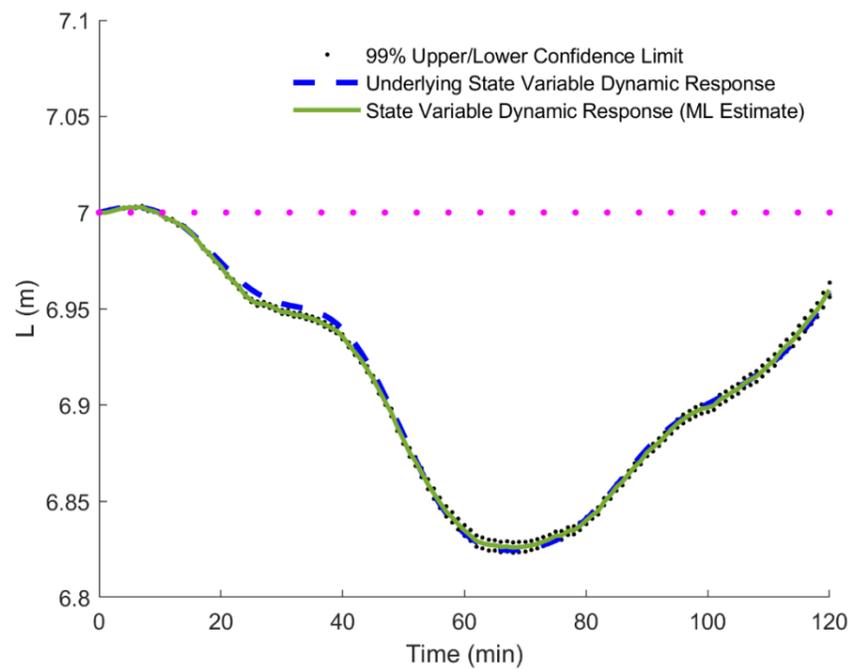


Figure 5.29: 99% confidence interval (Section 2.8) of the expected mean response of Equation 3.9. The dotted lines represent the lower and upper confidence limits.

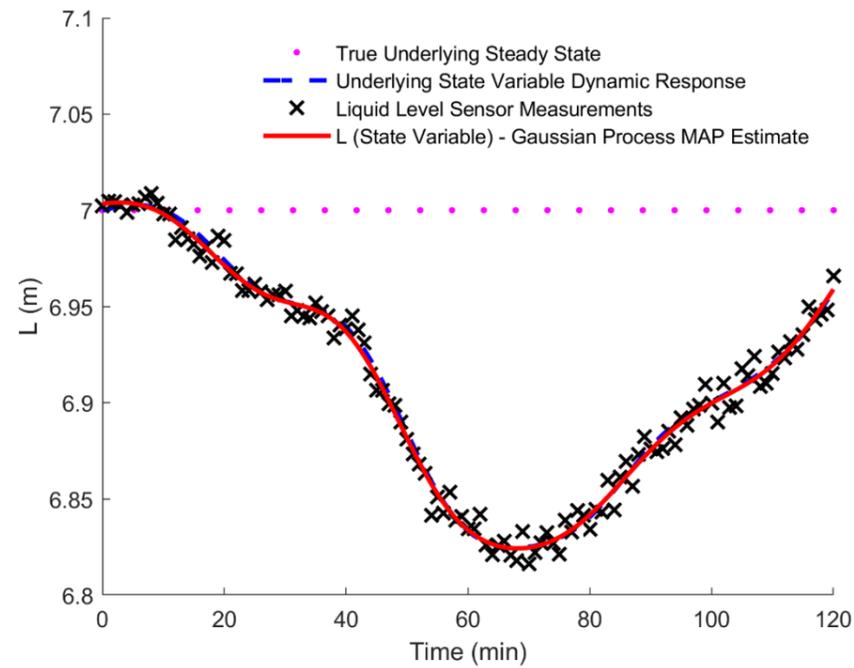


Figure 5.27: Gaussian process MAP estimate fit for the state variable L obtained from applying *Algorithm 4* (Section 4.5.2) to the synthetically generated sensor measurement data (Section 4.7.3). Note that the MAP fit track the underlying behaviour so closely that the curves lie above each other visually.

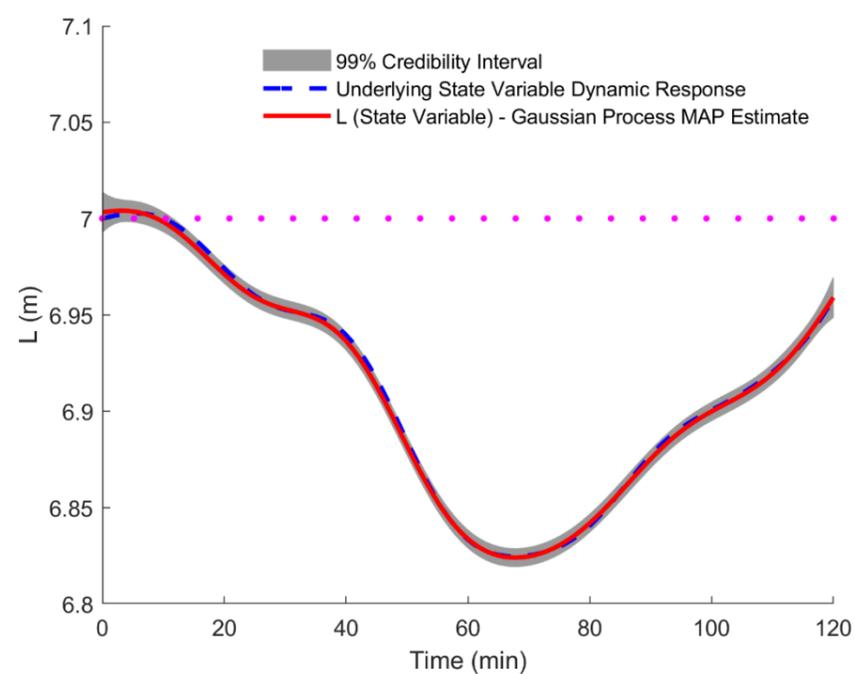


Figure 5.30: Gaussian process mean (MAP) estimate over the state variable L function values for Equation 3.9. The shaded region represents the 99% credibility interval (Section 2.8).

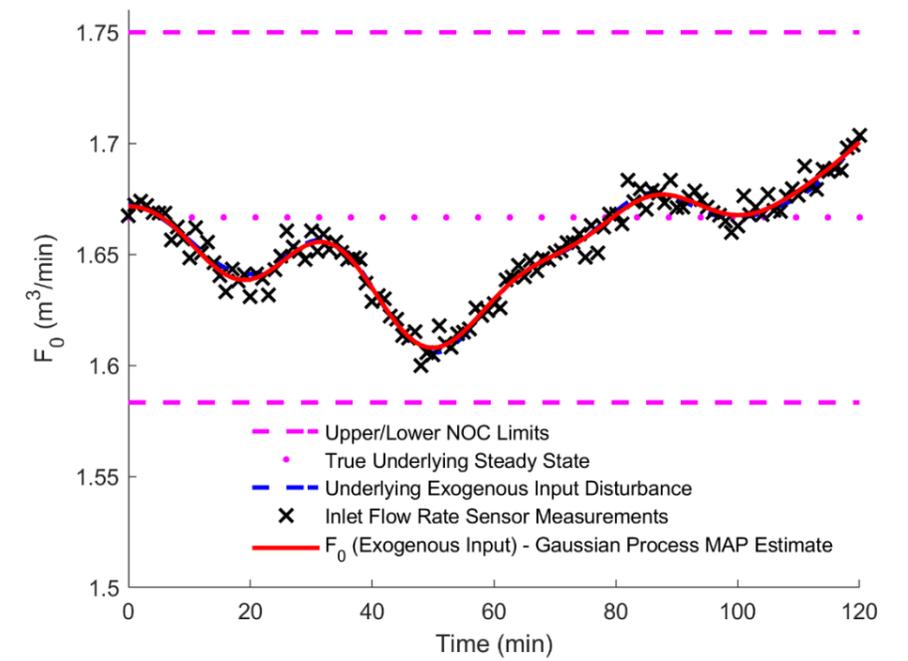


Figure 5.28: Gaussian process MAP estimate fit for the exogenous input F_0 obtained from applying *Algorithm 4* (Section 4.5.2) to the synthetically generated sensor measurement data (Section 4.7.3). Note that the MAP fit track the underlying behaviour so closely that the curves lie above each other visually.

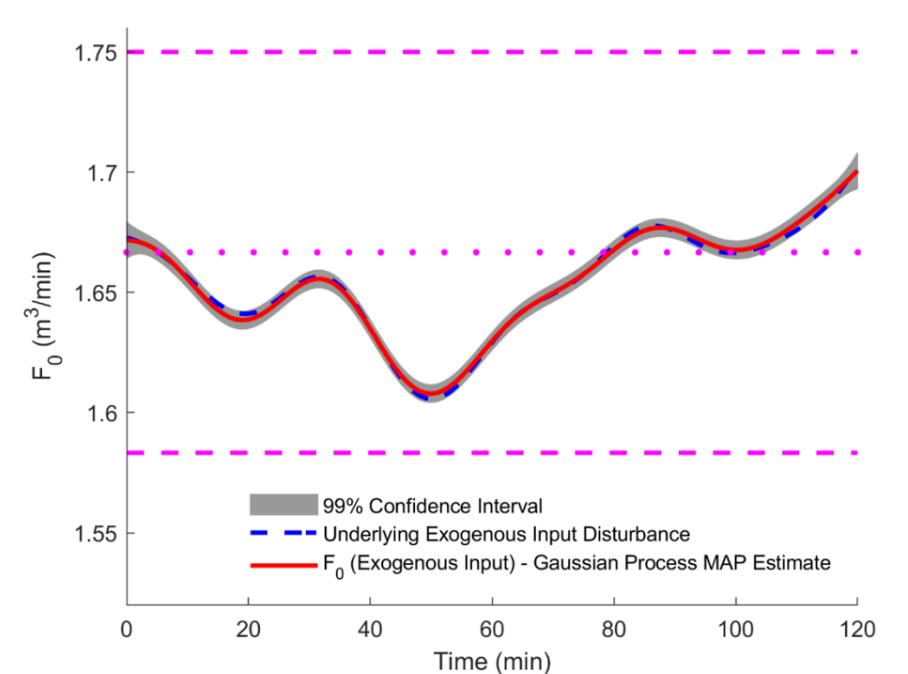


Figure 5.31: Gaussian process mean (MAP) estimate over the exogenous input F_0 function values for Equation 3.9. The shaded region represents the 99% credibility interval (Section 2.8).

Note that the liquid level NOC limits have been excluded from Figures 5.26, 5.27, 5.29 and 5.30, as well as the sensor measurements have been removed from Figures 5.29, 5.30 and 5.31, to aid in the visual interpretation of the results. Furthermore, note that all results are given for a single data set manifestation.

CHAPTER 5: RESULTS AND DISCUSSION

From Table 5.5 and Figures 5.21 through 5.31, one observes that the proposed Gaussian process based approach provides comparable results to the benchmark methodology for Case Study 3 (Equation 3.9) for a single data set manifestation. The 99% credibility interval of $1/A$ and k_v/A does not contain the simulation ground truth parameter values and the proposed Bayesian methodology takes longer to execute. However, the Bayesian methodology has an additional benefit in that it allows the engineer to estimate the exogenous input disturbance variable F_0 for which the characteristics and generation process is typically not known in a practical setting, while simultaneously producing a point estimate for the exogenous input disturbance sensor standard deviation parameter which is not possible with the benchmark approach.

5.3. Parameter Tracking Application Results

This section focuses on applying the proposed Gaussian process based Bayesian methodology outlined in Section 4.5.2 to the extended case studies presented in Sections 4.6.1 and 4.6.2, respectively, as well as the cost-benefit analysis associated with engineering decision making under uncertainty (Section 4.8 - performance criterion 6).

5.3.1. Extended Case Study 2: Isothermal CSTR with Catalyst Decay

Recall from Section 4.6.1 that the CSTR catalyst must be replaced when the proxy parameter k reaches a value of 0.0320 min^{-1} . However, the parameter k_0 is unknown. Algorithm 4, i.e. the Gaussian process based Bayesian methodology, can be applied to the ordinary differential equation model given by Equation 3.91, using the synthetically generated sensor measurement data (Section 4.7.2), to obtain a posterior distribution $p(\mathbf{w} | \boldsymbol{\mu}_{C_A}, \mathbf{d}_{C_A}, \mathbf{d}_{C_{A_0}})$ over the unknown ODE model parameters $\mathbf{w} = [F/V \ k_0]^T$. The engineer is specifically interested in the parameter k_0 since it can be used in combination with Equation 3.89 to predict the CSTR catalyst replacement time. One can obtain a marginal posterior distribution over k_0 by simply marginalising out the parameter F/V (Section 2.3.1) such that:

$$p(k_0) = \mathcal{N}(k_0 | [\mathbf{m}_N]_2, [\mathbf{S}_N]_{22}) \quad \text{Equation 4.10}$$

Since k_0 is Gaussian distributed the catalyst decay model output (Equation 3.89), which is a linear function of k_0 , is also Gaussian distributed. Thus, the engineer only requires the mean and covariance of $k(t_i, k_0)$ at the input time point t_i (Bishop, 2006). The mean can readily be obtained by evaluating the expected value (Sections 2.1.4 and 2.3.1) of $k(t_i, k_0)$ under the marginal posterior distribution $p(k_0)$ such that:

$$\mathbb{E}_{p(k_0)}[k|t_i] = \int_{-\infty}^{\infty} p(k_0) k_0 \exp\{-k_d t_i\} dk_0 \quad \text{Equation 4.11}$$

$$\mathbb{E}_{p(k_0)}[k|t_i] = [\mathbf{m}_N]_2 \exp\{-k_d t_i\} \quad \text{Equation 4.12}$$

The variance can readily be obtained by evaluating the expected values (Sections 2.1.4 and 2.3.1) given by Equation 4.13 under the marginal posterior distribution $p(k_0)$ such that:

CHAPTER 5: RESULTS AND DISCUSSION

$$\text{Var}[k|t_i] = \mathbb{E}_{p(k_0)}[(k|t_i)^2] - (\mathbb{E}_{p(k_0)}[k|t_i])^2 \quad \text{Equation 4.13}$$

$$\text{Var}[k|t_i] = \int_{-\infty}^{\infty} p(k_0) (k_0 \exp\{-k_d t\})^2 dk_0 - ([\mathbf{m}_N]_2 \exp\{-k_d t_i\})^2 \quad \text{Equation 4.14}$$

$$\text{Var}[k|t_i] = [\mathbf{S}_N]_{22} \exp\{-2k_d t_i\} \quad \text{Equation 4.15}$$

As a result, the probability of observing k at an arbitrary input time point t_i is given by:

$$p(k|t_i) = \mathcal{N}(k | [\mathbf{m}_N]_2 \exp\{-k_d t_i\}, [\mathbf{S}_N]_{22} \exp\{-2k_d t_i\}) \quad \text{Equation 4.16}$$

The result for Equation 4.16 is visually depicted in Figure 5.32.

