Model comparison and sparse learning of nonlinear physics-based models using Bayesian inference

by

Rimple Sandhu

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Civil Engineering

Carleton University
Ottawa, Ontario

©2019
Rimple Sandhu

Abstract

This thesis addresses the issue of overfitting while calibrating over-parameterized physical models with noisy and incomplete observations. A Bayesian inversion framework is augmented with model comparison and sparse learning algorithms to identify the optimal model nested under an over-parameterized model. The work is performed in three stages. First, the evidence-based Bayesian model comparison is implemented to rank competing models whereby the evidence is estimated using stationary samples from the posterior parameter probability density function (pdf) generated using a parallel and adaptive Markov Chain Monte Carlo (MCMC) sampler. Second, the concept of automatic relevance determination (ARD) is exploited to reformulate the model comparison problem into a sparse learning problem to alleviate the practical issues of 1) sensitivity of model evidence to prior pdf, and 2) overlooking nested models excluded in the candidate model set. ARD operates by assigning a zero-mean and unknown precision (hyperparameter) prior pdf to questionable model parameters. This ARD-based sparse learning approach is implemented using an MCMC-based evidence estimator and a gradient-free evidence optimizer to compute the optimal hyperparameters, which thereby picks out the optimal nested model. Third, a semi-analytical approach of nonlinear sparse Bayesian learning (NSBL) is proposed to alleviate the computational burden of the MCMC sampling within an optimization task. The analytical tractability of Bayesian entities is enabled by the Gaussian mixture-model approximation of the posterior parameter pdf without the ARD priors. A multistart Newton's method is designed to expedite the non-convex, unconstrained maximization of evidence using semi-analytically computed gradient and Hessian information of evidence. Each of the three proposed approaches is applied to model the complex physics behind a non-

linear aeroelastic oscillator observed in a wind-tunnel. Furthermore, several numerical studies are reported in this thesis to demonstrate the efficiency and robustness of NSBL in pruning redundant parameters from nonlinear physics-based models, leading to enhanced generalization capabilities of the predictive model.

*Dedicated to my wife, Manmeet,
and my son, Viraaj*

Acknowledgments

I want to express my sincere gratitude to my supervisors, Prof. Abhijit Sarkar, Prof. Dominique Poirel and Prof. Chris Pettit, for their unwavering guidance throughout my doctoral studies. They have provided me with the best possible mindset to perform this work, and I am thankful to them for that. I want to particularly thank Prof. Sarkar for his patience and dedication in supporting my quest for knowledge for the past nine years.

I am grateful to Dr. Mohammad Khalil for the stimulating weekly discussions that have produced several interesting ideas and has helped me hone this thesis. I am also deeply appreciative of Dr. Weixing Yuan for his support and guidance, and for providing me with the flexibility to work on my thesis while I was stationed at National Research Council. I also want to thank my colleagues, Philippe, Ajit, Brandon, Sudhi, Leandro, and Chaitanya, for their company and constant support. I am proud to call them my friends.

I would also like to thank my thesis committee: Prof. Eric Hernandez, Prof. Jason Nielsen, Prof. Mohammad Khalil, Prof. Thomas Walker and Prof. Ousmane Seidou for their insightful comments.

Last but not the least, I want to thank my family for sticking with me in this endeavour. My father has been a great confidence booster for me, and I want to thank him for the countless sacrifices he has made to make sure I succeed in life. I also want to thank my beloved wife, Manmeet, for her unfaltering support. I am forever indebted to her for her daily sacrifices in caring for our newborn son, Viraaj, while I pursued my doctoral work.

Citations

The contents of Chapter 2 are acquired from the following journal publication [1]:

“Bayesian inference of nonlinear unsteady aerodynamics from aeroelastic limit cycle oscillations”,
R. Sandhu, D. Poirel, C. Pettit, M. Khalil, A. Sarkar,
Journal of Computational Physics, 2016.

The contents of Chapter 3 are acquired from the following journal publication [2]:

“Bayesian model selection using automatic relevance determination for nonlinear dynamical systems”,
R. Sandhu, C. Pettit, M. Khalil, D. Poirel, A. Sarkar,
Computer Methods in Applied Mechanics and Engineering, 2017.

The contents of Chapter 4 are acquired from the following article that was recently submitted for publication:

“Nonlinear sparse Bayesian learning for physics-based models”,
R. Sandhu, M. Khalil, C. Pettit, D. Poirel, A. Sarkar,
Submitted to Journal of Computational Physics, 2019.

The basics of Bayesian parameter estimation are derived from the following PhD thesis [3]:

“Bayesian Inference for Complex and Large-scale Engineering Systems”,
M. Khalil,
Carleton University, 2013

Contents

1	Introduction	1
1.1	Focus of this thesis	1
1.2	Background	6
1.3	Review of Bayesian algorithms	11
1.4	Motivation	16
1.5	Problem statement	20
1.6	Research objectives	26
1.7	List of contributions and thesis organization	27
2	Bayesian model comparison for nonlinear dynamical systems	31
2.1	Bayesian inversion methodology	32
2.1.1	State-space representation of dynamical systems	32
2.1.2	Bayesian parameter estimation	34
2.1.3	Bayesian model comparison	36
2.2	Bayesian modelling of nonlinear aeroelasticity	38
2.2.1	Nonlinear aeroelastic oscillator	39
2.2.2	Brief review of unsteady aerodynamics	43
2.2.3	Nonlinear extension of Wagner’s model	48
2.3	Numerical investigations	51
2.3.1	Bayesian inversion of linear unsteady aerodynamics	51
2.3.2	Bayesian inversion of nonlinear unsteady aerodynamics	55
2.4	Bayesian model comparison using wind-tunnel data	64
3	Bayesian model reduction using automatic relevance determination	71
3.1	Bayesian model reduction methodology	72
3.1.1	Automatic relevance determination	72
3.1.2	Hyperparameter estimation algorithm	75
3.2	Numerical investigation: Nonlinear aeroelastic oscillator	79
3.2.1	Unidimensional model reduction	79
3.2.2	Multidimensional model reduction	85
3.2.3	Comparison of non-nested models	89
4	Nonlinear sparse Bayesian learning	96
4.1	Introduction	96
4.2	NSBL methodology	99

4.2.1	Hybrid prior pdf	100
4.2.2	Gaussian mixture-model approximation	101
4.2.3	Sparse learning optimization problem	103
4.2.4	Analytical calculation of Bayesian entitites	105
4.2.5	Numerical implementation details	118
4.2.6	Relation to SBL	122
4.3	Numerical investigations	125
4.3.1	Linear regression: Sparse PCE expansion of Ishigami function	125
4.3.2	Polynomial regression with a multimodal prior pdf	130
4.3.3	Three-dimensional mass-spring-damper system	136
4.3.4	Nonlinear aeroelastic oscillator	141
5	Conclusion and future research direction	157
5.1	Conclusion	157
5.2	Future research direction	159
	Appendices	163
A	Sparse learning in linear regression	164
A.1	Regularized least-square estimation	165
A.2	Bayesian linear regression	170
A.3	Sparse Bayesian learning	176
A.4	Bayesian compressive sensing	187
A.5	Relationship between LASSO, SBL, and BCS	193
A.6	Numerical study: Polynomial regression	197
B	Bayesian algorithms	211
B.1	Transitional MCMC	211
B.2	Delayed rejection adaptive Metropolis MCMC	214
B.3	The Chib-Jeliazkov method	217
B.4	Extended Kalman filter	219
B.5	Laplace approximation and BIC	220
C	Derivation of higher-order Wagner unsteady models	224
D	Application of NSBL to bimodal likelihood functions	228
E	PCE surrogate and Sobol sensitivity indices	235
F	Gaussian distribution identities	237

List of Figures

1.1	Overview of Bayesian inverse modelling.	9
1.2	Three-dof mass-spring-damper system.	21
1.3	Bayesian parameter estimation results for the over-parameterized model with $\phi = \{c_1, c_2, c_3\}$	23
1.4	Posterior parameter pdf for the data-generating model.	24
1.5	100 prediction realizations of velocity u_1	24
2.1	A general time-discretization scheme for stochastic models of dynamical systems shown in Eq. (2.1) and Eq. (2.2). Adopted from [1, 16, 86].	33
2.2	Schematic of the wing apparatus. Adopted from Khalil <i>et al.</i> [9].	39
2.3	Wind-tunnel pitch response and its PSD. Adopted from Sandhu <i>et al.</i> [1].	40
2.4	Laminar separation bubble. Adopted from Sandhu <i>et al.</i> [1].	41
2.5	Comparison of the existing Wagner function approximations with the values published by Wagner [107].	47
2.6	Phase-plane diagram for varying Wagner constant B , contrasted with the wind-tunnel data for $Re = 85,000$	50
2.7	Pitch response generated for varying forcing reduced frequency k_f	53
2.8	Posterior parameter statistics for varying forcing reduced frequency.	54
2.9	Synthetic LCO response generated using Eq. (2.22a) and Eq. (2.22b).	57
2.10	Marginal and joint posterior pdfs for the optimal model $\mathcal{M}_1^u$. The MAP estimates are shown in red-colored dotted lines. The ρ value shown in joint pdfs is the Pearson correlation coefficient.	63
2.11	Predicted Wagner function using the optimal model $\mathcal{M}_1^u$	63
2.12	Predicted Wagner function using the second-order model $\mathcal{M}_3^u$	64
2.13	Measured wind-tunnel LCO response. Re denotes the chord-wise Reynolds number, and k denotes the reduced frequency.	65
2.14	Filtered and trimmed wind-tunnel LCO response.	66
2.15	Reynolds number versus the posterior pdf of B for model $\mathcal{M}_2^u$	69
2.16	Static aerodynamic moment coefficient versus angle-of-attack for different Reynolds numbers for model $\mathcal{M}_2^u$	70
3.1	Laplace pdf versus Gaussian pdf, both having a precision of one.	73
3.2	Synthetic pitch measurements and its PSD. The data used for ARD computations is shown in red.	80
3.3	Bayesian model comparison entities for varying α , when e_3 is treated as questionable.	82

3.4	Marginal posterior pdf of e_3 parameter for various choices of prior pdf.	82
3.5	Bayesian model comparison entities for varying α , when e_5 is treated as questionable. In the bottom-left plot, the pdf bounds correspond to 99% probability mass centered around the mode.	84
3.6	Marginal posterior pdf of e_5 for various choices of prior pdf.	85
3.7	Results for ARD-based Bayesian model reduction of the first-order unsteady aerodynamic model $\mathcal{M}$ in Table 3.4.	87
3.8	Marginal and joint posterior pdf of e_3 and e_4 when assigning them 1) flat prior, or 2) the optimal ARD prior $p(\phi \alpha^{\text{MAP}})$	88
3.9	Results for the ARD-based Bayesian model reduction of the quasi-steady aerodynamic model $\mathcal{M}_1$ in Table 3.6.	91
3.10	Results for the ARD-based Bayesian model reduction of the second-order unsteady aerodynamic model $\mathcal{M}_3$ in Table 3.7.	92
3.11	Marginal posterior parameter pdf for the optimal model $\mathcal{M}_2^{\text{opt}}$	93
3.12	Selected joint posterior parameter pdf for the optimal model $\mathcal{M}_2^{\text{opt}}$	94
3.13	Updated state pdf for the optimal model $\mathcal{M}_2^{\text{opt}}$	95
4.1	Overview of NSBL.	97
4.2	Variation of sparse learning entities with increasing PCE order. NSBL is executed with γ_i tolerance set at 0.25, 0.50 and 0.75.	129
4.3	Noisy measurements versus the truth.	130
4.4	Inverse problem setup.	131
4.5	GMM representation for $p(\mathcal{D} \phi)p(\phi_{-\alpha})$ using multivariate KDE approximation.	133
4.6	NSBL results when initiating from $\alpha = \{6, 8\}$	134
4.7	NSBL results when initiating from $\alpha = \{-3, -3\}$	135
4.8	Model predictions using the log α value of a) $\{-\infty, -\infty\}$ (flat priors), b) $\{-1.42, 6.76\}$ (local optimum), and c) $\{6.24, -4.00\}$ (global optimum).	136
4.9	Inverse problem setup.	137
4.10	Marginal posterior pdf pertaining to the multivariate KDE approximation of $p(\mathcal{D} \phi)p(\phi_{-\alpha})$	139
4.11	Predicted versus true time-history of velocity $\dot{u}_2$	139
4.12	NSBL results for identifying sparsity among damping coefficients of the three-dof mass-spring-damper model.	140
4.13	Marginal posterior pdf for the relevant parameters of the three-dof mass-spring-damper model.	141
4.14	Predicted vs true time-history of velocity $\dot{u}_2$ using the optimal nested model.	142
4.15	Marginal parameter pdf obtained using 10 instances of 50 iid samples from $p(\mathcal{D} \phi)p(\phi_{-\alpha})$. The baseline (converged) pdf is obtained by using all of the 500 iid samples.	144
4.16	NSBL results for the relevance determination of an <i>a priori</i> relevant parameter e_3	145
4.17	Marginal parameter pdf obtained using 10 instances of 50 iid samples from $p(\mathcal{D} \phi)p(\phi_{-\alpha})$. The baseline (converged) pdf is obtained by using all of the 500 iid samples.	147
4.18	NSBL results for the relevance determination of an <i>a priori</i> irrelevant parameter e_5	148

4.19	NSBL results for the relevance determination of parameter e_3 and e_4 .	150
4.20	NSBL results for the relevance determination of parameter e_3 and e_5 .	151
4.21	NSBL results for the relevance determination of parameter e_5 and e_6 .	152
4.22	Marginal posterior pdf of model parameters before and after sparse learning, for the case when e_5 and e_6 are deemed irrelevant.	153
4.23	Marginal posterior pdf of model parameters before and after sparse learning, for the case when e_5 is deemed irrelevant and e_6 is deemed relevant.	153
4.24	NSBL results for the four-dimensional sparse learning.	155
4.25	Marginal posterior pdf of model parameters before the after the inclusion of optimal ARD prior.	155
A.1	Training data generated using $y_i = 1 + x_i^2 + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, 1/100)$.	197
A.2	Cross-validation prediction error for varying regularization coefficient λ .	200
A.3	Marginal posterior pdf of $\mathbf{w}$ for a fifth-order polynomial with prior $p(\mathbf{w}) = \mathcal{N}(\mathbf{w} \mathbf{0}, 10^3\mathbf{I}_N)$.	202
A.4	Model predictions using the posterior pdf from BLR, using MLE estimates, and using the true model.	202
A.5	Variation in GOF, EIG and log-evidence with varying prior variance of parameter w_0 and w_1 for a fifth-order polynomial model. All other model parameters are assigned a fixed prior $\mathcal{N}(w_i 0, 10^3)$.	205
A.6	Original SBL algorithm results for a fifth-order polynomial model.	208
A.7	Post-SBL marginal posterior pdf of $\mathbf{w}$ for a fifth-order polynomial.	208
A.8	Model predictions using joint posterior pdf of $\mathbf{w}$ before (pre-SBL) and after (post-SBL) the application of SBL.	209
A.9	Accelerated SBL algorithm results.	209
A.10	BCS results.	210
D.1	Bayesian entities pertaining to Case A from Table D.1.	229
D.2	NSBL results for Case A in Table D.1 when initiated from $\log \alpha = \{10, 10\}$.	230
D.3	NSBL results for Case A in Table D.1 when initiating at $\alpha = \{-5, -5\}$.	232
D.4	Variation in optimal Bayesian entities for varying hyperprior parameters $\log r_i$ or $\log s_i$.	232
D.5	Bayesian entities pertaining to Case B from Table D.1.	233
D.6	NSBL results for Case B in Table D.1 when initiating from $\log \alpha = \{10, 10\}$.	234

List of Tables

1.1	Existing sparse learning algorithms.	14
1.2	Bayesian model comparison results for varying prior variance. *Model $\mathcal{M}_3$ is the data-generating model.	26
2.1	Known parameters of the wind-tunnel apparatus.	43
2.2	Numerical values of Wagner function proposed by Wagner [107].	46
2.3	Existing Wagner function approximations.	46
2.4	Synthetic LCO: Prior parameter pdf for the questionable parameters of candidate nested models.	59
2.5	Bayesian model comparison results obtained using the synthetic LCO measurements. Note that Log-evidence is equal to GOF minus EIG.	61
2.6	Wind-tunnel LCO: Prior parameter pdf for the questionable parameters of candidate nested models.	66
2.7	Bayesian model comparison results obtained using the wind-tunnel LCO measurements at $Re = 75,000$	67
2.8	Bayesian model comparison results obtained using the wind-tunnel LCO measurements at $Re = 85,000$	67
2.9	Bayesian model comparison results obtained using the wind-tunnel LCO measurements at $Re = 90,000$	67
2.10	Bayesian model comparison results obtained using the wind-tunnel LCO measurements at $Re = 105,000$	68
2.11	Posterior estimates for Wagner constant B for varying Reynolds number.	68
3.1	Contrast between various Bayesian techniques. Adopted from Murphy [7].	77
3.2	Unidimensional model reduction setup where parameter e_3 is treated as questionable.	81
3.3	Unidimensional model reduction setup where parameter e_5 is treated as questionable.	83
3.4	Multidimensional model reduction setup where parameters e_3 , e_4 , e_5 , and e_6 are treated as questionable.	86
3.5	Reduced model and its optimal prior pdf, pertaining to the inverse problem defined in Table 3.4.	88
3.6	Model $\mathcal{M}_1$ and its associated prior pdf.	89
3.7	Model $\mathcal{M}_3$ and its associated prior pdf.	90
3.8	Candidate model set containing optimal nested models obtained from ARD-based model reduction.	92

3.9	Bayesian model comparison among optimal nested models.	93
4.1	Special cases of a Gamma hyperprior pdf.	104
4.2	Summary of the analytical Bayesian apparatus of NSBL.	118
4.3	NSBL setup for the polynomial regression model with a multimodal prior pdf for a_0	132
4.4	NSBL setup for the three-dof mass-spring-damper model.	138
4.5	Unidimensional sparse learning setup where parameter e_3 is treated as question- able.	143
4.6	Unidimensional sparse learning setup where parameter e_5 is treated as question- able.	147
4.7	Two-dimensional sparse learning setup where parameters e_3 and e_4 are treated as questionable.	150
4.8	Two-dimensional sparse learning setup where parameters e_3 and e_5 are treated as questionable.	150
4.9	Two-dimensional sparse learning setup where parameters e_5 and e_6 are treated as questionable.	151
4.10	Four-dimensional sparse learning setup.	155
5.1	Comparative review of Bayesian algorithms studied in this thesis.	162
A.1	Summary of frequentist estimation techniques for w	168
A.2	Summary of BLR.	175
A.3	Frequentist estimation results for w . (*optimal λ value computed by minimizing five-fold CV error)	200
A.4	Bayesian linear regression results for precision and correlation coefficient for a 5^{th} order polynomial model.	202
A.5	Bayesian model selection results for varying polynomial order. The data in parenthesis is the posterior model probability in percentage (rounded), without considering the true model.	205
D.1	Two cases of likelihood function considered for the application of NSBL. . . .	229
E.1	Sobol indices calculation from PCE surrogates through variance decomposition.	236

List of Algorithms

1	The ARD-based Bayesian model reduction algorithm	78
2	NSBL algorithm	123
3	Original SBL method by Tipping [78]	182
4	Accelerated SBL algorithm by Faul and Tipping [136]	188
5	BCS algorithm by Babacan <i>et al.</i> [136]	194

List of Acronyms

AIC Akaike Information Criterion

AM Adaptive Metropolis

ARD Automatic relevance determination

BCS Bayesian compressive sensing

BGR Brooks-Gelman-Rubin

BIC Bayesian Information Criterion

BLR Bayesian linear regression

COV Coefficient of variation

CV Cross-validation

DDM Domain decomposition method

dof degree-of-freedom

DR Delayed rejection

EIG Expected information gain

EM Expected maximization

EnKF Ensemble Kalman filter

FEM Finite element method

FSI Fluid-structure interaction

FVM Finite volume method

GOF Goodness-of-fit

HPC High performance computing

iid Independent and identically distributed

INCA Inter-chain communication

KDE Kernel density estimation

KL Kullback-Leibler

LASSO Least absolute shrinkage and selection operator

LCO Limit-cycle oscillation

LHS Latin hypercube sampling

LSB Laminar separation bubble

MAP *Maximum a posteriori*

MCMC Markov chain Monte Carlo

MH Metropolis-Hastings

MLE Maximum likelihood estimation

NN Neural networks

NSBL Nonlinear sparse Bayesian learning

ODE Ordinary differential equation

OLS Ordinary least-square

PCE Polynomial chaos expansion

PSD Power spectral density

pdf Probability density function

QOI Quantity-of-interest

RBF Radial basis function

RLS Regularized least-square

RMS Root-mean-square

RVM Relevance vector machine

RWM Random-walk Metropolis

SBL Sparse Bayesian learning

SNR Signal-to-noise ratio

SVM Support vector machine

TMCMC Transitional Markov chain Monte Carlo

UKF Unscented Kalman filter

UQ Uncertainty quantification

Chapter 1

Introduction

1.1 Focus of this thesis

This thesis is concerned with the inverse modelling of complex engineering systems. Inverse modelling is a necessary mathematical undertaking aimed at constructing a mathematical model of system physics using field measurements. Inverse modelling is referred to as *system identification* when dealing with dynamical systems and time-varying measurements [4, 5]. A mathematical model for a given static or dynamical mechanistic system consists of a model equation and a measurement equation. The model equation quantifies the system physics and is typically in the form of differential equations derived using conservation laws. The model equation relates the time-varying state variables (such as deflection, velocity, acceleration) with the unknown time-invariant parameters (such as stiffness and damping coefficients). The state variables (also known as latent variables) define the complete state of the dynamical system at any given time instant. The measurement equation relates the state variables with the observed quantity, and may also contain unknown time-invariant parameters. Hereafter, ‘model’ refers to the model and measurement equation pair, and ‘model parameters’ refers to the unknown time-invariant parameters contained in both model and measurement equations.

Unlike state variables, the unknown model parameters cannot be directly measured and,

therefore, inverse modelling is tasked at calibrating these parameters using the observed responses. For example, given a simplified, multidimensional mass-spring-damper model of a shear building frame, an inverse problem can be posed to estimate the column stiffness and damping coefficients in-between floors using the noisy deflection, velocity, and acceleration sensor data from multiple floors. Another example is an aircraft wing that is undergoing oscillatory motion in a wind-tunnel, and an inverse problem can be posed to model the unsteady aerodynamic loads by using a recorded noisy signal of wing oscillations. This thesis is concerned with such inverse problems aimed at constructing a predictive model that is representative of the underlying physical phenomena. This predictive model can then be used for purposes such as understanding the system physics, predicting unobserved physics, informing future data collection, and making policy and design decisions.

There exist an enormous amount of published literature regarding inverse modelling applications across all disciplines of engineering. To clearly define the scope of this thesis, we hereby enlist the salient attributes of the kind of inverse problems tackled in this thesis:

1. **Complex physics:** Here, ‘complex’ refers to potentially nonlinear physics, or the incomplete knowledge regarding plausible physics, or the apparent inability of existing laws of physics or empirical models to capture the system physics. For instance, such a scenario arises when modelling nonlinear interactions of fluid flow and structural motion in aeroelasticity wherein linear aerodynamic theories are inapplicable. The nonlinearity may exist in the model equation or the measurement equation or both. Hereafter, ‘nonlinearity’ or ‘nonlinear model’ refers to a nonlinear relation between the unknown model parameters and the observed quantity. This thesis is concerned with the inverse modelling algorithms that can handle nonlinearity in both model and measurement equations.
2. **Prior knowledge exists:** When dealing with engineering systems, some prior understanding always exists regarding the underlying physics. This understanding can originate from sources such as laws of physics (e.g. conservation laws), published literature or prior experimental studies. This prior knowledge often helps guide the proposed

model and the range of values expected for unknown parameters. This thesis focuses on inverse modelling algorithms that can incorporate any such prior knowledge. This physics-based modelling approach is in contrast to statistical modelling, where the entire onus of constructing a predictive model is on system measurements. The reader is referred to excellent books by Konishi and Kitagawa [6] and Murphy [7] for an exhaustive text on statistical modelling. The goal of this thesis is to corroborate or improve existing physics-based models using both measurements and prior knowledge.

3. **Noisy, incomplete, and sparse measurements:** The measurements acquired from the concerned system often possess the following properties. The data-acquisition apparatus introduces digital (non-physical) noise into the observed system response. Differentiating this non-physical noise from physics is a crucial aspect of inverse modelling. The incomplete nature of the measurements implies that the system state is only partially observed. For instance, when the measurements are only available for the deflection of the second floor in a three-storey mass-spring-damper model of a building frame. The measurements would be deemed complete when the deflection and velocity at all three floors are available. Sparse measurements imply a scarcity of data in the time domain. For example, measurements consisting of five samples per second are sparser than 20 samples per second, depending on the frequencies of oscillations of the system. In this thesis, we focus on the inverse modelling approaches that can handle such attributes of field measurements.
4. **Imperfect models:** A mathematical model represents an attempt to understand reality, which is of infinite dimension and infinite complexity. Therefore, mathematical models are an approximation by definition. In this thesis, we acknowledge the imperfectness of mathematical models and aim to quantify it during inverse modelling. Also, model imperfectness produces mismatch when fitting noisy, incomplete, and sparse measurements, giving rise to an ill-conditioned inverse problem. An inverse problem is ill-conditioned when the parameter estimate violates one of the following conditions: 1) existence, 2)

uniqueness, and 3) stability (i.e. the mapping between measurements and parameters is continuous) [8].

5. **Multiple models exist:** Traditionally, the inverse modelling task would only involve a parameter estimation (or model calibration) task since the model is always assumed to be adequate for its intended purpose. However, while dealing with complex physics and imperfect models, multiple models can exhibit a reasonable fit to the measurements. These models can result from competing physical laws, or from the addition of purely statistical elements to well-established physical laws, or when considering alternative approaches to reduced-order modelling. The existence of competing models is also fuelled by the growing desire to enhance the extrapolation capabilities of physics-based models under adverse conditions. This thesis is primarily aimed at addressing this issue of competing models through a model comparison or a sparse learning algorithm. Sparse learning is a special type of model comparison, where the goal is to identify the best model nested under an over-parameterized model.

Sec. 1.5 further elaborates on these attributes using an engineering mechanics example involving a multi-dimensional mass-spring-damper model of a multi-storey shear building frame. These attributes define the inverse modelling context of this thesis.

Given these attributes, inverse modelling would involve two essential tasks: 1) model comparison, and 2) parameter estimation. **The key focus of this thesis is on the design, implementation and application of a model comparison or a sparse learning algorithm that is probabilistic, robust and computationally efficient under the aforementioned circumstances.** The parameter estimation task is not the primary focus of this thesis, even though parameter estimates are usually available as a by-product of model comparison algorithms. See Khalil [9], Beck [5], Hernandez [10] and the references within for a comprehensive text on probabilistic parameter estimation techniques within the context of engineering mechanics.

Probabilistic model comparison algorithms often require the computation of high-dimensional integrals to ascertain the goodness (i.e. extrapolation capabilities) of competing models. This

task creates a computational bottleneck in the inversion process and often ends up dictating the choice of an appropriate model comparison algorithm. This computational challenge is particularly concerning for 1) high-dimensional models (a large number of model parameters or state variables), or 2) high-fidelity models (costly model/likelihood evaluation). A large model dimensionality is often the case for physical systems simulated using, for instance, finite-element or finite-volume based numerical solvers. A popular technique of domain decomposition method (DDM) is often employed to decompose an extensive numerical system into sub-systems with manageable dimensionality, which are then easily tackled using high-performance computing (HPC). Recent work by Desai [11] and Khalil [9] is aimed at designing efficient DDM-based solvers for solving forward/inverse problems in the presence of modelling uncertainties. The model comparison algorithms detailed in this thesis could be used in conjunction with DDM-based numerical models. However, this avenue is not pursued in this thesis. This thesis is primarily concerned with model comparison among models where storage/memory issues are absent or have been handled appropriately.