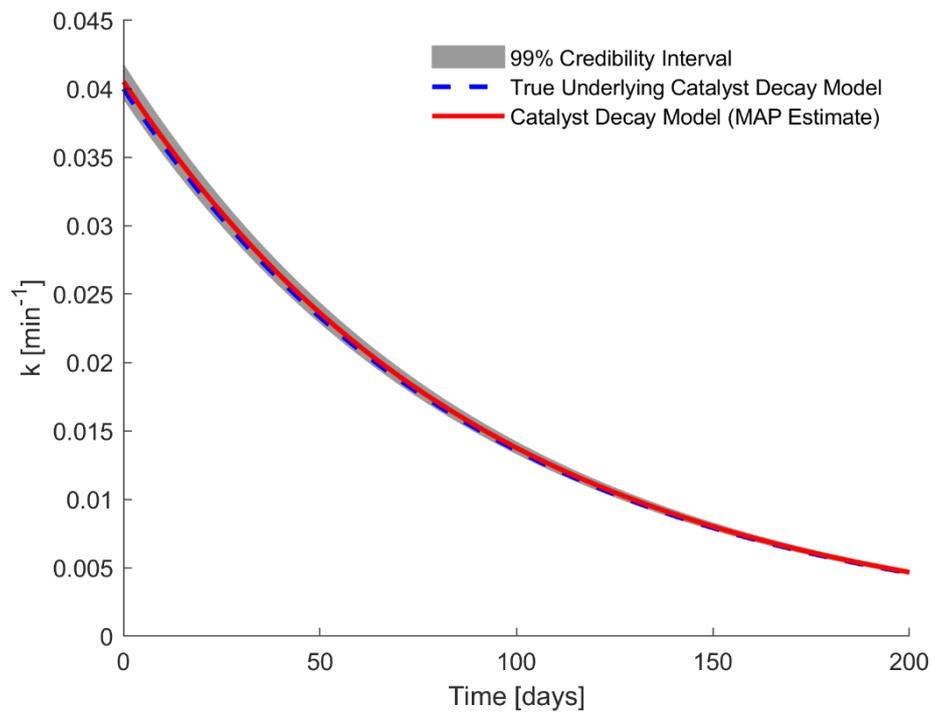


Figure 5.32: Catalyst decay expected mean response over a 200 day degradation period (*Algorithm 4* results). The shaded region represents the 99% credibility interval (Section 2.8). The results are given for a single data set manifestation.

The catalyst decay expected mean response over the 200 day degradation period obtained from applying the benchmark methodology (*Algorithm 2*), in conjunction with the ODE model given by Equation 3.91 and the synthetically generated sensor measurement data (Section 4.7.2), is visually depicted in Figure 5.33.

CHAPTER 5: RESULTS AND DISCUSSION

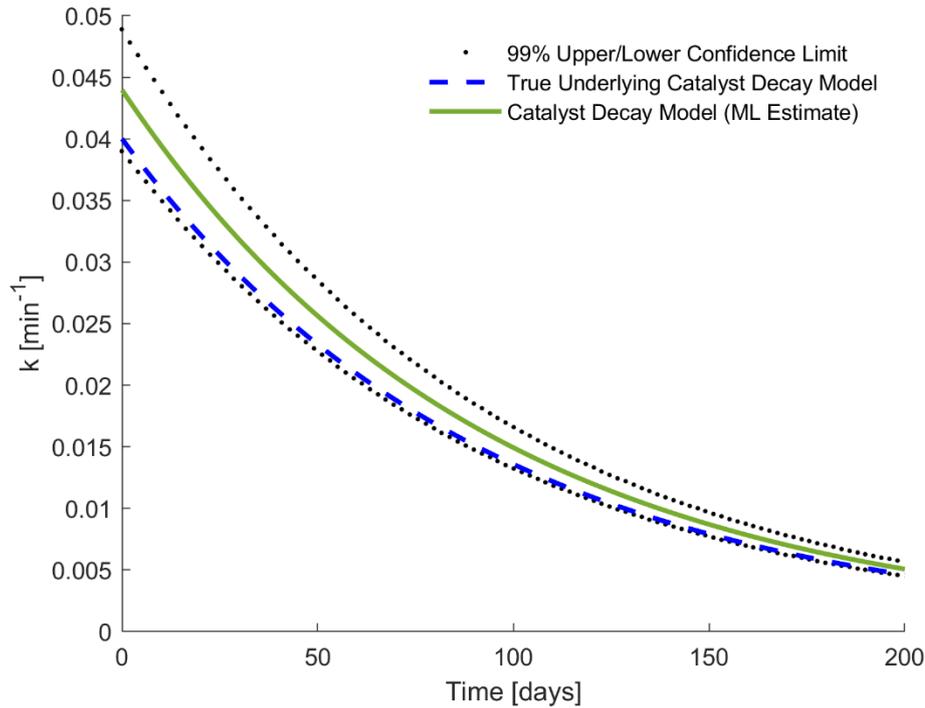


Figure 5.33: Catalyst decay expected mean response over a 200 day degradation period (*Algorithm 2* results). The dotted line represents the lower and upper confidence limits (Section 2.8). The results are given for a single data set manifestation.

Table 5.7 summarises the inference results for parameter k_0 obtained from *Algorithms 2* and *4*, respectively, as applied to the synthetically generated sensor measurement data. The results in Table 5.7 are presented for a single data set manifestation.

Table 5.7: Summary of the inferred lumped system ordinary differential equation model parameter k_0 . Note that these results are given for a single data set manifestation.

| Probabilistic Interpretation | Algorithm Number | Simulation Ground Truth | Inferred Value | % Error |
|------------------------------|------------------|---------------------------|---------------------------|---------|
| Frequentist | 2 | 0.0400 min^{-1} | 0.0440 min^{-1} | 9.92% |
| Bayesian (MAP estimate) | 4 | 0.0400 min^{-1} | 0.0407 min^{-1} | 1.77% |

Table 5.8 summarises the results obtained from applying *Algorithm 2* and *4* to the one hundred independently generated sensor measurement data sets (Sections 4.7 and 4.8 - results are given for performance criteria 7 through 10).

CHAPTER 5: RESULTS AND DISCUSSION

Table 5.8: Summary of the mean inferred parameter estimates, standard deviation, proportion of confidence/credibility intervals containing the simulation ground truth parameter values and average confidence/credibility interval width for the one hundred independently generated sensor measurement data sets. The units for all numerical entries in Table 5.8 are min^{-1} (excluding the Percentage Contained* column).

| Algorithm Number | Unknown Parameter | Simulation Ground Truth | Mean Inferred Value | Standard Deviation | Percentage Contained* | Average Interval Width |
|------------------|-------------------|-------------------------|---------------------|--------------------|-----------------------|------------------------|
| 2 | k_0 | 0.0400 | 0.0397 | 0.0020 | 100 % | 0.0085 |
| 4 | k_0 | 0.0400 | 0.0386 | 0.0023 | 30 % | 0.0024 |

*Percentage Contained refers to the proportion of the simulation ground truth parameter values that fall within the 99% marginal parameter confidence/credibility interval.

Observe from Table 5.8 that both the frequentist (*Algorithm 2*) and Bayesian (*Algorithm 4*) approaches provide consistent results when compared to the simulation ground parameter values. Note that 100% of the constructed frequentist confidence intervals (one for each of the independently generated sensor data sets) contain the simulation ground truth parameter value of k_0 whereas only 30% of the Bayesian credibility intervals contain the ground truth k_0 value. However, the average confidence interval width is approximately 3.5 times larger than the average credibility interval width. This indicates that the confidence interval estimates for k_0 , as obtained from *Algorithm 2*, are very conservative which explains why 100% of the constructed confidence intervals contain the true simulation value of k_0 .

Table 5.9 summarises the corresponding catalyst replacement times, as extracted from Figures 5.32 and 5.33, respectively, when the parameter proxy k reaches as value of 0.0320 min^{-1} . Note that the entries in Table 5.9 are relative to the end time point at which the regression procedure was performed. For the current case study, the sensor measurement data set was collected over a 120 minute period. Hence, all catalyst replacement time values reported in Table 5.9 are relative to the 120 minute regression mark. The corresponding value in brackets present the day of replacement relative to a 30 day period starting at day 0 where the start of day zero coincides with the first sensor measurement in the data set.

Table 5.9: Isothermal, constant volume CSTR catalyst replacement time obtained from applying *Algorithms 2* and *4*, respectively. Note that these results are given for a single data set manifestation.

| Probabilistic Interpretation | Lower Limit Catalyst Replacement Time | Mean Catalyst Replacement Time | Upper Limit Catalyst Replacement Time | Ground Truth Catalyst Replacement Time |
|------------------------------------|---------------------------------------|--------------------------------|---------------------------------------|--|
| Frequentist (<i>Algorithm 2</i>) | 18.30 days (18.38) | 29.34 days (29.42) | 39.21 days (39.29) | 20.58 days (20.66) |
| Bayesian (<i>Algorithm 4</i>) | 19.13 days (19.21) | 22.07 days (22.15) | 24.91 days (24.99) | 20.58 days (20.66) |

CHAPTER 5: RESULTS AND DISCUSSION

From Table 5.9 one observes that both the benchmark (frequentist) and the proposed Bayesian methodology contain the ground truth catalyst replacement time within the expected mean response lower and upper limits (Figure 5.32 and 5.33). However, the proposed Bayesian methodology provides a narrower interval between the lower and upper catalyst replacement limits. From an engineering perspective, if one has to choose a catalyst replacement time based on the results outlined in Table 5.9, the obvious engineering choice might be to replace the catalyst at the mean catalyst replacement time. If one follows the frequentist methodology, the catalyst should be replaced 29.34 days after performing the regression analysis (*Algorithm 2*) which is approximately 8.76 days later than the ground truth catalyst replacement time.

If the engineer decides on following the Bayesian methodology, the catalyst should be replaced 22.07 days after performing the regression analysis (*Algorithm 4*) which is approximately 1.50 days later than the ground truth catalyst replacement time. The problem that arises is that the selection of the catalyst replacement time has an effect on the profit associated with the CSTR process unit. If one replaces the catalyst too early, more money is spent on purchasing new catalyst and the existing active catalyst will be discarded prematurely. If the catalyst is replaced too late, the engineer is at risk of fouling/deactivating the catalyst, producing a product with a lower quality which in turn results in less income if the product does not meet the market-imposed quality specifications.

Based on the results outlined in Table 5.9 and Figures 5.32 and 5.33, it is difficult to select a single catalyst replacement time, regardless of the inference methodology used, especially given the effect of the decision making process on the cost associated with the isothermal CSTR process unit. In other words, there is a cost-benefit trade-off associated with the catalyst replacement time. Since the aim of any industrial process is to make as much profit as possible, while incurring the lowest operating and maintenance cost, the engineer can consider the effect of the estimated parameter k_0 on the isothermal CSTR process unit profit function to aid in the catalyst replacement time decision making process. Given that the assumed CSTR profit function (Equation 3.92) sufficiently describes the profit associated with the process unit, the engineer is interested in finding the time point t at which the maximum profit is reached.

From the frequentist perspective (*Algorithm 2*), the obvious choice would be to use $k_{0,ML}$ as a plug-in point estimate for Equation 3.92, followed by finding the time point t at which the CSTR profit function reaches a maximum value.

In other words,

$$t_F = \underset{t}{\operatorname{argmax}} [P_{CSTR}(t, k_{0,ML})] \quad \text{Equation 4.17}$$

With some algebraic manipulation, the engineer can show that, based on the plug-in point estimate $k_{0,ML}$, the frequentist methodology (*Algorithm 2*) provides the maximum profit time point at:

$$t_F = \frac{1400 - 5000 \left((k_{0,ML})^2 - 0.64k_{0,ML} + 0.0128 \right)}{100} \quad \text{Equation 4.18}$$

From the Bayesian perspective, instead of using a plug-in point estimate for k_0 , the engineer would like to evaluate the expected value of $P_{CSTR}(t, k_0)$ under the marginal posterior distribution $p(k_0)$ (Equation 4.10) followed by finding the time point t at which the CSTR profit function reaches a maximum value. In other words,

CHAPTER 5: RESULTS AND DISCUSSION

$$t_B = \operatorname{argmax}_t [\mathbb{E}_{p(k_0)}(P_{CSTR}(t, k_0))] \quad \text{Equation 4.19}$$

$$t_B = \operatorname{argmax}_t \left[\int_{-\infty}^{\infty} p(k_0) P_{CSTR}(t, k_0) dk_0 \right] \quad \text{Equation 4.20}$$

With some algebraic manipulation, the engineer can show that the Bayesian methodology (*Algorithm 4*) provides the maximum profit time point at:

$$t_B = \frac{1400 - 5000(([\mathbf{m}_N]_2)^2 + [\mathbf{S}_N]_{22} - 0.64([\mathbf{m}_N]_2) + 0.0128)}{100} \quad \text{Equation 4.21}$$

From Equation 4.19 through 5.21, observe that instead of using a plug-in point estimate for k_0 , the engineer is averaging over all values of k_0 supported by the marginal posterior distribution explicitly incorporating the uncertainty associated with the parameter k_0 . This is reflected by the presence of the marginal Gaussian posterior distribution $p(k_0)$ variance parameter $[\mathbf{S}_N]_{22}$ (*Algorithm 4*) in Equation 4.21 (cf. Equation 4.18). Table 5.10 outlines the time points at which the maximum profit is reached as obtained by evaluating Equations 5.18 (frequentist) and 5.21 (Bayesian), respectively. The results outlined in Table 5.10 are relative to a 30 day period starting at day zero. In other words, for a reported value of 20.70, it implies that the maximum profit time point is reached on day 20.70 within the 30 day period starting at day zero where day zero coincides with the first sensor measurement in the data set.

Figure 5.34 visually depicts the time points of maximum profit against the ground truth maximum profit time point. Note that the results outlined in Table 5.10 and Figure 5.34 are given for a single data set manifestation.

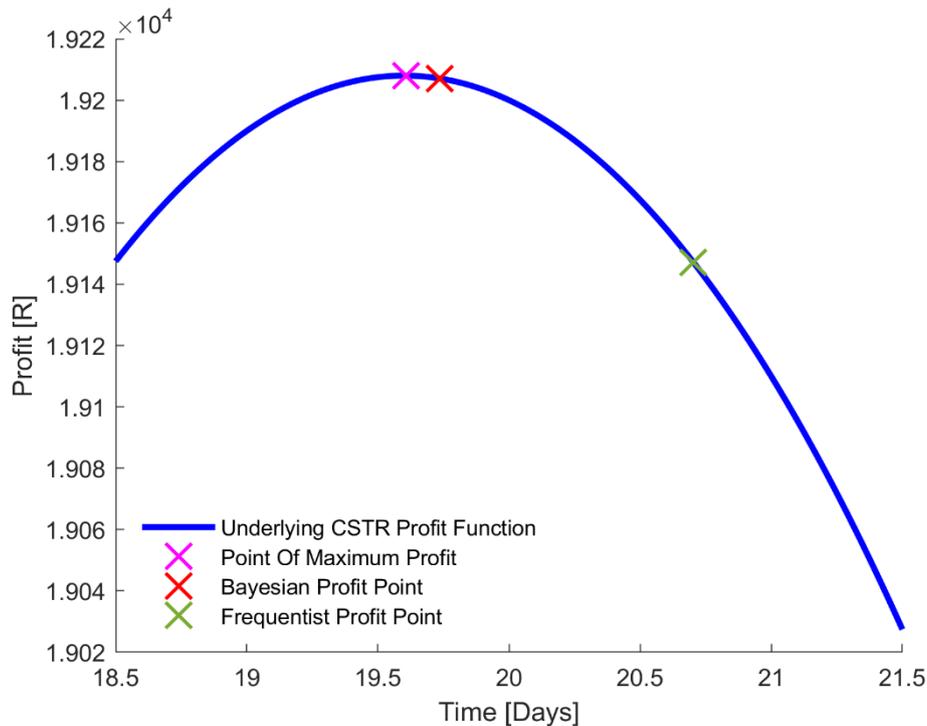


Figure 5.34: Visual depiction of the true underlying CSTR profit function given by Equation 3.92 with $k_0 = 0.040 \text{ min}^{-1}$ (ground truth value).

CHAPTER 5: RESULTS AND DISCUSSION

Table 5.10: Point of maximum profit for the isothermal CSTR profit function $P_{CSTR}(t, k_0)$ as obtained from applying *Algorithms 2* and *4*, respectively, to estimate parameter k_0 .

| Probabilistic Interpretation | Algorithm Number | Maximum Profit Time Point | Predicted Maximum Profit [R] | Ground Truth Maximum Profit Time Point | Ground Truth Maximum Profit [R] |
|------------------------------|------------------|---------------------------|------------------------------|--|---------------------------------|
| Frequentist | 2 | 20.70 | 19147 | 19.61 | 19208 |
| Bayesian | 4 | 19.80 | 19206 | 19.61 | 19208 |

From Tables 5.9 and 5.10, observe that the maximum profit time point obtained from both the frequentist and Bayesian methodologies fall between the lower and upper catalyst replacement limits. Table 5.11 summarises the reduction in profit associated with selecting the frequentist and Bayesian maximum profit time points as well as the reduction in profit associated with replacing the catalyst at the lower, mean and upper replacement time points (Table 5.9).