A costly model or likelihood evaluation is also a source of concern for an efficient implementation of the inversion process. This scenario is often present for system identification of dynamical systems, wherein likelihood function evaluation requires the estimation of unobserved time-varying state variables (more details in Sec. 2.1.2). This state estimation task is typically executed using Kalman filter for linear dynamical systems when both model and measurement equations are linear, and the error terms are additive and Gaussian [12, 13]. For nonlinear systems, the state estimation is performed using techniques such as extended Kalman filter (EKF) [12, 14, 15], unscented Kalman filter (UKF) [3], ensemble Kalman filter (EnKF) [3, 16, 17] and particle filter [3, 17–19]. See Hernandez [10], Hernandez and Bernal [20], and Khalil *et al.* [17] for an exhaustive review of state estimation techniques for civil engineering applications. This thesis is not concerned with the efficiency and execution of state estimation or likelihood computation tasks. Instead, we focus on designing efficient model comparison algorithms that require a minimum number of likelihood function evaluations from

each of the competing models.

1.2 Background

Inverse modelling plays a crucial role in enhancing the understanding of the physics of an engineering system. Choosing an appropriate inverse modelling algorithm often boils down to how the underlying modelling uncertainties are handled. Broadly speaking, uncertainties can originate either from the lack of knowledge (*epistemic*) or from the inherent randomness (*aleatory*). For example, a lack of knowledge regarding the system physics introduces epistemic uncertainty in the form of model discrepancy error that quantifies the mismatch between model and measurements. This uncertainty can be reduced by gathering more field data and improving the model, but it can never be eliminated due to the finiteness of information content in measurements and the inherently imperfect nature of models. In addition, epistemic uncertainty is introduced by the noise in measurements, which cannot be eliminated from the sensors. Also, epistemic uncertainty can originate from inherent randomness in the underlying physics, for example, say, due to the effect of airflow turbulence on aerodynamics around an oscillating wing. Incorporating these uncertainties during inverse modelling is necessary to produce robust probabilistic predictions of unobserved data [5].

Based on the interpretation of probability, inverse modelling approaches are considered as either 1) *frequentist* [6, 21], or 2) *Bayesian* [22–24]. The frequentist ideology views probability as the frequency of an event in the limit of infinite trials. The measurements are treated as random, while the unknown model parameters are treated as fixed (deterministic) quantities. Maximum likelihood estimation (MLE) is a popular frequentist parameter estimation approach that computes a point estimate for the unknown parameters by maximizing the likelihood function (i.e. probability of measurements conditioned on model parameters) [25]. The probability distribution of such point estimates, known as the sampling distribution, is computed using multiple instances (realizations) of the measured response. This parameter distribution is then

pushed through the calibrated model to obtain model predictions with a confidence bound.

A frequentist model comparison criterion typically consists of a data-fit term (such as the maximum log-likelihood) minus a penalty term (a monotonically increasing function of the number of model parameters). Maximizing this criterion leads to a model that provides the best balance between data-fit and model simplicity. This idea is referred to as the principle of parsimony or the bias-variance trade-off [21]. Some of the commonly used frequentist measures that adhere to this idea are the Akaike information criterion (AIC) [26], Bayesian information criterion (BIC) [27], Deviance information criterion [28], and Minimum description length [29]. See Konishi and Kitagawa [6] and Ding *et al.* [30] for an exhaustive text on such measures. Some recent applications of these frequentist measures include signal processing [31], machine learning [32], ecology and evolution [33], system identification of dynamical systems [34], and many others.

AIC and BIC are the two most commonly used frequentist measures for ranking competing models. See Burnham and Anderson [21] for a detailed theoretical and numerical contrast of AIC and BIC. AIC is based on information theory and is derived by minimizing the Kullback-Liebler (KL) divergence between the proposed model and the truth [35]. On the other hand, BIC is a simplification of Bayesian model comparison for the case of large size of measurement data. See Appendix B.5 for the derivation of BIC from Bayesian model comparison. Note that BIC is founded on the Bayesian ideology and its categorization as a frequentist measure is solely based on the notion that it relies on the MLE estimate and the number of model parameters to pursue the tradeoff between data-fit and model simplicity.

Contrary to the frequentist ideology, the Bayesian ideology is founded on the axioms of probability theory and measure theory, wherein probability is interpreted as the current state of knowledge [24, 36]. The measurements are treated as fixed (not random) quantities, while the unknown parameters are treated as random or probabilistic quantities with an associated probability density function (pdf). Fig. 1.1 provides a general overview of the Bayesian inverse modelling framework, which involves the following five key components:

1. **Candidate set of stochastic models:** When dealing with complex physics, a set of competing models with varying complexity is often constructed using the prior knowledge available to the modeller. Each model contains a model equation and a measurement equation. The stochasticity of candidate models pertain to the inclusion of 1) a model discrepancy error term into the model equation (Kennedy-O'Hagan framework [37]), and 2) a measurement error term in the measurement equation. These terms are usually modelled using parameterized probability distributions whose parameters could be treated as unknown and estimated alongside other model parameters. See Sec. 2.1.1 for mathematical details.
2. **Parameter prior pdfs:** Bayesian framework requires a prior pdf to be assigned to the unknown parameters of each candidate model. This prior pdf represents the current state of belief, and is typically constructed by converting the subjective prior knowledge into a known closed-form pdf. The inclusion of prior knowledge helps regularize the inversion process by restricting the posterior parameter space in regions that make physical sense and conform to the modeller's expectations.
3. **Bayesian model comparison:** Bayesian model comparison relies on the model evidence (type-II likelihood) to assign a posterior model probability to each candidate model. The evidence facilitates the trade-off between model complexity and data-fit to measurements (more details in Sec. 2.1.3). The model that receives the largest posterior probability is considered most likely to have produced the measurements. The goal of Bayesian model selection is not to identify the true model, but the best useful model among the candidate model set. This notion was echoed by Box [38] when he said: "All models are wrong, but some are useful." If all the candidate models are bad, Bayesian model comparison identifies the least bad model. See Sec. 2.1.3 for mathematical details.
4. **Bayesian parameter estimation:** The posterior pdf of unknown model parameters is computed by assimilating field measurements (through likelihood function) and prior

knowledge (through prior pdf) using Bayes' theorem. This posterior pdf quantifies the updated parameter knowledge obtained by combining the information contained in measurements and the prior knowledge. This Bayesian task is often referred to as *Bayesian inference* or *Bayesian model updating* in engineering and scientific literature.

5. **Bayesian model averaging:** For a given candidate model, the probabilistic prediction of a given quantity-of-interest is obtained by using the posterior parameter pdf computed during Bayesian parameter estimation. These predictions are then averaged across all candidate models where the prediction from each model is weighted according to its posterior model probability computed during Bayesian model comparison [39]. This approach of computing probabilistic predictions by incorporating the model parameter uncertainty and the model comparison uncertainty is referred to as Bayesian model averaging. These probabilistic predictions can then be exploited to achieve modelling goals such as optimal experimental design, structural design optimization, and parametric sensitivity studies.

Figure 1.1: Overview of Bayesian inverse modelling.

Both frequentist and Bayesian approaches have found widespread usage in solving inverse problems found in engineering practice. However, we believe the Bayesian inverse modelling approach offers the following advantages over the frequentist approach in the current setting:

- In the case of MLE, the sampling distribution of unknown model parameters is analytically available only in the particular case of a large number of measurements per instance (*Laplace approximation* [40]) or linear-in-parameter models with Gaussian likelihood functions [21]. Here, ‘linear-in-parameter’ means the observed quantity is linearly related to all the unknown

parameters. Such particular circumstances are rare in engineering practice. Also, while dealing with a limited number of measurements, multiple instances of MLE estimate needed for computing the sampling distribution are not available. In such a case, the parameter estimate and the resulting model prediction is without confidence bound. On the contrary, Bayesian parameter estimation produces a posterior parameter pdf irrespective of the number of measurements and the type of model. This property of the Bayesian approach allows for obtaining probabilistic predictions under any modelling circumstance.

- While dealing with nonlinear models and limited measurements, the likelihood function may exhibit multiple local optima. Based on the concept of *identifiability* [41, 42], point parameter estimates can be identified as either 1) globally identifiable (unique solution), 2) locally identifiable (multiple solutions), or 3) unidentifiable (infinite solutions, i.e. equally likely optima that form a low-dimensional manifold within the parameter space). The global unidentifiability or local identifiability of MLE estimates indicate the ill-conditioning of the inverse problem [41]. The existence of multiple likelihood peaks is problematic as the point predictions made using local MLE estimates cannot be dismissed in favour of predictions made using the global MLE estimate. Also, frequentist criteria such as AIC and BIC are rendered arbitrary in such cases since the maximum log-likelihood (the data-fit term) is also not unique (See Appendix B.5 for more details). The Bayesian approach is well-suited for globally unidentifiable scenarios since the entire posterior parameter pdf is used in executing model comparison and obtaining probabilistic predictions.
- Unlike the frequentist approaches of MLE and AIC/BIC, the Bayesian approach allows for the inclusion of prior parameter knowledge through prior pdfs. This regularization of posterior parameter space and model evidence by the prior pdf often helps restore the well-posedness or global identifiability of inverse problems encountered in engineering practice.
- The frequentist model comparison criteria such as AIC and BIC treat model complexity as a sole function of the number of unknown model parameters. This notion is valid for

regression-type models but is flawed for physics-based models containing nonlinear differential equations. Each parameter in physics-based model equations has a unique relation to the underlying physics. This uniqueness is lost when all parameters are treated equally for calculating model complexity by AIC and BIC. In contrast, the Bayesian framework computes model complexity as the KL divergence between the prior and the posterior pdf of model parameters. In other words, the Bayesian approach treats model complexity as the gain in information from measurements. More details regarding this property are provided in Sec. 2.1.3.

Due to these practical benefits, Bayesian methods have become a natural choice for performing inverse modelling in engineering mechanics applications, thereby producing robust probabilistic inferences that are informed by prior knowledge and modelling/measurement uncertainties [3, 5, 9, 36, 37, 41, 43–45]. Next, we provide a brief review of Bayesian inverse modelling algorithms, with a particular focus on model comparison and sparse learning algorithms.

1.3 Review of Bayesian algorithms

The recent advances in HPC and the associated scalable algorithms in computational mechanics and stochastic simulation algorithms was an excellent catalyst for the renewed interest in Bayesian methods. There exist a plethora of publications concerning Bayesian inversion algorithms for engineering applications [45–51]. However, this review of Bayesian algorithms focuses primarily on the model comparison aspect of Bayesian inverse modelling.

Bayesian model comparison relies on model evidence to facilitate the ranking of candidate models based on the trade-off achieved between data-fit and model simplicity. This trade-off property of model evidence is often referred to as the Ockham’s razor [52]. More details regarding this property are provided in Sec. 2.1.3. MacKay [53] first formalized the idea of Bayesian model comparison as a probabilistic Occam’s razor tool within the context of machine learning.

He also demonstrated the relationship between model evidence and MLE-based model comparison algorithms and advocated for a two-level Bayesian inference (i.e. parameter estimation and model comparison). Later, Beck [5] and colleagues [41, 44, 45, 54] pioneered the development of Bayesian inversion algorithms for engineering mechanics applications. Yuen [43] provided an exhaustive text on the application of Bayesian algorithms for civil engineering applications, including real-life case studies.

When dealing with dynamical systems and incomplete measurements (i.e. the entire state is not measured), the unknowns include time-invariant model parameters (like stiffness and damping coefficients) and time-varying state variables (like unobserved displacement and velocity components). The Bayesian estimation of all these parameters can be approached in two ways. In the first approach, the state is augmented with unknown time-invariant parameters, and the time-varying pdf of the augmented state is estimated through sequential assimilation of noisy measurements [19]. As Khalil *et al.* [19] pointed out, the augmented state space model has a stronger nonlinearity than the original state space model, which then demands the use of non-Gaussian sampling-based filters to achieve a reasonable accuracy in estimated state pdfs. The sampling-based filters such as the particle filter[3, 17–19] are computationally expensive since a large sample set is required to propagate a non-Gaussian state pdf through a highly nonlinear system [55].

On the contrary, in the second approach, the likelihood function is computed by marginalizing out the state variables to produce a likelihood that solely depends on the time-invariant parameters. The marginalization step requires an underlying state estimation task, which can be usually be executed using much faster algorithms such as the EKF due to the reduced nonlinearity [12, 14, 15]. Subsequently, the time-invariant parameters can be estimated using a Markov Chain Monte Carlo (MCMC) algorithm. See Sec. 2.1.2 for mathematical details regarding the marginalization process. This parameter estimation approach is also referred to as *combined state and parameter estimation* [19]. The reader is further referred to Hernandez [10], Hernandez and Bernal [20], and Khalil *et al.* [17] for an exhaustive review of sequential filtering algo-

rithms within the context of civil engineering applications. In recent work, Bisailon *et al.* [55] addressed the relationship between state estimation and parameter estimation for nonlinear dynamical systems. In this thesis, we employ the second approach, where the computation of state variables is considered encapsulated within the likelihood function.

The widespread use of Bayesian methods is fuelled by the recent developments in efficient MCMC algorithms to generate stationary samples from the unnormalized pdfs. See introductory texts by Gilks *et al.* [46], Robert and Casella [56] and Tierney [57] for more information on MCMC sampling. Two of the most commonly used MCMC algorithms are Metropolis-Hastings algorithm [58–60] and Gibbs sampler [61]. In particular, random-walk Metropolis (RWM) [51, 62] has been the most popular MCMC algorithms in engineering practice. Much of the recent research has been devoted towards developing efficient MCMC samplers for high-dimensional and non-Gaussian pdfs that require minimum model calls and limited to no prior knowledge regarding the plausible shape of the target pdf. Recent modifications to the MH algorithm aimed in this direction involve delayed rejection (DR) [63], adaptive Metropolis (AM) [64], adaptive proposal [51], inter-chain communication (INCA) [65]; among many others. See Appendix B.2 for further details regarding the DR, AM and INCA algorithms.

As the need for model comparison algorithms grew, many prominent MCMC algorithms were extended to include the estimation of model evidence using stationary samples from the posterior parameter pdf [48, 66–68]. Chib [68] proposed an evidence estimator that uses Gibbs sampling output to produce an evidence estimate. Chib and Jeliazkov [67] proposed an evidence estimator that exploits the stationary posterior samples generated using any variant of the MH algorithm. Other notable numerical methods of computing evidence include nested sampling [69], power posteriors [66], annealed importance sampling [70], harmonic mean estimator [71], Gauss-Hermite quadrature [72], and many others. See Friel and Wyse [47] for a comprehensive review of evidence estimation methods. Most notably, the transitional MCMC (TMCMC) algorithm has found widespread usage for sampling from multimodal posterior pdfs and estimating model evidence simultaneously [48]. TMCMC relies on sampling from a se-

quence of intermediate distributions that start with the prior distribution and approach the target posterior pdf in a finite number of steps. Recently, Betz *et al.* [73] provided modifications to TMCMC to ensure faster convergence and better accuracy of evidence estimates. See Appendix B.1 for mathematical details regarding TMCMC.

Sparse learning is a special type of model comparison task where the goal is to identify the best nested model. Table 1.1 provides an overview of the existing sparse learning algorithms. Least absolute shrinkage and selection operator (LASSO) [74] is a widely used frequentist algorithm that induces sparsity using regularized least-square (RLS) estimation with L_1 regularizer. The least angle regression (LARS) algorithm provides an efficient numerical implementation of LASSO [75]. More details about LASSO and its relation to other RLS techniques are discussed in Appendix A. From a Bayesian viewpoint, the LASSO solution equates to the MAP estimate when all the model parameters are assigned a zero-mean Laplace (double-exponential) prior pdf whose precision is proportional to the regularization coefficient used in LASSO. Park and Casella [76] proposed *Bayesian Lasso* as a Bayesian alternative to LASSO, where the hyperparameter is estimated by maximizing model evidence using Gibbs sampling and Monte Carlo expectation-maximization (EM) algorithm.

Method	Probabilistic viewpoint	Penalty/Prior type	Hyperparameter to parameter association	Hyperparameter estimation	Incorporates prior knowledge?
LASSO	Frequentist	L_1 -norm	Same-for-all	Cross-validation	No
Bayesian LASSO	Bayesian	Laplace	Same-for-all	Evidence	No
SBL	Bayesian	Student's-t	Individual	Evidence	No
BCS	Bayesian	Laplace	Individual	Evidence	No
Goal of this thesis	Bayesian	Any	Individual	Evidence	Yes

Table 1.1: Existing sparse learning algorithms.

Both LASSO and Bayesian LASSO are inherently data-based algorithms that disregard prior knowledge and treat all model parameters equally by associating each model parameter with the same hyperparameter (or regularization coefficient). The regularization coefficient is typically

estimated using cross-validation (CV) techniques [7]. Bergmeir and Benitez [77] provide a review of CV techniques for model selection among dynamical systems within the context of black-box modelling. In this thesis, we focus solely on probabilistic sparse learning algorithms that can handle the type of inverse problems defined in Sec. 1.1.

The Bayesian solution to sparse learning is obtained using the concept of automatic relevance determination (ARD). ARD was originally proposed in the form of relevance vector machine (RVM) or sparse Bayesian learning (SBL) [78] wherein the goal was to identify sparsity among linear-in-parameter models called the support vector machines (SVMs). ARD facilitates sparse learning by assigning each parameter a prior pdf with an unknown precision (hyperparameter), which is then estimated by maximizing model evidence [53, 78]. This sparse learning task is also referred to as *feature selection* or *variable selection* since most data-driven models are linear-in-parameter with each parameter associated with a unique basis or feature [40]. Nevertheless, SBL is not exclusive to SVM models and can be applied to any linear-in-parameter model. This linearity property and the use of Gaussian prior-posterior conjugacy relationships enabled an analytically tractable Bayesian analysis. The semi-analytical re-estimation algorithm of SBL ensured computational efficiency even for a large number of features/parameters. Bayesian compressive sensing (BCS) is another analytically tractable feature selection algorithm proposed in signal processing literature that exploits a multilevel ARD prior to induce parameter sparsity [79]. Some of the recent applications of SBL/BCS under these special conditions include gene expression data [80], sparse PCE surrogate for climate model [81] and earthquake warning system [82]. Appendix A details the mathematical foundation of SBL and BCS, along with a thorough numerical investigation on their performance for a polynomial regression example. The remaining attributes of the sparse learning algorithms listed in Table 1.1 are detailed in Appendix A.

1.4 Motivation

Predictive modelling of complex physics using field measurements is a challenging yet salient task in the engineering practice. Traditionally, there exist two types of approaches to predictive modelling: 1) data-based (black-box), and 2) physics-based. The data-based approaches rely on statistical or machine learning techniques to construct a predictive model by identifying salient trends in the measurements. The conservation laws concerning the underlying physical phenomena are not exploited in this approach. Examples of statistical models include linear regression, logistic regression, neural networks (NN), among others [6]. This approach has found limited usage in the engineering practice due to issues such as the poor extrapolation capability of statistical models; and the lack of closed-form relationship between the observed quantity, the hidden state variables and the physical parameters-of-concern. This approach is particularly challenging in the presence of limited data [6, 7].

In contrast, the physics-based approach relies solely on well-established laws of physics to propose a model for system physics. Examples of this approach include using the Navier-Stokes equation to model fluid flow, or using the three-dimensional elasticity equations to model beams. An appropriate numerical scheme such as the finite volume method (FVM) or the finite element method (FEM) is used to discretize these differential equations, leading to a high-fidelity numerical simulation model of the system. Subsequently, measurements are employed to calibrate the unknown model parameters. In contrast to a data-based approach, this approach is mathematically rigorous and achieves high accuracy in extrapolating the system physics. However, the calibration step requires multiple solve of the high-fidelity numerical solver, thereby making it computationally challenging.

There is a growing interest in the engineering community for predictive models that are computationally cheap to evaluate and calibrate and can provide an accurate and precise extrapolation of system physics. Such types of hybrid models are typically produced by augmenting low-fidelity physics-based models with appropriate statistical elements. These models are particularly appealing for modelling incompletely understood engineering systems where

the existing laws of physics are inadequate in representing the system physics. One such recent attempt includes physics-informed neural networks by Raissi *et al.* [83] where an NN is constrained/regularized by differential equations originating from concerned physical laws. Similarly, Khalil *et al.* [9] proposed adding appropriate nonlinear terms to linear quasi-steady aerodynamic models in an attempt to quantify the nonlinear oscillation of a freely pitching wing. Such hybrid models are computationally cheap to evaluate but are practically challenging to construct since the complexity of statistical elements can be varied to ensure a better fit to the measurements. This dilemma leads to a scenario where multiple competing models can be envisioned to represent system physics. Common engineering practice is to choose the most complex model as the model-of-choice. However, it has been well established that choosing the most complex model as the model-of-choice can lead to the issue of *overfitting* during unregularized model calibration [5, 6, 84]. An overfitted model lacks generalizability and is more likely to produce unwise predictions during extrapolation. On the other hand, a simpler model might lack the basic ability to represent the salient system physics. This performance of models with varying complexity is explained in detail in Sec. 1.5 using an engineering mechanics example. This dilemma in choosing an appropriate model necessitates a *model comparison* task to rank competing models based on their ability to quantify the underlying physics while ensuring minimum model complexity. This thesis is primarily focused on addressing this issue of overfitting among hybrid models through a model comparison or sparse learning algorithm. Research in this direction is required to enable robust prediction of unobserved behaviours of complex engineering systems.

Recently, a nonlinear aeroelastic system involving a self-sustained dynamic instability followed by a limit-cycle oscillation (LCO) of a rigid wing apparatus was observed in a wind-tunnel [85]. More details about the wind-tunnel set-up and the underlying physics are provided in Sec. 2.2.1. This instability was observed in flow conditions typical of unmanned aerial vehicles (UAVs). As a result, a need was realized to build a probabilistic predictive model to capture the potentially nonlinear and unsteady aerodynamics behind the oscillatory system. Given the

wind-tunnel response, a reduced-order model based on thin-airfoil theories was realized to be an efficient alternative to the high-fidelity approach of using the Navier-Stokes equation for modelling the aerodynamics. Previously, Sandhu *et al.* [86] and Khalil *et al.* [9] have attempted to construct such a reduced-order model, albeit the issue of competing aerodynamic models was not handled robustly and efficiently. In this thesis, we aim to resolve these shortcomings by pursuing an evidence-based Bayesian model comparison and sparse learning framework that is motivated by the need to build a probabilistic model for the nonlinear aeroelastic oscillator system. This predictive model can then be used for understanding the LCO aerodynamics, or for pursuing design optimization and control in UAVs, or for investigating energy harvesting feasibility for the self-sustained oscillations.

Recent advances in MCMC technology has expedited the use of Bayesian algorithms in constructing predictive models for complex engineering mechanics systems. Some of the recent advances in MCMC technology are exploited in this thesis to perform an efficient implementation of Bayesian model comparison using HPC. Despite these advances, Bayesian model comparison faces following two practical hurdles when dealing with over-parameterized models:

1. The issue of overfitting among over-parameterized models can be resolved by performing Bayesian model comparison on a candidate set consisting of nested models. However, for a high-dimensional model, the number of nested models is large. For example, a ten parameter model produces $\sum_{i=1}^{10} \binom{10}{i} = 1023$ nested models. Considering such a large number of candidate models makes the inversion process impractical and computationally inefficient. Even with a large subset of these nested models, Bayesian model comparison can miss out on a better model not included in the candidate set. This limitation of Bayesian model comparison in identifying the optimal nested model is illustrated in Appendix A using a polynomial regression example.
2. Bayesian framework requires the assignment of prior pdf to unknown parameters of each plausible model. However, assigning arbitrary prior pdfs to parameters whose relevance to the system physics is not known *a priori* can lead to difficulty in interpreting results

of Bayesian analysis. A typical engineering practice to this lack of prior knowledge is to assign priors with a broad support (large variance). The posterior parameter pdf remains unaffected by such priors, while the ranking of candidate models from Bayesian model comparison is highly dependent on the particular prior support. It has been well established that the Bayesian framework favours simpler models when assigning parameter priors with broad support [45, 82]. This arbitrariness in Bayesian model comparison is referred to as the Jeffreys-Lindley paradox [7]. This issue is further illustrated in Appendix A using an inverse problem involving polynomial regression.

In this thesis, we propose to resolve these practical hurdles by reformulating the problem of Bayesian model comparison among nested models as a sparse learning problem. Sparse learning is aimed at identifying the sparse set of parameters that are relevant to the system physics and eliminating the redundant parameters from the model. The machine learning concept of automatic relevance determination (ARD) [78, 87] plays a central role in facilitating sparse learning within the Bayesian framework. SBL [78, 88] and BCS [79] provide a semi-analytical implementation of ARD using Gaussian and Laplace prior pdfs, respectively. Despite their many desirable properties, the usage of SBL/BCS for sparse learning in the current context is hindered due to the following concerns:

1. SBL/BCS is designed for models where the unknown model parameters are linearly related to the measurements, or where the posterior pdf can be reasonably approximated as a Gaussian (*Laplace approximation* [40]). However, these conditions are rarely met when dealing with nonlinear models or limited measurement data. Most posterior pdfs encountered in engineering practice are non-Gaussian, and some are multimodal. SBL/BCS cannot be employed in such circumstances.
2. SBL/BCS assigns data-informed ARD priors to all the model parameters, with no regard to prior knowledge. However, while dealing with physical systems, we always possess some prior understanding regarding possible values for some (if not all) model

parameters. The inclusion of prior knowledge is critical for the regularization of posterior space. This data-based nature of SBL/BCS makes it impractical for sparse learning among physics-based models containing stochastic, nonlinear differential equations.

These issues with the application of SBL/BCS algorithms for sparse learning of physical models motivated us to explore alternative ways to enable ARD-based sparse learning within Bayesian inverse modelling. In this thesis, we propose a sampling-based and a semi-analytical approach to execute the ARD-based sparse learning. These algorithms are designed to handle sparsity in either of the following: 1) model equation (physical model), 2) measurement equation (relating measurements with state space), or 3) model and measurement error models (to quantify model discrepancy error or the measurement error). In this thesis, we only pursue sparsity in physical models.

Next, we describe the problem of model comparison and sparse learning by taking the example of an engineering mechanics system.

1.5 Problem statement

This thesis addresses model comparison and sparse learning problems for mechanistic systems modelled using nonlinear stochastic differential equations. In this section, we demonstrate the problem-at-hand by considering a dynamic structural system. Consider a multi-storey shear building frame with rigid floors being modelled as a three degree-of-freedom (dof) mass-spring-damper system, as shown in Fig. 1.2a. Given the displacement vector $\mathbf{u} = \{u_1, u_2, u_3\}$, the equation-of-motion for the mass-spring-damper system is given by the following ordinary differential equation (ODE):

$$\mathbf{M}\ddot{\mathbf{u}} + \mathbf{C}\dot{\mathbf{u}} + \mathbf{K}\mathbf{u} = \mathbf{f}(t), \quad (1.1)$$

where $\mathbf{M}$ is the mass matrix, $\mathbf{C}$ is the damping matrix, $\mathbf{K}$ is the stiffness matrix, and $\mathbf{f}(t)$ is the external forcing. For the current numerical study we consider the case of free-vibration, i.e. $\mathbf{f}(t) = 0$. Given the state vector $\mathbf{u} = \{\mathbf{u}, \dot{\mathbf{u}}\}$, Eq. (1.1) is represented in the state space form as

$\dot{\mathbf{u}} = \mathbf{A}\mathbf{u}$, where matrix $\mathbf{A} = [\mathbf{0}, \mathbf{I}; -\mathbf{M}^{-1}\mathbf{K}, -\mathbf{M}^{-1}\mathbf{C}]$. Given the initial state $\mathbf{u}_0$, the solution to the linear first-order differential equation $\dot{\mathbf{u}} = \mathbf{A}\mathbf{u}$ is obtained as $\mathbf{u}(t) = e^{\mathbf{A}t}\mathbf{u}_0$. Following values are used for simulating the three-dof system for this study: $\mathbf{u}_0 = \{0, 0, 0, 0, 10, 0\}$, $m_1 = m_2 = m_3 = 1.0$, $k_1 = k_2 = k_3 = 1000.0$, $c_1 = 0$, $c_2 = 0$, $c_3 = 20$. Notice that the damping is only present between the second and the third storey. Fig. 1.2b shows a noisy time-history of the third-floor (u_3) displacement obtained by corrupting the simulated (true) response with Gaussian white noise process that has a zero mean and a variance of 0.01. The measurements consist of 160 points spread over four seconds. The measurements are deemed incomplete since the displacement at the first and second floor is not observed.

Figure 1.2: Three-dof mass-spring-damper system.