Table 5.11: Isothermal, constant volume CSTR cost-benefit analysis results obtained from applying *Algorithms 2* and *4*, respectively (results given for a single data set manifestation).

| Probabilistic Interpretation | Reduction In Profit At Maximum Time Point [R] | Underlying Catalyst Decay At Maximum Profit Time Point | Reduction In Profit At: [R] | | |
|------------------------------|---|--|----------------------------------|---------------------------------|----------------------------------|
| | | | Lower Catalyst Replacement Limit | Mean Catalyst Replacement Limit | Upper Catalyst Replacement Limit |
| Frequentist | 61 | 0.0318 min^{-1} | 74 | 4822 | 19385 |
| Bayesian | 2 | 0.0323 min^{-1} | 8 | 325 | 1453 |

From Table 5.11 one observes that if the engineer follows the frequentist decision rule (Equation 4.17) in conjunction with the plug-in point estimate $k_{0,ML}$ obtained from *Algorithm 2*, then the reduction in profit associated with the CSTR profit function $P_{CSTR}(t_F, k_{0,ML})$ is R61 relative to the ground truth profit. Furthermore, based on the frequentist maximum profit time point, the catalyst is allowed to decay slightly more than the recommended value of 0.0320 min^{-1} used in the current work.

From the Bayesian perspective, the reduction in profit associated with the CSTR profit function is R2 relative to the ground truth profit. Based on the Bayesian maximum profit time point, the CSTR catalyst decay is slightly less than the recommended value of 0.0320 min^{-1} . One observes that the Bayesian methodology results in a smaller reduction in profit. The smaller reduction in profit is due to averaging over all possible values of the parameter k_0 supported by marginal posterior distribution $p(k_0)$ to obtain the maximum profit time point instead of resorting to a single parameter point estimate for k_0 . Thus, the uncertainty captured by the marginal posterior distribution $p(k_0)$ is explicitly incorporated into the engineering decision making process to determine the maximum profit time point such that one can trade-off between the reduction in profit associated with the maximum profit time point and the reduction in profit associated with the lower, mean and upper catalyst replacement time points. If the aim is to minimise the reduction in profit, based on the results outlined in Table 5.11, it is recommended to replace the catalyst at the Bayesian point of maximum profit.

5.3.2. Extended Case Study 3: Draining Tank –Valve Degradation

Recall from Section 4.6.2 that the liquid draining tank valve must be replaced when the proxy parameter k_v reaches a value of $1.2k_{vo}$. However, similar to the catalyst decay case study, the parameter k_{vo} is unknown. One can estimate the parameter k_{vo} by applying Algorithm 4 to the ordinary differential equation described by Equation 3.95, using the synthetically generated sensor measurement data (Section 4.7.3), to obtain a posterior distribution $p(\mathbf{w}|\boldsymbol{\mu}_L', d_L, d_{F_0})$ over the unknown ODE model parameters $\mathbf{w} = \left[\frac{1}{A} \frac{k_{vo}}{A} \frac{A_v}{A} \right]^T$. The engineer is specifically interested in the two parameters k_{vo}/A and A_v/A since these parameters can be used with Equation 3.93 to predict the draining tank valve remaining use life (RUL). One can obtain a joint posterior distribution over k_{vo}/A and A_v/A by simply marginalising out the parameter $1/A$ (Section 2.3.1) such that:

$$p\left(\frac{k_{vo}}{A}, \frac{A_v}{A}\right) = \mathcal{N}\left(\frac{k_{vo}}{A}, \frac{A_v}{A} \mid \begin{bmatrix} [\mathbf{m}_N]_2 \\ [\mathbf{m}_N]_3 \end{bmatrix}, \begin{bmatrix} [\mathbf{S}_N]_{22} & [\mathbf{S}_N]_{23} \\ [\mathbf{S}_N]_{32} & [\mathbf{S}_N]_{33} \end{bmatrix}\right) \quad \text{Equation 4.22}$$

Note that Equation 4.22 is Gaussian distributed, and, as a result, the valve degradation model output (Equation 3.93), which is a linear function of k_{vo} and A_v , is also Gaussian distributed. For inference purposes, Equation 3.93 is scaled by the cross-sectional area of the liquid draining tank since *Algorithm 4* infers the ratio of k_{vo} to A and A_v to A . Thus, the engineer only requires the mean and covariance of $k_v/A(t_i; k_{vo}/A; A_v/A)$ at the input time point t_i (Bishop, 2006). The mean can readily be obtained by evaluating the expected value (Sections 2.1.4 and 2.3.1) of $k_v/A(t_i; k_{vo}/A; A_v/A)$ under the joint posterior distribution $p\left(\frac{k_{vo}}{A}, \frac{A_v}{A}\right)$ such that:

$$\mathbb{E}_{p\left(\frac{k_{vo}}{A}, \frac{A_v}{A}\right)}\left[\frac{k_v}{A} \mid t_i\right] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p\left(\frac{k_{vo}}{A}, \frac{A_v}{A}\right) \left(\frac{k_{vo}}{A} + \frac{A_v}{A} t_i^{b_v}\right) d\frac{k_{vo}}{A} d\frac{A_v}{A} \quad \text{Equation 4.23}$$

$$\mathbb{E}_{p\left(\frac{k_{vo}}{A}, \frac{A_v}{A}\right)}\left[\frac{k_v}{A} \mid t_i\right] = \begin{bmatrix} [\mathbf{m}_N]_2 & [\mathbf{m}_N]_3 \end{bmatrix} \begin{bmatrix} 1 \\ t_i^{b_v} \end{bmatrix} \quad \text{Equation 4.24}$$

The variance can readily be obtained by evaluating the expected values (Sections 2.1.4 and 2.3.1) given by Equation 4.25 under the joint posterior distribution $p(k_{vo}/A; A_v/A)$ such that:

$$\text{Var}\left[\frac{k_v}{A} \mid t_i\right] = \mathbb{E}_{p\left(\frac{k_{vo}}{A}, \frac{A_v}{A}\right)}\left[\left(\frac{k_v}{A} \mid t_i\right)^2\right] - \left(\mathbb{E}_{p\left(\frac{k_{vo}}{A}, \frac{A_v}{A}\right)}\left[\frac{k_v}{A} \mid t_i\right]\right)^2 \quad \text{Equation 4.25}$$

$$\text{Var}\left[\frac{k_v}{A} \mid t_i\right] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p\left(\frac{k_{vo}}{A}, \frac{A_v}{A}\right) \left(\frac{k_{vo}}{A} + \frac{A_v}{A} t_i^{b_v}\right)^2 d\frac{k_{vo}}{A} d\frac{A_v}{A} - \left(\begin{bmatrix} [\mathbf{m}_N]_2 & [\mathbf{m}_N]_3 \end{bmatrix} \begin{bmatrix} 1 \\ t_i^{b_v} \end{bmatrix}\right)^2 \quad \text{Equation 4.26}$$

$$\text{Var}\left[\frac{k_v}{A} \mid t_i\right] = \begin{bmatrix} 1 & t_i^{b_v} \end{bmatrix} \begin{bmatrix} [\mathbf{S}_N]_{22} & [\mathbf{S}_N]_{23} \\ [\mathbf{S}_N]_{32} & [\mathbf{S}_N]_{33} \end{bmatrix} \begin{bmatrix} 1 \\ t_i^{b_v} \end{bmatrix} \quad \text{Equation 4.27}$$

CHAPTER 5: RESULTS AND DISCUSSION

As a result, the probability of observing k_v/A at an arbitrary input time point t_i is given by:

$$p\left(\frac{k_v}{A} | t_i\right) = \mathcal{N}\left(\frac{k_v}{A} \mid \begin{bmatrix} [\mathbf{m}_N]_2 & [\mathbf{m}_N]_3 \end{bmatrix} \begin{bmatrix} 1 \\ t_i^{b_v} \end{bmatrix}, \begin{bmatrix} 1 & t_i^{b_v} \end{bmatrix} \begin{bmatrix} [\mathbf{S}_N]_{22} & [\mathbf{S}_N]_{23} \\ [\mathbf{S}_N]_{32} & [\mathbf{S}_N]_{33} \end{bmatrix} \begin{bmatrix} 1 \\ t_i^{b_v} \end{bmatrix}\right) \quad \text{Equation 4.28}$$

The result given by Equation 4.28 is visually depicted in Figure 5.35.

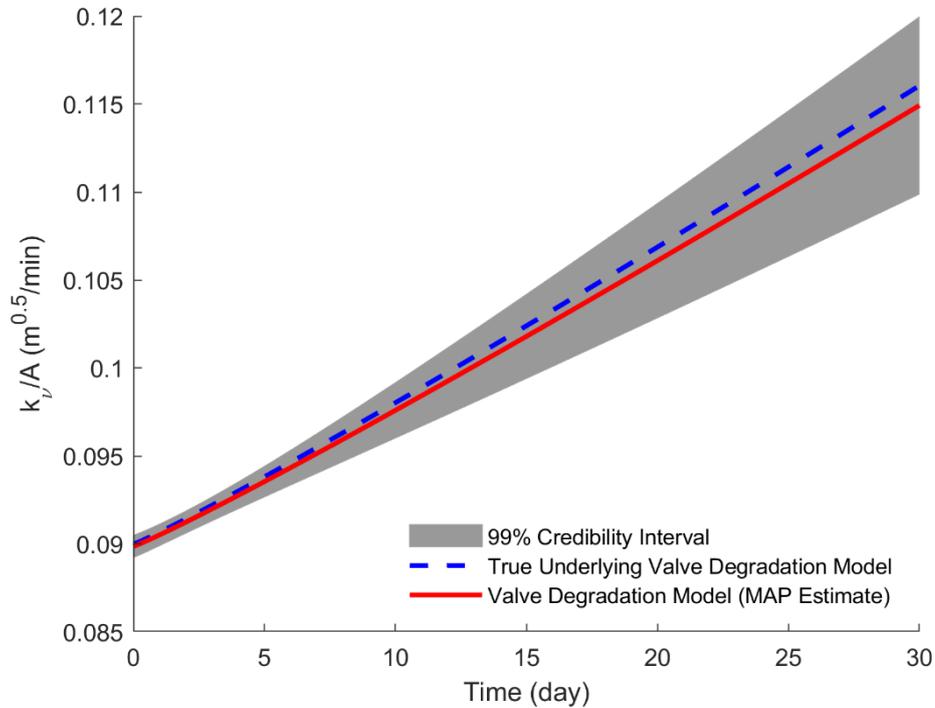


Figure 5.35 Valve degradation expected mean response over a 30 day degradation period (*Algorithm 4* results). The shaded region represents the 99% credibility interval (Section 2.8). The results are given for a single data set manifestation.

The valve degradation expected mean response over the 30 day degradation period obtained from applying the benchmark methodology (*Algorithm 2*), to the ODE model given by Equation 3.95, using the synthetically generated sensor measurement data (Section 4.7.3), is visually depicted in Figure 5.36.

CHAPTER 5: RESULTS AND DISCUSSION

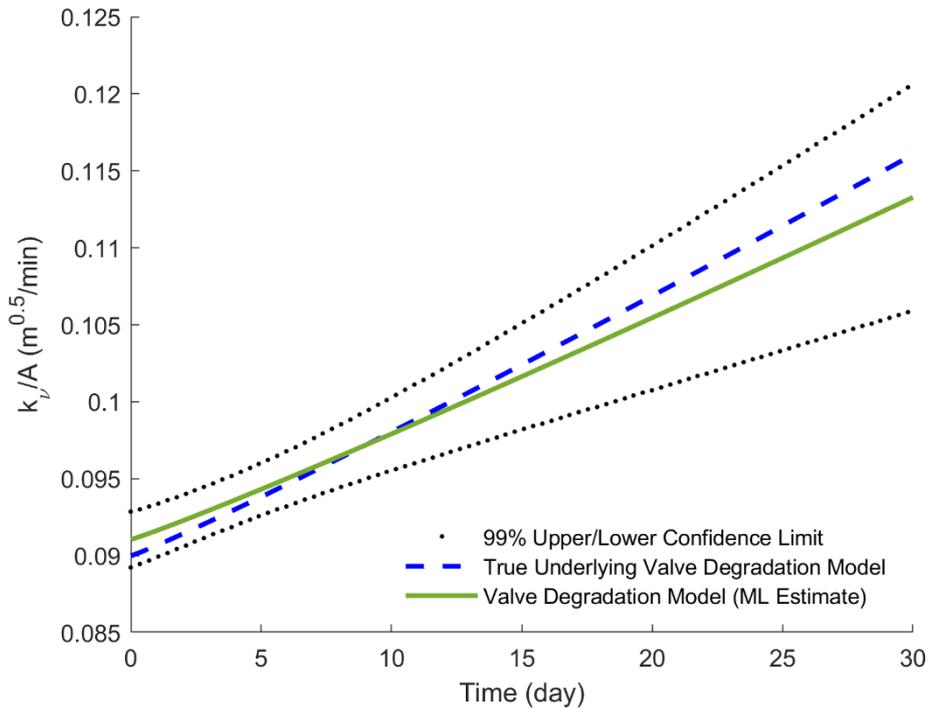


Figure 5.36: Valve degradation expected mean response over a 30 day degradation period (*Algorithm 2* results). The dotted line represents the lower and upper confidence limits (Section 2.8). The results are given for a single data set manifestation.

Table 5.12 summarises the results for parameters k_{v_0}/A and A_v/A obtained from *Algorithms 2* and *4*, respectively, as applied to the synthetically generated sensor measurement data. The results in Table 5.12 are presented for a single data set manifestation.

Table 5.12: Summary of the inferred lumped system ordinary differential equation model parameters k_{v_0}/A and A_v/A . Note that these results are given for a single data set manifestation.

| Probabilistic Interpretation | Algorithm Number | Unknown Parameter | Simulation Ground Truth | Inferred Value | % Error |
|------------------------------|------------------|---------------------|---|--|---------|
| Frequentist | 2 | $\frac{k_{v_0}}{A}$ | $0.0900 \frac{m^{0.5}}{min}$ | $0.0911 \frac{m^{0.5}}{min}$ | 1.18% |
| | | $\frac{A_v}{A}$ | $2.857 \times 10^{-7} \frac{m^{0.5}}{min^{2.07}}$ | $2.44 \times 10^{-7} \frac{m^{0.5}}{min^{2.07}}$ | 14.71% |
| Bayesian (MAP estimate) | 4 | $\frac{k_{v_0}}{A}$ | $0.0900 \frac{m^{0.5}}{min}$ | $0.0898 \frac{m^{0.5}}{min}$ | 0.24% |
| | | $\frac{A_v}{A}$ | $2.857 \times 10^{-7} \frac{m^{0.5}}{min^{2.07}}$ | $2.77 \times 10^{-7} \frac{m^{0.5}}{min^{2.07}}$ | 3.04% |

Table 5.13 summarises the results obtained from applying *Algorithm 2* and *4* to the one hundred independently generated sensor measurement data sets (Sections 4.7 and 4.8 - results are given for performance criteria 7 through 10).

CHAPTER 5: RESULTS AND DISCUSSION

Table 5.13: Summary of the mean inferred parameter estimates, standard deviation, proportion of confidence/credibility intervals containing the simulation ground truth parameter values and average confidence/credibility interval width for the one hundred independently generated sensor measurement data sets. The units for all numerical values of k_{v_0}/A and A_v/A in Table 5.13 are $(m^{0.5}/min)$ and $(m^{0.5}/min^{2.07})$, respectively (excluding the Percentage Contained* column).

| Algorithm Number | Unknown Parameter | Simulation Ground Truth | Mean Inferred Value | Standard Deviation | Percentage Contained* | Average Interval Width |
|------------------|---------------------|-------------------------|------------------------|------------------------|-----------------------|------------------------|
| 2 | $\frac{k_{v_0}}{A}$ | 0.0900 | 0.0898 | 0.0018 | 77 % | 0.0035 |
| | $\frac{A_v}{A}$ | 2.857×10^{-7} | 2.826×10^{-7} | 1.272×10^{-7} | 48 % | 1.708×10^{-7} |
| 4 | $\frac{k_{v_0}}{A}$ | 0.0900 | 0.0895 | 0.0016 | 23 % | 0.0016 |
| | $\frac{A_v}{A}$ | 2.857×10^{-7} | 2.850×10^{-7} | 1.293×10^{-7} | 37 % | 1.400×10^{-7} |

* Percentage Contained refers to the proportion of the simulation ground truth parameter values that fall within the 99% marginal parameter confidence/credibility interval.

Observe from Table 5.13 that both the frequentist (*Algorithm 2*) and Bayesian (*Algorithm 4*) approaches provide consistent results when compared to the simulation ground parameter values. Note that only 23% of the constructed credibility intervals (one for each of the independently generated data sets) contain the ground truth parameter value for k_{v_0}/A . Furthermore, the average confidence interval width of k_{v_0}/A is approximately 2.2 times larger than the average credibility interval width. This indicates that the confidence interval estimate for k_{v_0}/A is more conservative which explains why 77% of the constructed confidence intervals contain the true value of k_{v_0}/A . Similar behaviour is observed for the average confidence interval width of parameter A_v/A when compared to the average credibility interval width.

Table 5.14 summarises the valve remaining useful life (RUL), as extracted from Figures 5.35 and 5.36, respectively, when the proxy parameter k_v reaches a value of $1.2k_{v_0}$. Note that the valve RUL is relative to the end time point at which the regression procedure was performed. For the current case study, the sensor measurement data set was collected over a 300 minute period.

Hence the RUL values outlined in Table 5.14 are relative to the 300 minute regression analysis mark. The corresponding value in brackets present the day of valve replacement relative to a 30 day period starting at day 0 where the start of day zero coincides with the first sensor measurement in the data set.

CHAPTER 5: RESULTS AND DISCUSSION

Table 5.14: Remaining useful life for the liquid draining tank valve obtained from applying *Algorithms 2* and *4*, respectively. Note that these results are given for a single data set manifestation.

| Probabilistic Interpretation | Lower Limit Valve RUL | Mean Valve RUL | Upper Valve RUL | Ground Truth Valve RUL |
|------------------------------|-----------------------|--------------------|--------------------|------------------------|
| Frequentist (Algorithm 2) | 19.20 days (19.41) | 24.70 days (24.91) | 35.38 days (35.59) | 21.01 days (21.23) |
| Bayesian (Algorithm 4) | 18.47 days (18.68) | 21.87 days (22.08) | 26.93 days (27.14) | 21.01 days (21.23) |

From Table 5.14 one observes that both the benchmark (frequentist) and the proposed Bayesian methodology contain the ground truth valve RUL within the expected mean response lower and upper limits (Figures 5.35 and 5.36). Again, the proposed Bayesian methodology provides a narrower valve RUL margin between the lower and upper valve replacement limits. From an engineering perspective, if one has to choose a valve RUL time point based on the results outlined in Table 5.14, the obvious engineering choice would be to replace the valve at the mean RUL time point. If one follows the frequentist methodology, the valve should be replaced 24.70 days after performing the regression analysis (*Algorithm 2*) which is approximately 3.69 days later than the ground truth RUL time point. If the engineer decides on following the Bayesian methodology, the valve should be replaced 21.87 days after performing the regression analysis (*Algorithm 4*) which is approximately 0.86 days later than the ground truth RUL time point.

Similar to the previous CSTR case study with catalyst decay, the challenge is that the selection of the valve RUL time point has an effect on the profit function associated with the liquid draining tank. If the valve is replaced too early, the engineer is prematurely disposing of hardware that is still functional which can be expensive to replace. However, if the valve is replaced too late, the engineer is at risk of losing more product stored in the tank due to valve wear/degradation, which can also subsequently influence downstream processing and product quality.

Based on the results outlined in Table 5.14 and Figures 5.35 and 5.36, it is difficult to select a single valve RUL replacement time point, regardless of the inference methodology used, especially given the effect of the decision making process on the cost associated with the liquid draining tank process unit.

The engineer can consider the effect of the estimated ODE model parameter k_{v0}/A on the liquid draining tank process unit profit function to aid in the valve RUL time point decision making process. It is also possible to include the effect of the parameter A_v/A , however, for the current work the parameter A_v/A is excluded from the draining tank profit function.

Given that the assumed draining tank profit function (Equation 3.96) sufficiently describes the profit associated with the process unit, the engineer is interested in finding the time point t as which the maximum profit is reached. From the frequentist perspective (*Algorithm 2*), the obvious choice would be to use $(k_{v0}/A)_{ML}$ as a plug-in point estimate for Equation 3.96 followed by finding the time point t at which the draining tank profit function reaches a maximum value. In other words,

CHAPTER 5: RESULTS AND DISCUSSION

$$t_F = \operatorname{argmax}_t \left[P_{Tank} \left(t, \left(\frac{k_{vo}}{A} \right)_{ML} \right) \right] \quad \text{Equation 4.29}$$

This leads to the maximum profit time point at:

$$t_F = \frac{1400 - 5000 \left(\left(\left(\frac{k_{vo}}{A} \right)_{ML} \right)^2 - 1.44 \left(\left(\frac{k_{vo}}{A} \right)_{ML} \right) + 0.0648 \right)}{100} \quad \text{Equation 4.30}$$

From the Bayesian perspective, the engineer would like to evaluate the expected value of $P_{Tank}(t, k_{vo}/A)$ under the marginal posterior distribution $p(k_{vo}/A)$ followed by finding the time point t at which the liquid draining tank profit function reaches a maximum value. In other words,

$$t_B = \operatorname{argmax}_t \left[\mathbb{E}_{p(k_{vo}/A)} P_{Tank} \left(t, \frac{k_{vo}}{A} \right) \right] \quad \text{Equation 4.31}$$

$$t_B = \operatorname{argmax}_t \left[\int_{-\infty}^{\infty} p \left(\frac{k_{vo}}{A} \right) P_{Tank} \left(t, \frac{k_{vo}}{A} \right) d \frac{k_{vo}}{A} \right] \quad \text{Equation 4.32}$$

The distribution $p(k_{vo}/A)$ is obtained by simply marginalising out the parameter A_v/A (Section 2.3.1) from Equation 4.22 such that:

$$p \left(\frac{k_{vo}}{A} \right) = \mathcal{N} \left(\frac{k_{vo}}{A} \mid [\mathbf{m}_N]_2, [\mathbf{S}_N]_{22} \right) \quad \text{Equation 4.33}$$

The engineer can show that the Bayesian methodology (*Algorithm 4*) provides the maximum profit time point at:

$$t_B = \frac{1400 - 5000 \left(([\mathbf{m}_N]_2)^2 + [\mathbf{S}_N]_{22} - 1.44([\mathbf{m}_N]_2) + 0.0648 \right)}{100} \quad \text{Equation 4.34}$$

Table 5.15 outlines the time points at which the maximum profit is reached as obtained by evaluating Equation 4.30 (frequentist) and Equation 4.34 (Bayesian), respectively. The results outlined in Table 5.15 are relative to a 30 day period starting at day zero. In other words, for a reported value of 16.90, it implies that the maximum profit time point is reached on day 16.90 within the 30 day period starting at day zero where day zero coincides with the first sensor measurement in the data set.

Figure 5.37 visually depicts the time points of maximum profit against the ground truth maximum profit time point. Note that the results outlined in Table 5.15 and Figure 5.37 are given for a single data set manifestation.

CHAPTER 5: RESULTS AND DISCUSSION

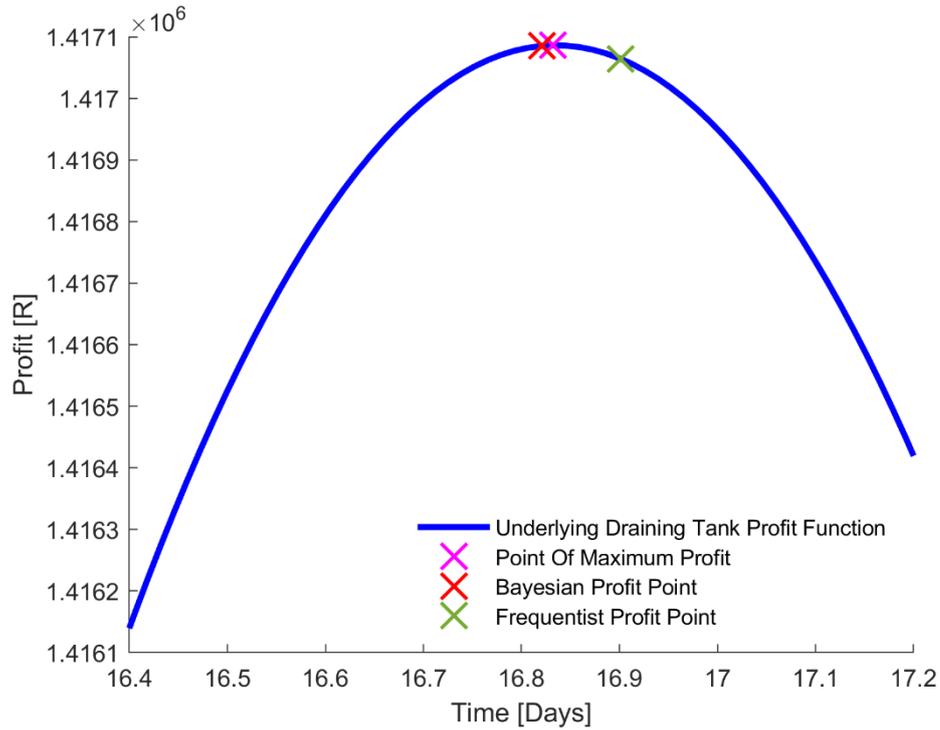


Figure 5.37: Visual depiction of the true underlying liquid draining tank profit function given by Equation 3.96 with parameter $k_{vo}/A = 0.0900 \text{ m}^{0.5}/\text{h}$ (ground truth value).

Table 5.15: Point of maximum profit for the liquid draining tank profit function as obtained by applying *Algorithms 2 and 4*, respectively, to estimate parameter k_{vo}/A .

| Probabilistic Interpretation | Algorithm Number | Maximum Profit Time Point | Predicted Maximum Profit [Rx100] | Ground Truth Maximum Profit Time Point | Ground Truth Maximum Profit [Rx100] |
|------------------------------|------------------|---------------------------|----------------------------------|--|-------------------------------------|
| Frequentist | 2 | 16.90 | 14170.64 | 16.83 | 14170.86 |
| Bayesian | 4 | 16.82 | 14170.85 | 16.83 | 14170.86 |

From Tables 5.14 and 5.15, observe that the maximum profit time point, as obtained from both the frequentist and Bayesian methodologies, does not fall within the lower and upper valve RUL time point limits. Table 5.16 summarises the reduction in profit associated with selecting the frequentist and Bayesian maximum profit time points as well the reduction in profit associated with replacing the valve at the lower, mean and upper RUL time points (Table 5.14).

CHAPTER 5: RESULTS AND DISCUSSION

Table 5.16: Liquid draining tank cost-benefit analysis results obtained from applying *Algorithms 2* and *4* (results given for a single data set manifestation).

| Probabilistic Interpretation | Reduction In Profit At Maximum Profit Time Point [R] | Reduction In Profit At: [Rx100] | | |
|------------------------------|--|---------------------------------|----------------------|-----------------------|
| | | Lower Valve RUL Limit | Mean Valve RUL Limit | Upper Valve RUL Limit |
| Frequentist | 22 | 331.53 | 3260.28 | 17587.50 |
| Bayesian | 1 | 170.20 | 1375.50 | 5309.65 |

From Table 5.16 one observes that if the engineer follows the frequentist methodology in conjunction with the plug-in point estimate $(k_{vo}/A)_{ML}$ obtained from *Algorithm 2*, then the reduction in profit associated with the liquid draining tank cost function $P_{Tank}(t, k_{vo}/A)$ is R22 relative to the ground truth profit. From the Bayesian perspective, the reduction in profit associated with the liquid draining tank profit function is R1 relative to the ground truth profit. The smaller reduction in profit is due to averaging over all possible values of the model parameter k_{vo}/A supported by the marginal posterior distribution $p(k_{vo}/A)$ to obtain the maximum profit time point instead of resorting to a single parameter point estimate for k_{vo}/A .

Thus, the uncertainty captured by the marginal posterior distribution $p(k_{vo}/A)$ is explicitly incorporated into the engineering decision making process to determine the maximum profit time point such that one can trade-off between the reduction in profit associated with the maximum profit time point and the reduction in profit associated with the lower, mean and upper valve RUL time points. If the aim is to minimise the reduction in profit, based on the results outlined in Table 5.16, it is recommended to replace the liquid draining tank valve at the Bayesian point of maximum profit even though the Bayesian point of maximum profit does not fall within the lower and upper valve RUL replacement margin.

5.4. Comment on ‘Conservative’ Estimates and Coverage Frequencies

The author would like to explicitly point out that the use of the word ‘conservative’ in the preceding discussions (Section 5.2 and 5.3) should not be interpreted in any adverse sense. The use of the word ‘conservative’ is simply to emphasise that throughout the different case studies and repeated experimentation results, the frequentist confidence intervals (from an empirical simulation-based perspective) seem to be consistently larger in interval length/width when compared to the Bayesian credibility interval. From the simulation-based results, it appears as if the frequentist confidence intervals are more ‘cautions’ (hence the use of the word ‘conservative’) in the sense that the interval is larger relative to the Bayesian credibility interval.

Furthermore, observe that for all the case studies, the proportion of the simulation ground truth parameter values that fall within the 99% marginal parameter credibility interval (Bayesian) is significantly lower than the expected 99% coverage frequency (possibly excluding the coverage frequency for the parameter K_p credibility interval in Table 5.2). When considering *Algorithm 3*, the primary reason the proposed Bayesian methodology displays such a low coverage frequency (excluding the argument that one is using a first order Taylor approximation) stems

CHAPTER 5: RESULTS AND DISCUSSION

from using variational inference (Section 2.6) to address the nonlinear least squares problem. It is well known and documented in the literature that variational inference underestimates the variance of the true posterior distribution (Bishop (2006), Fox and Roberts (2012), Murphy (2012) and Blei, Kucukelbir and McAuliffe (2018)).

Thus, even though the proposed algorithm converges close to the true model parameters in parameter space, the resulting approximate variational Gaussian distribution underestimates the variance of the true parameter posterior distribution. For example, when considering Figures 5.4 and 5.5, observe that the mean of the approximate Gaussian distribution is close to the true parameter setting. However, due to the variance being underestimated as a result of the variational approximation, the 99% marginal credibility interval does not contain the true model parameter value. This phenomenon possibly gives rise to the low coverage frequency observed for the process time constant parameter τ . Recall that the CSTR outlet concentration is nonlinear in the parameter τ but linear in the parameter K_p . Overall, the low coverage frequency for the parameter τ is a combination of exploiting a first order Taylor approximation and variational inference.

When considering *Algorithm 4*, the primary reason the coverage frequency is low stems from the Gaussian process hyperparameter optimisation routine. Recall that the current work uses gradient ascent to maximise the log marginal likelihood (Section 4.5.2). However, the log marginal likelihood is a nonconvex function implying there are multiple local hyperparameter maxima. Thus, even after performing multiple optimisation routines on a single data set manifestation, there is no guarantee that the optimisation routine will converge to a parameter setting that sufficiently describes the data. If the optimisation routine converges to a hyperparameter setting that poorly describes the observational data, the corresponding likelihood function used to infer the ODE model parameters will poorly describe the observational data. As a result, the engineer will infer the incorrect ODE model parameters despite having a well-defined regression problem. Consequently, the constructed ODE model parameter credibility interval will not contain the ground truth parameter value resulting in a low coverage frequency for independently repeated experimentation.

5.5. Summary

Chapter 5 started by presenting the different case study (Table 3.3) parameter inference results, as obtained from applying *Algorithms 1* through *4* (Sections 4.4 and 4.5) to the synthetically generated sensor measurement data (Section 4.7), and the frequentist and Bayesian results were compared based on the performance criteria outlined in Section 4.8. What is interesting to note is that for the *single data set manifestation* for each case study, no noteworthy difference is observed between the frequentist and Bayesian parameter inference results when considering performance criteria 1 through 5. From the results outlined in Section 5.2, it appears as if there is no benefit in obtaining a posterior distribution over the algebraic or ODE dynamic model parameters. As a result, one can argue that there is no need to approach the parameter estimation procedure from a Bayesian perspective. The more common frequentist approach is sufficient.