Given the measurements in Fig. 1.2b, we pose an inverse problem to estimate the inter-storey damping coefficients. For simplicity, the remaining model parameters (stiffness, mass, and measurement noise strength) are assumed to be known. To imitate the real-life circumstances, we assume that we have no prior knowledge of the underlying damping representation. In other words, it is not known apriori that the damping is only active between the second (having displacement u_2) and third (having displacement u_3) floor. Under such conditions, the damping is assumed to be present in-between all floors for the sake of inverse modelling. The unknown parameter vector for this model is $\phi = \{c_1, c_2, c_3\}$. This model is hereby referred to as the three-parameter model. This model has an over-parameterized representation of the inter-storey damping.

Bayesian parameter estimation is executed wherein the TMCMC algorithm is employed to generate 2500 samples from the posterior pdf of ϕ assuming a flat prior pdf for all the damping coefficients of the three-parameter model. See Appendix B.1 for details regarding the TMCMC algorithm. This choice of a noninformative prior pdf is a reasonable choice to make since we don't have any prior knowledge regarding the probable values of damping coefficients. The marginal posterior pdf and the stationary posterior samples are shown in Fig. 1.3. The pdf plotted in Fig. 1.3a is computed using kernel density estimation (KDE) [89].

In Fig. 1.3a, the posterior pdf of parameter c_1 is centered around the value of -4.0 whereas the true value of c_1 was zero. These negative c_1 values are non-physical since a structural damping coefficient cannot be negative. Similarly, the posterior for c_2 centered around the value of 8.0, whereas its true value was zero. This behaviour of the posterior pdf for c_1 and c_2 demonstrate the issue of overfitting as these parameters are attaining non-physical values to ensure a better fit to the noisy realization shown in Fig. 1.2b. In other words, the negative values of c_1 are trying to negate the effect of positive values of c_2 while trying to fit the noise buried in measurements. Also, notice the non-Gaussian nature of the posterior pdf, although the model equation in Eq. (1.1) represents a linear dynamical ODE. This non-Gaussian behaviour is introduced by the nonlinear relation between the observed quantity ($\mathbf{u}$) and the damping coefficients ($\mathbf{C}$), as seen from the solution $\mathbf{u}(t) = e^{At}\mathbf{u}_0$ of Eq. (1.1).

This type of system identification problem is common in structural health monitoring applications wherein damage detection is performed by monitoring the change in model parameter (stiffness/damping) values of the undamaged structure (e.g. using the incoming sensor data from a high-rise building or an aircraft wing). In such applications, the posterior parameter pdf of an overfitted model (such as the one shown in Fig. 1.3) will indicate damage on all floors. However, this finding is inaccurate since the data-generating model had non-zero damping only between the second and third floors. This aspect of the traditional Bayesian inversion approach demonstrates the issue of overfitting when using over-parameterized models to model noisy and incomplete measurements.

Figure 1.3: Bayesian parameter estimation results for the over-parameterized model with $\phi = \{c_1, c_2, c_3\}$.

To demonstrate the effect of overfitting on predictions, we consider a model which is same as the true model, and contains only one unknown parameter, i.e. $\phi = \{c_3\}$. This model is hereby referred to as the one-parameter model. Next, the TMCMC algorithm is executed to generate 2500 samples from the posterior pdf of ϕ using a flat prior for c_3 . Fig. 1.4 shows the KDE approximation of the posterior pdf of c_3 . Notice the reduction in the posterior variance of c_3 in comparison to Fig. 1.3a wherein parameters c_1 and c_2 were also estimated alongside c_3 . Also, the posterior pdf of c_3 in Fig. 1.4 is closely Gaussian and centered around the true value of 20.0. This reduction in posterior variance for a simpler model leads to reduction in variance among predictions. Fig. 1.5 shows the predicted realizations for velocity ($\dot{u}_1$) at the first floor obtained using 100 iid samples randomly selected from the set of 2500 posterior samples generated previously using TMCMC. Notice the reduction in prediction variance for the true model (Fig. 1.5b) in comparison to the over-parameterized model (Fig. 1.5b).

It should be noted that the severity of overfitting depends on the strength of measurement noise (or signal-to-noise ratio) and the size of measurement data. For clean measurements

Figure 1.4: Posterior parameter pdf for the data-generating model.

(a) Over-parameterized model with $\phi = \{c_1, c_2, c_3\}$

(b) Data-generating model with $\phi = \{c_3\}$

Figure 1.5: 100 prediction realizations of velocity u_1 .

(low measurement noise) and/or large measurement data, the posterior parameter pdf is more likely to be close to the true value with a low uncertainty. This behaviour of the posterior pdf and the resulting predictions were confirmed through numerical experiments, and the results are not presented here for brevity. Overfitting is present even in the case of low measurement noise data due to the finiteness of the measurement data. Also, overfitting is different from unidentifiability, which depends on the shape and multimodality of the objective function being optimized to estimate the parameters. However, as demonstrated in Chapter 4, both overfitting

and unidentifiability can be remedied by assigning appropriate prior parameter pdfs.

A natural solution to this issue of overfitting among over-parameterized models is to execute Bayesian model comparison among candidate nested models. As mentioned in Sec. 1.4, this approach runs into issues such as the sensitivity of Bayesian model comparison to the prior pdf support (or variance). To demonstrate this issue, consider the MCMC-based Bayesian model comparison algorithm detailed in Chapter 2.1. Next, consider that all the models nested under the three-parameter model are treated as candidate models. Table 1.2 shows the damping representation for each of the seven candidate models. Table 1.2 also lists the posterior model probability (in percentage) computed using the measurements shown in Fig. 1.2b. The results are reported for three different choices of uniformly distributed prior pdf chosen for each of the unknown damping coefficients. Following inferences are made based on results in Table 1.2:

- The data-generating model ($\mathcal{M}_3$) is selected as the most optimal nested model for all choices of prior pdf, albeit the model probability decreases with the shrinking of prior support, i.e. going from $\mathcal{U}(-100, 100)$ to $\mathcal{U}(-25, 25)$. This identification of the data-generating model while accounting for modelling/measurement uncertainties reinforces the confidence in the evidence-based Bayesian model comparison approach.
- The posterior model probability is sensitive to the choice of prior pdf. Although these broad priors did not affect posterior pdfs, the Bayesian model comparison results are altered by varying the support of uniform prior pdfs.
- The posterior model probability of models that contain c_3 parameter is higher than those that do not contain c_3 . This behaviour demonstrates redundancy or sparsity within the ϕ space, i.e. parameters c_1 and c_2 are potentially irrelevant to the underlying physics.

Despite these concerns, we also acknowledge that considering all the nested models in the candidate set is not always possible. As a result, an efficient sparse learning algorithm is preferable over Bayesian model comparison among nested models to deal with sparsity in over-parameterized models (like $\mathcal{M}_7$ in Table 1.2).

$\mathcal{M}_i$	c_1	c_2	c_3	p($\mathbf{D} \mathcal{M}_i$) using prior pdf:		
				$\mathcal{U}(-25, 25)$	$\mathcal{U}(-50, 50)$	$\mathcal{U}(-100, 100)$
$\mathcal{M}_1$	•			0.00	0.00	0.00
$\mathcal{M}_2$		•		0.00	0.00	0.00
$\mathcal{M}_3^*$			•	88.6	94.5	97.3
$\mathcal{M}_4$	•	•		0.00	0.00	0.00
$\mathcal{M}_5$	•		•	4.51	2.41	1.24
$\mathcal{M}_6$		•	•	4.65	2.48	1.27
$\mathcal{M}_7$	•	•	•	2.24	0.60	0.15

Table 1.2: Bayesian model comparison results for varying prior variance.

*Model $\mathcal{M}_3$ is the data-generating model.

1.6 Research objectives

The objective of this research is to design an efficient Bayesian inversion framework that can automatically (without manually choosing candidate models) reduce an over-parameterized model (like the three-parameter model in Fig. 1.3) to an optimal model (like the one-parameter model in Fig. 1.4) that is free from overfitting and provides an optimal trade-off between data-fit and model simplicity. Although the proposed algorithms are motivated by the need to build a probabilistic predictive model for the nonlinear aeroelastic system observed in wind-tunnel, these algorithms can be applied to any general inverse problem possessing attributes defined in Sec. 1.1.

The algorithms proposed this thesis fall into the following two categories:

1. **Model comparison:** Consider that a candidate model set is available using prior knowledge regarding system physics. Subsequently, we pose the following question: ‘**Given the candidate model set and the measurements, what is the optimal physical model that provides the best trade-off between data-fit and model simplicity in a Bayesian setting?**’. This problem is tackled using an evidence-based Bayesian model comparison algorithm, which is designed using recent advances in MCMC algorithms and HPC.
2. **Sparse learning:** The model comparison problem among nested models can be reposed as a sparse learning problem to alleviate issues such as the need to 1) manually select candidate nested models, and 2) assign prior pdfs to model parameters with limited knowl-

edge. Hence, we pose the following question: ‘**Given the measurements, what is the optimal model nested under an over-parameterized model, assuming that models are nonlinear and not all parameters are questionable?**’. Chapter 3 and Chapter 4 provide more details regarding the ARD-based Bayesian algorithms proposed to answer this question, and its application to the nonlinear oscillator, along with other numerical investigations.

1.7 List of contributions and thesis organization

To this end, the following contributions are reported in this thesis (listed chapter-wise):

Ch. 1: We present a computationally efficient Bayesian model comparison algorithm powered by the recent advances in MCMC algorithms. This algorithm is motivated by the need to model the nonlinear aeroelastic system observed in the wind-tunnel. The following specific contributions are reported in this chapter:

- An efficient RWM MCMC algorithm is proposed to sample from unnormalized non-gaussian posterior parameter pdfs by exploiting the concepts of DR [63], AM [64], and INCA [65]. This parallel and adaptive MCMC sampler is augmented with the Chib-Jeliazkov method of computing model evidence needed for Bayesian model comparison.
- To model the nonlinear aeroelastic system, several nonlinear and unsteady models of LCO aerodynamics are constructed based on Wagner’s linear unsteady theory. A detailed mathematical derivation is provided for converting the integro-differential equations of Wagner’s theory into an ODE form. This linear ODE is modified to model the nonlinear LCO aerodynamics through the careful addition of nonlinear terms. Such models fall into the category of reduced-order models that contain both physics-based and data-based elements. We also report a detailed feasibility study into the applicability of such reduced-order models for modelling LCOs.

- A thorough numerical investigation is undertaken to identify the optimal candidate model for a synthetically generated LCO resembling the wind-tunnel LCO. This numerical study is intended to investigate the performance of the proposed Bayesian algorithm while dealing with nonlinear stochastic ODEs.
- The MCMC-powered Bayesian algorithm is exploited to model the complex aerodynamics of the nonlinear aeroelastic oscillator using noisy and incomplete wind-tunnel measurements for varying flow conditions. Note that the author did not conduct the wind-tunnel tests. Nevertheless, the parameter estimation and model comparison results for the wind-tunnel LCO are reported and discussed in length. The probabilistic predictions obtained using the optimal model are contrasted with the published literature.

Ch. 2: In this chapter, we report an ARD-based sparse learning algorithm designed for identifying the optimal nested model under an over-parameterized model. Following specific contributions are reported in this chapter:

- A hybrid prior pdf scheme is presented, wherein the parameters with available prior knowledge are assigned known priors, and the parameters with limited prior knowledge are assigned ARD priors. The sparsity inducing property of Laplace and Gaussian ARD priors is also investigated, along with their effects on posterior parameter pdfs.
- The proposed ARD algorithm is validated using a synthetically generated LCO, resembling the wind-tunnel aeroelastic system. The data-generating model is identified for all choices of over-parameterized models, demonstrating the applicability of the ARD algorithm in an engineering setting.

Ch. 3: A semi-analytical Bayesian algorithm is proposed as an efficient alternative to sampling-based sparse learning of over-parameterized nonlinear physical models. We call this new algorithm *nonlinear sparse Bayesian learning* (NSBL) since it is partially motivated by SBL, but applies to nonlinear physics-based models. Following specific contributions are

reported in this chapter:

- A detailed mathematical derivation into the semi-analytical apparatus of NSBL is presented, including the semi-analytical calculation of posterior parameter pdf, model evidence, and the gradient and Hessian information of evidence. A multistart Newton's method is proposed to optimize model evidence and efficiently identify parameter sparsity. It is also shown (analytically and numerically) that SBL is a special case of NSBL in a linear regression setting.
- For validating NSBL against SBL and BCS, we present a detailed numerical study in a linear regression setting, wherein a sparse polynomial chaos expansion (PCE) surrogate is constructed for the strongly-nonlinear Ishigami function. The computational efficiency of NSBL is contrasted with SBL and BCS with varying model dimensionality (number of model parameters).
- A linear regression setting wherein the posterior parameter pdf and the model evidence are both multimodal is considered for validating NSBL. These special circumstances are synthetically generated using a multimodal prior pdf on a regression parameter. A detailed numerical investigation is undertaken to assess the applicability of NSBL to physics-based inverse problems involving multimodality in either the parameter or the hyperparameter space. SBL and BCS do not apply to such cases.
- A structural dynamics example consisting of a three-dof mass-spring-damper model of a shear building frame is considered for validating NSBL. This example is similar to the one reported in Sec. 1.5. NSBL is applied to identify the sparse damping structure of the three-dof system using a noisy and incomplete realization of the floor displacement during free vibration.
- NSBL is applied to explore sparsity among ODEs proposed to model the aerodynamics of a synthetically generated nonlinear aeroelastic oscillator. Several cases of 1D, 2D and multidimensional sparse learning are presented to investigate the performance of

NSBL while dealing with over-parameterized nonlinear dynamical models. It is demonstrated that NSBL offers a significant computational advantage over the sampling-based sparse learning framework reported in Chapter 3.

Chapter 2

Bayesian model comparison for nonlinear dynamical systems

In this chapter, we report an evidence-based Bayesian model comparison algorithm that exploits recent advances in MCMC algorithms to perform the efficient computation of model evidence for nonlinear dynamical systems. The proposed algorithm is driven by the need to build a probabilistic predictive model for the nonlinear aeroelastic system using wind-tunnel measurements. Several plausible reduced-order aeroelastic models with increasing complexity are envisioned to model the nonlinear and unsteady aerodynamics behind the oscillator. The models are in the form of nonlinear stochastic ODEs and are constructed based on existing aerodynamic theories. The goal of this work is to identify an optimal model among plausible models that provide an optimal trade-off between model simplicity and data-fit, as facilitated by model evidence.

This chapter is structured as follows. In Sec. 2.1, we detail a general Bayesian inversion approach for executing the parameter estimation and model comparison tasks for nonlinear dynamical systems such as the wind-tunnel oscillator. In Sec. 2.2, we describe the nonlinear aeroelastic system and provide a brief review of the previous attempts aimed at modelling this system. In Sec. 2.3, we report and discuss the numerical investigations undertaken to validate the proposed Bayesian algorithm using a synthetically generated nonlinear oscillator. In

Sec. 2.4, we present the results for the application of the Bayesian approach to the wind-tunnel data.

2.1 Bayesian inversion methodology

Consider that an inverse problem is posed to model a complex mechanistic system using the acquired measurement data $\mathcal{D}$. When dealing with complex or incompletely-understood physics, multiple plausible models can be envisioned to fit the measurement. Consider that a candidate model set $\mathcal{M}$ consisting of $N_{\mathcal{M}}$ number of competing models is constructed based on the modeller's prior understanding of the physics. Next, given $\mathcal{M}$ and $\mathcal{D}$, we detail the Bayesian inversion algorithm aimed at constructing a predictive model for unobserved physics.

2.1.1 State-space representation of dynamical systems

A typical mathematical model $\mathcal{M}$ describing a dynamical system can be represented in the general discrete stochastic state-space form [9, 14, 16]

$$\mathbf{u}_{j+1} = \mathbf{g}_j(\mathbf{u}_j, \mathbf{f}_j, \mathbf{q}_j; \boldsymbol{\phi}), \quad (2.1)$$

where $\mathbf{u} \in \mathbb{R}^{N_u}$ is the time-varying state vector, $\mathbf{g} \in \mathbb{R}^{N_u}$ is the discrete model operator, $\mathbf{f} \in \mathbb{R}^{N_f}$ is the external deterministic (known) forcing, $\mathbf{q} \in \mathbb{R}^{N_q}$ is the model discrepancy error, and $\boldsymbol{\phi} \in \mathbb{R}^{N_\phi}$ is the unknown time-invariant parameter vector. The error term $\mathbf{q}_j$ is aimed at quantifying unmodelled dynamics and unknown random input not captured by the model operator $\mathbf{g}$. Quantifying the model discrepancy is essential to account for the inherent imperfectness of physical models, especially reduced-order models.

The measured entity $\mathbf{y} \in \mathbb{R}^{N_y}$ is mapped to the state vector through the measurement equation

$$\mathbf{y}_k = \mathbf{h}_k(\mathbf{u}_{j(k)}, \boldsymbol{\epsilon}_k; \boldsymbol{\phi}), \quad (2.2)$$

where $\mathbf{h} \in \mathbb{R}^{N_y}$ is the discrete measurement operator and $\epsilon \in \mathbb{R}^{N_\epsilon}$ denotes the measurement noise stemming from the data-acquisition system. The measurement data $\mathcal{D} = \{\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_{N_d}\}$ considered for the inversion purpose pertains to a realization of $\mathbf{y}$ captured in the field. The model integration time step and the measurement arrival time are allowed to differ and are therefore indexed by j and $j(k)$, respectively. Fig. 2.1 illustrates the time indexing scheme.

Figure 2.1: A general time-discretization scheme for stochastic models of dynamical systems shown in Eq. (2.1) and Eq. (2.2). Adopted from [1, 16, 86].

Note that both Eq. (2.1) and Eq. (2.2) are jointly referred to as the *model*. In this thesis, both model and measurement operators are allowed to be nonlinear, and the unknown parameter vector ϕ may contain parameters from both the equations.

The inclusion of $\mathbf{q}_j$ in Eq. (2.1) and ϵ_k in Eq. (2.2) introduces stochasticity to the model and enables the solution of the inverse problem in a statistical (Bayesian) sense. Throughout this thesis, we assume that $\mathbf{q}_j \sim \mathcal{N}(\bar{\mathbf{q}}_j, \mathbf{Q}_j)$ and $\epsilon_k \sim \mathcal{N}(\bar{\epsilon}_k, \mathbf{\Gamma}_k)$ are mutually independent and Gaussian, where $\bar{\mathbf{q}}_j \in \mathbb{R}^{N_q}$ and $\bar{\epsilon}_k \in \mathbb{R}^{N_\epsilon}$ are the mean vectors and $\mathbf{Q}_j \in \mathbb{R}^{N_q \times N_q}$ and $\mathbf{\Gamma}_k \in \mathbb{R}^{N_\epsilon \times N_\epsilon}$ are the covariance matrices. This assumption is supported by the maximum entropy principle [5, 90], which states that a Gaussian distribution provides the maximum uncertainty to a random variable with a finite mean and covariance. Hence, a Gaussian ϵ_k and $\mathbf{q}_j$ introduces minimal interference into the inversion process from the modeller's perspective. Alternatively, a more refined colored-noise model (additive or multiplicative) can be adopted

for ϵ_k and $\mathbf{q}_j$ and the associated parameter estimation can be carried out as per the Kennedy-O'Hagan framework [37]. However, this thesis is concerned with the refinement of the physical aspect of the model and the error model refinement is addressed elsewhere [91].

2.1.2 Bayesian parameter estimation

Given the measurements $\mathcal{D}$ and the candidate model set $\mathcal{M}$, the posterior pdf $p(\phi|\mathcal{D}, \mathcal{M})$ of unknown model parameters ϕ for each model $\mathcal{M} \in \mathcal{M}$ is obtained using the Bayes' theorem as

$$p(\phi|\mathcal{D}, \mathcal{M}) = \frac{p(\mathcal{D}|\phi, \mathcal{M})p(\phi|\mathcal{M})}{p(\mathcal{D}|\mathcal{M})} \propto p(\mathcal{D}|\phi, \mathcal{M})p(\phi|\mathcal{M}), \quad (2.3)$$

where $p(\mathcal{D}|\phi, \mathcal{M})$ is the likelihood function, $p(\phi|\mathcal{M})$ is the prior pdf and $p(\mathcal{D}|\mathcal{M})$ is the model evidence (also known as marginal likelihood or type-II likelihood). The likelihood $p(\phi|\mathcal{D}, \mathcal{M})$ incorporates the information from measurements while the prior $p(\phi|\mathcal{M})$ incorporates the prior parameter knowledge into the Bayesian framework. The evidence $p(\mathcal{D}|\mathcal{M})$ acts as a normalization constant in Eq. (2.3) for the sake of parameter estimation. The maximum a posteriori (MAP) estimate $\phi^{\text{MAP}} = \arg \max\{p(\phi|\mathcal{D}, \mathcal{M})\}$ is the mode of the posterior parameter pdf.

Given the state-space representation from Sec. 2.1.1, the likelihood function $p(\mathcal{D}|\phi, \mathcal{M})$ in Eq. (2.3) is expanded as [9]

$$p(\mathcal{D}|\phi, \mathcal{M}) = p(\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_k, \dots, \mathbf{d}_{N_d}|\phi, \mathcal{M}) \quad (2.4a)$$

$$= \prod_{k=1}^{N_d} p(\mathbf{d}_k|\mathbf{d}_{1:k-1}, \phi, \mathcal{M}) \quad (2.4b)$$

$$= \prod_{k=1}^{N_d} \int p(\mathbf{d}_k, \mathbf{u}_{j(k)}|\mathbf{d}_{1:k-1}, \phi, \mathcal{M}) d\mathbf{u}_{j(k)} \quad (2.4c)$$

$$= \prod_{k=1}^{N_d} \int p(\mathbf{d}_k|\mathbf{u}_{j(k)}, \phi) p(\mathbf{u}_{j(k)}|\mathbf{d}_{1:k-1}, \phi, \mathcal{M}) d\mathbf{u}_{j(k)}, \quad (2.4d)$$

where $p(\mathbf{d}_k|\mathbf{u}_{j(k)}, \phi)$ is available using the model equation Eq. (2.1) and $p(\mathbf{u}_{j(k)}|\mathbf{d}_{1:k-1}, \phi, \mathcal{M})$ is the forecast state pdf given the data history $\mathbf{d}_{1:k-1}$. Hence, computing the likelihood $p(\mathcal{D}|\phi, \mathcal{M})$

for a given ϕ requires a sequential filtering procedure for estimating the forecast state pdf at each time instant the measurement is available. The Bayesian model comparison and sparse learning algorithms considered in this thesis are independent of the technique employed for computing the likelihood function.

For nonlinear dynamical systems, algorithms such as EKF, UKF, EnKF and particle filter provide a reasonable solution to the non-Gaussian state estimation problem. Khalil *et al.* [17] provides an exhaustive contrast between these algorithms for nonlinear oscillatory systems. EKF assumes a Gaussian state pdf (in local sense) whose parameters are updated by linearizing the nonlinear model and measurement operator around the current state mean using the first-order Taylor series expansion. EKF is computationally efficient but only works if the measurements are dense enough to hold the assumption of linearity in-between measurements. In contrast, sampling-based filters such as the particle filter can handle sparse measurements since they use Monte Carlo samples to represent and propagate the non-Gaussian state pdf through the nonlinear model. Nevertheless, the Bayesian inversion framework presented in this thesis relies on the likelihood function being computable at any given ϕ value and is not connected to the filtering algorithm. For the wind-tunnel oscillator considered in this thesis, the measurements were found to be dense enough to apply EKF for performing the filtering task needed in Eq. (2.4). See Appendix B.4 for mathematical details regarding EKF and the resulting expression for $p(\mathcal{D}|\phi, \mathcal{M})$. For a high-dimensional state space (order of millions), EKF becomes computationally impractical in terms of memory requirement as it involves storing the high-dimensional covariance matrix. In such cases, sampling-based filters such as EnKF provides a computationally efficient alternative which involves just storing the ensemble matrix whose size depends on the state space dimension and the number of samples in the ensemble (assuming it is less than the dimension of state space).

Once the forecast pdf $p(\mathbf{u}_{j(k)}|\mathbf{d}_{1:k-1}, \phi, \mathcal{M})$ is available at each data acquisition time $t_{j(k)}$, the resulting (unnormalized) posterior parameter pdf in Eq. (2.3) is sampled using an MCMC algorithm [60, 92]. In particular, we employ the RWM MCMC algorithm having a Gaussian

proposal pdf centered at the current MCMC sample. The proposal covariance governs the length of the burn-in period (i.e. chain convergence rate) and the quality (i.e. autocorrelation) of stationary MCMC samples. Therefore, a good choice of the proposal covariance is necessary for efficient Bayesian parameter estimation, especially while dealing with numerous competing models with potentially non-Gaussian posterior parameter pdfs. In this work, we resolve this issue by using an adaptive and parallel RWM algorithm exploiting the following improvements of RWM MCMC: 1) DRAM algorithm [49] and 2) INCA [93].

DRAM algorithm is constructed by combining the concepts of DR [63, 94] and AM [64]. The AM algorithm performs automatic tuning of the RWM proposal covariance based on the sampling history, while DR involves a multi-level sampling procedure for the case of a rejected Markov chain move. The proposal tuning ensures faster chain convergence and efficient exploration of the posterior space. DR produces more accepted MCMC samples in the initial stages of sampling, thereby expediting the proposal tuning. Further, DRAM is implemented in parallel using INCA, whereby MCMC chains are allowed to communicate among themselves after fixed sample intervals. This sharing of information expedites the burn-in as it allows for faster proposal tuning. In a nutshell, a poor choice of the starting point and the proposal pdf is automatically remedied by the proposed MCMC algorithm. See Appendix B.2 for mathematical and implementation details regarding the DRAM algorithm. Next, we detail the Bayesian model comparison algorithm proposed to ascertain the goodness of plausible models.

2.1.3 Bayesian model comparison

Bayesian model comparison exploits Bayes' theorem to calculate the posterior model probability $p(\mathcal{M}|\mathcal{D}, \mathcal{M})$ of each model $\mathcal{M}$ in the candidate model set $\mathcal{M}$ as [5, 44, 45, 95]

$$P(\mathcal{M}|\mathcal{D}, \mathcal{M}) = \frac{p(\mathcal{D}|\mathcal{M})P(\mathcal{M}|\mathcal{M})}{p(\mathcal{D}|\mathcal{M})}, \quad (2.5)$$

where $P(\mathcal{M}|\mathcal{M})$ is the prior model probability, $p(\mathcal{D}|\mathcal{M})$ is the model evidence (from Eq. (2.3)) and $p(\mathcal{D}|\mathcal{M})$ is a normalization constant for a given $\mathcal{D}$ and $\mathcal{M}$. Considering all candidate models are equally likely a priori, the posterior model probability is then solely dictated by the evidence. The evidence is calculated using Eq. (2.3) as

$$p(\mathcal{D}|\mathcal{M}) = \int p(\mathcal{D}|\phi, \mathcal{M})p(\phi|\mathcal{M})d\phi. \quad (2.6)$$

The evidence $p(\mathcal{D}|\mathcal{M})$ is known as a quantitative Ockham's razor that performs an explicit trade-off between the data-fit and the model simplicity [5]. This fact is revealed when log-evidence is expanded as [54]

$$\underbrace{\ln p(\mathcal{D}|\mathcal{M})}_{\text{Log-evidence}} = \underbrace{E[\ln p(\mathcal{D}|\phi, \mathcal{M})]}_{\text{Data-fit}} - \underbrace{E\left[\ln \frac{p(\phi|\mathcal{D}, \mathcal{M})}{p(\phi|\mathcal{M})}\right]}_{\text{Model complexity}} \quad (2.7)$$

where the expectation $E[\cdot]$ is with respect to the posterior pdf $p(\phi|\mathcal{D}, \mathcal{M})$. A detailed derivation of Eq. (2.7) is provided in Appendix A.2. The first term of the right-hand side of Eq. (2.7) quantifies the data-fit capability of the model as it relates to the likelihood function. This term is also referred to as the goodness-of-fit (GOF) [54]. The second term quantifies the KL divergence between the prior and the posterior pdf, and is also referred to as the expected information gain (EIG) [54]. Note that an over-parametrized model is more likely to overfit noisy $\mathcal{D}$ by mistaking the measurement noise as a part of the system physics. Consequently, the posterior is vastly different from the prior due to the overfitting caused by the unrestricted complexity of the model. In turn, EIG is higher for over-parameterized models, which then penalizes the model by reducing the log-evidence according to Eq. (2.7). This competing nature of GOF and EIG imparts model evidence the quantitative Ockham's razor property. Also, this probabilistic viewpoint of model complexity is in contrast with techniques such as AIC and BIC where model complexity is a sole function of the number of unknown parameters (N_ϕ) [6]. As demonstrated through Eq. (2.7), the Bayesian notion of model complexity relies both on the model form (through

$p(\phi|\mathcal{D}, \mathcal{M})$) and the prior knowledge (through $p(\phi|\mathcal{M})$).

The evidence $p(\mathcal{D}|\mathcal{M})$ in Eq. (2.6) is analytically intractable for any general nonlinear dynamical model. Laplace’s method provides an efficient way of estimating evidence for a nearly Gaussian posterior pdf $p(\phi|\mathcal{D}, \mathcal{M})$ [96]. However, the assumption of Gaussianity only holds for ‘*large and dense*’ measurement data $\mathcal{D}$. This property of $\mathcal{D}$ is difficult to establish in practice and can only be ascertained by sampling the posterior parameter pdf. See Appendix B.5 for mathematical details regarding the Laplace’s method. Recent advances in MCMC methods and HPC has enabled an efficient exploration of non-Gaussian posterior pdfs and the subsequent computation of model evidence. Some notable evidence estimation algorithms include nested sampling [69], power posteriors [66], annealed importance sampling [70], harmonic mean estimator [71], Gauss-Hermite quadrature [72], transitional MCMC [48], the Chib method (Gibbs sampling) [68], the Chib-Jeliazkov method (MH sampling) [67], and many others. See Freil and Wyse [47] for a comprehensive and recent review of evidence computation methods. In this work, we employ the Chib-Jeliazkov method for estimating evidence that uses the stationary MCMC samples generated from $p(\phi|\mathcal{D}, \mathcal{M})$ using the DRAM algorithm [67]. See Appendix B.3 for mathematical details regarding the Chib-Jeliazkov method.

2.2 Bayesian modelling of nonlinear aeroelasticity

The application of Bayesian inference to aeroelastic systems is part of a growing appreciation of the scientific and engineering community for the importance of “the role of uncertainty quantification (UQ) in efforts to understand the physics of nonlinear aeroelasticity and to certify aeroelastic stability” [97]. The goal of this work is concerned with the former motivation, which is an understanding of the physics through Bayesian inference, in particular choosing the most plausible model, and then trying to infer the nature of the physics from the identified model. In this section, we introduce the nonlinear aeroelastic oscillator that is the motivation behind the proposed Bayesian inversion methodology. Subsequently, we discuss the existing

unsteady aerodynamic theory and provide a detailed mathematical and numerical exposition into the proposed extension of this theory to model the nonlinear aeroelastic oscillator.

2.2.1 Nonlinear aeroelastic oscillator

The aeroelastic oscillator is located in the low-speed wind-tunnel at the Royal Military College of Canada and consists of a rigid wing with a uniform NACA0012 airfoil cross-section. The NACA0012 airfoil has a thin and uncambered (symmetric) cross-section. The wing apparatus is shown in Fig. 2.2. Although the wing apparatus is capable of undergoing both plunge (translation) and pitch (rotation) motions, the aeroelastic oscillator studied in this thesis is a pure-pitch (single dof) phenomenon with the plunge dof held fixed. The wing is placed vertically to eliminate the effects of gravity. End plates are located on each side of the wing to minimize aerodynamic 3D effects such that the flow can effectively be considered to be 2D. The axis-of-rotation (elastic axis) is located at 18.6% chord ahead of the leading edge. See Poirel *et al.* [85] for further details regarding the wind-tunnel apparatus and the testing conditions.

Figure 2.2: Schematic of the wing apparatus. Adopted from Khalil *et al.* [9].

The wing is observed to undergo a self-sustained dynamic instability when placed in the transitional Reynolds number ranging $5.0 \times 10^4 \leq Re \leq 1.3 \times 10^5$ (Re is the chord-based Reynolds number) [85]. Fig. 2.3 shows the observed pitch response and its power spectral density (PSD)

at $Re = 85,000$. Notice the diverging pitch amplitude and the subsequent stabilization to LCO with a constant amplitude of around 5° . Also, notice the peak at the third superharmonic frequency in the PSD plot in Fig. 2.3. The superharmonic at 30 Hz is due to the wind-tunnel motor and is not related to the LCO physics. Nevertheless, the key feature of the observed LCOs in the transitional Re regime remain the low amplitude (less than 5.5°), which differentiates this instability from the stall flutter (large-amplitude flutter). See Poirel *et al.* [85] for a detailed experimental investigation into the oscillator.

Figure 2.3: Wind-tunnel pitch response and its PSD. Adopted from Sandhu *et al.* [1].

We first look towards the published literature to gain an insight into the physics behind the observed LCO. Poirel *et al.* [85] first reported this aeroelastic phenomenon and attributed its origin to the unique flow conditions present in the transitional Re regime. Reynolds number is a non-dimensional entity that quantifies the ratio of inertial and viscous forces, which in turn dictates the type of flow (laminar or turbulent or transitional) in the boundary layer attached to the wing surface. Transitional Re flows have been well-known to display numerous complex boundary layer characteristics such as laminar separation, transition of the laminar shear layer and re-attachment of the turbulent boundary layer which leads to the formation of a *laminar separation bubble* (LSB) on the wing surface [98–101]. Fig. 2.4 shows a schematic of LSB.

Subsequent studies have shown that the aeroelastic instability is caused by the laminar boundary layer trailing-edge separation, while the stabilization to LCO is powered by the presence of LSB in a feedback relationship with the free pitching wing [102]. As a result, the LCO in Fig. 2.3 is also known as the *laminar separation flutter* [3]. The effect of unsteady aerodynamics was speculated to be behind the stabilization to LCO, although this was not verified during experimental investigations [85]. The reduced frequency k is often used as a dimensionless indicator of unsteadiness, whereby $0 < k < 0.05$ indicates quasi-steady aerodynamics and $k > 0.1$ indicates (potentially) unsteady aerodynamics. The reduced frequency for the observed LCOs was in the range 0.1-0.2, which indicates that unsteady aerodynamics might be contributing to the LCO physics.

Figure 2.4: Laminar separation bubble. Adopted from Sandhu *et al.* [1].

The richness and complexity of this aeroelastic phenomenon have the potential of impacting the safe operation of flight vehicles such as the unmanned aerial vehicles. Hence, understanding the aerodynamics behind the aeroelastic oscillator is critical to the robust control and prevention of such destructive instabilities. This understanding needs to be perfected by constructing a representative predictive model of the aerodynamic loading on the wing using the wind-tunnel response shown in Fig. 2.3. This model should accurately identify and quantify the potentially nonlinear and unsteady aerodynamics behind the LCO while acknowledging the inherent modelling uncertainties. There is a growing interest among the aeroelasticity community on the use of Bayesian methods to solve such inverse problems [3, 97, 103, 104]. The model obtained using Bayesian inverse modelling should be free from overfitting, and thus be capable of pre-

dicting the wing response beyond the realm of wind-tunnel measurements (i.e. extrapolation).

In this work, we focus solely on the inverse modelling of aerodynamic loading while the structural properties of the wing apparatus are considered as known. The structural properties in the pure-pitch dof has been estimated previously using measurements from a free-decay motion under no airflow conditions [105]. The equation-of-motion for the single-dof pitching motion is available as

$$I_{EA}\theta'' + D_l\theta' + K_l\theta + K_{nl}\theta^3 = D_{nl}\text{sign}(\theta') + M_{EA}(\theta, \theta', \theta''; \phi) , \quad (2.8)$$

where $\theta = \theta(t)$ is the pitch displacement, θ', θ'' are derivatives with respect to time t , I_{EA} is the mass moment of inertia, D_l is the linear viscous damping coefficient, K_l is the linear stiffness coefficient, K_{nl} is the nonlinear stiffness coefficient, D_{nl} is the nonlinear friction damping coefficient, ϕ is the unknown time-invariant parameter vector, $M_{EA} = qc^2sC_M$ is the aerodynamic moment about the elastic axis, $q = \rho U^2/2$ is the dynamic pressure, U is the free-stream airflow velocity, c is the wing chord (constant-chord), s is the wing span, and $C_M(\cdot)$ is the aerodynamics moment coefficient. Table 2.1 lists the estimated value (the MAP estimate) of the structural parameters of Eq. (2.8) and other geometrical properties of the wing. Notice that, in addition to the linear terms, a small amount of dry friction was observed, as well as nonlinearity in stiffness for larger amplitudes; these effects show up in the D_{nl} and K_{nl} terms, respectively. The uncertainty in structural parameters is ignored for the sake of this study as we focus solely on the accurate modelling of the LCO aerodynamics. Although not pursued in this thesis, the proposed Bayesian algorithms can handle the simultaneous estimation of both structural and aerodynamic parameters as the method is general.

The structural model in Eq. (2.8) is non-dimensionalized by replacing time t with the non-dimensional time $\tau = 2Ut/c$. This ensures a consistent comparison of results among LCOs captured at different free-stream velocities U . Subsequently, Eq. (2.8) is re-written as

$$\ddot{\theta} = c_1\text{sign}(\dot{\theta}) + c_2\theta + c_3C_M + c_4\dot{\theta} + c_5\theta^3, \quad (2.9)$$

Parameter	Value	Parameter	Value (Estimated [105])
I	1.40e-03 kg.m ²	K_l	3.39e-01 kg.m ² /s ²
ρ	1.19 kg/m ³	K_{nl}	3.03e-00 kg.m ² /(s ² rad ²)
s	0.61 m	D_l	1.11e-03 kg.m ² /s
c	0.156 m	D_{nl}	1.14e-03 kg.m ² .rad/s

Table 2.1: Known parameters of the wind-tunnel apparatus.

where $\theta = \theta(\tau)$ is the pitch displacement as a function of non-dimensional time τ , and the derivatives $\dot{\theta}, \ddot{\theta}$ are with respect to τ . The new parameters c_1, c_2, c_3, c_4, c_5 in Eq. (2.9) can be calculated using Table 2.1. The values thus obtained are $c_1 = -6.875\text{e-}05$, $c_2 = -2.038\text{e-}02$, $c_3 = 3.819\text{e-}02$, $c_4 = -7.275\text{e-}03$ and $c_5 = 1.824\text{e-}01$. An inverse problem is posed to model C_M (the aerodynamic loading) in Eq. (2.9) using the wind-tunnel response shown in Fig. 2.3. Next, we review existing unsteady aerodynamic theories and discuss how these theories can be extended to model the potentially nonlinear and unsteady aerodynamics of the LCO.

2.2.2 Brief review of unsteady aerodynamics

The formulation of linear unsteady aerodynamics of thin airfoils in subsonic conditions dates back to Birnbaum and Wagner [106]. Wagner [107] proposed a time-domain approach for calculating the unsteady circulatory aerodynamic loads on a 2D thin airfoil using an indicial response function under ideal and incompressible flow conditions. Wagner obtained an analytical expression for the lift force for a step-change in angle-of-attack, which was applied to an arbitrary airfoil motion by using Duhamel's integral. A decade later, Theoderson [108] provided an analytical expression in the frequency domain for the unsteady lift force for a sinusoidal pitching airfoil through a transfer function. Garrick [109] then showed that Theoderson's transfer function and Wagner function form a Fourier transform pair. Dowell [110] proposed a first-order approximation to Wagner function wherein the Wagner constants were estimated from the exact frequency-domain data through a least-squares approach.

Wagner's theory considers flow unsteadiness by incorporating the effect of trailing-edge vortices in the modelling of circulatory aerodynamic loads acting on the airfoil [107]. Consider

that an airfoil at a non-zero angle-of-attack starts impulsively from rest to a uniform velocity U . According to Kelvin's circulation theorem, the net circulation around an airfoil is a conserved quantity for an inviscid, incompressible flow. Hence, a starting vortex is shed from the trailing edge having circulation equal and opposite to the change in circulation around the airfoil. The quasi-steady assumption implies that the lift force adjusts instantly to conform to the new flow conditions and, therefore, ignores the effect of the starting vortex. In contrast, Wagner revealed that only half of the change in lift force is felt at the instant the circulation changes. The lift force gradually reaches its steady state value as the starting vortex moves downstream. This delay in lift force reaching the steady state is due to the presence of the starting vortex in the wake. This temporal variation of the lift force due to a step-change in circulation is quantified in the form of an indicial response function, also called the Wagner function $\varphi(t)$. Thus, the Wagner function introduces memory effects in the aerodynamic load by considering not only the current state of the airfoil but also the history of its motion.

Next, we present a brief mathematical analysis on Wagner's unsteady theory for the case of 1) pure pitch motion, 2) symmetric and thin airfoil, and 3) small angle-of-attack ($\sin \theta \approx \theta$). Wagner's theory relies on the effective angle-of-attack $\alpha_{\text{eff}}(t)$ of the airfoil, which represents the relative angle the airfoils sees when pitching with an angular velocity θ' . The expression for α_{eff} is available from the quasi-steady thin-airfoil theory as [106]

$$\alpha_{\text{eff}}(t) = \frac{w_{3c/4}(t)}{U} = \theta + \frac{c(0.5 - a_h)}{2U} \theta' \quad (2.10)$$

where a_h (positive aft of mid-chord) is the distance between mid-chord and elastic axis normalized by $c/2$, and $w_{3c/4}(t)$ is the effective downwash velocity calculated at the three-quarter chord position from the leading edge. Note that $\theta(t) > 0$ implies nose-up. The net circulatory lift force acting on the airfoil is computed as $L = qcsC_L$, where the lift coefficient $C_L = 2\pi\alpha_{\text{eff}}$ for symmetric and thin airfoils.

In Wagner's theory, the incremental lift force $dL(t)$ due to an incremental step change in

$\alpha_{\text{eff}}(t)$ is computed using Eq. (2.10) as

$$dL(t) = qcs \cdot dC_L \cdot \varphi(t) = qcs \cdot 2\pi d\alpha_{\text{eff}} \cdot \varphi(t) = \frac{2\pi qcs}{U} \cdot dw_{3c/4} \cdot \varphi(t), \quad (2.11)$$

where $\varphi(t) \in [0.5, 1.0)$ is the Wagner function that incorporates the unsteady (memory) effect into $dL(t)$. Notice that $\varphi(t) = 1.0$ implies quasi-steady aerodynamics. For an arbitrary pitching motion, the net circulatory lift is obtained by integrating Eq. (2.11) using Duhamel's integral as [111]

$$L(t) = \frac{2\pi qcs}{U} \left[w_{3c/4}(0)\varphi(t) + \int_0^t \varphi(t-t_0) \frac{dw_{3c/4}(t_0)}{dt_0} dt_0 \right]. \quad (2.12)$$

The aerodynamic moment $M_{\text{ea}}(t)$ at the elastic axis is obtained as $M_{\text{ea}}(t) = eL(t) + M_{\text{ac}}(t)$, where $e = 0.5c(0.5 + a_h)$ is the distance between the aerodynamic center and the elastic axis and $M_{\text{ac}}(t)$ is the moment at the aerodynamic center. For symmetric and thin airfoils (such as NACA0012), the aerodynamic center is located at the quarter-chord point from the leading edge, and $M_{\text{ac}}(t)$ is zero [111]. For the wind-tunnel LCO, $a_h = -0.63$ and $e = -0.01014$, meaning the elastic axis is located between the aerodynamic center and the leading edge.

The aerodynamic moment can also be written as $M_{\text{ea}}(t) = qscC_M$, where C_M is the moment coefficient. Using Eq. (2.12) and $M_{\text{ea}}(t) = eL(t)$, an expression for C_M is obtained in terms of the dimensionless time $\tau = 2Ut/c$ as

$$C_M(\tau) = \frac{2\pi e}{cU} \left[w_{3c/4}(0)\varphi(\tau) + \int_0^\tau \frac{dw_{3c/4}(\tau_0)}{d\tau_0} \varphi(\tau - \tau_0) d\tau_0 \right], \quad (2.13)$$

where $w_{3c/4}(\tau) = U\{\theta + (0.5 - a_h)\dot{\theta}\}$.

Table 2.2 shows the values of $\varphi(\tau)$ provided by Wagner [107] for an increasing τ . Following Wagner's work, many analytical approximations of the Wagner function have been proposed, as listed in Table 2.3. Fig. 2.5 shows the contrast between these analytical approximations of $\varphi(\tau)$ and the values published by Wagner (Table 2.2). Notice that all the $\varphi(\tau)$ approximations satisfy the constraints: 1) $\varphi(\tau = 0.0) = 0.5$, and 2) $\varphi(\tau \rightarrow \infty) = 1.0$. In other words, the lift force is

half of its steady state value at the instant when the downwash or the effective angle-of-attack changes. As τ increases, $\varphi(\tau)$ asymptotically approaches the steady state value of $\varphi(\tau) = 1$ since the flow reaches stationarity with the starting vortex moving farther downstream. Another way to understanding Wagner function is by realizing that $\tau=1$ gives $U = c/(2t)$, meaning the starting vortex travels a distance of $c/2$ for a unit change in τ . From Table 2.2, at $\tau = 20$, the starting vortex has traveled ten chord-lengths in the downstream and the circulatory lift force is 93.2% governed by quasi-steady aerodynamics and 6.8% by the unsteady aerodynamics.

τ	$\varphi(\tau)$	τ	$\varphi(\tau)$
< 0.0	0.0000	6.0	0.8125
0.0	0.5000	7.0	0.8325
0.5	0.5557	8.0	0.8485
1.0	0.6006	9.0	0.8625
2.0	0.6693	10.0	0.8745
3.0	0.7195	20.0	0.9321
4.0	0.7582	∞	1.0000
5.0	0.7880		

Table 2.2: Numerical values of Wagner function proposed by Wagner [107].

Reference	Wagner function $\varphi(\tau)$
Wagner [107]	Numerical values (Table 2.2)
Garrick [109]	$\frac{\tau + 2}{\tau + 4}$
Dowell [110]	$1 - 0.5 e^{-0.2\tau}$
R.T. Jones [112]	$1 - 0.165 e^{-0.0455\tau} - 0.335 e^{-0.3\tau}$
W.P. Jones [113]	$1 - 0.165 e^{-0.041\tau} - 0.335 e^{-0.32\tau}$
Venkatesan and Friedmann [114]	$1.0 - 0.203 e^{-0.072\tau} - 0.236 e^{-0.261\tau} - 0.06 e^{-0.8\tau}$

Table 2.3: Existing Wagner function approximations.

The exponential form of Wagner functions in Table 2.3 is preferred since it can be easily manipulated using Laplace transforms. A general expression for a J^{th} -order exponential form

Figure 2.5: Comparison of the existing Wagner function approximations with the values published by Wagner [107].

of the Wagner function is defined as

$$\varphi^J(\tau) = 1.0 - \sum_{j=1}^J A_j e^{-B_j \tau}, \quad \sum_{j=1}^J A_j = 0.5, \quad B_j > 0 \quad \forall j, \quad (2.14)$$

where A_j, B_j are the Wagner constants [107, 113]. The higher the number of exponential terms (or the order J), the higher the accuracy of the Wagner function in relation to the values in Table 2.2 in the sense of mean-square error. As noted by Dowell [110] and Vepa [115], the Wagner constants A_j, B_j are independent of the characteristics of the force and motion and depend solely on the aerodynamics.

Next, we look to transform the integro-differential form of Eq. (2.13) into an ODE form, which can then be extended to model C_M in Eq. (2.9) for the wind-tunnel oscillator. Defining Laplace transforms $\mathcal{L}\{\theta(\tau)\} = \Theta(r)$ and $\mathcal{L}\{\dot{\theta}(\tau)\} = \mathcal{X}(r)$, we take the Laplace transform of $C_M(\tau)$ in Eq. (2.13) as

$$\mathcal{L}\{C_M(\tau)\} = \frac{2\pi e}{cU} \left[w_{3c/4}(0) \mathcal{L}\{\varphi(\tau)\} + \mathcal{L}\left\{ \frac{dw_{3c/4}(\tau)}{d\tau} \right\} \mathcal{L}\{\varphi(\tau)\} \right] \quad (2.15a)$$

$$= \frac{2\pi e}{cU} [s \mathcal{L}\{w_{3c/4}(\tau)\} \mathcal{L}\{\varphi(\tau)\}], \quad (2.15b)$$

where $\mathcal{L}\{w_{3c/4}(\tau)\} = U\{\Theta(r) + (0.5 - a_h)\mathcal{X}(r)\}$. We consider a first-order Wagner function,

i.e. $\varphi(\tau) = 1 - Ae^{-B\tau}$. Using Eq. (2.10), we perform the Laplace transform of $\varphi(\tau)$ as

$$\mathcal{L}\{\varphi(\tau)\} = \frac{1}{r} - \frac{A}{r+B} = \frac{r(1-A) + B}{r(r+B)}. \quad (2.16)$$

Next, we substitute $\mathcal{L}\{\varphi(\tau)\}$ from Eq. (2.16) in Eq. (2.15) to obtain

$$\begin{aligned} r\mathcal{L}\{C_M(\tau)\} + B\mathcal{L}\{C_M(\tau)\} &= \frac{2\pi e}{c}[B\Theta(r) + B(0.5 - a_h)\mathcal{X}(r) \\ &\quad + (1-A)r\Theta(r) + (1-A)(0.5 - a_h)r\mathcal{X}(r)] \end{aligned} \quad (2.17)$$

Taking the inverse Laplace transform of the above equation, we get

$$\frac{\dot{C}_M}{B} + C_M = \frac{2\pi e}{c} \left\{ \theta + \left(\frac{1}{2} - a_h + \frac{1-A}{B} \right) \dot{\theta} + \left(\frac{1}{2} - a_h \right) \left(\frac{1-A}{B} \right) \ddot{\theta} \right\}. \quad (2.18)$$

Note that the Wagner constants A and B must satisfy the first-order Wagner function constraints of $A = 0.5$ and $B > 0$. From Eq. (2.18), a manifest physical interpretation of the parameter B is that it is inversely proportional to the strength of the unsteadiness; a large B corresponds to quasi-steady flow as $\dot{C}_M$ and $\ddot{\theta}$ terms in Eq. (2.18) can be neglected. In that case, we are left with the familiar quasi-steady expression $C_M = 2\pi e/c(\theta + (0.5 - a_h)\dot{\theta})$. The corresponding derivation for a second-order and a third-order Wagner function is provided in Appendix C.

Despite its fundamental soundness, the linear ODE obtained in Eq. (2.18) using Wagner's unsteady theory is not applicable to model a nonlinear separated flow phenomenon. Next, we investigate the suitability of a nonlinear extension of Eq. (2.18) to model the nonlinear and unsteady aerodynamics of the wind-tunnel LCO from Sec. 2.2.1.

2.2.3 Nonlinear extension of Wagner's model

Under quasi-steady flow assumptions, the nonlinear aerodynamic loading (C_M) of a flutter phenomenon is typically modelled by the so-called generalized Duffing-van der Pol oscilla-

tor [9, 86, 102]

$$C_M = e_1\theta + e_2\dot{\theta} + e_3\theta^3 + e_4\theta^2\dot{\theta}. \quad (2.19)$$

where e_i 's are the unknown time-invariant parameters. The Duffing terms (θ, θ^3) quantify the aerodynamic stiffness, while the van der Pol terms $(\dot{\theta}, \theta^2\dot{\theta})$ account for the aerodynamic damping. In order to emulate the diverging oscillations of the wind-tunnel aeroelastic system, the linear aerodynamic damping term $e_2\dot{\theta}$ in Eq. (2.19) needs to be larger (and opposite sign) than the viscous structural damping term $c_4\dot{\theta}$ in Eq. (2.9). As the oscillation amplitude grows, the $e_4\theta^2\dot{\theta}$ term becomes dominant and the LCO is achieved (for a negative e_4). Despite these useful features, Eq. (2.19) fails to capture the unsteady aerodynamics of the LCO.

We propose to combine Eq. (2.18) using Eq. (2.19) to obtain the following empirically derived ODE for modelling the nonlinear unsteady aerodynamic moment:

$$\frac{\dot{C}_M}{B} + C_M = e_1\theta + e_2\dot{\theta} + e_3\theta^3 + e_4\theta^2\dot{\theta} + \frac{c_6}{B}\ddot{\theta}. \quad (2.20)$$

Note that $c_6 = 0.5\pi(0.25 - a_h^2)$ is known from Eq. (2.18). The form of Eq. (2.20) is conceptually similar to the so-called time-linearized nonlinear modelling used successfully in transonic flutter [116], except for the additional nonlinear van der Pol damping needed to capture the LCO. Similar types of fluid oscillator models have been used in the context of bluff body aerodynamics [117–119]. However, in these cases an additional $\ddot{C}_M$ (inertia) term is generally included to enable self-sustained vortex shedding, which leads to a vortex-induced vibrations of an elastic structure. The origin of the aeroelastic LCO is not vortex-induced vibrations but negative aerodynamic damping [85, 102]. Hence, there is no need to include a fluid inertia term. Nevertheless, such a term appears naturally as a consequence of using a two-state representation of Wagner's function [1].

The ODE in Eq. (2.20) can then be coupled to the structure equation-of-motion in Eq. (2.8) to complete the full set of coupled equations of motion for the aeroelastic system. For numerical illustrations, we generate a synthetic LCO using the coupled system for different values of B

and fixed value of $A = 0.5$. We substitute the following values for aerodynamic coefficients: $e_1 = -1.25$, $e_2 = 1.00$, $e_3 = 100.0$ and $e_4 = -500.0$. These values were chosen such that the resulting LCO is similar to the wind-tunnel LCO. Fig. 2.6 shows the phase-plane plot for the steady state LCO response for various values of B , in contrast to the filtered wind-tunnel data (steady state) at $Re = 85,000$. The qualitative closeness of the phase-plane plots between the synthetic and wind-tunnel data further justifies the appropriateness of Eq. (2.8) and Eq. (2.20) in capturing the unsteady and nonlinear aerodynamics in the wind-tunnel tests.

Figure 2.6: Phase-plane diagram for varying Wagner constant B , contrasted with the wind-tunnel data for $Re = 85,000$.

In Fig. 2.6, as B is increased, the system response approaches the quasi-steady characteristics. Another way to look at this behaviour is through Wagner function. For higher B , the Wagner function $\varphi(\tau) = 1 - 0.5e^{-B\tau}$ behaves like a step function from 0.5 to 1.0, meaning the aerodynamics adjusts almost instantly to any change in the airfoil motion, leading to quasi-steady flow conditions.

Based on Eq. (2.19) and Eq. (2.20), several plausible quasi-steady and unsteady aerodynamic models with varying nonlinearity are proposed to quantify the interplay between the LCO response and the aerodynamic loading. Subsequently, Bayesian model comparison is used to

select an optimal aerodynamic model using the wind-tunnel measurements.

2.3 Numerical investigations

2.3.1 Bayesian inversion of linear unsteady aerodynamics

The goal of this study is to understand the relation between the reduced frequency (a quantitative indicator of unsteadiness) and Wagner function for a linear unsteady model similar to Eq. (2.18). Also, we want to validate the capability of Bayesian parameter estimation for inverting linear unsteady aerodynamics, before undertaking the Bayesian inversion of nonlinear unsteady models.

The measurements $\mathcal{D}$ required for posing the inverse problem are generated by the coupled simulation of a linear equation-of-motion and the linear unsteady model in Eq. (2.18). Note that Eq. (2.18) was derived using a first-order Wagner function approximation. In order to artificially sustain a harmonic pitch response, a sinusoidal forcing is applied to Eq. (2.18). Specifically, the data-generating model equation consists of the following coupled ODEs:

$$\ddot{\theta} + 2k_n\zeta\dot{\theta} + k_n^2\theta = a_s C_M, \quad (2.21a)$$

$$\frac{\dot{C}_M}{B} + C_M = c_1\theta + c_2\dot{\theta} + c_3\ddot{\theta} + a_f \sin(k_f\tau) + \sigma\xi(\tau), \quad (2.21b)$$

where ζ is the structural damping ratio, $k_n = \omega c/(2U)$ is the reduced natural frequency, k_f is the reduced forcing frequency, σ^2 is the variance of the model discrepancy error and $\xi(\tau)$ is a Gaussian white noise process. The model discrepancy term $\sigma\xi(\tau)$ is included in Eq. (2.21b) to introduce non-physical attributes in the simulated pitch response, including the turbulent flow effects. The wing structure is assumed to have the same properties as the wind-tunnel wing. Using the wing parameters listed in Table 2.1, c_i 's are known from Eq. (2.18) and $a_s = c^4 \rho s / (8I_{EA})$ is known from Eq. (2.8). Notice that the coupled ODEs in Eq. (2.21) are written in terms of the dimensionless time $\tau = 2Ut/c$.