Furthermore, based on the results obtained from applying performance criteria 7 through 10 (Section 4.8) to the additional one hundred independently generated data sets, both the frequentist (*Algorithm 1* and *2*) and proposed Bayesian parameter estimation methodologies (*Algorithm 3* and *4*) provide consistent results across the multiple data sets.

CHAPTER 5: RESULTS AND DISCUSSION

This further emphasises that the more common frequentist approach is sufficient for parameter estimation purposes. On average, it was observed that the frequentist confidence interval estimate is more conservative when compared to the Bayesian credibility interval such that across multiple repeated experiments, the frequentist confidence interval contained a higher proportion of the simulation ground truth parameter values.

However, the advantage of the posterior distribution was illustrated with a simple cost-benefit analysis for each of the extended case studies (Section 5.3) where, by explicitly propagating the uncertainty captured by the posterior distribution, it was possible to obtain narrower lower and upper limits for the catalyst replacement time (Section 5.3.1) and liquid draining tank valve RUL (Section 5.3.2), compared to the more common frequentist methodology. Nevertheless, the selection of the catalyst replacement time and draining tank valve RUL has an effect on the corresponding process unit profit.

By further propagating the uncertainty captured by the parameter posterior distribution to engineering decision making informed by the corresponding process unit profit function, it was also possible to more accurately predict the point of maximum profit for the selected process unit case studies (for a *single data set manifestation* for each case study). In extended Case Studies 2 and 3, the predicted point of maximum profit (Tables 5.11 and 5.16) controls the catalyst replacement time and valve RUL, respectively, if the goal is to make as much profit as possible while incurring the lowest operating and maintenance cost. When the predicted maximum profit time point is reached, the engineer should replace the CSTR catalyst or the draining tank valve (depending on what case study is considered) since the aforementioned time point coincides with the smallest reduction in profit. Selecting any other time point, excluding the ground truth maximum profit time point, will result in a larger loss in profit. This behaviour stems from the structure of the process unit profit functions outlined in Sections 4.6.1 and 4.6.2 over the selected 30 day period.

Chapter 6

Conclusions

“The only use I know for a confidence interval is to have confidence in it”

- Leonard Jimmie Savage, (-)

Recall that the current work focuses on estimating the parameters of lumped system algebraic and ODE models that describe a physical system. The equations governing the dynamics of the physical system stem from expert engineering knowledge, however, the principles underpinning the model development phase do not always provide insight into selecting suitable values for all the model parameters. Furthermore, it is not always possible to measure these model parameters directly from the system. Consequently, one has to estimate the model parameters from noise-corrupted time series data while simultaneously quantifying how reliable the parameter estimates are. This work investigated the estimation of model parameters from both the frequentist and Bayesian statistical inference perspective and attempted to evaluate the merit of applying Bayesian probabilistic techniques in the chemical engineering setting.

6.1. Frequentist vs Bayesian

From the case study parameter inference results outlined in Section 5.2, one observes that there is no noteworthy difference between the results obtained from the (frequentist) benchmark (Section 4.4) and the proposed Bayesian methodologies derived in Sections 4.5.1 and 4.5.2. Both the benchmark and the proposed methodologies provide comparable results for a single data set manifestation (Section 4.7). Arguably, from a pragmatic engineering perspective, there is no reason to use the benchmark methodology over the proposed Bayesian methodologies besides the fact that the benchmark occasionally outperforms the proposed Bayesian methodologies and vice versa when considering the results for performance criteria 1 to 5 (Sections 4.8, 5.2.1, 5.2.2 and 5.2.3).

The aforementioned point is further emphasised when considering the results of performance criteria 7 through 10 (Section 4.8), as applied to the multiple independently generated data sets, where it was shown that both the frequentist (*Algorithm 1* and *2*) and Bayesian (*Algorithm 3* and *4*) parameter estimation methodologies provide consistent results. However, from a statistical perspective, the fundamental difference between the benchmark and the proposed Bayesian methodologies is the way in which the parameter estimation problem is approached and the interpretation associated with the confidence/credibility interval/region results (Section 2.8). Furthermore, in a practical setting, the engineer might not always have access to multiple data sets and repeated experimentation might not be a feasible option.

In both the Bayesian and frequentist viewpoints, the likelihood function plays an important role, however, the way in which it is used for inference purposes differs. The frequentist benchmark relies on the maximum likelihood heuristic to select suitable values for the algebraic or ODE model parameters, whereas the proposed Bayesian methodologies infer a posterior distribution over the unknown model parameters. The author would like to explicitly point out

CHAPTER 6: CONCLUSIONS

that both proposed methodologies derived in Sections 4.5.1 and 4.5.2 are grounded on the fundamental rules of probability theory (sum and product rule, Section 2.1.1) and rely on operations such as marginalisation and conditioning to make conclusions about the unknown model parameters from noise-corrupted time series data. These operations are natural steps to follow once the engineer has decided to model the unknown parameters as random variables. The proposed Bayesian methodologies avoid choosing between *ad hoc* heuristics such as maximum likelihood to estimate the model parameters.

The impact of the fundamental difference between the (frequentist) benchmark and proposed Bayesian methodologies becomes apparent in Section 5.3 (parameter tracking application results) when one considers the procedure of propagating uncertainty and engineering decision making under uncertainty (Section 4.8 – performance criterion 6). One of the most desirable benefits of the proposed methodologies is the principled way in which uncertainty is incorporated into the inference procedure. The experimentalist starts by positing a prior distribution, incorporates the evidence (via the likelihood function) provided by the sensor measurements and obtains a revised distribution - the parameter posterior distribution - which captures all possible values (hence the explicit modeling of uncertainty) of the algebraic or ODE model parameters supported by the statistical model (Table 1.1).

Even though there is no distinct difference between the case study parameter inference results obtained from the benchmark and proposed Bayesian methodologies in Section 5.2, the value of the Bayesian parameter posterior distribution shows up when one considers the subsequent application of the inferred parameters in day-to-day engineering tasks such as when to replace the reactor catalyst or draining tank valve – here it was possible to obtain more accurate estimates for the catalyst replacement time and valve RUL due to propagating the uncertainty captured by the posterior – which inevitably has an impact on the process unit profit margin.

Thus, by considering the effect of the uncertainty about the parameters, it was possible for the engineer to make more informed decisions compared to the more traditional frequentist decision making process which routinely relies on a plug-in point estimate for the model parameters. From the frequentist perspective, it is not always obvious how one should go about propagating uncertainty to the subsequent decision making process in a principled way. Furthermore, the notion of a confidence interval (Section 2.8) does not provide guidance as to which value of the unknown parameter the engineer should make use, making the maximum likelihood parameter setting the only obvious choice (Table 1.1). Contrast this with the Bayesian approach which averages over all possible settings of the unknown parameters, supported by the parameter posterior distribution, for decision making purposes (Section 5.3). However, it is important to note that these conclusions are based on a single data set manifestation.

Based on the similarity of the case study parameter inference results outlined in Section 5.2 and the parameter tracking application results outlined in Section 5.3, it is worth further exploring the benefit of probabilistic techniques and explicitly modeling with uncertainty, i.e. Bayesian statistical inference, in the chemical engineering setting as the aforementioned sections serve as proof of concept that there is indeed value in following the Bayesian paradigm of statistical inference. Furthermore, Bayesian inference allows the engineer to sensibly incorporate prior knowledge which can be a valuable tool in scenarios where prior knowledge

CHAPTER 6: CONCLUSIONS

is available, and the engineer can move beyond exploiting conjugate priors with the current Bayesian inference software packages available.

However, when critically evaluating Bayesian statistical inference, one can argue that even the proposed Bayesian methodologies use *ad hoc* ideas. For example, the choice of prior is completely subjective and can heavily influence the posterior distribution results. Furthermore, it might be difficult to convince other experimentalists of the selected prior distribution and the attained posterior distribution results. Setting the prior hyperparameters (to reflect the experimentalist's initial belief about the unknown model parameters) are also subjective. For most models of practical interest, one has to resort to approximation schemes to approximate the intractable posterior distribution and the choice of approximation scheme can also be regarded as *ad hoc*. Despite Bayesian methods becoming mainstream in recent years, there are still a lot of barriers that prohibit the widespread application of Bayesian parameter estimation methods.

Bayesian methods require a high level of critical thinking which immediately makes it challenging to any individual with a limited conceptual understanding. However, to build a conceptual understanding, one must spend time and engage with Bayesian inference which can be problematic in an engineering environment where the practitioner does not have time to develop an intuition for the Bayesian school of thought. Bayesian techniques require a high level of mathematical involvement which might be beyond the scope of most employed practitioners. Furthermore, most practitioners are not concerned with uncertainty estimates, thus, they do not consider the application of Bayesian inference as an option. Ultimately, the choice of statistical inference is a preference and depends on the application and needs of the practitioner. Bayesian approaches are beneficial (as illustrated in Section 5.3), however, non-Bayesian methods (which are often more accessible to practitioners) have proven to be versatile and powerful tools (Section 4.4).

6.2. Review of Objectives

Recall that the objectives of the current study are outlined in Section 1.5.2. The first objective is concerned with the dynamic modeling and simulation of case study process units using the forward modeling approach with the addition of sensor noise. This objective was predominantly addressed in Sections 4.2, 4.6 and 4.7 where the author introduced the isothermal CSTR and liquid draining tank case studies, the extended case studies and the data generation process for each case study. The second objective required the proposal and application of various Bayesian inference techniques for estimating the parameters of lumped system dynamic models which were addressed in Sections 4.5.1 and 4.5.2.

Following the proposal and application of the various Bayesian inference techniques, objective three required benchmarking the proposed Bayesian techniques against traditional parameter estimation methods. The current work selected the Gauss-Newton nonlinear least squares methodology as the benchmark approach (Section 4.4) and compared the results obtained from the Bayesian methodologies to the benchmark by considering the performance criteria outlined in Section 4.8. The corresponding results are given in Section 5.2. The final objective focused on applying the proposed Bayesian and benchmark techniques to parameter tracking applications. Here the extended case studies presented in Section 4.6 were used as vehicles for

CHAPTER 6: CONCLUSIONS

illustrative purposes with the corresponding parameter tracking results outlined in Section 5.3. Thus, all the objective outlined in Section 1.5.2 were successfully completed.

6.3. Novelty and Contribution

Based on the existing literature outlined in Table 3.2 (Section 3.4), the current work provides an extension of the Gaussian process based ODE parameter estimation procedures to include ODEs with an arbitrary exogenous input disturbance structure (Sections 1.5.4, 3.5 and 4.5.2) which, to the best of the author's knowledge, has not been addressed in the current Gaussian process ODE parameter estimation literature.

Furthermore, the current work also contributes open-source software (Section 1.5.4) for each simulation case study with the corresponding code illustrating the proposed algorithm's implementation (Sections 4.4 and 4.5), and the obtained results. The open-source software is available at:

Open-source software URL:

<https://gitlab.com/pleased/bayesian-ode-parameter-estimation>

Chapter 7

Recommendations

“Probability is expectation founded upon partial knowledge. A perfect acquaintance with all the circumstances affecting the occurrence of an event would change expectation into certainty, and leave nether room nor demand for a theory of probabilities.”

- George Boole, (1951)

This chapter focuses on presenting several recommendations, potential pitfalls, and limitations as well as ideas for future work and closing remarks that pertain to the Bayesian methodologies developed in this thesis.

7.1. Considering more Complex Systems

Extension to a single nonlinear (in the parameters) ODE. Recall that the current thesis restricts attention to a single lumped system continuous-time dynamic model with time-invariant parameters (Section 1.5.3). Furthermore, all the proposed Bayesian methodologies developed in the current work pertain to continuous-time models that can be written as a linear combination of the unknown model parameters (Sections 4.5.1 and 4.5.2). However, several physical system case studies may arise in which the governing ordinary differential equation describing the state variable contains model parameters that form part of a nonlinear function. For example, refer back to the extended case studies in Section 4.6 where it was not possible to infer the parameters k_d and b_v using the existing Gaussian process based methodology developed in Section 4.5.2.

In these types of scenarios, it would be desirable to have an inference procedure that is capable of inferring the parameters of ordinary differential equations that can not be written as a linear combination of the models parameters. Here one possible approach is to infer the ODE state variable derivative function values using the approach outlined in Section 4.5.2. With the derivative information available, the engineer can linearise the ODE about the mode of the parameter posterior distribution and use the variational Bayesian nonlinear regression approach outlined in Section 4.5.1 to iteratively update the ODE model parameters until the ELBO converges.

Extension to a system of linear (in the parameter) ODEs. Most physical systems of practical interest to chemical engineers are described by multiple ordinary differential equations. For a subset of these systems, the resulting ODEs are all linear in the model parameters. If one desires to jointly infer the parameters of the system of ordinary differential equations, then it is worthwhile extending the Bayesian methodology outlined in Section 4.5.2 such that it can be applied to a system of ODEs. The extension procedure in itself is straightforward. However, for a system of ODEs, the parameter vector \mathbf{w} would now become a parameter matrix \mathbf{W} to account for the parameters associated with each ODE. This implies that the multivariate Gaussian distribution associated with \mathbf{w} would generalise to its matrix counterpart, namely, the matrix Gaussian distribution (also referred to as the matrix Normal distribution).

CHAPTER 7: RECOMMENDATIONS

Extension to a system of nonlinear (in the parameter) ODEs. If the physical system under consideration contains multiple ODEs that are nonlinear in the model parameters, then it is still possible to adapt the proposed Bayesian methodology outlined in Section 4.5.2 for inference purposes. Note that one will run into intractability problems since it is no longer possible to write the system of differential equations as a linear combination of the unknown ODE model parameters. To circumvent this problem, the author recommends linearising the system of ODEs about the parameter posterior distribution mode \mathbf{W} whereby variational inference can be used to iteratively update the posterior distribution mode until the ELBO converges. To account for the parameters associated with each ODE, one would have to generalise the multivariate Gaussian distribution to its matrix counterpart. In order to apply variational inference, the author further recommends using the mean-field variational family whereby the joint posterior distribution over all ODE model parameters (for the linearised system of ODEs) are decomposed into a product of marginal posterior distributions where each marginal posterior distribution pertains to the ODE model parameters for a single ODE.

7.2. Bayesian Practicalities

Applications to online learning (sequential estimation). Appendix A briefly addressed the application of online learning for the mean parameter of a univariate Gaussian distribution from data. However, the idea of using Bayes' rule for sequential estimation is much more general. For instance, in Sections 5.2 and 5.3, the regression analysis was performed after collecting the sensor measurements, i.e. offline conclusions were made about the model parameters, valve degradation, etc., based on a batch of collected data. However, one can use the posterior distribution as the new prior distribution in a sequential manner as new data observations become available to re-estimate the unknown model parameters. The re-estimated parameters can then be used to make revised decisions about when to replace the reactor catalyst, draining tank valve, etc.

Non-conjugate priors. The proposed Bayesian methodologies derived in Sections 4.5.1 and 4.5.2 all rely on exploiting conjugate priors such that one can obtain closed-form solutions (update equations) for the parameter posterior probability distribution. The selected prior distributions are completely valid choices since there is no 'correct' way of selecting a prior distribution. However, a conjugate prior might not always be able to express the experimentalist's initial belief or prior knowledge about the unknown model parameters and for certain inference scenarios it might be more beneficial to use a non-conjugate prior. The choice of prior does depend on the type of application considered. In general, the posterior distribution results can be heavily influenced by the choice of prior. From a practical engineering perspective, it might be difficult to convince other experimentalists of the selected prior distribution and the attained posterior distribution results. Thus, the experimentalist must carefully consider the choice of prior distribution used.

Constrain ODE model parameter to be positive. One drawback of the proposed Bayesian methodologies derived in Sections 4.5.1 and 4.5.2 stem from the use of the Gaussian conjugate priors (Section 2.3.3, Table 1.1). Recall from Section 2.3.1 that the univariate Gaussian distribution provides support for $x \in \mathbb{R}$. In other words, the univariate Gaussian distribution, as well as its multivariate counterpart, allow the model parameters to be negative. This is strictly speaking not correct for all of the model parameters. For example, when one considers

CHAPTER 7: RECOMMENDATIONS

Case Study 2 (Section 5.2.2, Table 5.3) it does not make sense to allow the parameter F/V to take on any negative values.