A simulated pitch response is generated from Eq. (2.21) using the following values: $\zeta = 0.5$, $k_n = 0.5$, $a_f = 1.0$ and $\sigma = 0.01$. The parameters of the first-order Wagner function $\varphi(\tau) = 1 - Ae^{-B\tau}$ are taken to be $A = 0.5$ and $B = 0.2$, based on the work by Dowell [110]. The measured response $\mathcal{D} = \{d_k; k = 1, 2, \dots, n\}$ is obtained by corrupting the simulated response using $d_k = \theta(\tau_k) + \epsilon_k$, where $\epsilon_k \sim \mathcal{N}(0, \gamma^2)$ are independent and identically distributed (iid) Gaussian random variables. To ensure consistent comparison of results, the variance γ^2 of the measurement noise is chosen such that the signal-to-noise ratio (SNR) is held constant across all k_f values. This condition is necessary since the amplitude of the response varies with k_f . For this study, we choose $\text{SNR} = 20$ (in decibel scale). Fig. 2.7 shows the measured and simulated (true) pitch response for $k_f = 0.3, 0.2, 0.1$ and 0.02 . Also, in Fig. 2.7, we show the quasi-steady counterpart of these responses computed by eliminating terms $\dot{C}_M/B$ and $c_3\ddot{\theta}$ in Eq. (2.21b). This quasi-steady setting is also equivalent to setting the Wagner constant B as infinity, i.e $\varphi(\tau) = 1$.

Given $\mathcal{D}$, an inverse problem is posed for estimating the unknown parameter vector $\phi = \{B, \sigma\}$ of the data-generating aerodynamic model in Eq. (2.21b). The Wagner constant B appears in parameter c_2 and c_3 in Eq. (2.21). The remaining parameters of Eq. (2.21) and the measurement noise strength is assumed as known for the sake of simplicity. Also, $A = 0.5$ is a constraint for a first-order Wagner function $\varphi(\tau)$, and is therefore considered as known [110].

The Bayesian parameter estimation algorithm detailed in Sec. 2.1.2 is employed to compute the posterior pdf of ϕ given $\mathcal{D}$ in Fig. 2.7. Given the lack of information regarding the joint statistics of B and σ , the joint prior pdf of ϕ is assumed to be the product of the marginal pdfs of B and σ . The prior parameter independence is not an inherent property of ϕ but a probabilistic representation of our information at hand that ensures minimum interference with the posterior $p(\phi|\mathcal{D})$. We assume a log-normal prior pdf for B with the median at 0.2 (true value) and coefficient of variation (COV) of 50%. The modelling error strength σ is assigned a log-normal distribution with the median at 0.01 (true value) and COV of 50%. The choice of a log-normal prior pdf was based on the fact that both B and σ are strictly positive entities. The

Figure 2.7: Pitch response generated for varying forcing reduced frequency k_f .

state estimation involved in the computation of the likelihood function is performed using the Kalman filter due to the linearity of the model and measurement equations [14]. Subsequently, the adaptive MCMC algorithm from Sec. 2.1.2 is employed to compute the posterior $p(\phi|\mathcal{D})$.

Fig. 2.8 show the 95 % credible intervals computed using the posterior pdf of B and σ for various k_f values. As shown in Fig. 2.8, the Wagner constant B is estimated with a reasonable accuracy for all the reduced frequencies. The uncertainty in B increases as the reduced frequency decreases due to a lower unsteadiness in the overall aerodynamics encapsulated in

$\mathcal{D}$. This behavior of the posterior pdf of B reinforces the notion that B quantifies the unsteady effects.

Figure 2.8: Posterior parameter statistics for varying forcing reduced frequency.

Fig. 2.8 also plots the posterior statistics of σ for the quasi-steady and unsteady models. Since the measurements were generated using the unsteady aerodynamic model, the MAP estimate for σ is naturally higher for the quasi-steady model. Also, the MAP estimate of σ for the unsteady model is reasonably close to the true value, i.e. $\log \sigma = -4.605$. This successful inversion of linear unsteady aerodynamics in the presence of modelling and measurement uncertainties enforces strong confidence in the Bayesian apparatus. Next, a nonlinear extension of Eq. (2.21b) is studied for modelling a synthetically generated LCO.

2.3.2 Bayesian inversion of nonlinear unsteady aerodynamics

An extensive numerical investigation is undertaken to ascertain the feasibility of nonlinear and unsteady aerodynamic models of the form of Eq. (2.20). Also, we aim to validate the Bayesian model comparison algorithm detailed in Sec. 2.1 for stochastic nonlinear ODE based models, and noisy and incomplete measurements.

2.3.2.1 Synthetic aeroelastic oscillator

A synthetic aeroelastic oscillator is constructed using the coupled ODEs (in terms of the dimensionless time τ)

$$\ddot{\theta} = c_1 \text{sign}(\dot{\theta}) + c_2 \theta + c_3 C_M + c_4 \dot{\theta} + c_5 \theta^3, \quad (2.22a)$$

$$\frac{\dot{C}_M}{B} + C_M = e_1 \theta + e_2 \dot{\theta} + e_3 \theta^3 + e_4 \theta^2 \dot{\theta} + \frac{c_6}{B} \ddot{\theta} + \sigma \xi(\tau). \quad (2.22b)$$

The ODE in Eq. (2.22a) is rewritten from Eq. (2.9) and it represents the structural equation of motion for the freely pitching wing located in the wind-tunnel. The ODE in Eq. (2.22b) is rewritten from Eq. (2.20) and it models the nonlinear and unsteady aerodynamic loading (C_M) acting on the wing.

A synthetic LCO response, resembling the wind-tunnel LCO, is generated using the following values in Eq. (2.22b): $B = 0.2$, $e_1 = -1.25$, $e_2 = -1.0$, $e_3 = 100.0$, $e_4 = -500.0$ and $\sigma = 0.002$. The parameters of Eq. (2.22a) are provided in Sec. 2.2.1, and parameter c_6 is known from Eq. (2.18). Next, the coupled ODE system in Eq. (2.22) is converted into a state-space form using the state vector $\mathbf{u} = \{\theta, \dot{\theta}, C_M\}$. The continuous-time stochastic ODEs in Eq. (2.22) are converted into the discrete-time state-space form of Eq. (2.1) using the Euler-Maruyama discretization scheme [120]. The forward time integration of the discretized system is initiated from $\{\theta, \dot{\theta}, C_M\} = \{0, 0, 0\}$, resembling the wind-tunnel conditions.

Fig. 2.9a shows the simulated (true) pitch response and its PSD. The dynamic instability is induced by the synthetic turbulence introduced by the modelling error term $\sigma \xi(\tau)$ in Eq. (2.22b).

The dynamic instability is triggered at different time instants due to the varying seed of the random number generator used in generating the white-noise process $\xi(\tau)$ in Eq. (2.22b). Once the wing oscillations are observed, the simulated pitch displacement θ_k is captured with a time interval of $\Delta t = 1\text{e-}03$. Notice the amplitude modulations in the LCO response due to the model error term $\sigma\xi(\tau)$ in Eq. (2.22b).

Fig. 2.9b shows the noisy pitch response obtained by corrupting the simulated response from Fig. 2.9a with iid Gaussian noise having pdf $\mathcal{N}(0, \gamma^2)$. We choose $\gamma = 0.002$ in this study. This additive noise mimics the digital noise induced by the data-acquisition system in wind-tunnel tests. See Poirel *et al.* [85] for a detailed description of the wind-tunnel data-acquisition system.

For inverse modelling purpose, the noisy pitch response is filtered using a low-pass filter with a cut-off frequency of 25 Hz. Fig. 2.9c shows the filtered pitch response and its PSD. The cut-off frequency of 25 Hz ensures that the super-harmonics lower and equal to $7f$ are maintained in the pitch response, where f is the fundamental frequency of the LCO, while the high-frequency measurement noise is removed. As explained later in Sec. 2.4, this filtering was necessary to eliminate the 30 Hz wind-tunnel motor noise from the captured LCO response shown in Fig. 2.3. The filtering of synthetic LCO response is performed to maintain the same inverse modelling setup as is used for modelling the wind-tunnel LCO.

The measurement data $\mathcal{D}$ needed for inversion is obtained by trimming the filtered response, as shown in Fig. 2.9c. The trimming keeps the data size N_d to a minimum, while ensuring that the essential aerodynamics is well represented in $\mathcal{D}$. In this case, $N_d = 5000$.

2.3.2.2 Bayesian inversion results

The Bayesian inversion methodology detailed in Sec. 2.1 is applied to model the LCO aerodynamics using $\mathcal{D}$ shown in Fig. 2.9c. For inversion purpose, the structural equation of motion is considered known from Eq. (2.22a) with known coefficients c_1, c_2, c_3, c_4 , and c_5 . The measurement equation and the measurement noise strength ($\gamma = 0.002$) are also considered known for the sake of inversion. Based on the prior knowledge regarding the underlying LCO, the

(a) True pitch response and its PSD

(b) Noisy pitch response and its PSD

(c) Filtered pitch response and its PSD

Figure 2.9: Synthetic LCO response generated using Eq. (2.22a) and Eq. (2.22b).

following over-parameterized ODE is proposed to model the aerodynamic moment coefficient C_M in Eq. (2.22a):

$$\begin{aligned} \frac{\ddot{C}_M}{B_1 B_2} + \frac{(B_1 + B_2)\dot{C}_M}{B_1 B_2} + C_M = e_1 \theta + e_2 \dot{\theta} + e_3 \theta^3 + e_4 \theta^2 \dot{\theta} + e_5 \theta^5 \\ + \frac{(2c_6 c_7 + 0.5)\ddot{\theta}}{B_1 B_2} + \frac{c_6 \ddot{\theta}}{B_1 B_2} + \sigma \xi(\tau). \end{aligned} \quad (2.23)$$

This ODE is obtained by adding the nonlinear terms θ^3 , $\theta^2 \dot{\theta}$ and θ^5 to the second-order Wagner unsteady ODE from Appendix C.1. The coupled ODE system of Eq. (2.22a) and Eq. (2.23) is hereby referred to as model $\mathcal{M}_G$, with an unknown parameter vector $\phi = \{A_1, B_1, B_2, e_1, e_2, e_3, e_4, e_5, \sigma\}$. Notice that the model error strength σ in Eq. (2.23) is considered an unknown and is estimated during Bayesian inversion. From Appendix C.1, we deduce that $A_2 = 0.5 - A_1$ and $c_7 = (B_1(1 - A_1) + B_2(1 - A_2))$. Hence, there is no need to include A_2 and c_7 in ϕ . Also, from Appendix C.1, parameter $c_6 = 0.5\pi(0.25 - a_h^2)$ depends solely on the wing structural properties and is therefore considered known for the sake of inversion. Note that Wagner unsteady parameters must satisfy the constraints of 1) $A_1 + A_2 = 0.5$ and 2) $B_1, B_2 > 0$.

The goal of this inversion exercise is to investigate the following two aspects of the LCO aerodynamics: 1) nonlinearity (dictated by e_3, e_4 and e_5), and 2) unsteady aerodynamics (dictated by A_1, B_1, B_2). In general, the inversion of nonlinear physics relies on the forcing and/or initial conditions as nonlinearity captured in system response depends on the forcing amplitude. In the case of LCO, the system is self-excited and its amplitude is dictated by the freestream air-speed. Note that the LCO is a phenomenon induced by the nonlinearity (inherently nonlinear) and controlled by the interaction of structural motion and aerodynamic loading.

In order to pursue this goal, a candidate model set $\mathcal{M}$ containing a mix of quasi-steady and unsteady aerodynamic models nested under $\mathcal{M}_G$ is constructed. These nested models are proposed using the following thought process. Based on the prior understanding of LCO aerodynamics, parameters e_1, e_2, e_3 and e_4 are deemed salient for capturing C_M in Eq. (2.23). Therefore, these parameters are considered present in all the candidate nested models, with

their prior pdfs chosen as: $e_1 \sim \mathcal{U}(-2, 0)$, $e_2 \sim \mathcal{U}(-2, 0)$ for unsteady models, $e_2 \sim \mathcal{U}(0, 2)$ for quasi-steady models, $e_3 \sim \mathcal{U}(-250, 250)$ and $e_4 \sim \mathcal{U}(-600, 0)$. Here, $\mathcal{U}(a, b)$ denotes a uniform pdf with support $[a, b]$ and $b > a$.

In contrast to the salient parameters, parameters A_1 , B_1 , B_2 and e_5 in Eq. (2.23) may or may not be considered necessary for capturing C_M . based on our limited prior understanding of the LCO aerodynamics. We hereby refer to these parameters as the *questionable* parameters. Hence, the candidate nested models are constructed by iteratively removing these questionable parameters from $\mathcal{M}_G$. This removal is facilitated by the prior parameter pdf $p(\phi|\mathcal{M}_G)$. Table 2.4 lists the prior pdf of questionable parameters used for constructing a candidate set $\mathcal{M} = \{\mathcal{M}_1^q, \mathcal{M}_2^q, \mathcal{M}_1^u, \mathcal{M}_2^u, \mathcal{M}_3^u, \mathcal{M}_4^u, \mathcal{M}_5^u, \mathcal{M}_6^u\}$, where the superscript ‘ q ’ denotes a quasi-steady model and the superscript ‘ u ’ denoted an unsteady model. In Table 2.4, $\mathcal{L}_n(a, b)$ denotes a log-normal distribution with median at ‘ a ’ and a COV of ‘ b ’. Note that choosing a Dirac-delta prior pdf $\delta(z)$ for any parameter equates to fixing its value at z . For instance, the prior $\delta(0)$ for e_5 in $\mathcal{M}_1^u$ means the term $e_5\theta^5$ is absent from the model. The choice of the log-normal prior pdf $B_1 \sim \mathcal{L}_n(0.2, 0.5)$ in $\mathcal{M}_1^u, \mathcal{M}_2^u$ is motivated by the work of Dowell [110]. The choice of the log-normal prior pdf $B_1 \sim \mathcal{L}_n(0.32, 0.5)$ and $B_2 \sim \mathcal{L}_n(0.04, 0.5)$ in $\mathcal{M}_3^u$ and $\mathcal{M}_4^u$ is adopted from Jones [113].

$\mathcal{M} \in \mathcal{M}$	A_1	B_1	B_2	e_5
$\mathcal{M}_1^q$	$\delta(0)$	$\delta(\infty)$	$\delta(\infty)$	$\delta(0)$
$\mathcal{M}_2^q$	$\delta(0)$	$\delta(\infty)$	$\delta(\infty)$	$\mathcal{U}(-1e+4, 1e+4)$
$\mathcal{M}_1^u$	$\delta(0.5)$	$\mathcal{L}_n(0.2, 0.5)$	$\delta(\infty)$	$\delta(0)$
$\mathcal{M}_2^u$	$\delta(0.5)$	$\mathcal{L}_n(0.2, 0.5)$	$\delta(\infty)$	$\mathcal{U}(-1e+4, 1e+4)$
$\mathcal{M}_3^u$	$\mathcal{U}(0, 0.5)$	$\mathcal{L}_n(0.32, 0.5)$	$\mathcal{L}_n(0.04, 0.5)$	$\delta(0)$
$\mathcal{M}_4^u$	$\mathcal{U}(0, 0.5)$	$\mathcal{L}_n(0.32, 0.5)$	$\mathcal{L}_n(0.04, 0.5)$	$\mathcal{U}(-1e+4, 1e+4)$
$\mathcal{M}_5^u$	$\delta(0.335)$	$\mathcal{L}_n(0.32, 0.5)$	$\mathcal{L}_n(0.04, 0.5)$	$\delta(0)$
$\mathcal{M}_6^u$	$\mathcal{U}(0, 0.5)$	$\mathcal{L}_n(0.2, 0.5)$	$\mathcal{L}_n(0.2, 0.5)$	$\delta(0)$

Table 2.4: Synthetic LCO: Prior parameter pdf for the questionable parameters of candidate nested models.

The ODEs pertaining to each of the nested aerodynamic model defined in Table 2.4 can be

written as

$$\mathcal{M}_1^q : C_M = e_1\theta + e_2\dot{\theta} + e_3\theta^3 + e_4\theta^2\dot{\theta} + \sigma\xi(\tau) \quad (2.24a)$$

$$\mathcal{M}_2^q : C_M = e_1\theta + e_2\dot{\theta} + e_3\theta^3 + e_4\theta^2\dot{\theta} + e_5\theta^5 + \sigma\xi(\tau) \quad (2.24b)$$

$$\mathcal{M}_1^u : \frac{\dot{C}_M}{B} + C_M = e_1\theta + e_2\dot{\theta} + e_3\theta^3 + e_4\theta^2\dot{\theta} + \frac{c_6}{B}\ddot{\theta} + \sigma\xi(\tau) \quad (2.24c)$$

$$\mathcal{M}_2^u : \frac{\dot{C}_M}{B} + C_M = e_1\theta + e_2\dot{\theta} + e_3\theta^3 + e_4\theta^2\dot{\theta} + e_5\theta^5 + \frac{c_6}{B}\ddot{\theta} + \sigma\xi(\tau) \quad (2.24d)$$

$$\begin{aligned} \mathcal{M}_3^u, \mathcal{M}_5^u, \mathcal{M}_6^u : & \frac{\ddot{C}_M}{B_1B_2} + \frac{(B_1+B_2)\dot{C}_M}{B_1B_2} + C_M = e_1\theta + e_2\dot{\theta} + e_3\theta^3 + e_4\theta^2\dot{\theta} \\ & + \frac{(2c_6c_7+0.5)\ddot{\theta}}{B_1B_2} + \frac{c_6\ddot{\theta}}{B_1B_2} + \sigma\xi(\tau) \end{aligned} \quad (2.24e)$$

$$\begin{aligned} \mathcal{M}_4^u : & \frac{\ddot{C}_M}{B_1B_2} + \frac{(B_1+B_2)\dot{C}_M}{B_1B_2} + C_M = e_1\theta + e_2\dot{\theta} + e_3\theta^3 + e_4\theta^2\dot{\theta} + e_5\theta^5 \\ & + \frac{(2c_6c_7+0.5)\ddot{\theta}}{B_1B_2} + \frac{c_6\ddot{\theta}}{B_1B_2} + \sigma\xi(\tau). \end{aligned} \quad (2.24f)$$

Note that this candidate set is by no means exhaustive as it considers only eight of the 26 ($\sum_{i=0}^{i=4} \binom{4}{i}$) models nested under $\mathcal{M}_G$. Hence, the performance of Bayesian model comparison is conditioned on this candidate set. Also, models $\mathcal{M}_3^u$, $\mathcal{M}_5^u$ and $\mathcal{M}_6^u$ have the same ODE form and are differentiated only in terms of the prior pdf assigned to the questionable parameters (as shown in Table 2.4). Hence, in a Bayesian setting, a model is defined by both the model/measurement equation and the prior parameter pdf. Changing any one of these produces a different model with a different calibration and prediction characteristics.

Models $\mathcal{M}_1^q$ and $\mathcal{M}_2^q$ are both quasi-steady models but differ in terms of the degree of aerodynamic nonlinearity. Models $\mathcal{M}_1^u$ and $\mathcal{M}_2^u$ are the first-order unsteady counterparts of $\mathcal{M}_1^q$ and $\mathcal{M}_2^q$, while $\mathcal{M}_3^u$ and $\mathcal{M}_4^u$ are the corresponding second-order unsteady models. Model $\mathcal{M}_5^u$ is an extension of $\mathcal{M}_3^u$ where $A_1 = 0.335$ and $A_2 = 0.115$ are fixed parameters, based on the work by Jones [113]. Model $\mathcal{M}_6^u$ is also an extension of $\mathcal{M}_3^u$ where the prior pdfs for B_1 and B_2 are unbiased, i.e. $B_1 \sim \mathcal{L}_n(0.2, 0.5)$ and $B_2 \sim \mathcal{L}_n(0.2, 0.5)$. Note that the parametric structure of the model $\mathcal{M}_1^u$ in Eq. (2.24c) is the same as the data-generating model in Eq. (2.22b). Hence,

we should expect it to be selected as the most probable model if the data is dense (informative) and the priors include the appropriate parameter ranges.

Next, we execute the Bayesian inversion of LCO aerodynamics using $\mathcal{D}$ from Fig. 2.9c and the candidate model set $\mathcal{M}$ defined in Eq. (2.24). Given the prior pdfs in Table 2.4, we sample the parameter posterior pdf $p(\phi|\mathcal{D}, \mathcal{M})$ for all $\mathcal{M} \in \mathcal{M}$ using the MCMC algorithm detailed in Sec. 2.1.2. The state estimation task required to compute the likelihood function is executed using EKF. See Appendix B.4 for mathematical details regarding EKF. Following the MCMC sampling of the posterior parameter pdf, the statistical entities pertaining to Bayesian model comparison (evidence, EIG, GOF) are estimated using the Chib-Jeliazkov method detailed in Sec. 2.1.3. See Eq. (2.7) for the mathematical relation between log-evidence, GOF (data-fit) and EIG (model complexity). Table 2.5 list the estimated value of these entities for all $\mathcal{M} \in \mathcal{M}$.

$\mathcal{M} \in \mathcal{M}$	GOF Data-fit (27700+)	EIG Model complexity	$\ln p(\mathcal{D} \mathcal{M})$ Log-evidence (27700+)	$P(\mathcal{M} \mathcal{D}, \mathcal{M})$ Posterior prob.
$\mathcal{M}_1^q$	35.7	19.3	16.4	0
$\mathcal{M}_2^q$	36.2	20.5	15.7	0
$\mathcal{M}_1^u$	132.7	19.1	113.6	0.845
$\mathcal{M}_2^u$	132.5	20.5	112.0	0.155
$\mathcal{M}_3^u$	123.8	44.7	79.1	0
$\mathcal{M}_4^u$	123.7	45.6	78.1	0
$\mathcal{M}_5^u$	123.9	54.6	69.3	0
$\mathcal{M}_6^u$	121.3	40.8	80.5	0

Table 2.5: Bayesian model comparison results obtained using the synthetic LCO measurements. Note that Log-evidence is equal to GOF minus EIG.

The results shown in Table 2.5 reinforce the argument that complex models have higher information gain (or EIG) as compared to the simpler models. For instance, the second-order unsteady models $\mathcal{M}_3^u$, $\mathcal{M}_4^u$, $\mathcal{M}_5^u$ and $\mathcal{M}_6^u$ have a comparatively higher EIG than the first-order models $\mathcal{M}_1^u$ and $\mathcal{M}_2^u$. The higher information gain is due to the fact that complex models try to model the measurement noise and model error (e.g. amplitude modulations) in $\mathcal{D}$ as a part of the LCO physics. Although the information gain for quasi-steady models $\mathcal{M}_1^q$ and $\mathcal{M}_2^q$ is comparable with those from their unsteady counterparts $\mathcal{M}_1^u$ and $\mathcal{M}_2^u$, their lower goodness-of-fit ensures their lower log-evidence value. On the other hand, a higher goodness-of-fit overcomes

the information gain term for unsteady models $\mathcal{M}_1^u$ and $\mathcal{M}_2^u$ resulting in a higher evidence. Between $\mathcal{M}_1^u$ and $\mathcal{M}_2^u$, a relatively complex model $\mathcal{M}_2^u$ has a higher goodness-of-fit but the increase in its information gain leads to its lower posterior probability than that of $\mathcal{M}_1^u$.

As expected, the most probable model is $\mathcal{M}_1^u$ having the maximum posterior model probability of 0.845. Model $\mathcal{M}_1^u$ has the same parametric structure as the data-generating model in Eq. (2.22b). The marginal and joint posterior pdfs along with the MAP estimates for $\mathcal{M}_1^u$ are shown in Fig. 2.10. As evident from Fig. 2.10, the MAP estimates are close to the true parameter values used in generating the synthetic LCO data. Fig. 2.11 shows the comparison of the predicted and the true Wagner function along with the credible intervals computed using the posterior MCMC samples of B_1 for model $\mathcal{M}_1^u$. In Fig. 2.11, the Wagner function is predicted with a reasonable accuracy using $\mathcal{M}_1^u$. On the other hand, the predictions made using the over-parameterized model $\mathcal{M}_3^u$ shown in Fig. 2.12 have high uncertainty. The model comparison step therefore helps avoid choosing an over-parametrized or a simple model by promoting a trade-off between data-fit and model simplicity. In addition, these results validate the capability of Bayesian parameter estimation and model comparison in identifying the true nonlinear and unsteady LCO aerodynamics while dealing with nonlinear coupled ODE style models and noisy measurements.

Note that probabilistic predictions can also be made using two or more models with Bayesian model averaging [39, 86]. This might be desirable when the true model structure is unknown (as always in the case of experimental data) and multiple candidate models achieve a non-zero posterior probability. In this study, both model $\mathcal{M}_1^u$ and $\mathcal{M}_2^u$ could be used to obtain predictions using Bayesian model averaging, wherein the predictions from each model is weighted according to its posterior model probability listed in Table 2.5.

Figure 2.10: Marginal and joint posterior pdfs for the optimal model $\mathcal{M}_1^u$. The MAP estimates are shown in red-colored dotted lines. The ρ value shown in joint pdfs is the Pearson correlation coefficient.

Figure 2.11: Predicted Wagner function using the optimal model $\mathcal{M}_1^u$.

Figure 2.12: Predicted Wagner function using the second-order model $\mathcal{M}_3^u$.

2.4 Bayesian model comparison using wind-tunnel data

In this section, we present the Bayesian model comparison results obtained using the wind-tunnel LCO measurements. Fig. 2.13 shows the measured wind-tunnel LCO response for four values of Reynolds number. The Bayesian inversion process followed in this section is the same as in Sec. 2.3.2.

For the sake of inversion, the measurements in Fig. 2.13 are first filtered using a low-pass filter with a cut-off frequency of 25 Hz. The cut-off frequency is chosen such that the frequencies up to $5f$ (f being the fundamental frequency of LCOs) are retained in the filtered data for all the Re cases considered here. The filtered response is trimmed to remove the initial low amplitude response because it has a low SNR and it contains little information about the nonlinear aerodynamics. Note that a low-amplitude (around 0.8°) LCO is also visible in Fig. 2.13 for the cases of $\text{Re} = 75,000$ and $\text{Re} = 105,000$. We neglect this apparent low-amplitude LCO around 1° up to the point where the pitch oscillations starts to diverge with a significant rate. The filtered and trimmed LCO response (denoted as $\mathcal{D}$) is shown in Fig. 2.14, to be used for Bayesian model comparison. For consistency, the size N_d of measurements $\mathcal{D}$ is kept constant at $N_d = 5000$ for all the Re cases.

The candidate model set $\mathcal{M}$ from Sec. 2.3.2 is considered to model the aerodynamic loading (C_M) of the wind-tunnel oscillator. Eq. (2.24) list the ODEs proposed to model C_M for each

Figure 2.13: Measured wind-tunnel LCO response. Re denotes the chord-wise Reynolds number, and k denotes the reduced frequency.

candidate model. The prior pdfs for e_1, e_2, e_3 and e_4 are same for all $\mathcal{M} \in \mathcal{M}$ and Re cases, and are chosen as: $e_1 \sim \mathcal{U}(-2.0, 0.0)$, $e_2 \sim \mathcal{U}(-2.0, 2.0)$ for unsteady models, $e_2 \sim \mathcal{U}(0.0, 2.0)$ for quasi-steady models, $e_3 \sim \mathcal{U}(-500.0, 500.0)$ and $e_4 \sim \mathcal{U}(-1000.0, 0.0)$. The prior distribution assigned to the remaining (questionable) model parameters is listed in Table 2.6.

Tables 2.7 to 2.10 show the Bayesian model comparison results for the four Re cases. The model selected is $\mathcal{M}_2^u$ for all the Re cases. The over-parameterized models like $\mathcal{M}_3^u$, $\mathcal{M}_4^u$, $\mathcal{M}_5^u$ and $\mathcal{M}_6^u$ have a much higher information gain that penalizes the log-evidence and the posterior model probability. On the contrary, the simpler models like the quasi-steady models $\mathcal{M}_1^q$ and $\mathcal{M}_2^q$ and the unsteady model $\mathcal{M}_1^u$ have a comparable information gain but a lower data-fit, which ends up driving their posterior probability to zero. The posterior parameter pdf was also

Figure 2.14: Filtered and trimmed wind-tunnel LCO response.

$\phi(\text{true})$	A_1	B_1	B_2	e_5
$\mathcal{M}_1^q$	$\delta(0)$	$\delta(\infty)$	$\delta(\infty)$	$\delta(0)$
$\mathcal{M}_2^q$	$\delta(0)$	$\delta(\infty)$	$\delta(\infty)$	$\mathcal{U}(-1.0\text{e}+5, 1.0\text{e}+5)$
$\mathcal{M}_1^u$	0.5	$\mathcal{L}_n(0.2, 0.5)$	$\delta(\infty)$	$\delta(0)$
$\mathcal{M}_2^u$	0.5	$\mathcal{L}_n(0.2, 0.5)$	$\delta(\infty)$	$\mathcal{U}(-1\text{e}+5, 1\text{e}+5)$
$\mathcal{M}_3^u$	$\mathcal{U}(0, 0.5)$	$\mathcal{L}_n(0.32, 0.5)$	$\mathcal{L}_n(0.04, 0.5)$	$\delta(0)$
$\mathcal{M}_4^u$	$\mathcal{U}(0, 0.5)$	$\mathcal{L}_n(0.32, 0.5)$	$\mathcal{L}_n(0.04, 0.5)$	$\mathcal{U}(-1\text{e}+5, 1\text{e}+5)$
$\mathcal{M}_5^u$	0.335	$\mathcal{L}_n(0.32, 0.5)$	$\mathcal{L}_n(0.04, 0.5)$	$\delta(0)$
$\mathcal{M}_6^u$	$\mathcal{U}(0, 0.5)$	$\mathcal{L}_n(0.2, 0.5)$	$\mathcal{L}_n(0.2, 0.5)$	$\delta(0)$

Table 2.6: Wind-tunnel LCO: Prior parameter pdf for the questionable parameters of candidate nested models.

computed for the optimal model $\mathcal{M}_2^u$ for all Re cases.

The high posterior probability of model $\mathcal{M}_2^u$ highlights the importance of unsteady aerody-

$\mathcal{M} \in \mathcal{M}$	GOF Data-fit (27700+)	EIG Model complexity	$\ln p(\mathcal{D} \mathcal{M})$ Log-evidence (27700+)	$P(\mathcal{M} \mathcal{D}, \mathcal{M})$ Posterior prob.
$\mathcal{M}_1^q$	-10.2	20.2	-30.4	0
$\mathcal{M}_2^q$	-8.8	22.7	-31.5	0
$\mathcal{M}_1^u$	-1.1	22.0	-23.1	0.029
$\mathcal{M}_2^u$	4.6	23.9	-19.3	0.968
$\mathcal{M}_3^u$	1.5	37.9	-36.4	0
$\mathcal{M}_4^u$	5.8	41.0	-35.2	0
$\mathcal{M}_5^u$	1.5	37.9	-36.4	0
$\mathcal{M}_6^u$	2.4	27.7	-25.3	0.002

Table 2.7: Bayesian model comparison results obtained using the wind-tunnel LCO measurements at $Re = 75,000$.

$\mathcal{M} \in \mathcal{M}$	GOF Data-fit (27700+)	EIG Model complexity	$\ln p(\mathcal{D} \mathcal{M})$ Log-evidence (27700+)	$P(\mathcal{M} \mathcal{D}, \mathcal{M})$ Posterior prob.
$\mathcal{M}_1^q$	-12.5	20.4	-32.9	0
$\mathcal{M}_2^q$	-12.7	23.1	-35.8	0
$\mathcal{M}_1^u$	34.9	21.7	13.2	0.004
$\mathcal{M}_2^u$	42.5	23.9	18.6	0.805
$\mathcal{M}_3^u$	43.2	37.1	6.1	0.
$\mathcal{M}_4^u$	54.9	40.1	14.8	0.018
$\mathcal{M}_5^u$	43.3	37.1	6.2	0
$\mathcal{M}_6^u$	44.1	27.0	17.1	0.174

Table 2.8: Bayesian model comparison results obtained using the wind-tunnel LCO measurements at $Re = 85,000$.

$\mathcal{M} \in \mathcal{M}$	GOF Data-fit (27600+)	EIG Model complexity	$\ln p(\mathcal{D} \mathcal{M})$ Log-evidence (27600+)	$P(\mathcal{M} \mathcal{D}, \mathcal{M})$ Posterior prob.
$\mathcal{M}_1^q$	45.9	20.5	25.4	0
$\mathcal{M}_2^q$	47.1	23.1	24.0	0
$\mathcal{M}_1^u$	66.9	21.7	45.2	0.022
$\mathcal{M}_2^u$	72.9	23.8	49.1	0.966
$\mathcal{M}_3^u$	70.9	37.4	33.5	0
$\mathcal{M}_4^u$	82.6	40.6	42.0	0.001
$\mathcal{M}_5^u$	71.0	37.5	33.5	0
$\mathcal{M}_6^u$	71.7	27.2	44.5	0.011

Table 2.9: Bayesian model comparison results obtained using the wind-tunnel LCO measurements at $Re = 90,000$.

namics in the wind-tunnel LCO. The posterior statistics of B is listed in Table 2.11 and plotted in Fig. 2.15 for varying RE value. We realize that the MAP estimate of B is of the same order of

$\mathcal{M} \in \mathcal{M}$	GOF Data-fit (27500+)	EIG Model complexity	$\ln p(\mathcal{D} \mathcal{M})$ Log-evidence (27500+)	$P(\mathcal{M} \mathcal{D}, \mathcal{M})$ Posterior prob.
$\mathcal{M}_1^q$	30.0	20.3	9.7	0.003
$\mathcal{M}_2^q$	35.7	22.9	12.8	0.083
$\mathcal{M}_1^u$	32.2	31.8	0.4	0
$\mathcal{M}_2^u$	48.0	32.7	15.3	0.914
$\mathcal{M}_3^u$	29.0	62.1	-33.1	0
$\mathcal{M}_4^u$	47.6	64.3	-16.7	0
$\mathcal{M}_5^u$	29.1	75.2	-46.1	0
$\mathcal{M}_6^u$	30.6	46.7	-16.1	0

Table 2.10: Bayesian model comparison results obtained using the wind-tunnel LCO measurements at $Re = 105,000$.

magnitude as Dowell's one-state Wagner's model for all RE cases except for $Re=105,000$ [110]. This is consistent with previous observations and analysis which compared the LCO fundamental frequency with the calculated (linear) aeroelastic natural frequency modelled with Wagner's unsteady aerodynamics [85]. It was observed that the LCO and natural frequencies matched for Re values up to approximately 85,000 and started to depart significantly for larger Reynolds numbers, thus indicating two regimes of oscillations. Table 2.11 and Fig. 2.15 indicate a marked change in the posterior statistics of B for Re value higher than 90,000. This behaviour of B implies that the unsteady effects are less important for $Re=105,000$. Moreover, Table 2.10 shows that $Re=105,000$ is the only case where the two quasi-steady models $\mathcal{M}_1^q$ and $\mathcal{M}_2^q$ have a non-zero posterior probability.

Re	MAP estimate of B from $\mathcal{M}_2^u$	LCO frequency (Hz)	LCO amplitude (deg)	Reduced frequency
75,000	0.345	3.02	5.05	0.1974
85,000	0.230	3.22	5.27	0.1847
90,000	0.258	3.35	4.90	0.1811
105,000	1.252	3.95	4.13	0.1837

Table 2.11: Posterior estimates for Wagner constant B for varying Reynolds number.

Note that since the data-generating model is unknown in this case, validating the optimal model is a cumbersome yet crucial task. One way to conduct such validation is to perform

Figure 2.15: Reynolds number versus the posterior pdf of B for model $\mathcal{M}_2^u$.

additional wind-tunnel experiments, as reported in Khalil *et al.* [9]. If experiments can be repeated numerous times, techniques such as cross-validation (CV) can be employed to ascertain the generalizability of the chosen model [7]. These techniques demand numerous solves of the inverse problem. When dealing with real-life engineering systems such as the wind-tunnel LCO, the measurements are limited and the inversion process is computationally challenging. Due to these reasons, the application of CV is not pursued in this thesis.

In this case, we attempt to ascertain the predictive capability of the optimal model $\mathcal{M}_2^u$ using the published literature. Nonlinear effects are best captured by plotting the aerodynamic moment coefficient in static conditions, as shown in Fig. 2.16. This is done by retaining only the static terms of the chosen model and using the MAP estimates for the e_i parameters. Note that the range of angles shown is limited to the LCO amplitude for each Re, as the reliability of the model decreases significantly outside this range. Also plotted is the 2D thin-airfoil theory case (about the same axis at 0.18c), which is equivalent to an infinitely large Re. In addition, results from a static test performed by Huang *et al.* [100] are also shown. A detailed comparison between these results and ours is not possible due to the differences in test conditions. Huang's results are for a cantilevered NACA0012, thus having 3D flow effects, and the moment is taken about a different axis (at 0.25c). However, both sets of results are very

similar in shape and magnitude, thus providing added confidence in the optimal model obtained using Bayesian model comparison.

Figure 2.16: Static aerodynamic moment coefficient versus angle-of-attack for different Reynolds numbers for model $\mathcal{M}_2^u$

Chapter 3

Bayesian model reduction using automatic relevance determination

In this chapter, we detail a Bayesian model reduction (sparse learning) algorithm that exploits the concept of ARD to identify an optimal model nested under an over-parameterized physical model. This model reduction approach offers two advantages over the Bayesian model comparison algorithm detailed in Chapter 2 when dealing with nested models. First, this approach alleviates the need to assign prior pdf to a parameter with limited prior knowledge. Second, ARD facilitates implicit comparison of all nested models, thereby eliminating the need to propose a finite number of candidate nested models. This model reduction approach is powered by a parallel and adaptive MCMC sampler and is validated using a synthetically generated nonlinear oscillator similar to the wind-tunnel oscillator reported in Sec. 2.2.1.

The chapter is structured as follows. In Sec. 3.1, we present the mathematical details behind the ARD-based Bayesian inversion algorithm. In Sec. 3.2, the Bayesian algorithm is validated using synthetic measurements from a nonlinear aeroelastic oscillator, similar to the wind-tunnel oscillator. We report several studies involving unidimensional and two-dimensional evidence to highlight the sparse learning capabilities of the ARD algorithm in an engineering setting.

3.1 Bayesian model reduction methodology

Consider that a mathematical model $\mathcal{M}$ with an unknown parameter vector $\phi \in \mathbb{R}^{N_\phi}$ is proposed to model a physical system using noisy and incomplete measurements $\mathcal{D}$. The model $\mathcal{M}$ is considered to be over-parameterized, meaning ϕ contains many redundant parameters. The mathematical notations used here were previously defined in Sec. 2.1. Next, we detail the ARD setup proposed to identify redundancies in ϕ .

3.1.1 Automatic relevance determination

The concept of ARD was first presented in the machine learning literature as an effective tool for feature selection among regression and classification models [40, 53, 121, 122]. ARD operates by pruning redundant features or terms in an over-parametrized model through a data-dependent regularization of the solution space. The regularization is achieved by assigning a parametrized prior distribution to model parameters. The unknown parameters of ARD priors are then estimated using the measurement data [122]. This approach of assigning prior pdf using the data is also known as the parametric empirical Bayes (EB) method [40, 123, 124]. The EB or ARD approach is in contrast to the standard Bayesian statistics where known prior distributions have to be assigned by the modeller, and also in contrast to the frequentist statistics where no prior is assigned.

In machine learning applications, the ARD prior is assigned to all model parameters and the subsequent inferences are entirely data-dependent and are free from prior beliefs (if any). As demonstrated in Appendix A.3, the Bayesian algorithms of SBL and BCS follow this approach. This modelling approach is not well-suited for engineering systems where some *a priori* information about system physics is always available. The prior parameter knowledge often helps regularize the posterior parameter pdf and, therefore, should be incorporated during Bayesian model reduction. In this work, we implement a hybrid approach of assigning ARD priors that allows automatic pruning of redundant parameters while permitting the inclusion of

prior information about certain model parameters. This idea is implemented by decomposing $\phi = \{\phi_\alpha, \phi_{-\alpha}\}$ (in that order), where $\phi_\alpha \in \mathbb{R}^{N_\alpha}$ contains parameters with no prior knowledge (and hence questionable), and $\phi_{-\alpha} \in \mathbb{R}^{N_\phi - N_\alpha}$ contains parameters with a known prior pdf $p(\phi_{-\alpha})$. We hereby refer ϕ_α as the questionable parameter vector, and $\phi_{-\alpha}$ as the *a priori* relevant parameter vector.

Next, sparsity in the questionable parameter vector ϕ_α is identified by assigning each $\phi_i \in \phi_\alpha$ an ARD prior. In general, any unbounded, symmetric probability distribution that can asymptotically approach a Dirac-delta function can be chosen as an ARD prior. In this chapter, we consider the following two choices: 1) Gaussian ARD prior $\mathcal{N}(\phi|0, \alpha^{-1})$, and 2) Laplace ARD prior $\mathcal{L}_p(\phi|0, \alpha^{-1})$, where α is the unknown prior precision (hyperparameter). Fig. 3.1 contrasts the Laplace and Gaussian pdf having a precision of one. In contrast to Gaussian pdf, Laplace pdf has thicker tails and the majority of the probability mass is concentrated close to zero. In Sec. 3.2.1, we investigate how this difference in Gaussian and Laplace pdfs affects their sparsity inducing capabilities for nonlinear ODEs. As demonstrated in Appendix A.5, the marginal prior pdf $p(\phi)$ obtained by marginalizing $p(\phi, \alpha)$ is a Student's t-distribution, which is similar to the Laplace distribution as it is peaked sharply at the mode (zero in this case).

Figure 3.1: Laplace pdf versus Gaussian pdf, both having a precision of one.

Assuming prior independence among ϕ_α and $\phi_{-\alpha}$, the joint prior pdf for ϕ can be written as $p(\phi|\alpha) = p(\phi_{-\alpha})p(\phi_\alpha|\alpha)$, where $\alpha \in \mathbb{R}^{N_\alpha}$ is the unknown hyperparameter vector. The known

prior pdf $p(\phi_{-\alpha})$ of $\phi_{-\alpha}$ can take any pdf form, and is constructed using the prior parameter knowledge. The joint ARD prior $p(\phi_{\alpha}|\alpha)$ is chosen as the product of individual ARD priors as

$$\text{Gaussian ARD: } p(\phi_{\alpha}|\alpha) = \prod_{\phi_i \in \phi_{\alpha}} \mathcal{N}(\phi_i|0, \alpha_i^{-1}) = \prod_{\phi_i \in \phi_{\alpha}} \sqrt{\frac{\alpha_i}{2\pi}} \exp\left\{-\frac{\alpha_i \phi_i^2}{2}\right\}, \quad (3.1a)$$

$$\text{Laplace ARD: } p(\phi_{\alpha}|\alpha) = \prod_{\phi_i \in \phi_{\alpha}} \mathcal{L}_p(\phi_i|0, \alpha_i^{-1}) = \prod_{\phi_i \in \phi_{\alpha}} \sqrt{\frac{\alpha_i}{2}} \exp\left\{-\sqrt{2\alpha_i}|\phi_i|\right\}. \quad (3.1b)$$

Note that the use of multivariate ARD priors (i.e. non-diagonal covariance matrix) can also be considered for revealing additional information regarding the interplay between questionable parameters [125]. However, this multivariate ARD prior contains additional hyperparameters in the form of non-diagonal terms of the covariance matrix, thereby increasing the dimensionality of evidence with respect to hyperparameters. In the end, the benefit from incorporating additional hyperparameters has to be weighed against the computational hardship faced in maximizing a high-dimensional evidence integral. In this thesis, we consider only the independent ARD priors. The use of multidimensional ARD prior in the engineering context will be explored in future research.

In this thesis, we assume that fixing a model parameter at zero eliminates its contribution from the model. This is the reason ARD priors in Eq. (3.1) are zero-mean. An ARD prior with a non-zero mean can be exploited if the parameter needs to be fixed at a nonzero value during sparse learning. This can be easily pursued within the current Bayesian apparatus by modifying Eq. (3.1). However, such a modification is not required for the physical models considered in this thesis. In addition, we will demonstrate through numerical investigations that a relevant parameter with a non-zero true value can be reasonably estimated (i.e. low bias) using a zero-mean prior.

In Eq. (3.1), the hyperparameter α_i controls the contribution of parameter ϕ_i in the model since setting α_i to infinity would force both the prior and posterior pdf of ϕ_i to be a Dirac-delta function centered at zero, i.e. $\mathcal{N}(\phi_i|0, 0)$ or $\mathcal{L}_p(\phi_i|0, 0)$. Therefore, α controls the complexity

of the model. This property of ARD priors implies that, for a given prior $p(\phi_{-\alpha})$, all models nested under the proposed model $\mathcal{M}$ can be obtained by varying α_i values in Eq. (3.1). Therefore, identifying an optimal α equates to selecting an optimal nested model under $\mathcal{M}$. Stated differently, the ARD-based model reduction approach enables implicit comparison of all nested models by transforming the model comparison problem from a discrete model domain (as in the candidate model set $\mathcal{M}$ in Sec. 2.3.2) into a continuous hyperparameter (α) domain.

Given the prior pdf in Eq. (3.1), the posterior parameter pdf from Eq. (2.3) can be rewritten as

$$p(\phi|\mathcal{D}, \alpha) = \frac{p(\mathcal{D}|\phi, \alpha)p(\phi|\alpha)}{p(\mathcal{D}|\alpha)} \propto p(\mathcal{D}|\phi)p(\phi_{-\alpha})p(\phi_{\alpha}|\alpha) \quad (3.2)$$

where both the posterior pdf $p(\phi|\mathcal{D}, \alpha)$ and the model evidence $p(\mathcal{D}|\alpha)$ are now a function of α for a given $\mathcal{D}$. Next, we detail the algorithm for computing an optimal α using $\mathcal{D}$.

3.1.2 Hyperparameter estimation algorithm

The goal of Bayesian model reduction (or sparse learning) is to find the optimal nested model under $\mathcal{M}$ by pruning redundant parameters from ϕ . In the current Bayesian setup, the hyperparameter vector α dictates the relevance of questionable parameters ϕ_{α} . The model reduction problem thus reduces to estimating α using the measurements $\mathcal{D}$. Using the hierarchical Bayes approach, the hyperparameter posterior pdf $p(\alpha|\mathcal{D})$ is obtained as [7]

$$p(\alpha|\mathcal{D}) = \frac{p(\mathcal{D}|\alpha)p(\alpha)}{p(\mathcal{D})} \propto p(\mathcal{D}|\alpha)p(\alpha), \quad (3.3)$$

where $p(\mathcal{D}|\alpha)$ is the model evidence from Eq. (3.2), $p(\alpha)$ is the hyperprior (prior for hyperparameter α), and $p(\mathcal{D})$ is just a normalization constant for a fixed $\mathcal{D}$. The evidence integral from Eq. (2.6) can be rewritten for the ARD setup as

$$p(\mathcal{D}|\alpha) = \int p(\mathcal{D}|\phi)p(\phi|\alpha)d\phi. \quad (3.4)$$

Note that $p(\mathcal{D}|\alpha)$ is also referred to as type-II likelihood, whereas the type-I likelihood is $p(\mathcal{D}|\phi)$. The type-I likelihood is a function of parameters and the type-II likelihood is a function of hyperparameters. As discussed in Sec. 2.1.3, the integral in Eq. (3.4) is analytically intractable for a general nonlinear dynamical system, and is typically computed using stochastic sampling techniques such as MCMC.

Our interest in the sparsity of ϕ_α requires us to obtain an optimal α value and not the entire posterior distribution $p(\alpha|\mathcal{D})$. In a Bayesian setting, the MAP estimate α^{MAP} provides such an optimal choice. Employing the empirical Bayes approach, α^{MAP} is computed by maximizing $p(\alpha|\mathcal{D})$ as

$$\alpha^{\text{MAP}} = \arg \max_{\alpha} \{p(\alpha|\mathcal{D})\} = \arg \max_{\alpha} \{p(\mathcal{D}|\alpha)p(\alpha)\}. \quad (3.5)$$

For a sparse ϕ_α , many α_i^{MAP} will approach infinity, thereby forcing both the prior and the posterior pdf of ϕ_i to be a Dirac-delta function centered at zero. A finite α_i^{MAP} will imply a relevant parameter $\phi_i \in \phi_\alpha$.

Similar to SBL [78], the hyperparameters in α are assumed to be *a priori* independent and the marginal hyperprior pdf is chosen as $p(\log \alpha_i) \propto 1$. This flat hyperprior (in logarithm space) is also known as the Jeffrey's prior [126]. See Sec. 4.2.3 for additional details on choosing hyperpriors. For the current choice of $p(\log \alpha_i) \propto 1$, Eq. (3.5) reduces to an evidence maximization problem. This approach of identifying the optimal nested model by maximizing model evidence is also known as the type-II maximum likelihood (type-II ML) [7]. Table 3.1 highlights the contrast among various Bayesian estimation techniques. Notice that a full Bayes approach would involve estimating the joint posterior pdf $p(\phi, \psi|\mathcal{D})$ of parameters and hyperparameters. This pdf can provide additional information regarding the interplay between relevant and irrelevant parameters, but is insignificant for sparsity identification purposes. In other words, the posterior uncertainty in hyperparameters is not generally needed when making a binary decision whether a parameter is relevant or not. In general, the MAP estimate of hyperparameters is sufficient (the full pdf is not needed) to make such a binary decision.

The resulting Bayesian model reduction algorithm is summarized in Algorithm 1. Given

method	Definition
MLE or type-I ML	$\phi^{\text{MLE}} = \arg \max_{\phi} p(\mathcal{D} \phi)$
MAP	$\phi^{\text{MAP}} = \arg \max_{\phi} p(\phi \mathcal{D}, \alpha) = \arg \max_{\phi} p(\mathcal{D} \phi)p(\phi \alpha)$
Empirical Bayes or type-II ML	$\alpha^{\text{MAP}} = \arg \max_{\alpha} p(\alpha \mathcal{D}) = \arg \max_{\alpha} p(\mathcal{D} \alpha)p(\alpha)$
Full Bayes	$p(\phi, \alpha \mathcal{D}) \propto p(\mathcal{D} \phi)p(\phi \alpha)p(\alpha)$

Table 3.1: Contrast between various Bayesian techniques. Adopted from Murphy [7].

$\mathcal{D}$ and the prior pdf $p(\phi|\alpha)$, the evidence $p(\mathcal{D}|\alpha)$ is maximized using a gradient-free optimizer (line-search) to compute α^{MAP} . For each α^{try} value attempted during maximization, the evidence $p(\mathcal{D}|\alpha^{\text{try}})$ in Eq. (3.4) is computed using the Chib-Jeliazkov method detailed in Appendix B.3. The Chib-Jeliazkov method requires stationary (post burn-in) MCMC samples generated using an MH algorithm from the unnormalized posterior pdf $p(\phi|\mathcal{D}, \alpha^{\text{try}})$ in Eq. (3.2). This MCMC sampling task is executed using the parallel DRAM algorithm [49] detailed in Appendix B.2. The MAP estimate ϕ^{MAP} is kept up-to-date with each MCMC sampling of the posterior $p(\phi|\mathcal{D}, \alpha^{\text{try}})$. Initiating the DRAM algorithm with the latest ϕ^{MAP} further shortens the MCMC burn-in period since $p(\phi|\mathcal{D}, \alpha)$ does not change significantly with a slight change in α . This observation is based on the notion that model evidence is more sensitive to changes in parameter prior pdf than the posterior parameter pdf.

The evidence optimizer is considered converged once the change in $p(\phi|\mathcal{D}, \alpha^{\text{MAP}})$ is below some predefined tolerance. The evidence computed using the Chib-Jeliazkov method possess some sampling variability as it relies on a finite number of stationary MCMC samples. This sampling variability is taken into consideration in determining the convergence of the evidence maximizer. Further, the sampling variability is lowered by making the following two modifications. First, the evidence is computed by trimming the stationary MCMC chain, which then lowers the sample autocorrelation and the variance in evidence estimates. Second, the evidence estimate produced from each parallel chain is averaged to obtain a mean estimate, which is then maximized to obtain the optimal hyperparameters.

The MCMC sampling of the posterior pdf $p(\phi|\mathcal{D}, \alpha^{\text{try}})$ in Eq. (3.2) relies on the availability

Algorithm 1 The ARD-based Bayesian model reduction algorithm

- 1: Gather measurements $\mathcal{D}$ and propose a model with an unknown parameter vector ϕ
 - 2: Decompose $\phi = \{\phi_\alpha, \phi_{-\alpha}\}$ based on prior knowledge
 - 3: Construct a prior parameter pdf $p(\phi|\alpha) = p(\phi_{-\alpha})p(\phi_\alpha|\alpha)$
 - 4: Initiate α^{MAP} and ϕ^{MAP}
 - 5: **while** evidence optimizer not converged **do**
 - 6: Propose α^{try}
 - 7: Initiate parallel DRAM MCMC sampling of the posterior pdf $p(\phi|\mathcal{D}, \alpha^{\text{try}})$ using ϕ^{MAP}
 - 8: Extract stationary (post burn-in) MCMC samples from $p(\phi|\mathcal{D}, \alpha^{\text{try}})$ and update ϕ^{MAP}
 - 9: Compute model evidence $p(\mathcal{D}|\alpha^{\text{try}})$ using the Chib-Jeliazkov method
 - 10: **if** $p(\mathcal{D}|\alpha^{\text{try}}) > p(\mathcal{D}|\alpha^{\text{MAP}})$ **then** Set $\alpha^{\text{MAP}} = \alpha^{\text{try}}$
 - 11: **end if**
 - 12: **end while**
 - 13: Compute the posterior pdf $p(\phi|\mathcal{D}, \alpha^{\text{MAP}})$ using the optimal ARD prior $p(\phi|\alpha^{\text{MAP}})$
-

of the likelihood function $p(\mathcal{D}|\phi)$ for each ϕ value attempted by the sampler. When dealing with stochastic dynamical systems and noisy and incomplete measurements, the likelihood computation requires the estimation of state pdf at each measurement arrival time. See Sec. 2.1.2 for mathematical details behind this likelihood computation approach. For the nonlinear aeroelastic oscillator considered in this work, the state estimation task is executed using EKF. The measurements $\mathcal{D}$ considered in the numerical investigations (and in wind-tunnel) were dense enough to allow EKF to reasonably propagate a Gaussian state pdf through the nonlinear system. See Appendix B.4 for details regarding EKF.

The ARD-based algorithm in Algorithm 1 is implemented in C++ where the Armadillo library [127] is used to perform linear algebra operations while the Message Passing Interface (MPI) [128] library is used for interprocessor communication. The simulations were performed on 16 Intel Xeon E5540 cores provided by the SCINET consortium [129].

Once α^{MAP} is available, the optimal nested model is obtained by computing the posterior pdf $p(\phi|\mathcal{D}, \alpha^{\text{MAP}})$ using the optimal ARD prior $p(\phi|\alpha^{\text{MAP}})$ in Eq. (3.2). The posterior pdf for a redundant parameter is a Dirac-delta function centered at zero. The optimal nested model obtained through evidence maximization possess the best trade-off between model simplicity and data-fit. This property of the evidence is elaborated in Sec. 2.1.3.

3.2 Numerical investigation: Nonlinear aeroelastic oscillator

In this section, we investigate the applicability of the ARD-based Bayesian algorithm as an automatic model reduction tool for stochastic, nonlinear ODE based models. The synthetically-generated nonlinear aeroelastic oscillator from Sec. 2.3.2.1 is considered for this purpose. The oscillator resembles the wind-tunnel system detailed in Sec. 2.2.1. The aeroelastic system consists of a rigid NACA0012 wing undergoing pure-pitch LCO in the transitional Reynolds numbers regime [85].