However, if the posterior distribution is sharply peaked around the true parameter values, i.e. the posterior distribution assigns high probability mass in parameter space close to the true setting of the model parameters, then the subsequent use of the Gaussian posterior distribution for decision making is not detrimental. However, if the posterior distribution did not converge close to the true parameter values (which can happen if the engineer does not have enough sensor measurements), the subsequent decisions based on the posterior distribution can be misleading. In order to remedy the situation, the author recommends replacing the Gaussian distribution with a log-Normal distribution. The log-Normal distribution provides support for $x \in (0, \infty)$. Alternatively, one can also consider using the exponential distribution which provides support for $x \in [0, \infty)$

Incorporating decision making under uncertainty. Although Bayesian statistical inference provides the engineer with a consistent mathematical framework for manipulating and quantifying the model parameter uncertainty, it does not provide a framework for making decisions under uncertainty. However, the engineer can use decision theory/analysis, in conjunction with probability theory, to make optimal decisions in situations involving uncertainty. Although decision theory is not addressed in the current work, the interested reader is referred to Bishop (2006) and Murphy (2012) for an introduction to elements of decision theory/analysis and its application to the field of Machine Learning. For a more detailed treatment of Bayesian decision theory/analysis, refer to Smith (2010).

7.3. Gaussian Process Considerations

Gaussian process hyperparameter learning. The current work uses gradient-based optimisation (gradient ascent with $\psi_i > 0$) to maximise the Gaussian process log marginal likelihood to obtain point estimates for the Gaussian process hyperparameters, the state variable as well as the exogenous input disturbance sensor variance parameters from noise corrupted time-series data. Recall that this type of technique is known as a type 2 maximum likelihood procedure. Another gradient-based optimisation alternative is to use the method of conjugate gradients to optimise the log marginal likelihood. Conjugate gradients ensure that the Gaussian process covariance matrix remain symmetric positive-definite after each iteration of the optimisation routine which is not guaranteed with gradient ascent. If the second derivative of the log marginal likelihood is available, one can also consider applying the reduced Newton's method to maximise the log marginal likelihood.

However, one can take a fully Bayesian approach to infer the Gaussian process hyperparameters, the state variable as well as the exogenous input disturbance sensor variance parameters by specifying prior distributions for the aforementioned unknowns. Due to the nonlinear relationship between the unknown parameters, the kernel function and the inverse covariance matrix, one typically can not exploit conjugacy. Furthermore, depending on which of the unknown parameters one considers, different prior distributions can be specified depending on the properties of the unknown parameter (negative values allowed, positive values only, etc.). In a typical setting, an MCMC algorithm is used to infer the aforementioned unknown parameters for noise-corrupted time series data after specifying the appropriate prior distributions.

CHAPTER 7: RECOMMENDATIONS

Gaussian process kernel function selection. Recall from Section 4.5.2 that the author relaxed the assumption that the structure, characteristics and generation process of the exogenous input disturbance is known precisely, but constrained the exogenous input to exhibit global and local smooth behaviour via the choice of kernel function. A similar argument was used to encode the expected behaviour of the state variable dynamic response. However, this was a form of prior knowledge that the author encoded that might not necessarily reflect the expected underlying behaviour of all types of physical systems. If the engineer expects that the state variable and exogenous input disturbance do not display such global and local smooth behaviour, the expected behaviour can simply be encoded by specifying a different type of kernel function. However, the engineer should ensure that the specification of the kernel function results in a symmetric positive-definite covariance matrix. The ideal scenario would be to incorporate automatic model selection such that the specific form of the kernel function is automatically learned from the noise-corrupted time series data. The interested reader is referred to the work of Duvenaud (2014) which addresses the problem of automatic model selection/construction with Gaussian processes.

Gaussian process scalability. The central problem that arises with Gaussian process regression is the inversion of the $N \times N$ matrix \mathbf{C}_{dd} which requires $O(N^3)$ computations (Section 2.5.2). This computational cost prohibits the use of Gaussian process regression for large data sets. Thus, if one considers the extension of the proposed Bayesian methodology derived in Section 4.5.2 to systems of linear or nonlinear (in the parameter) ODE models, the dimensions of matrix \mathbf{B}_{GP} (Equation 3.50) increase with each new ODE added to the system of ODEs. For example, for Case Study 2 (which is described by a single ODE) with a total of $N = 121$ sensor measurements for the state variable and exogenous input disturbance, respectively, the matrix \mathbf{B}_{GP} is 242×242 , i.e. $2N \times 2N$. If the engineer incorporates an additional ODE (for a total of two ODEs describing the physical system) where the additional ODE predicts a single state variable and has a single exogenous input disturbance, the matrix \mathbf{B}_{GP} increases in size to 484×484 , i.e. $4N \times 4N$.

Although a matrix this size is easily inverted using standard methods, for large systems consisting of hundreds of ODEs, the matrix inversion process becomes prohibitively slow. Thus, the engineer requires techniques that will allow Gaussian processes to scale to big data settings. The reader is referred to Titsias (2009) which provides a variational inference based approximation for Gaussian process regression based on the idea of inducing variables which allow Gaussian processes to scale to big data settings. Furthermore, Quíñonero-Candela and Rasmussen (2005) provide a comprehensive overview of various other techniques that are typically used to scale Gaussian processes to big data settings. Among these techniques is the popular fully independent training conditional (FITC) Gaussian process approximation approach (Snelson and Ghahramani, 2005).

7.4. Algorithm 4 Practicalities

Caution should be taken when implementing *Algorithm 4*. If the discrepancy given by Equation 3.83 at each time point t_i is too large, i.e. the Gaussian process is not interpolating the underlying state variable derivative function values accurately enough, the ODE parameter estimates obtained from *Algorithm 4* can be misleading. This can be due to (1) extremely noise-corrupted time series data, (2) the gradient ascent optimisation routine might converge to a local optimum for the Gaussian process hyperparameters that do not describe the data set

CHAPTER 7: RECOMMENDATIONS

(Section 5.4) or (3) the selected kernel functions might not sufficiently describe the underlying function values. The aforementioned practicalities should be taken in to consideration when implementing *Algorithm 4*.

Also, recall that one hundred additional independent data sets were generated with the same exogenous input disturbance signal and sensor variance parameters to determine whether the Bayesian (*Algorithm 4*) and the frequentist (*Algorithm 2*) approach provides consistent results when compared to the simulation ground truth parameter values, based on performance criteria 7 through 10 (Section 4.8). However, one can also consider varying the exogenous input disturbance signal and sensor variance parameter values to determine whether *Algorithm 4* will still provide consistent results across different experimental starting conditions and measuring devices.

By using the building blocks provided in Chapter 2, Sections 4.5.1 and 4.5.2, as well as the recommendations provided in this chapter, it is possible for the engineer to develop a bespoke solution for estimating the parameters of a system of ODEs describing some physical system with the additional benefit of explicitly modeling the uncertainty associated with each model parameter.

References

- Ackermann, E. R., De Villiers, J. P. and Cilliers, P. J. (2011) ‘Nonlinear dynamic systems modeling using Gaussian processes: Predicting ionospheric total electron content over South Africa’, *Journal of Geophysical Research: Space Physics*, 116, pp. 1–13.
- Attias, H. (1999) ‘Inferring Parameters and Structure of Latent Variable Models by Variational Bayes’, *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pp. 21–30. Available at: <http://arxiv.org/abs/1301.6676>.
- Au, S. K. (2012) ‘Connecting Bayesian and frequentist quantification of parameter uncertainty in system identification’, in *Mechanical Systems and Signal Processing*.
- Barber, D. (2012) *Bayesian Reasoning and Machine Learning*. Cambridge, New York: Cambridge University Press.
- Barber, D. and Wang, Y. (2014) ‘Gaussian Processes for Bayesian Estimation in Ordinary Differential Equations’, *Proceedings of the 31st International Conference on Machine Learning*, pp. 1485–1493.
- Bernardo, J. M. and Smith, A. F. M. (1994) *Bayesian theory*. Chichester, West Sussex: Wiley.
- Bishop, C. M. (2006) *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer.
- Blei, D. M., Kucukelbir, A. and McAuliffe, J. D. (2018) ‘Variational Inference : A Review for Statisticians’, *Journal of the American Statistical Association*, 112, pp. 1–41.
- Box, G. E. P. and Tiao, G. C. (2011) *Bayesian Inference in Statistical Analysis*. Wiley.
- Calderhead, B., Girolami, M. and Lawrence, N. D. (2009) ‘Accelerating Bayesian Inference over Nonlinear Differential Equations with Gaussian Processes’, in *Advances in Neural Information Processing Systems 21*, pp. 217–224.
- Campbell, D. and Steele, R. J. (2012) ‘Smooth functional tempering for nonlinear differential equation models’, *Statistics and Computing*, 22, pp. 429–443.
- Chappell, M. A. *et al.* (2009) ‘Variational Bayesian Inference for a Nonlinear Forward Model’, *IEEE Transactions on Signal Processing*, 57, pp. 223–236.
- Cox, R. T. (1946) ‘Probability, Frequency and Reasonable Expectation’, *American Journal of Physics*, 14, pp. 1–13.
- Cressie, N. A. C. (1993) *Statistics for spatial data*. Revised Ed. New York: Wiley.
- Croaze, A., Pittman, L. and Reynolds, W. (2012) ‘Solving Nonlinear Least-Squares Problems With the Gauss-Newton and Levenberg-Marquardt Methods’, pp. 1–16. Available at: <https://www.math.lsu.edu/system/files/MunozGroup1 - Paper.pdf>.
- Dondelinger, F. *et al.* (2013) ‘ODE Parameter Inference Using Adaptive Gradient Matching With Gaussian Processes’, *Sixteenth International Conference on Artificial Intelligence and Statistics*, pp. 216–228.
- Duvenaud, D. K. (2014) *Automatic Model Construction with Gaussian Processes, PhD Thesis*. Available at: <https://www.cs.toronto.edu/~duvenaud/thesis.pdf>.

REFERENCES

- Edwards, W., Lindman, H. and Savage, L. J. (1963) 'Bayesian Statistical Inference for Psychological Research', *Psychological Review*, 70, pp. 193–242.
- Englezos, P. and Kalogerakis, N. (2001) *Applied Parameter Estimation for Chemical Engineers*. CRC Press.
- Fogler, H. S. (2006) *Elements of Chemical Reaction Engineering*. 4th ed. Upper Saddle River, NJ: Prentice Hall PTR.
- Fox, C. W. and Roberts, S. J. (2012) 'A Tutorial on Variational Bayesian Inference', *Artificial Intelligence Review*, 38, pp. 85–95.
- Gelman, A. *et al.* (2004) *Bayesian Data Analysis*. 2nd ed. Chapman & Hall/CRC.
- Ghahramani, Z. and Beal, M. J. (2000) 'Variational Inference for Bayesian Mixtures of Factor Analyses', in *Advances in Neural Information Processing Systems 12*, pp. 449–455.
- Ghasemi, A., Yacout, S. and Ouali, M.-S. (2009) 'Parameter Estimation for Condition Based Maintenance with Proportional Hazard Model', in *International Conference on Industrial Engineering and Systems Management*, pp. 1–7.
- Ghasemi, A., Yacout, S. and Ouali, M. S. (2010) 'Parameter Estimation Methods for Condition-Based Maintenance With Indirect Observations', *IEEE Transactions on Reliability*, 59, pp. 426–439.
- Gorbach, N. S., Bauer, S. and Buhmann, J. M. (2017) 'Scalable Variational Inference for Dynamical Systems', *31st Conference on Neural Information Processing Systems*.
- Hastie, T. *et al.* (2006) *An Introduction to Statistical Learning*. Springer Texts.
- Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. 2nd ed. Springer.
- Hazra, A. (2017) 'Using the Confidence Interval Confidently', *Journal of Thoracic Disease*, 9, pp. 4125–4130.
- Hinton, G. E. (2002) 'Training Products of Experts by Minimizing Contrastive Divergence', *Neural Computation*, 14, pp. 1771–1800.
- Isermann, R. (1984) 'Process Fault Detection Based on Modelling and Estimation Methods - A survey', *Automatica*, 20, pp. 387–404.
- Isermann, R. (1985) 'Process Fault Diagnosis with Parameter Estimation Methods', *IFAC Proceedings Volumes*, 18, pp. 51–60.
- Jardine, A. K. S., Lin, D. and Banjevic, D. (2006) 'A Review on Machinery Diagnostics and Prognostics Implementing Condition-Based Maintenance', *Mechanical Systems and Signal Processing*, 20, pp. 1483–1510.
- Jaynes, E. T. (2003) *Probability Theory : The Logic of Science*. Edited by G. L. Bretthorst. Cambridge University Press.
- Jordan, M. I. *et al.* (1999) 'An Introduction to Variational Methods for Graphical Models', *Machine Learning*, 37, pp. 183–233.
- Kocijan, J. *et al.* (2005) 'Dynamic systems identification with Gaussian processes', *Mathematical and Computer Modelling of Dynamical Systems*, 11, pp. 411–424.

REFERENCES

- Kocijan, J. (2016) *Modelling and Control of Dynamic Systems Using Gaussian Process Models*. Springer.
- Kullback, S. and Leibler, R. A. (1951) ‘On Information and Sufficiency’, *Annals of Mathematical Statistics*, 22, pp. 79–86.
- Lai, W. H., Kek, S. L. and Tay, K. G. (2017) ‘Solving Nonlinear Least Squares Problem Using Gauss-Newton Method’, *International Journal of Innovative Science, Engineering & Technology*, 4, pp. 258–262.
- Ljung, L. (1999) *System Identification (Second Edition): Theory for the User*. Upper Saddle River, NJ, USA: Prentice Hall PTR.
- Macdonald, B., Higham, C. and Husmeier, D. (2015) ‘Controversy in Mechanistic Modelling with Gaussian Processes’, in *Proceedings of the 32nd International Conference on Machine Learning*, PMLR, pp. 1539–1547.
- MacKay, D. J. C. (1996) ‘Hyperparameters: Optimize, or Integrate Out?’, in Heidbreder, G. R. (ed.) *Maximum Entropy and Bayesian Methods*. Springer, pp. 43–59.
- MacKay, D. J. C. (2004) *Information Theory, Inference, and Learning algorithms*. Cambridge, U.K.: Cambridge University Press.
- Marlin, T. E. (2000) *Process Control : Designing Processes and Control Systems for Dynamic Performance*. 2nd ed. Boston: McGraw-Hill.
- Martin, K. F. (1994) ‘A Review by Discussion of Condition Monitoring and Fault Diagnosis in Machine Tools’, *International Journal of Machine Tools and Manufacture*, 34, pp. 527–551.
- Migal, Y. F. (2008) ‘Inverse Problem in XANES Theory’, *Journal of Structural Chemistry*, 49, pp. 92–101.
- Minka, T. P. (2001) ‘Pathologies of Orthodox Statistics’, pp. 1–7. Available at: <https://www.microsoft.com/en-us/research/publication/pathologies-orthodox-statistics/>.
- Murphy, K. P. (2012) *Machine Learning: A Probabilistic Perspective*. The MIT Press.
- Neyman, J. (1937) ‘Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability’, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 236, pp. 333–380.
- Van Noordwijk, J. M. and Pandey, M. D. (2003) ‘A Stochastic Deterioration Process for Time-Dependent Reliability Analysis’, *Proceedings of the Eleventh IFIP WG 7.5 Working Conference on Reliability and Optimization of Structural Systems*, pp. 259–265.
- Papoulis, A. and Pillai, S. U. (2002) *Probability, Random Variables, and Stochastic Processes*. Fourth Ed. Boston: McGraw Hill.
- Poshtan, J., Doraiswami, R. and Stevenson, M. (1997) ‘A Real-Time Fault Diagnosis and Parameter Tracking Scheme’, *Proceedings of the American Control Conference*, pp. 483–487.
- Quiñonero-Candela, J. and Rasmussen, C. E. (2005) ‘A Unifying View of Sparse Approximate Gaussian Process Regression’, *Journal of Machine Learning Research*, 6, pp. 1939–1959.
- Rasmussen, C. E. and Williams, C. K. I. (2006) *Gaussian Processes for Machine Learning*. The MIT Press.