The coupled ODEs used for the generating the synthetic LCO are rewritten from Eq. (2.22) as (in terms of the dimensionless time τ)

$$\ddot{\theta} = c_1 \text{sign}(\dot{\theta}) + c_2 \theta + c_3 C_M + c_4 \dot{\theta} + c_5 \theta^3, \quad (3.6a)$$

$$\frac{\dot{C}_M}{B} + C_M = e_1 \theta + e_2 \dot{\theta} + e_3 \theta^3 + e_4 \theta^2 \dot{\theta} + \frac{c_6}{B} \ddot{\theta} + \sigma \xi(\tau). \quad (3.6b)$$

The ODE in Eq. (3.6a) represents the structural equation of motion for the freely pitching wing located in the wind-tunnel. The ODE in Eq. (3.6b) models the nonlinear and unsteady aerodynamic loading (C_M) on the wing. The measurements $\mathcal{D}$ generated using these ODEs in Sec. 2.3.2.1 is considered for the inversion purpose. Fig. 3.2 shows the measurements $\mathcal{D}$ and its PSD, re-plotted here from Fig. 2.9c. See Sec. 2.3.2.1 for parameter values and other settings used for generating the measurements $\mathcal{D}$ shown in Fig. 3.2.

3.2.1 Unidimensional model reduction

To begin with, we consider an inverse problem where only one of the unknown model parameters is deemed questionable and is, therefore, assigned an ARD prior. In this case, the hyperparameter vector α is a scalar. This unidimensional setting is considered to reveal the numerical underpinnings of the ARD-based sparse learning, as well as validate the proposed algorithm in a simple plottable setting.

Figure 3.2: Synthetic pitch measurements and its PSD. The data used for ARD computations is shown in red.

3.2.1.1 ARD prior for a parameter present in the data-generating model

The proposed model is chosen to be the data-generating model consisting of the coupled ODE system shown in Eq. (3.6). The parameters of the structural model in Eq. (3.6a) and the parameter c_6 in Eq. (3.6b) are treated as known for the sake of inversion. Therefore, the unknown parameter vector for the proposed model is $\phi = \{B, e_1, e_2, e_3, e_4, \sigma\}$.

For illustration purpose, we assume that sufficient *a priori* information exists about parameters B, e_1, e_2, e_4 , and σ such that an informative prior pdf can be assigned to these parameters. On the contrary, it is assumed that limited information is available about e_3 and, therefore, the relevance of $e_3\theta^3$ term to the LCO aerodynamics is put into question. Note that executing Bayesian model comparison using an arbitrarily chosen diffuse (large support) prior for e_3 can inadvertently alter the model evidence leading to unreliable posterior model probabilities [95]. ARD offers a practical solution in such circumstances by allowing the data to determine the prior pdf and the relevance of questionable parameters. Based on the ARD methodology outlined in Sec. 3.1.1, ϕ is decomposed into the *a priori* relevant parameter vector $\phi_{-\alpha} = \{B, e_1, e_2, e_4, \sigma\}$ and the questionable parameter $\phi_{\alpha} = \{e_3\}$. Table 3.2 summarizes the inverse problem, and lists the known prior pdfs assigned to $\phi_{-\alpha}$. Both Laplace and Gaussian ARD prior are considered for e_3 .

Aerodynamic model	$\frac{\dot{C}_M}{B} + C_M = e_1\theta + e_2\dot{\theta} + e_3\theta^3 + e_4\theta^2\dot{\theta} + \frac{c_6}{B}\ddot{\theta} + \sigma\xi(\tau)$
ϕ decomposition	$\phi_\alpha = \{e_3\}$, $\phi_{-\alpha} = \{B, e_1, e_2, e_4, \sigma\}$
ARD prior, $p(\phi_\alpha \alpha)$	$\mathcal{N}(e_3 0, \alpha^{-1})$ or $\mathcal{L}_p(e_3 0, \alpha^{-1})$; $\alpha = \{\alpha\}$
Known prior, $p(\phi_{-\alpha})$	$\mathcal{L}_n(B 0.2, 0.5)\mathcal{U}(e_1 -2, 0)\mathcal{U}(e_2 -2, 0)\mathcal{U}(e_4 -600, 0)\mathcal{L}_n(\sigma 0.002, 0.5)$

Table 3.2: Unidimensional model reduction setup where parameter e_3 is treated as questionable.

Fig. 3.3 shows GOF, EIG, log-evidence, and the posterior and prior bounds (99% probability mass centered at ϕ^{MAP}) of e_3 for varying α . Fig. 3.3 shows data from both Gaussian and Laplace ARD prior. The numerical implementation details are provided in Sec. 3.1.2. As seen in Fig. 3.3, the log-evidence $\ln p(\mathcal{D}|\alpha)$ has a unique optimum $\alpha^{\text{MAP}} = 1\text{e-}04$. Hence, following evidence maximization, the optimal ARD prior is obtained as $\mathcal{N}(e_3|0, 10^4)$ or $\mathcal{L}_p(e_3|0, 10^4)$. Notice that both Gaussian and Laplace ARD priors provide the same solution for α . At this optimum, the posterior pdf of e_3 is dictated by $\mathcal{D}$ and not the ARD prior pdf, as seen in Fig. 3.3. In other words, the cubic aerodynamic stiffness ($e_3\theta^3$) is deemed relevant to the LCO aerodynamics.

In Fig. 3.3, notice the sharp decrease in GOF when α is increased beyond the optimum $\alpha^{\text{MAP}} = 1\text{e-}04$. This is because a prior pdf with high precision (α) overcomes the likelihood function in dictating the posterior parameter pdf. A zero-mean prior with high precision pulls the posterior space toward zero, as seen in the bottom-right plot in Fig. 3.3. Beyond $\alpha = 1\text{e+}1$, the prior pdf and the posterior pdf of e_3 are a Dirac-delta function at zero. Since e_3 is salient to the LCO aerodynamics, this forced removal of e_3 leads to the sharp decrease in GOF (data-fit). Since log-evidence is GOF minus EIG, this decrease in GOF is responsible for the decrease in log-evidence. Note that the decrease in EIG is minuscule in comparison to the decrease in GOF, and log-evidence remains dictated by GOF.

Fig. 3.4 shows the marginal posterior pdf of e_3 computed using flat prior $p(e_3) \propto 1$, optimum Gaussian ARD prior $\mathcal{N}(e_3|0, 10^4)$, and optimum Laplace ARD prior $\mathcal{L}_p(e_3|0, 10^4)$. The remaining inversion details remain the same for all the three cases, as listed in Table 3.2. The

Figure 3.3: Bayesian model comparison entities for varying α , when e_3 is treated as questionable.

marginal posterior pdf and the MAP estimate of e_3 show a close resemblance to the true value $e_3 = 100.0$ for all three cases. However, choosing a flat prior renders Bayesian model comparison arbitrary. Therefore, ARD offers a practical solution to this issue by eliminating the responsibility of assigning prior pdfs to questionable parameters such as e_3 .

Figure 3.4: Marginal posterior pdf of e_3 parameter for various choices of prior pdf.

It has been argued that choosing a zero-mean ARD prior can bias the posterior parameter statistics. It is evident from Fig. 3.4 that, when compared to the flat prior case, the bias in the posterior statistic of e_3 is higher for Laplace ARD than Gaussian ARD. This slight increase in

the bias can be attributed to the thicker tails of Laplace pdf. Nevertheless, the bias for Laplace and Gaussian ARD priors is manageable, i.e. 100 (true) vs 97.2 (Flat) vs 96.9 (Gaussian) vs 96.4 (Laplace). In other words, the posterior parameter statistics of e_3 is mostly dictated by the likelihood $p(\mathcal{D}|\phi)$ and the prior pdf $p(\phi_{-\alpha})$ of *a priori* relevant parameters (in case of strong correlation among ϕ_α and $\phi_{-\alpha}$). Note that this conclusion can only be generalized for ‘*large and dense*’ data $\mathcal{D}$ case and may not hold good when the information contents in likelihood and prior are comparable.

3.2.1.2 ARD prior for a parameter absent in the data-generating model

For inversion purpose, the proposed model is chosen to be the coupled ODE system shown in Eq. (3.6) but with an additional term $e_5\theta^5$ added to the right-hand side of the aerodynamic model equation in Eq. (3.6b). The unknown parameter vector for this model is $\phi = \{B, e_1, e_2, e_3, e_4, e_5, \sigma\}$. For illustration purpose, we assume that limited *a priori* knowledge is available about e_5 and its relevance to the LCO dynamics is regarded questionable. Hence, we decompose ϕ as $\phi_\alpha = \{e_5\}$ and $\phi_{-\alpha} = \{B, e_1, e_2, e_3, e_4, \sigma\}$. Table 3.3 summarizes the inverse problem, and also lists the known prior pdf $p(\phi_{-\alpha})$ assigned to $\phi_{-\alpha}$. Both Laplace and Gaussian ARD prior are considered for e_5 .

Aerodynamic model	$\frac{\dot{C}_M}{B} + C_M = e_1\theta + e_2\dot{\theta} + e_3\theta^3 + e_4\theta^2\dot{\theta} + e_5\theta^5 + \frac{c_6}{B}\ddot{\theta} + \sigma\xi(\tau)$
ϕ decomposition	$\phi_\alpha = \{e_5\}, \quad \phi_{-\alpha} = \{B, e_1, e_2, e_3, e_4, \sigma\}$
ARD prior, $p(\phi_\alpha \alpha)$	$\mathcal{N}(e_5 0, \alpha^{-1})$ or $\mathcal{L}_p(e_5 0, \alpha^{-1})$; $\alpha = \{\alpha\}$
Known prior, $p(\phi_{-\alpha})$	$\mathcal{L}_n(B 0.2, 0.5)\mathcal{U}(e_1 -2, 0)\mathcal{U}(e_2 -2, 0)\mathcal{U}(e_3 -250, 250)$ $\mathcal{U}(e_4 -600, 0)\mathcal{L}_n(\sigma 0.002, 0.5)$

Table 3.3: Unidimensional model reduction setup where parameter e_5 is treated as questionable.

Fig. 3.5 shows GOF, EIG, log-evidence, and the posterior and prior pdf bounds of e_5 for varying α , pertaining to both Gaussian and Laplace ARD priors. The implementation details are supplied in Sec. 3.1.2. As evident from Fig. 3.5, log-evidence is maximum at $1e-5$ and it remains

unchanged for all $\alpha > 1e-5$. Also, the posterior and prior pdf for e_5 is a Dirac-delta function for all $\alpha > 1e-05$. The maximum log-evidence pertains to $\alpha^{\text{MAP}} = \infty$ (or any value greater than $1e-5$), thereby indicating that e_5 is irrelevant to the LCO aerodynamics. Increasing α beyond $1e-5$ has no impact on GOF, EIG and log-evidence since e_5 has already been eliminated from Eq. (3.6b) at $\alpha = 1e-5$.

Figure 3.5: Bayesian model comparison entities for varying α , when e_5 is treated as questionable. In the bottom-left plot, the pdf bounds correspond to 99% probability mass centered around the mode.

In Fig. 3.5, GOF (data-fit) remains unaffected during the removal of e_5 since $e_5\theta^5$ is not contributing much towards the understanding of LCO. This behaviour of GOF indicates that e_5 is a redundant parameter. The decrease in EIG during sparse learning ($10^{-12} < \alpha < 1e-5$) demonstrates the notion that EIG quantifies model complexity. The increase in log-evidence is dictated by the decrease in EIG with no change in GOF for all α values. This behaviour of GOF, EIG and log-evidence demonstrates the quantitative Occam's razor property of model evidence that aims to find the optimal trade-off between data-fit and model simplicity, for a given $\mathcal{D}$ and $p(\phi_{-\alpha})$. This property of model evidence was previously revealed in Sec. 2.1.3, and is also

discussed in detail in Appendix A.2.

Fig. 3.6 shows the marginal posterior pdf of e_5 computed using flat prior $p(e_5) \propto 1$, optimum Gaussian ARD prior $\mathcal{N}(e_5|0, 10^2)$, and optimum Laplace ARD prior $\mathcal{L}_p(e_5|0, 10^2)$. The remaining inversion details remain the same for all the three cases, as listed in Table 3.2. Although all three pdfs are centered around zero (the true value of e_5), a key difference in these pdfs is the posterior variance. For instance, the posterior bound of e_5 while using Gaussian ARD prior is roughly $(-0.3, 0.3)$, whereas it is $(-3000, 5000)$ while using flat prior. Hence, the posterior pdf pertaining to the maximum evidence can be deemed as a Dirac-delta function in comparison to the posterior pdf obtained using a flat prior. This elimination of e_5 through ARD leads to the data-generating model in Eq. (3.6), wherein the posterior pdf of the rest of the parameters ($\phi_{-\alpha}$) is available following evidence maximization. Although not shown here, the posterior pdf of $\phi_{-\alpha}$ closely resembles the true parameter values used in generating $\mathcal{D}$.

Figure 3.6: Marginal posterior pdf of e_5 for various choices of prior pdf.

Notice that, for both the unidimensional cases studied in this section, Gaussian and Laplace ARD prior produces similar relevance determination characteristics. Hence, only Gaussian ARD priors are considered from now onwards.

3.2.2 Multidimensional model reduction

In this section, we investigate the performance of a multidimensional Gaussian ARD prior in determining relevance of multiple questionable parameters. Table 3.4 shows the proposed aerodynamic model, the parameter decomposition $\phi = \{\phi_{\alpha}, \phi_{-\alpha}\}$, and the known prior pdf for

$\phi_{-\alpha}$. The proposed model has the additional terms $e_5\theta^5$ and $e_6\theta^4\dot{\theta}$ in comparison to the data-generating aerodynamic model in Eq. (3.6b). It is assumed that the relevance of parameters e_3 , e_4 , e_5 , and e_6 is unknown and needs to be determined using the ARD-based Bayesian algorithm. In this case, the evidence $p(\mathcal{D}|\alpha)$ is a four-dimensional function of $\alpha = \{\alpha_1, \alpha_2, \alpha_3, \alpha_4\}$. Note that the model $\mathcal{M}$ in Table 3.4 corresponds to a first-order Wagner function representation of the unsteady aerodynamics.

Aerodynamic model	$\frac{\dot{C}_M}{B} + C_M = e_1\theta + e_2\dot{\theta} + e_3\theta^3 + e_4\theta^2\dot{\theta} + e_5\theta^5 + e_6\theta^4\dot{\theta} + \frac{c_6}{B}\ddot{\theta} + \sigma\xi(\tau)$
ϕ decomposition	$\phi_\alpha = \{e_3, e_4, e_5, e_6\}, \quad \phi_{-\alpha} = \{B, e_1, e_2, \sigma\}$
ARD prior, $p(\phi_\alpha \alpha)$	$\mathcal{N}(e_3 0, \alpha_1^{-1})\mathcal{N}(e_4 0, \alpha_2^{-1})\mathcal{N}(e_5 0, \alpha_3^{-1})\mathcal{N}(e_6 0, \alpha_4^{-1})$, $\alpha = \{\alpha_1, \alpha_2, \alpha_3, \alpha_4\}$
Known prior, $p(\phi_{-\alpha})$	$\mathcal{L}_n(B 0.2, 0.5)\mathcal{U}(e_1 -2, 0)\mathcal{U}(e_2 -2, 0)\mathcal{L}_n(\sigma 0.002, 0.5)$

Table 3.4: Multidimensional model reduction setup where parameters e_3 , e_4 , e_5 , and e_6 are treated as questionable.

Algorithm 1 is executed to compute the optimal ARD prior (or α^{MAP}). The evidence maximization algorithm is initiated from $\alpha = \{1\text{e-}10, 1\text{e-}10, 1\text{e-}10, 1\text{e-}10\}$. Fig. 3.7a shows GOF, EIG, and log-evidence for successful evidence maximization iterations. Fig. 3.7b shows the corresponding trend in marginal posterior and prior pdf of the questionable parameters. The evidence optimizer converges at $\alpha^{\text{MAP}} = \{1\text{e-}04, 1\text{e-}05, \infty, \infty\}$, which implies that e_3 and e_4 are deemed relevant to the LCO aerodynamics while e_5 and e_6 are deemed redundant. This conclusion is in agreement with the data-generating model from Eq. (3.6b).

Several observations can be made from Fig. 3.7a and Fig. 3.7b. The increase in log-evidence is largely dictated by the decrease in EIG, whereas GOF remains unaffected during the maximization. The decrease in EIG (or model complexity) is due to the removal of e_5 and e_6 from the model. This removal is facilitated by the Gaussian ARD prior. The parameters e_5 and e_6 are driven towards zero by the zero-mean ARD prior with a large precision (α_3 and α_4 , respectively). This removal of redundancies has a minimal effect on the data-fit property (or GOF) of the model. Also, from Fig. 3.7b, the posterior pdf of relevant parameters e_3 and e_4 remain

(a) Bayesian model comparison entities

(b) Posterior and prior marginal pdf of questionable parameters. The pdf bounds (solid lines) correspond to 99% probability mass centered around the mode (dashed lines).

Figure 3.7: Results for ARD-based Bayesian model reduction of the first-order unsteady aerodynamic model $\mathcal{M}$ in Table 3.4.

unaltered by the ARD priors, except that there is a small decrease in the posterior variance of e_3 and e_4 following the removal of e_5 and e_6 . This decrease in variance is an indication that the over-parameterized model in Eq. (3.6b) is being alleviated of overfitting. The apparent noise in log-evidence in Fig. 3.7a is because the evidence is computed using a finite set of MCMC samples. This sampling variability in log-evidence is taken into account when determining the convergence of the evidence maximization algorithm in Algorithm 1.

The optimal nested model, denoted as $\mathcal{M}^{\text{opt}}$, and the optimized prior pdf $p(\phi|\alpha^{\text{MAP}})$ is listed in Table 3.5. Note that $\mathcal{M}^{\text{opt}}$ is the same as the data-generating model in Eq. (3.6b). Fig. 3.8

shows the marginal and joint posterior pdf of relevant parameters a_3 and a_4 for the following to cases: 1) flat priors ($p(e_3, e_4) \propto 1$), and 2) optimal ARD prior $p(\phi|\alpha^{\text{MAP}})$ from Table 3.5. For both these cases, the prior pdf of other parameters is assigned as per Table 3.5. The closeness of the posterior pdfs for the two choices of prior pdf further reinforce the fact that zero-mean ARD prior does not bias the posterior parameter statistics when the information content in $\mathcal{D}$ dominates that of the prior pdf (a ‘large’ data case). In summary, an over-parameterized model consisting of a coupled nonlinear ODE system is reduced to an optimally fitted model without the need to 1) propose candidate nested models, and 2) assign prior pdfs to parameters with limited prior knowledge (like e_5 and e_6 in Eq. (3.6b)).

Optimal log-evidence; $\ln p(\mathcal{D} \alpha^{\text{MAP}}) = 2.78127\text{e}+04$	
$\mathcal{M}^{\text{opt}}$	$\frac{\dot{C}_M}{B} + C_M = e_1\theta + e_2\dot{\theta} + e_3\theta^3 + e_4\theta^2\dot{\theta} + \frac{c_6}{B}\ddot{\theta} + \sigma\xi(\tau)$
ϕ	$\{B, e_1, e_2, e_3, e_4, \sigma\}$
$p(\phi \alpha^{\text{MAP}})$	$\mathcal{L}(B 0.2, 50) \mathcal{U}(e_1 -2, 0) \mathcal{U}(e_2 -2, 0) \mathcal{N}(e_3 0, 1\text{e}+4) \mathcal{N}(e_4 0, 1\text{e}+5) \mathcal{L}(\sigma 0.002, 0.5)$

Table 3.5: Reduced model and its optimal prior pdf, pertaining to the inverse problem defined in Table 3.4.

Figure 3.8: Marginal and joint posterior pdf of e_3 and e_4 when assigning them 1) flat prior, or 2) the optimal ARD prior $p(\phi|\alpha^{\text{MAP}})$.

3.2.3 Comparison of non-nested models

In this section, we consider the comparison of competing models that cannot be considered as nested under a common encompassing model. This scenario arises when the models have different state-space composition. For instance, a quasi-steady model has the state vector $\mathbf{u} = \{\theta, \dot{\theta}\}$, while a first-order unsteady model (like Eq. (3.6b)) has the state vector $\mathbf{u} = \{\theta, \dot{\theta}, C_M\}$. Note that exploiting ARD for inducing sparsity in the state space involves complications. For instance, the state estimator needs to be modified during evidence optimization as the dimensionality of the state space changes during such type of sparse learning. This issue resembles the case of multimodel inference [130], wherein the governing model varies (discretely) depending on the regime of measurements. For instance, when dealing with multiple LCO attractors, the LCO model will change depending on the measured LCO amplitude. The use of ARD in such cases remains an open question, and this thesis focuses on inverse problems where the state space remains unchanged for all regimes of observations.

In this section, we consider three such over-parameterized (allegedly) models: 1) quasi-steady model $\mathcal{M}_1$ in Table 3.6, 2) first-order unsteady model $\mathcal{M}_2$ in Table 3.2 (previously referred to as $\mathcal{M}$), and 3) second-order unsteady model $\mathcal{M}_3$ in Table 3.7. These models are obtained by adding the nonlinear terms $e_3\theta^3$, $e_4\theta^2\dot{\theta}$, $e_5\theta^5$ and $e_6\theta^4\dot{\theta}$ into the corresponding linear ODEs derived using thin-airfoil aerodynamic theories. For instance, the second-order model $\mathcal{M}_3$ in Table 3.7 is based on the Wagner linear unsteady model derived in Appendix C.1 by assuming a second-order approximation of Wagner function.

Aerodynamic model	$C_M = e_1\theta + e_2\dot{\theta} + e_3\theta^3 + e_4\theta^2\dot{\theta} + e_5\theta^5 + e_6\theta^4\dot{\theta} + \sigma\xi(\tau)$
ϕ decomposition	$\phi_\alpha = \{e_3, e_4, e_5, e_6\}$, $\phi_{-\alpha} = \{e_1, e_2, \sigma\}$
ARD prior, $p(\phi_\alpha \alpha)$	$\mathcal{N}(e_3 0, \alpha_1^{-1})\mathcal{N}(e_4 0, \alpha_2^{-1})\mathcal{N}(e_5 0, \alpha_3^{-1})\mathcal{N}(e_6 0, \alpha_4^{-1})$, $\alpha = \{\alpha_1, \alpha_2, \alpha_3, \alpha_4\}$
Known prior, $p(\phi_{-\alpha})$	$\mathcal{U}(e_1 -2, 0)\mathcal{U}(e_2 0, 2)\mathcal{L}_n(\sigma 0.002, 0.5)$

Table 3.6: Model $\mathcal{M}_1$ and its associated prior pdf.

Aerodynamic model	$\frac{\ddot{C}_M}{B_1 B_2} + \frac{(B_1+B_2)\dot{C}_M}{B_1 B_2} + C_M = e_1\theta + e_2\dot{\theta} + e_3\theta^3 + e_4\theta^2\dot{\theta} + e_5\theta^5 + e_6\theta^4\dot{\theta}$ $+ \frac{(2c_6c_7+0.5)\ddot{\theta}}{B_1 B_2} + \frac{c_6\ddot{\theta}}{B_1 B_2} + \sigma\xi(\tau)$
ϕ decomposition	$\phi_\alpha = \{e_3, e_4, e_5, e_6\}$, $\phi_{-\alpha} = \{A, B_1, B_2, e_1, e_2, \sigma\}$
ARD prior, $p(\phi_\alpha \alpha)$	$\mathcal{N}(e_3 0, \alpha_1^{-1})\mathcal{N}(e_4 0, \alpha_2^{-1})\mathcal{N}(e_5 0, \alpha_3^{-1})\mathcal{N}(e_6 0, \alpha_4^{-1})$, $\alpha = \{\alpha_1, \alpha_2, \alpha_3, \alpha_4\}$
Known prior, $p(\phi_{-\alpha})$	$\mathcal{U}(A 0, 0.5) \mathcal{L}(B_1 0.32, 0.5) \mathcal{L}(B_2 0.04, 0.5)$ $\mathcal{U}(e_1 -2, 0) \mathcal{U}(e_2 -2, 0) \mathcal{L}(\sigma 0.002, 0.5)$

Table 3.7: Model $\mathcal{M}_3$ and its associated prior pdf.

The Bayesian model comparison of non-nested over-parameterized models $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$ is executed as following, First, ARD-based model reduction is executed to identify the optimal model $\mathcal{M}_i^{\text{opt}}$ nested under each over-parameterized model $\mathcal{M}_i$. These optimal nested models correspond to the optimal representation of nonlinear aerodynamics for a given assumption of quasi-steady or unsteady aerodynamics. Next, Bayesian model comparison is executed on the optimal nested model set $\mathcal{M}^{\text{opt}} = \{\mathcal{M}_1^{\text{opt}}, \mathcal{M}_2^{\text{opt}}, \mathcal{M}_3^{\text{opt}}\}$ using the corresponding optimal evidence values. This model comparison task is aimed at identifying the nature of unsteady aerodynamics. The optimal model identified through Bayesian model comparison thus have the highest evidence among all models nested under the three over-parameterized models: $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$. In other words, the optimal model corresponds to an optimal representation of unsteady and nonlinear aerodynamics of the LCO.

The ARD-based model reduction results for the model $\mathcal{M}_2$ has already been reported and discussed in Sec. 3.2.2. Next, we execute model reduction of $\mathcal{M}_1$ and $\mathcal{M}_3$ using Algorithm 1. The prior pdf $p(\phi|\alpha)$ and the ϕ decomposition is reported in Table 3.6 for $\mathcal{M}_1$, and in Table 3.7 for $\mathcal{M}_3$. Fig. 3.9 shows the trend in GOF, EIG, log-evidence, and posterior pdf for successful evidence maximization iterations for model $\mathcal{M}_1$. Notice that parameter e_3 and e_4 are deemed relevant, while parameter e_5 and e_6 are deemed irrelevant and can be pruned from $\mathcal{M}_1$. Fig. 3.10 shows the trend in GOF, EIG, log-evidence, and posterior pdf for successful evidence maximization iterations for model $\mathcal{M}_3$. Notice that parameter e_3 , e_4 and e_6 are deemed relevant,

while parameter e_5 is deemed irrelevant and can be pruned from $\mathcal{M}_3$.

(a) Bayesian model comparison entities

(b) Posterior and prior marginal pdf of questionable parameters. The pdf bounds (solid lines) correspond to 99% probability mass centered around the mode (dashed lines).

Figure 3.9: Results for the ARD-based Bayesian model reduction of the quasi-steady aerodynamic model $\mathcal{M}_1$ in Table 3.6.

The model set $\mathcal{M}^{\text{opt}} = \{\mathcal{M}_1^{\text{opt}}, \mathcal{M}_2^{\text{opt}}, \mathcal{M}_3^{\text{opt}}\}$ containing optimal nested models is listed in Table 3.8. Consequently, Bayesian model comparison is executed on $\mathcal{M}^{\text{opt}}$ using the optimal evidence value $p(\phi|\alpha^{\text{MAP}})$ available from the ARD algorithm. Hence, no additional computational effort is needed to execute this model comparison task. Table 3.9 shows the Bayesian model comparison entities, including GOF and EIG. Note that $\mathcal{M}_2^{\text{opt}}$ is selected as the optimal model in $\mathcal{M}^{\text{opt}}$ with a posterior model probability of one. $\mathcal{M}_2^{\text{opt}}$ is the same as the data-generating aerodynamic model in Eq. (3.6b).

(a) Bayesian model comparison entities

(b) Posterior and prior marginal pdf of questionable parameters. The pdf bounds (solid lines) correspond to 99% probability mass centered around the mode (dashed lines).

Figure 3.10: Results for the ARD-based Bayesian model reduction of the second-order unsteady aerodynamic model $\mathcal{M}_3$ in Table 3.7.

$\mathcal{M}^{\text{opt}}$	Optimal stochastic ODE for modelling C_M in Eq. (3.6a)
$\mathcal{M}_1^{\text{opt}}$	$C_M = e_1\theta + e_2\dot{\theta} + e_3\theta^3 + e_4\theta^2\dot{\theta} + \sigma\xi(\tau)$
$\mathcal{M}_2^{\text{opt}}$	$\frac{\dot{C}_M}{B} + C_M = e_1\theta + e_2\dot{\theta} + e_3\theta^3 + e_4\theta^2\dot{\theta} + \frac{c_6}{B}\ddot{\theta} + \sigma\xi(\tau)$
$\mathcal{M}_1^{\text{opt}}$	$\frac{\ddot{C}_M}{B_1B_2} + \frac{(B_1+B_2)\dot{C}_M}{B_1B_2} + C_M = e_1\theta + e_2\dot{\theta} + e_3\theta^3 + e_4\theta^2\dot{\theta} + e_6\theta^4\dot{\theta} + \frac{(2c_6c_7+0.5)\ddot{\theta}}{B_1B_2} + \frac{c_6\ddot{\theta}}{B_1B_2} + \sigma\xi(\tau)$

Table 3.8: Candidate model set containing optimal nested models obtained from ARD-based model reduction.