REFERENCES

- Särkkä, S. (2019) ‘The Use of Gaussian Processes in System Identification’, in *(to appear in) Encyclopedia of Systems and Control, 2nd edition*, pp. 1–13. Available at: <https://arxiv.org/abs/1907.06066>.
- Van de Schoot, R. *et al.* (2014) ‘A Gentle Introduction to Bayesian Analysis: Applications to Developmental Research’, *Child Development*, 85, pp. 842–860.
- Seborg, D. E. *et al.* (2011) *Process Dynamics and Control (International Student Version)*. 3rd ed. Hoboken, N.J.: John Wiley & Sons.
- Shannon, C. E. (1948) ‘A Mathematical Theory of Communication’, *Bell System Technical Journal*, 27, pp. 379–423.
- Sivia, D. S. and Skilling, J. (2012) *Data Analysis: A Bayesian Tutorial*. 2nd ed. Oxford Science Publications.
- Smith, J. Q. (2010) *Bayesian Decision Analysis: Principles and Practice*. Cambridge University Press.
- Snelson, E. and Ghahramani, Z. (2005) ‘Sparse Gaussian processes using pseudo-inputs’, in *Advances in Neural Information Processing Systems*, pp. 1257–1264.
- Tarantola, A. (2005) *Inverse Problem Theory and Methods for Model Parameter Estimation*. Society for Industrial and Applied Mathematics, Philadelphia.
- Titsias, M. K. (2009) ‘Variational learning of inducing variables in sparse Gaussian processes’, in *Proceedings of Machine Learning Research*, pp. 567–574.
- Wenk, P. *et al.* (2018) ‘Fast Gaussian Process Based Gradient Matching for Parameter Identification in Systems of Nonlinear ODEs’, *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*.
- Woolrich, M. W. and Behrens, T. E. (2006) ‘Variational Bayes Inference of Spatial Mixture Models for Segmentation’, *IEEE Transactions on Medical Imaging*, 25, pp. 1380–1391.
- Zhiqiang, G. and Zhihuan, S. (2013) *Multivariate Statistical Process Control Process Monitoring Methods and Applications*. Springer.

Appendix A

Illustrative Example

“We cannot understand the scientific method without a previous investigation of the different kinds of probability”

- Bertrand Russel, 1948

Appendix A is an expansion of Section 2.1.3 with the aim of introducing the reader to Bayes’ rule for parameter estimation problems which is central to understanding the work presented in the current thesis. The subsequent discussion is kept relatively informal in order to avoid unnecessary complexity.

Suppose, as the reader, you are working on an industrial plant and a senior engineer approaches you with a data set of $N = 3$ draining tank liquid level measurements. The senior engineer tells you that she models the steady state behaviour of the liquid draining tank using a univariate Gaussian distribution (Section 2.3.1, Equation 1.24) and that the three draining tank liquid level measurements are generated from the Gaussian distribution. Your first task, as a new junior engineer, is to find the mean of the Gaussian distribution the senior engineer is using given that she provided you with the 3 generated liquid level measurements. The senior engineer asks you to report back to her before the end of the working day, however, before she leaves for her coffee break, she hints that the mean of the Gaussian distribution is around $\mu_L = 7.0 \text{ m}$ but the variance is definitely $\sigma_L^2 = 0.04 \text{ m}^2$.

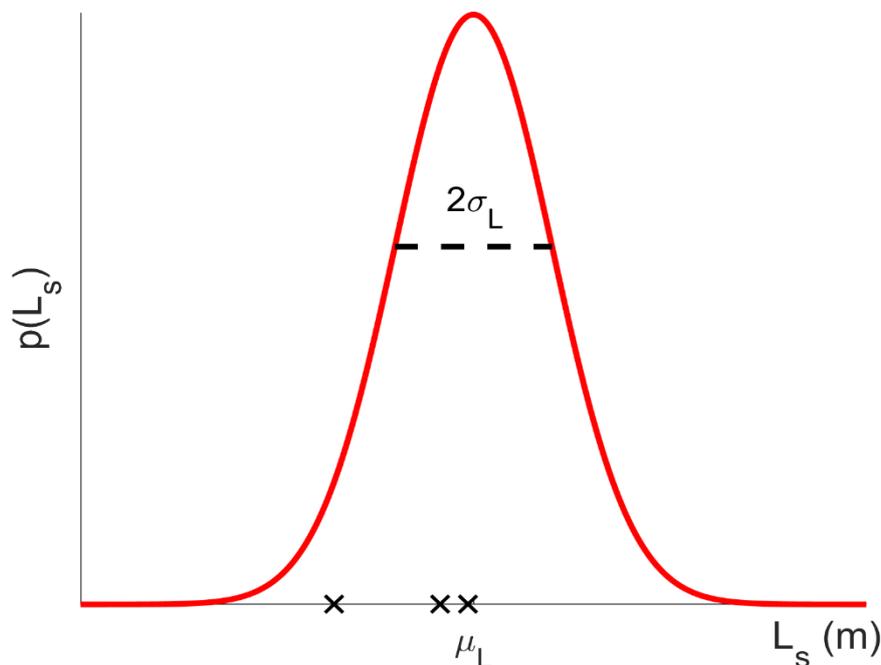


Figure A.1: Plot of the steady state liquid level univariate Gaussian showing the mean μ_L (unknown) and the standard deviation σ_L (known). The three crosses represent the generated liquid level measurements obtained from the senior engineer.

APPENDIX A

After the senior engineer departs, you receive an email from her with the three steady state liquid level L_s measurements (Table A.1) and Figure A.1. Curiously, you ask around about the senior engineer and finds out she has 15 year of experience on the plant.

Table A.1: Generated steady state liquid level measurements provided by the senior engineer

| Liquid Level Measurement L_s | Measurement Value (m) |
|--------------------------------|---------------------------|
| 1 | 6.92 |
| 2 | 6.65 |
| 3 | 6.99 |

As a junior engineer, how would you address this problem?

It is common to assume that data is generated in an independent manner. In the outlined scenario, the junior engineer can assume that the liquid level measurements are generated independently from the univariate Gaussian distribution used by the senior engineer (Figure A.1). Data points that are drawn independently from the same distribution are referred to as *independent and identically distributed*, abbreviated i.i.d., data. Because the liquid level measurements are i.i.d., the joint probability of the data set, given μ_L , can be written as the product of marginal probabilities (Equation 1.4). In other words:

$$p(L_{s1}, L_{s2}, L_{s3} | \mu_L) = \prod_{i=1}^3 \mathcal{N}(L_{si} | \mu_L, \sigma_L^2) \quad \text{Equation A.1}$$

For notational convenience, group the liquid level measurements into a 3×1 vector such that $\mathbf{L}_s = [L_{s1} \ L_{s2} \ L_{s3}]^T$. Equation A.1 can now be expressed as follows:

$$p(\mathbf{L}_s | \mu_L) = \prod_{i=1}^3 \mathcal{N}(L_{si} | \mu_L, \sigma_L^2) \quad \text{Equation A.2}$$

Note that the quantity $p(\mathbf{L}_s | \mu_L)$ is evaluated for the liquid level measurements and, when viewed as a function of the unknown parameter μ_L (recall σ_L^2 is known), is referred to as the *likelihood function*. Depending on what literature sources are used, the likelihood function may also be denoted by $\mathcal{L}(\mu_L; L_{s1}, L_{s2}, L_{s3})$ or $\mathcal{L}(\mu_L; \mathbf{L}_s)$ (Bishop, 2006; Hastie et al., 2006; Hastie, Tibshirani and Friedman, 2009; Barber, 2012; Murphy, 2012).

It is important to note that while the likelihood function expresses how probable the liquid level measurements are for different settings of the parameter μ_L , the likelihood function is *not* a probability distribution over μ_L . This is because the integral of the likelihood function with respect to μ_L does not (necessarily) equal 1. In both the Bayesian and frequentist viewpoints, the likelihood function given by Equation A.2 plays an important role, however, the way in which it is used differs.

A.1. Frequentist Viewpoint

From the frequentist viewpoint, the parameter μ_L is considered to be fixed but unknown and the value of μ_L is determined using an ‘estimator’. An estimator can informally be thought of as a way of ‘selecting’ the value of μ_L based on the liquid level measurements. A popular and widely used estimator in frequentist statistics stems from the *Maximum Likelihood* heuristic which states that the maximum likelihood estimator for μ_L , denoted by $\mu_{L,ML}$, should have the property that (Minka, 2001):

$$\mu_{L,ML} = \operatorname{argmax}_{\mu} (p(\mathbf{L}_s|\mu)) \quad \text{Equation A.3}$$

$$\mu_{L,ML} = \operatorname{argmax}_{\mu} \left(\prod_{i=1}^3 \mathcal{N}(L_{si}|\mu_L, \sigma_L^2) \right) \quad \text{Equation A.4}$$

Note that the maximum likelihood estimator $\mu_{L,ML}$ is often also denoted as $\hat{\mu}_L$ in the statistics literature. However, the author uses the subscript notation $\mu_{L,ML}$ throughout the thesis. Based on the maximum likelihood heuristic, the junior engineer should select the value of μ_L that maximises the likelihood function given by Equation A.4. In order to analytically evaluate Equation A.4, it is easier to work with the natural logarithm of the likelihood function. Since the natural logarithm is a monotonic transformation, maximising the natural logarithm of the likelihood function is equivalent to maximising the likelihood function. In other words,

$$\mu_{L,ML} = \operatorname{argmax}_{\mu} \left(\prod_{i=1}^3 \mathcal{N}(L_{si}|\mu_L, \sigma_L^2) \right) = \operatorname{argmax}_{\mu} \left(\ln \left\{ \prod_{i=1}^3 \mathcal{N}(L_{si}|\mu_L, \sigma_L^2) \right\} \right) \quad \text{Equation A.5}$$

The natural logarithm of the likelihood function is typically referred to in the machine learning literature as the *log likelihood*. Writing out the log likelihood results in the following:

$$\ln p(\mathbf{L}_s|\mu) = -\frac{1}{2\sigma_L^2} \sum_{i=1}^3 (L_{si} - \mu_L)^2 - \frac{3}{2} \ln \sigma_L^2 - \frac{3}{2} \ln 2\pi \quad \text{Equation A.6}$$

Maximising Equation A.6 with respect to the μ_L simply amounts to taking the derivative of $\ln p(\mathbf{L}_s|\mu)$, equating it to zero and solving for the maximum likelihood estimate $\mu_{L,ML}$. The junior engineer can show that:

$$\mu_{L,ML} = \frac{1}{3} \sum_{i=1}^3 L_{si} \quad \text{Equation A.7}$$

In other words, based on the maximum likelihood heuristic, the setting of the parameter μ_L that the junior engineer should use simply corresponds to the sample mean of the generated liquid level measurements provided by the senior engineer. In general, the junior engineer can show that for N observations, the maximum likelihood setting for μ_L corresponds to:

$$\mu_{L,ML} = \frac{1}{N} \sum_{i=1}^N L_{si} \quad \text{Equation A.8}$$

APPENDIX A

Equation A.8 provides a convenient opportunity to introduce the concept of sequential estimation from the frequentist viewpoint. Sequential estimation allows the engineer to process measurements one at a time, after which the measurement may be discarded or retained, and is important for on-line applications in an engineering setting. Following the discussion in Bishop (2006), if the junior engineer inspects Equation A.8 closely, they can show that the contribution of the final liquid level measurement L_{s3} can be obtained from:

$$\mu_{L,ML}^{(3)} = \mu_{L,ML}^{(2)} + \frac{1}{3} \left(L_{s3} - \mu_{L,ML}^{(2)} \right) \quad \text{Equation A.9}$$

The superscript notation $\mu_{L,ML}^{(2)}$ and $\mu_{L,ML}^{(3)}$ refers to the maximum likelihood estimate for μ_L when it is based on 2 and 3 liquid level measurements, respectively. Equation A.9 has the following interpretation: after observing two data points, the junior engineer can estimate μ_L with $\mu_{L,ML}^{(2)}$. Once the new liquid level measurement becomes available, the junior engineer can obtain a revised estimate for μ_L by moving the old estimate, proportional to the factor $1/3$, in the direction of the ‘error signal’ given by $(L_{s3} - \mu_{L,ML}^{(2)})$. In general, the sequential update equation for the maximum likelihood estimate of μ_L for N liquid level measurements is given by:

$$\mu_{L,ML}^{(N)} = \mu_{L,ML}^{(N-1)} + \frac{1}{N} \left(L_{sN} - \mu_{L,ML}^{(N-1)} \right) \quad \text{Equation A.10}$$

From Equation A.10 one observes that as N increases the contribution of each successive liquid level measurement to the estimate of μ_L decreases.

A.2. Bayesian Viewpoint

From the frequentist viewpoint, one observes that the inference procedure for μ_L only requires the likelihood function given by Equation A.2. The parameter μ_L was treated as unknown but fixed and estimated from the maximum likelihood heuristic. From the Bayesian viewpoint, the parameter μ_L is treated as unknown and thus μ_L is modeled as a random variable. Therefore, the junior engineers is uncertain about its value and defines a distribution over values for μ_L .

Recall from Equation 1.10 in Section 2.1.3 that Bayes’ rule for estimating the parameters of a lumped system dynamic model is given by:

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})}$$

However, the junior engineer is not interested in making conclusions about lumped system dynamic model parameters. Rather, the junior engineer wants to make conclusions about μ_L from the three liquid level measurements that the senior engineer generated from the univariate Gaussian distribution depicted in Figure A.1. Consequently, the junior engineer can simply rewrite Equation 1.10 for the inference problem at hand such that:

$$p(\mu_L|\mathcal{D}) = \frac{p(\mathcal{D}|\mu_L)p(\mu_L)}{p(\mathcal{D})} \quad \text{Equation A.11}$$

APPENDIX A

The notation \mathcal{D} simply refers to the liquid level measurements that the senior engineer provided, i.e. $\mathcal{D} = \{L_{si}\}_{i=1}^N$. The quantity $p(\mathcal{D}|\mu_L) = p(L_{s1}, L_{s2}, L_{s3}|\mu_L)$ is identical to Equation A.1 and, therefore, represents the likelihood function. Again, the author emphasises that the likelihood function is *not* a probability distribution over μ_L . The quantity $p(\mu_L)$ is referred to as the *prior* distribution over parameter μ_L and expresses the junior engineer's initial belief about what values the parameter μ_L can take and their relative plausibility before observing any liquid level measurements. By combining the prior distribution over parameter μ_L with the likelihood function (Equation A.1), the junior engineer can obtain an updated belief about the values of μ_L in the form of the conditional distribution $p(\mu_L|\mathcal{D})$, also referred to as the *posterior* distribution (Bishop, 2006).

The quantity $p(\mathcal{D})$ in the denominator of Equation A.11 is the normalisation constant which ensures that the posterior distribution $p(\mu_L|\mathcal{D})$ is a valid probability distribution, i.e. integrates to one over the domain of μ_L . The normalisation constant is also referred to as the *evidence* or *marginal likelihood* in machine learning literature (Bishop, 2006; Hastie et al., 2006; Murphy, 2012).

From the sum (Equation 1.2) and product (Equation 1.3) rules for continuous random variables, as well as the univariate Gaussian discussed in Section 2.3.1, one can show that the evidence can be calculated as follows:

$$p(\mathcal{D}) = \int_{-\infty}^{\infty} p(\mathcal{D}|\mu_L)p(\mu_L)d\mu_L \quad \text{Equation A.12}$$

Since the junior engineer made the assumption that the three liquid level measurements provided by the senior engineer are generated in an i.i.d. manner (see Equation 1.4) from the univariate Gaussian depicted in Figure A.1, Bayes' rule for inferring a posterior distribution over μ_L (Equation A.11) can be rewritten such that:

$$p(\mu_L|\mathcal{D}) = \frac{\prod_{i=1}^3 \mathcal{N}(L_{si}|\mu_L, \sigma_L^2) p(\mu_L)}{\int_{-\infty}^{\infty} \prod_{i=1}^3 \mathcal{N}(L_{si}|\mu_L, \sigma_L^2) p(\mu_L)d\mu_L} \quad \text{Equation A.13}$$

The question that remains unanswered is: “*How does the junior engineer select the prior distribution $p(\mu_L)$?*” Within the fields of statistics and machine learning, it is common to exploit *conjugacy*. Although conjugacy is not directly addressed here, the reader is referred to Section 2.3.3 where the concept of conjugacy is explained. The main idea of conjugacy is as follows: the junior engineer can specify a specific functional form for the prior distribution over μ_L such that when they combine it with the likelihood function given by Equation A.1, the posterior distribution will have the exact same functional form as the prior. Since the posterior distribution $p(\mu_L|\mathcal{D})$ will have the exact same functional form as the prior distribution, there is no need to evaluate the denominator in Equation A.13. The posterior distribution can simply be normalised by looking up the appropriate normalisation constant which will stem from the choice of prior distribution. Also, conjugacy allows the junior engineer to obtain closed-form update solutions for the posterior distribution. The conjugate prior for the likelihood function given by Equation A.1 is another univariate Gaussian distribution such that:

APPENDIX A

$$p(\mu_L) = \mathcal{N}(\mu_L | \mu_0, \sigma_0^2) \quad \text{Equation A.14}$$

The parameters μ_0 and σ_0 are known as *hyperparameters* and express the junior engineer's initial belief about what values the parameter μ_L is likely to take on. Since the junior engineer exploited conjugacy, the posterior distribution over μ_L is proportional to the product of the prior and the likelihood function. In other words,

$$p(\mu_L | \mathcal{D}) \propto p(\mathcal{D} | \mu_L) p(\mu_L) \quad \text{Equation A.15}$$

$$p(\mu_L | \mathcal{D}) \propto \prod_{i=1}^3 \mathcal{N}(L_{si} | \mu_L, \sigma_L^2) \mathcal{N}(\mu_L | \mu_0, \sigma_0^2) \quad \text{Equation A.16}$$

By completing the square in the exponent when expanding Equation A.16, it can be shown that the posterior distribution over μ_L is another Gaussian distribution such that:

$$p(\mu_L | \mathcal{D}) = \mathcal{N}(\mu_L | \mu_3, \sigma_3^2) \quad \text{Equation A.17}$$

$$\text{where,} \quad \mu_3 = \frac{\sigma_L^2}{3\sigma_0^2 + \sigma_L^2} \mu_0 + \frac{3\sigma_0^2}{3\sigma_0^2 + \sigma_L^2} \mu_{ML} \quad \text{Equation A.18}$$

$$\frac{1}{\sigma_3^2} = \frac{1}{\sigma_0^2} + \frac{3}{\sigma_L^2} \quad \text{Equation A.19}$$

In general, the junior engineer can show that:

$$\mu_N = \frac{\sigma_L^2}{N\sigma_0^2 + \sigma_L^2} \mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma_L^2} \mu_{ML} \quad \text{Equation A.20}$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma_L^2} \quad \text{Equation A.21}$$

From Equation A.20, one observes that the posterior distribution mean μ_N is a compromise between the prior distribution mean μ_0 and the frequentist maximum likelihood estimate for μ_L . If the junior engineer had no liquid level measurements, i.e. $N = 0$, the posterior mean reduces back to the prior mean. For the case where $N \rightarrow \infty$, the posterior mean converges to frequentist maximum likelihood estimate for μ_L . From Equation A.21, observe that the posterior distribution variance is naturally expressed in terms of the precision, i.e. the inverse variance. This is also the way in which most introductory machine learning texts represent the variance.

Furthermore, note from Equation A.21 that the precisions are additive. The first precision contribution stems from the prior distribution over μ_L while the second 'data' precision contribution is obtained from the liquid level measurements. Specifically, each liquid level measurement contributes one count. As the number of liquid level measurements increase, the posterior distribution variance will decrease. When no liquid level measurements are available, the posterior variance reduces to the prior variance. When $N \rightarrow \infty$, the posterior variance goes

APPENDIX A

to zero and the posterior distribution becomes infinitely peaked around the frequentist maximum likelihood estimate for μ_L . Therefore, one observes that the frequentist maximum likelihood results for μ_L can be recovered from the Bayesian viewpoint in the limit of an infinite number of liquid level measurements (Bishop, 2006).

Bayes' rule is inherently online. As new observations become available, the current posterior distribution can simply be used as the prior distribution for future analysis. Thus, Bayes' rule is ideal for sequential parameter estimation in an online setting. Since the junior engineer now has two different frameworks for making conclusions about the parameter μ_L , they can provide the senior engineer with a value for the parameter μ_L . However, if the junior engineer chooses to work in the Bayesian framework, they must decide on values for the hyperparameters for the prior distribution over μ_L . After giving it some thought, the junior engineer remembers that the senior engineer said she thinks the mean parameter is around $\mu_L = 7.0 \text{ m}$. Also, the senior engineer has 15 year of experience on the plant. As a result, the junior engineer decides to set the prior distribution hyperparameters (Equation A.14) to $\mu_0 = 7.0 \text{ m}$ and $\sigma_0^2 = 0.01 \text{ m}^2$. Figure A.2 depicts the junior engineer's prior distribution over μ_L .

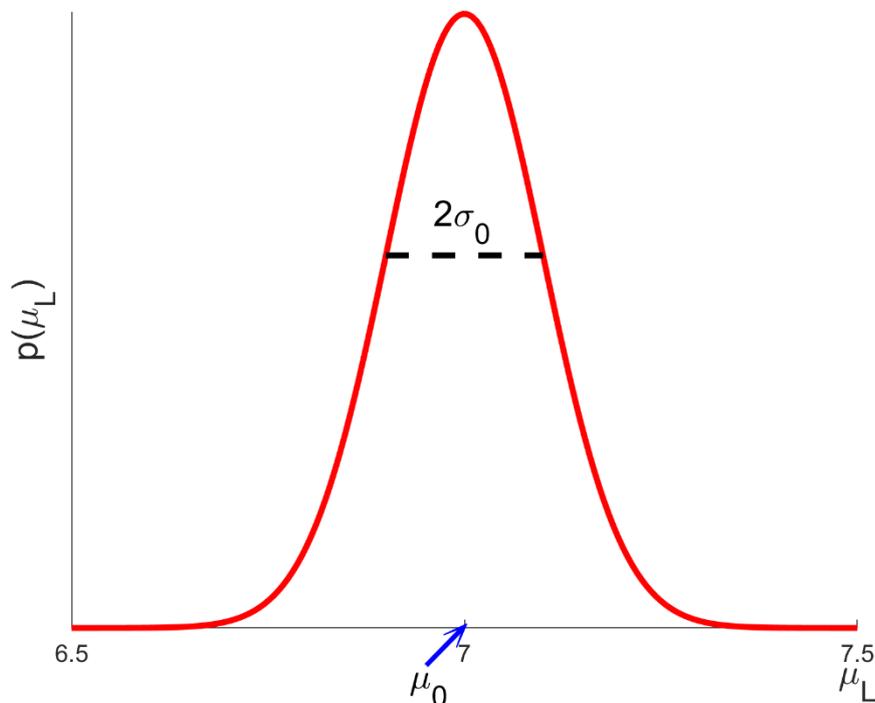


Figure A.2: Depiction of the junior engineer's prior belief about the values of the parameter μ_L . The distribution is centred at $\mu_0 = 7.0 \text{ m}$ with a variance of $\sigma_0^2 = 0.01 \text{ m}^2$.

Intuitively, the prior distribution in Figure A.2 can be interpreted as assigning high probability to the range of values between 6.5 m and 7.5 m for the parameter μ_L before observing any liquid level measurements. However, this is not strictly correct since the univariate Gaussian distribution provides support for $\mu_L \in \mathbb{R}$. The selection of the prior hyperparameters stem from knowledge the junior engineer acquired before performing inference, i.e. the junior engineer constructed the prior belief based on the information obtained from the senior engineer and other work colleagues. This is why the Bayesian inference framework is sometimes also referred to as the *subjective inference* framework since the framework is subjective to the

APPENDIX A

experimentalist, i.e. the junior engineer for the outlined scenario (Bernardo and Smith, 1994; Gelman et al., 2004) Based on the selected prior hyperparameters, the posterior distribution hyperparameters can be evaluated from Equation A.18 and A.19, respectively, such that:

$$\mu_3 = 6.94 \text{ m} \quad \text{Equation A.22}$$

$$\sigma_3^2 = 0.006 \text{ m}^2 \quad \text{Equation A.23}$$

The resulting posterior distribution over the unknown parameter μ_L is illustrated in Figure A.3.

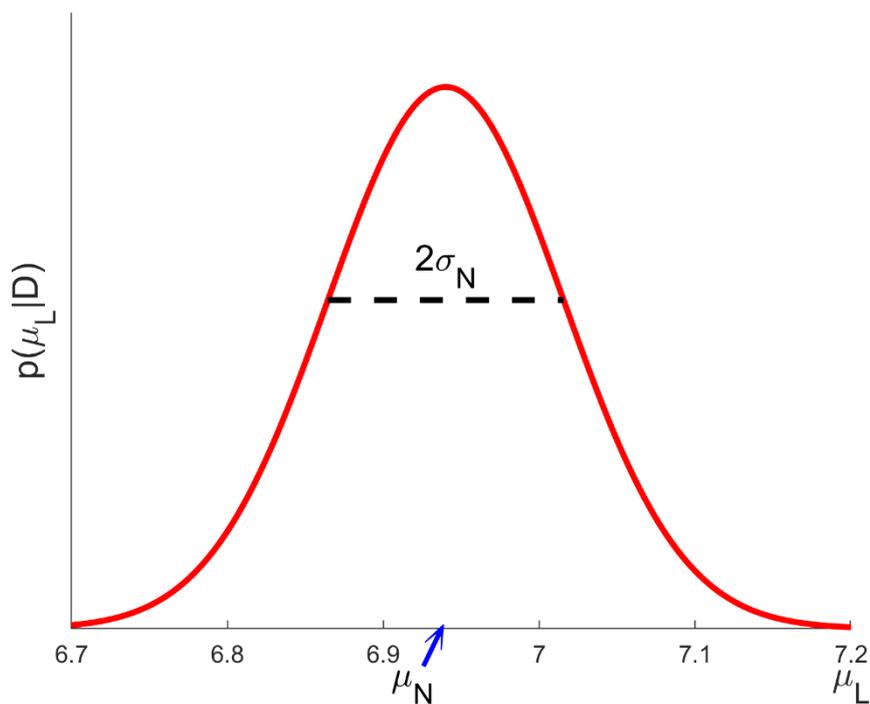


Figure A.3: Depiction of the junior engineer's posterior belief about the values of the parameter μ_L after observing three liquid level measurements. The distribution is centred at $\mu_N = 6.94 \text{ m}$ with a variance of $\sigma_N^2 = 0.006 \text{ m}^2$ (cf. – Figure A.2 prior distribution).

Suppose the junior engineer has to provide the senior engineer with a single value for μ_L . The problem the junior engineer now faces is selecting a value for μ_L from the posterior distribution $p(\mu_L | \mathcal{D})$. A common choice is to select the setting of the parameter μ_L that maximises the posterior distribution. This setting of the parameter μ_L is commonly referred to as the *maximum a posteriori*, abbreviated MAP, estimate. However, the junior engineer could have selected any other value from the posterior distribution or work with the posterior distribution as a whole. Typically, the selection of μ_L is determined by using decision theory which will allow the junior engineer to make the optimal decision for μ_L in situations involving uncertainty. However, for now, the MAP estimate for μ_L will suffice. Note from Section 2.3.1 that the maximum of a univariate Gaussian is its mode which coincides with the distribution mean (Bishop, 2006).

APPENDIX A

Thus, the MAP estimate of Equation A.17, for the three observed liquid level measurements, corresponds to its mean, i.e. Equation A.22. Table A.2 summarises the parameter estimation results from both the frequentist (Equation A.7) and Bayesian viewpoints.

Table A.2: Inference results for the Gaussian mean parameter μ_L obtained from the three generated liquid level measurements provided by the senior engineer.

| Inference Viewpoint | Symbol Denoting Estimate | Estimate Value | Ground Truth |
|---------------------|--------------------------|----------------|--------------|
| Frequentist | $\mu_{L,ML}$ | 6.85 m | 7.00 m |
| Bayesian | $\mu_{L,MAP}$ | 6.94 m | 7.00 m |

Note from Table A.2 that the Bayesian viewpoint resulted in a slightly better estimate for μ_L , given the junior engineer selects the posterior MAP estimate, due to the Bayesian methodology allowing the junior engineer to incorporate prior knowledge. In other words, the junior engineer could explicitly incorporate the fact that the senior engineer hinted that the true distribution mean is around $\mu_L = 7.0$ m. Furthermore, the junior engineer could also incorporate the ‘trustworthiness’ of the senior engineers’ statement about $\mu_L = 7.0$ m via the Gaussian prior distribution variance parameter. The junior engineer believes that the senior engineer is a skilled expert and, as a result, allowed high probability for a small range of parameter μ_L values around $\mu_L = 7.0$ m via the prior distribution (Figure A.2). Here the idea of subjective inference arises again (Bernardo and Smith, 1994; Gelman et al., 2004).

A.3. On-Line Learning/Sequential Estimation

Suppose the junior engineer emails the senior engineer Table A.2 to check whether the estimates for the mean of the distribution in Figure A.1 are reasonable. The senior engineer replies and states that the estimates in Table A.2 are acceptable, but that she forgot to add one of the liquid level measurements she generated from the distribution in Figure A.1. Attached in the email is a fourth liquid level measurement that corresponds to 7.05 m with instructions to re-estimate the parameter μ_L .

Using the sequential update rule given by Equation A.10, the junior engineer can re-estimate the frequentist maximum likelihood estimate for the parameter μ_L as follows:

$$\mu_{L,ML}^4 = \mu_{L,ML}^3 + \frac{1}{4}(L_{s4} - \mu_{L,ML}^3) \quad \text{Equation A.24}$$

$$\mu_{L,ML}^4 = 6.85m + \frac{1}{4}(7.05m - 6.85m) \quad \text{Equation A.25}$$

$$\mu_{L,ML}^4 = 6.90 m \quad \text{Equation A.26}$$

From the Bayesian perspective, the posterior distribution hyperparameters given by Equation A.22 and A.23, as obtained from the initial three liquid level measurements, can be used as the prior distribution hyperparameters for the new liquid level measurement.

APPENDIX A

The additional liquid level measurement result in the following posterior distribution hyperparameters:

$$\mu_4 = 6.95 \text{ m} \quad \text{Equation A.27}$$

$$\sigma_4^2 = 0.005 \text{ m}^2 \quad \text{Equation A.28}$$

The MAP estimate for μ_L corresponds to $\mu_{L,MAP} = 6.95 \text{ m}$. Table A.3 summarises the estimates for the mean parameter μ_L based on the additional liquid level measurement.

Table A.3: Updated inference results for the Gaussian distribution mean parameter μ_L based on the additional generated liquid level measurement.

| Inference Viewpoint | Symbol Denoting Estimate | Estimate Value | Ground Truth |
|---------------------|--------------------------|----------------|--------------|
| Frequentist | $\mu_{L,ML}$ | 6.90 m | 7.00 m |
| Bayesian | $\mu_{L,MAP}$ | 6.95 m | 7.00 m |

Furthermore, note that an additional benefit of obtaining a posterior distribution $p(\mu_L|\mathcal{D})$ over the unknown parameter μ_L is that the junior engineer can directly extract the uncertainty about the parameter μ_L from the posterior distribution itself.

For example, if the junior engineer wants to construct 99% credibility intervals (Section 2.8) for the parameter μ_L from the Gaussian posterior distribution $p(\mu_L|\mathcal{D})$, the junior engineer simply has to evaluate the expression:

$$\mu_{L,Lower} \leq \mu_L \leq \mu_{L,Upper} \quad \text{Equation A.29}$$

where

$$\mu_{L,Upper} = \mu_N + 2.576 \sqrt{\sigma_N^2} \quad \text{Equation A.30}$$

$$\mu_{L,Lower} = \mu_N - 2.576 \sqrt{\sigma_N^2} \quad \text{Equation A.31}$$

From the frequentist maximum likelihood perspective, it is not obvious from the estimation procedure how the junior engineer should go about quantifying the uncertainty for the estimate $\mu_{L,ML}$. Typically, frequentist confidence intervals are constructed from the *sampling distribution* (Section 2.8) (Bishop, 2006; Murphy, 2012; Van de Schoot et al., 2014).