Fig. 3.11 shows the marginal pdfs, and Fig. 3.12 shows selected joint pdfs, both pertaining to the posterior parameter pdf $p(\phi|\mathcal{D}, \alpha^{\text{MAP}})$ for $\mathcal{M}^{\text{opt}}$. These pdfs are generated using KDE

$\mathcal{M}^{\text{opt}}$	Unsteady aero.	Nonlinear aero.	Data-fit GOF (+2.77e4)	Complexity EIG	Log-evidence $\log p(\mathcal{D} \alpha^{\text{MAP}})$ (+2.77e4)	Post. prob. $p(\mathcal{M}_i^{\text{opt}} \mathcal{D})$
$\mathcal{M}_1^{\text{opt}}$	Quasi-steady	$\theta^3, \theta^2\dot{\theta}$	37.8	19.5	18.3	0.0
$\mathcal{M}_2^{\text{opt}}$	1st-ord unsteady	$\theta^3, \theta^2\dot{\theta}$	135.5	22.8	112.7	1.0
$\mathcal{M}_3^{\text{opt}}$	2nd-ord unsteady	$\theta^3, \theta^2\dot{\theta}, \theta^4\dot{\theta}$	129.4	48.2	81.2	0.0

Table 3.9: Bayesian model comparison among optimal nested models.

approximation of stationary MCMC samples. Again, no additional computation effort is needed to obtain these pdfs since stationary MCMC samples are already available from the evidence maximization algorithm. As seen in Fig. 3.11 and Fig. 3.12, the posterior pdf $p(\phi|\mathcal{D}, \alpha^{\text{MAP}})$ is in close agreement with the true ϕ value. A small bias can be noticed between the true value and the MAP estimate for some parameters. The bias in the model error strength σ is larger than the bias in other parameters. This bias is partially due to the finiteness of the measurement data, and, therefore, this bias could change with varying realization of the model error and the measurement noise. The bias in model error strength may also be attributed to the filtering (low-pass filter) of noisy measurement data prior to Bayesian inversion.

Figure 3.11: Marginal posterior parameter pdf for the optimal model $\mathcal{M}_2^{\text{opt}}$.

From Fig. 3.12, the strong correlation between e_1 and e_3 indicates that these parameters are

attempting to capture similar physics, in this case nonlinear aerodynamic stiffness. The low correlation between e_3 and e_4 makes sense since e_3 quantifies aerodynamic stiffness and e_4 quantifies aerodynamic damping. Also, noise strength σ appears to be uncorrelated to physical parameters such as B_1 and e_1 . This behaviour of σ indicates that the model error term $\sigma\xi(\tau)$ largely quantifies the non-physical aspect of the LCO aerodynamics, such as the effect of turbulence.

Figure 3.12: Selected joint posterior parameter pdf for the optimal model $\mathcal{M}_2^{opt}$.

Fig. 3.13 shows the state pdf (updated) of pitch displacement $\theta(\tau)$ for $\mathcal{M}_2^{opt}$ computed using EKF. The updated state mean is in close agreement with the measurements, demonstrating the performance of EKF which relies on a linearized model equation to propagate a Gaussian state pdf. Notice that the state variance decays as more measurements are assimilated. The state variance never decays to zero due to the presence of measurement noise. Nevertheless, in case EKF fails, sampling-based algorithms such as the particle filter can be employed to execute the state estimation task.

In closing, ARD offers a principled approach to deal with Bayesian model comparison among nested models by transforming the model domain from a discrete model space $\mathcal{M}$ to a continuous hyperparameter α space. This transformation enlarges the model domain and increases the chances of obtaining a better model than Bayesian model comparison among a finite

Figure 3.13: Updated state pdf for the optimal model $\mathcal{M}_2^{opt}$.

number of models.

Chapter 4

Nonlinear sparse Bayesian learning

4.1 Introduction

The ARD-based model reduction approach detailed in Chapter 3 involves the following two numerical tasks: 1) Bayesian model evidence computation, and 2) evidence maximization (type-II maximum likelihood) [121]. In Algorithm 1, the model evidence was estimated using stationary MCMC samples from the posterior parameter pdf, followed by a gradient-free maximization of the evidence estimate. This MCMC sampling within an optimization step hinders the computational efficiency of the resulting algorithm, especially while dealing with nonlinear dynamical systems wherein the likelihood function computation involves a state estimation task [17]. In addition, the accuracy of the ARD algorithm relies on the quality (i.e. sample auto-correlation) of stationary MCMC samples. As the posterior parameter pdf varies with a changing ARD prior (hyperparameter), the tuning and monitoring of the RWM MCMC algorithm during evidence maximization becomes increasingly cumbersome. These practical issues are addressed in this chapter by proposing a semi-analytical sparse learning algorithm of NSBL that 1) eliminates the need for sampling-within-optimization task, and 2) involves an efficient gradient-based evidence optimizer.

Fig. 4.1 provides an overview of NSBL. The central idea behind NSBL is that the entity

consisting of the product of likelihood function and the known prior pdf of *a priori* relevant parameters remains fixed during sparse learning (i.e it is independent of hyperparameters). This entity also equates to the unnormalized posterior parameter pdf while the questionable parameters are assigned flat priors (i.e non-informative priors). As a result, NSBL operates by building a Gaussian mixture-model (GMM) approximation [40, 131] of this entity using stationary samples from the unnormalized pdf. This GMM approximation offers two advantages in the current context. First, the GMM approximation can handle non-Gaussian posterior pdfs encountered during Bayesian inversion of nonlinear models using sparse/noisy/incomplete measurements. Second, the GMM approximation enables analytical scrutiny of the Bayesian framework, allowing for semi-analytical computation (in terms of Gaussian kernels) of model evidence and its gradient and Hessian information with respect to hyperparameters. Consequently, the sampling-within-optimization step of the previous ARD-based approach (Chapter 3) is replaced with a Newton’s iteration that exploits this gradient and Hessian information to expedite evidence maximization. Also, the posterior parameter pdf of the resulting sparse model is available analytically (i.e. in terms of Gaussian kernels).

Figure 4.1: Overview of NSBL.

NSBL is in contrast to the previous Bayesian inversion approaches reported in the literature involving cheap surrogates for the computationally intensive model or the likelihood function or the posterior pdf [132–135]. Most notably, Marzouk *et al.* [132] exploited an intrusive stochastic spectral technique to reformulate the governing equations using a prior-informed PCE of model parameters, and then sampling the PCE coefficients instead of model parameters. Also, Galbally *et al.* [135] proposed the projection of a high-fidelity model on a reduced subspace using proper orthogonal decomposition to expedite MCMC sampling in a high-dimensional parameter space. The underlying goal of these approaches has been to compute the posterior parameter pdf faster. In contrast, NSBL is primarily aimed at identifying sparse parameter representation among physics-based models. The fact that the posterior parameter pdf is available analytically through NSBL is an added benefit. Also, we do not offer NSBL as a replacement of SBL or BCS for linear-in-parameter models. NSBL is primarily aimed towards physics-based applications where SBL and BCS are inapplicable. In fact, NSBL can be seen as an extension of SBL to nonlinear models or non-Gaussian likelihood function cases where prior parameter knowledge is significant and needs to be incorporated.

This chapter is structured as following:

- In Sec. 4.2 we provide a detailed mathematical derivation into the semi-analytical apparatus of NSBL, followed by the numerical implementation details regarding GMM construction and Newton’s method for optimizing evidence. We also show analytically that SBL is special case of NSBL for linear-in-parameter models while Gaussian ARD priors is assigned to all unknown parameters.
- Following numerical investigations are reported in Sec. 4.3 to demonstrate the performance of NSBL:
 - In Sec. 4.3.1, we consider a linear regression setting involving the construction of a sparse PCE surrogate for the strongly-nonlinear Ishigami function. This numerical exercise is aimed at validating the semi-analytical formulation of NSBL against SBL since both the

algorithms are expected to produce similar results in a linear regression setting. In addition, computational efficiency of NSBL is contrasted with SBL and BCS when dealing with high-dimensional models.

- In Sec. 4.3.2, we consider a linear regression setting wherein the posterior parameter pdf and the model evidence are both multimodal. These special circumstances are synthetically generated using a multimodal prior pdf on a regression model parameter. This numerical investigation highlights the applicability of NSBL to physics-based inverse problems involving multimodality in both parameter and hyperparameter space. SBL and BCS are inapplicable to such cases.
- In Sec. 4.3.3, we consider the structural dynamics example from Sec. 1.5 involving a three-dof mass-spring-damper model of shear building frame. NSBL is applied to identify the sparse damping structure of the three-dof system using a sparse, noisy and incomplete realization of floor displacement during free vibration. This numerical exercise is aimed at investigating the performance of NSBL in alleviating overfitting during inverse modelling of mechanistic systems.

4.2 NSBL methodology

Consider that a system model $f : \phi \mapsto \mathbf{y}$ is proposed to model a physical system, where the model operator f maps the unknown model parameter vector $\phi \in \mathbb{R}^{N_\phi}$ to the observed entity $\mathbf{y} \in \mathbb{R}^{N_y}$. The N_d number of measurements of $\mathbf{y}$ are denoted as $\mathcal{D}$. We acknowledge that the proposed model f is over-parameterized, and measurements in $\mathcal{D}$ are noisy, sparse, and incomplete (i.e. the entire system state is not measured). Also, it is assumed that the likelihood function $p(\mathcal{D}|\phi)$ is known for any ϕ value, and the estimation of hidden variables (or unobserved state variables) is encapsulated within the likelihood function. Next, we detail the Bayesian setup that empowers the semi-analytical apparatus of NSBL.

4.2.1 Hybrid prior pdf

Bayesian inference of unknown model parameters ϕ requires the modeller to assign a prior parameter pdf for ϕ . For an over-parameterized model, the prior pdfs can usually be fabricated for a few parameters that are known to be relevant to the system physics. Rest of the parameters in ϕ are treated as questionable, whose relevance to the system physics is determined through sparse learning. NSBL algorithm incorporates this idea by decomposing $\phi = \{\phi_\alpha, \phi_{-\alpha}\}$ (in that order), where $\phi_\alpha \in \mathbb{R}^{N_\alpha}$ contains parameters with no prior knowledge (and hence questionable), and $\phi_{-\alpha} \in \mathbb{R}^{N_\phi - N_\alpha}$ contains parameters with a known prior pdf $p(\phi_{-\alpha})$. This ϕ decomposition approach was also exploited in Chapter 3.

Next, the concept of ARD is exploited to identify sparsity in the questionable parameter vector ϕ_α . Following the similar ARD approach as SBL, ϕ_α is assigned a Gaussian ARD prior $p(\phi_\alpha|\alpha) = \mathcal{N}(\phi_\alpha|\mathbf{0}, \mathbf{A}^{-1})$, where $\mathbf{A} = \text{Diag}(\alpha)$ is the prior precision matrix, and $\alpha \in \mathbb{R}^{N_\alpha}$ is the unknown hyperparameter vector. Notice that the matrix $\mathbf{A}$ is diagonal, indicating the prior independence among questionable parameters. The marginal ARD prior pdf for parameter $\phi_i \in \phi_\alpha$ can be written as $p(\phi_i|\alpha_i) = \mathcal{N}(\phi_i|0, \alpha_i^{-1})$. Note that hyperparameter α_i controls the contribution of parameter ϕ_i in the model since setting $\alpha_i = \infty$ would force both the prior and posterior pdf of ϕ_i to be a Dirac-delta pdf $\mathcal{N}(\phi_i|0, 0)$ centered at zero. In other words, α controls the complexity of the model.

NSBL employs a Gaussian ARD prior due to two reasons. First, Gaussian priors enable the analytical computation of Bayesian entities in the proposed NSBL setup (explained later). Second, using the principle of maximum entropy, a Gaussian pdf contains minimum information or maximum entropy for a random variable with finite mean and finite variance [7]. This property of a Gaussian ARD prior ensures minimum interference from the modeller on the posterior distribution of sparse relevant parameters identified through sparse learning. This property is desired since a zero-mean prior pdf tends to pull the posterior parameter space towards the origin. Employing a Gaussian ARD prior minimizes this pull.

The joint prior pdf of ϕ chosen in NSBL is summarized as

$$p(\phi|\alpha) = p(\phi_{-\alpha})p(\phi_{\alpha}|\alpha) = p(\phi_{-\alpha})\mathcal{N}(\phi_{\alpha}|\mathbf{0}, \mathbf{A}^{-1}). \quad (4.1)$$

This hybrid prior pdf enables sparse learning of questionable parameters in ϕ_{α} through ARD prior $p(\phi_{\alpha}|\alpha)$ while incorporating prior knowledge about $\phi_{-\alpha}$ through $p(\phi_{-\alpha})$.

4.2.2 Gaussian mixture-model approximation

Given measurements $\mathcal{D}$ and hyperparameter vector α , the parameter posterior pdf $p(\phi|\mathcal{D}, \alpha)$ is obtained using Bayesian inference as

$$p(\phi|\mathcal{D}, \alpha) = \frac{p(\mathcal{D}|\phi)p(\phi|\alpha)}{p(\mathcal{D}|\alpha)} = \frac{p(\mathcal{D}|\phi)p(\phi_{-\alpha})\mathcal{N}(\phi_{\alpha}|\mathbf{0}, \mathbf{A}^{-1})}{p(\mathcal{D}|\alpha)}, \quad (4.2)$$

where $p(\mathcal{D}|\phi)$ is the likelihood function, $p(\mathcal{D}|\alpha)$ is the model evidence (or marginal likelihood or type-II likelihood), and $p(\phi|\alpha)$ is the joint prior pdf from Eq. (4.1). NSBL operates by building a GMM approximation for the entity $p(\mathcal{D}|\phi)p(\phi_{-\alpha})$ in Eq. (4.2) as

$$p(\mathcal{D}|\phi)p(\phi_{-\alpha}) \approx \sum_{k=1}^K a^{(k)} \mathcal{N}(\phi|\mu^{(k)}, \Sigma^{(k)}), \quad (4.3)$$

where K is the total number of kernels, $a^{(k)} \in \mathbb{R}$ is the kernel coefficient ($a^{(k)} > 0$) and $\mathcal{N}(\phi|\mu^{(k)}, \Sigma^{(k)})$ is a Gaussian pdf with mean vector $\mu^{(k)} \in \mathbb{R}^{N_{\phi}}$ and covariance matrix $\Sigma^{(k)} \in \mathbb{R}^{N_{\phi} \times N_{\phi}}$.

Note that only one kernel ($K = 1$) is sufficient in Eq. (4.3) under the special circumstances of 1) linear regression with Gaussian likelihood $p(\mathcal{D}|\phi)$ and Gaussian prior $p(\phi_{-\alpha})$ (same setup as SBL), or 2) ‘large’ number of measurements in $\mathcal{D}$ such that $p(\mathcal{D}|\phi)p(\phi_{-\alpha})$ can be approximated as a Gaussian (Laplace approximation [40]). For any other case, the approximation in Eq. (4.3) requires more than one Gaussian kernel and unknown entities $a^{(k)}$, $\mu^{(k)}$ and $\Sigma^{(k)}$ need to be estimated numerically. For instance, KDE with Gaussian kernels can be employed to construct the

GMM in Eq. (4.3). Sec. 4.2.5 provides specific implementation details behind Eq. (4.3).

The GMM approximation in Eq. (4.3) offers the following benefits for sparse learning among physics-based models:

- The use of Gaussian kernels in Eq. (4.3), in combination with Gaussian ARD priors, enables analytical evaluation of parameter posterior pdf $p(\phi|\mathcal{D}, \alpha)$, model evidence $p(\mathcal{D}|\alpha)$, and the gradient and Hessian of evidence with respect to hyperparameters α . The semi-analytical apparatus of NSBL is powered by this analytical tractability of Bayesian entities.
- Entity $p(\mathcal{D}|\phi)p(\phi_{-\alpha})$ in Eq. (4.3) is much well-behaved (in terms of identifiability [41]) than the likelihood function $p(\mathcal{D}|\phi)$ due to the regularization effect of the known prior $p(\phi_{-\alpha})$. In other words, the known prior $p(\phi_{-\alpha})$ helps restrict $p(\mathcal{D}|\phi)p(\phi_{-\alpha})$ in the ϕ space that makes physical sense and complies with the prior knowledge. This property is desired when generating stationary samples from $p(\mathcal{D}|\phi)p(\phi_{-\alpha})$ for the sake of constructing a GMM. This is one of the key differences between NSBL and purely data-based techniques that solely rely on the likelihood function.
- Since $p(\mathcal{D}|\phi)p(\phi_{-\alpha})$ is independent of α , the GMM approximation in Eq. (4.3) is only needed to be built once for the sake of sparse learning.
- The kernel-based GMM approximation in Eq. (4.3) allows for representing non-Gaussian or multimodal $p(\mathcal{D}|\phi)p(\phi_{-\alpha})$ encountered in engineering applications. This property of the GMM allows for the handling of multimodal or skewed likelihood function $p(\mathcal{D}|\phi)$, or a non-Gaussian prior pdf $p(\phi_{-\alpha})$.
- Eq. (4.3) facilitates direct handling of the likelihood function instead of the model $f(\phi)$. This approach of constructing a surrogate for the likelihood function is in contrast to the approach where a surrogate for the model is constructed. This model-free property of NSBL is desired for inverse problems where the model $f(\phi)$ is only available as a black-box numerical solver. In other words, NSBL algorithm only needs access to the likelihood function, and a closed-form expression for the physical model is not required. Also, Constructing a surrogate

Next, we define the sparse learning problem in mathematical terms.

4.2.3 Sparse learning optimization problem

Using the hierarchical Bayes approach, the hyperparameter posterior pdf $p(\alpha|\mathcal{D})$ is obtained as [7]

$$p(\alpha|\mathcal{D}) = \frac{p(\mathcal{D}|\alpha)p(\alpha)}{p(\mathcal{D})} \propto p(\mathcal{D}|\alpha)p(\alpha), \quad (4.4)$$

where $p(\mathcal{D}|\alpha)$ is the model evidence from Eq. (4.2), $p(\alpha)$ is the hyper-prior (prior for hyperparameter α), and $p(\mathcal{D})$ is just a normalization constant for a fixed $\mathcal{D}$. Note that our interest in the sparsity of ϕ_α requires us to obtain an optimal α value and not the entire posterior distribution $p(\alpha|\mathcal{D})$ from Eq. (4.4). The MAP estimate α^{MAP} provides such an optimal choice, obtained by maximizing the hyperparameter posterior $p(\alpha|\mathcal{D})$ from Eq. (4.4) as

$$\alpha^{\text{MAP}} = \arg \max_{\alpha} \{p(\alpha|\mathcal{D})\} = \arg \max_{\alpha} \{p(\mathcal{D}|\alpha)p(\alpha)\}. \quad (4.5)$$

For a sparse ϕ_α , many α_i^{MAP} approach infinity, thereby forcing the marginal posterior pdf of ϕ_i to be a Dirac-delta function centered at zero. A finite α_i^{MAP} implies a relevant parameter $\phi_i \in \phi_\alpha$.

Similar to SBL [78], the hyperparameters in α are assumed to be *a priori* independent and the marginal hyperprior pdf $p(\alpha_i)$ is chosen to be a Gamma distribution. The joint hyperprior $p(\alpha)$ is written as

$$p(\alpha) = \prod_{i=1}^{N_\alpha} p(\alpha_i) = \prod_{i=1}^{N_\alpha} \mathcal{G}(\alpha_i|r_i, s_i) = \prod_{i=1}^{N_\alpha} \frac{s_i^{r_i}}{\Gamma(r_i)} \alpha_i^{r_i-1} e^{-s_i \alpha_i}, \quad (4.6)$$

where $\mathcal{G}(\alpha_i|r_i, s_i)$ denotes a univariate Gamma distribution with shape parameter $r_i > 0$ and rate parameter $s_i > 0$. Aside from enforcing positivity constraint for the precision parameter α_i , a Gamma distribution can be reduced to many simplified informative or non-informative distributions by varying r and s values, as detailed in Table 4.1. In this work, we employ

Jeffrey’s prior (flat prior over $\log \alpha_i$) for all $\alpha_i \in \alpha$, which is obtained by using values of $s \approx 0$ and $r \approx 0$ in Eq. (4.6). When using a non-informative prior for α (such as Jeffrey’s prior), the optimal value of α^{MAP} from Eq. (4.5) is solely dictated by model evidence $p(\mathcal{D}|\alpha)$. Nevertheless, the effect of hyperprior parameters (r and s) on sparse learning capabilities of NSBL is investigated in Appendix D.

Prior type	Parameters in Eq. (4.6)	Type of pdf	$p(\alpha)$
Informative	$r = 1, s = \lambda$	Exponential	$\lambda e^{-\lambda\alpha}$
	$r > 0, s \rightarrow \infty$	Dirac-delta	$\delta(\alpha - r/s)$
Non-informative	$r \rightarrow 0^+, s \rightarrow 0^+$	Jeffery’s prior	$p(\alpha) \propto 1/\alpha$ or $p(\log \alpha) \propto 1$
	$r \rightarrow 1^+, s \rightarrow 0^+$	Flat prior	$p(\alpha) \propto 1$ or $p(\log \alpha) \propto \alpha $

Table 4.1: Special cases of a Gamma hyperprior pdf.

Rewriting Eq. (2.7) in the current setting, it is realized that the evidence facilitates the optimal trade-off between data-fit and model simplicity as

$$\underbrace{\ln p(\mathcal{D}|\alpha)}_{\text{Log-evidence}} = \underbrace{\mathbb{E}[\ln p(\mathcal{D}|\phi)]}_{\text{Data-fit}} - \underbrace{\mathbb{E}\left[\ln \frac{p(\phi|\mathcal{D}, \alpha)}{p(\phi|\alpha)}\right]}_{\text{Complexity}}, \quad (4.7)$$

where the expectation is with respect to the posterior $pr(\phi|\mathcal{D}, \alpha)$.

Note that Eq. (4.27) is a non-convex optimization problem for any given combination of likelihood function $p(\mathcal{D}|\phi)$, prior pdf $p(\phi|\alpha)$ and hyperprior pdf $p(\alpha)$ [40]. In the SBL/RVM setup of linear regression models with Gaussian errors, Faul and Tipping [136] showed analytically that the log-evidence $\log p(\mathcal{D}|\alpha)$ has a unique global optimum with respect to an individual hyperparameter α_i (not the entire α vector). SBL and BCS exploited this property to propose a semi-analytical re-estimation procedure derived by setting the gradient of log-evidence with respect to α_i to zero. Faul and Tipping [136] also showed that the optimal α obtained through this re-estimation procedure is a joint optimum for all α_i . However, the uniqueness of this optimal α was put in question by Faul and Tipping [136], pointing towards the non-convex nature of the evidence optimization. Nevertheless, the models encountered in engineering mechan-

ics are far from linear regression, where SBL and BCS are inapplicable and a unique global optimum of log-evidence with respect to α_i or α is not guaranteed. We also demonstrate this non-uniqueness of α^{MAP} (computed using Eq. (4.5)) through a numerical example in Sec. 4.3.1.

NSBL operates by employing Newton's method to solve the non-convex optimization of $p(\mathcal{D}|\alpha)p(\alpha)$ in Eq. (4.27). This choice is motivated by the fact that the model evidence and its gradient and Hessian information is analytically tractable in the current Bayesian setup. Next, we provide a detailed mathematical exposition into the analytical calculation of Bayesian entities. Further details regarding the numerical aspect of Newton's method and the resulting NSBL algorithm are provided in Sec. 4.2.5.

4.2.4 Analytical calculation of Bayesian entities

4.2.4.1 Model evidence

Given the the joint prior pdf in Eq. (4.1), the model evidence in Eq. (4.2) is written as

$$p(\mathcal{D}|\alpha) = \int p(\mathcal{D}|\phi)p(\phi|\alpha)d\phi = \int p(\mathcal{D}|\phi)p(\phi_{-\alpha})p(\phi_{\alpha}|\alpha)d\phi. \quad (4.8)$$

An estimate $\hat{p}(\mathcal{D}|\alpha)$ of model evidence is obtained by substituting Eq. (4.3) and ARD prior $p(\phi|\alpha) = \mathcal{N}(\phi_{\alpha}|\mathbf{0}, \mathbf{A}^{-1})$ in Eq. (4.8) as

$$\begin{aligned} \hat{p}(\mathcal{D}|\alpha) &= \int \left\{ \sum_{k=1}^K a^{(k)} \mathcal{N}(\phi|\boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}) \right\} \mathcal{N}(\phi_{\alpha}|\mathbf{0}, \mathbf{A}^{-1})d\phi \\ &= \sum_{k=1}^K a^{(k)} \int \mathcal{N}(\phi|\boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)})\mathcal{N}(\phi_{\alpha}|\mathbf{0}, \mathbf{A}^{-1})d\phi. \end{aligned} \quad (4.9)$$

Note that the integral in Eq. (4.9) involves the product of two multivariate Gaussian pdfs with different dimensions (ϕ_{α} vs ϕ).

Using the parameter decomposition $\phi = \{\phi_{\alpha}, \phi_{-\alpha}\}$, the Gaussian kernel $\mathcal{N}(\phi|\boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)})$ in

Eq. (4.9) is rewritten in an expanded form as

$$\mathcal{N}(\phi|\boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}) = \mathcal{N}\left(\left\{\begin{array}{c} \phi_\alpha \\ \phi_{-\alpha} \end{array}\right\} \middle| \left\{\begin{array}{c} \boldsymbol{\mu}_\alpha^{(k)} \\ \boldsymbol{\mu}_{-\alpha}^{(k)} \end{array}\right\}, \left[\begin{array}{cc} \boldsymbol{\Sigma}_\alpha^{(k)} & \mathbf{C}^{(k)} \\ (\mathbf{C}^{(k)})^T & \boldsymbol{\Sigma}_{-\alpha}^{(k)} \end{array}\right]\right), \quad (4.10)$$

where $\boldsymbol{\mu}_\alpha^{(k)} \in \mathbb{R}^{N_\alpha}$ and $\boldsymbol{\Sigma}_\alpha^{(k)} \in \mathbb{R}^{N_\alpha \times N_\alpha}$ pertain to the questionable parameters ϕ_α ; $\boldsymbol{\mu}_{-\alpha}^{(k)} \in \mathbb{R}^{N_\phi - N_\alpha}$ and $\boldsymbol{\Sigma}_{-\alpha}^{(k)} \in \mathbb{R}^{(N_\phi - N_\alpha) \times (N_\phi - N_\alpha)}$ pertain to the *a priori* relevant parameters $\phi_{-\alpha}$; and $\mathbf{C}^{(k)} \in \mathbb{R}^{N_\alpha \times (N_\phi - N_\alpha)}$ is the cross-covariance matrix of ϕ_α and $\phi_{-\alpha}$. Using the conditional distribution relations for multivariate Gaussian pdfs from Appendix F.1, the Gaussian kernel in Eq. (4.10) is re-written as (using $p(X_1, X_2) = p(X_1|X_2)p(X_2)$)

$$\mathcal{N}(\phi|\boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}) = \mathcal{N}(\phi_{-\alpha}|\tilde{\boldsymbol{\mu}}_{-\alpha}^{(k)}, \tilde{\boldsymbol{\Sigma}}_{-\alpha}^{(k)})\mathcal{N}(\phi_\alpha|\boldsymbol{\mu}_\alpha^{(k)}, \boldsymbol{\Sigma}_\alpha^{(k)}), \quad (4.11a)$$

$$\tilde{\boldsymbol{\mu}}_{-\alpha}^{(k)} = \boldsymbol{\mu}_{-\alpha}^{(k)} + (\mathbf{C}^{(k)})^T (\boldsymbol{\Sigma}_\alpha^{(k)})^{-1} (\phi_\alpha - \boldsymbol{\mu}_\alpha^{(k)}), \quad (4.11b)$$

$$\tilde{\boldsymbol{\Sigma}}_{-\alpha}^{(k)} = \boldsymbol{\Sigma}_{-\alpha}^{(k)} - (\mathbf{C}^{(k)})^T (\boldsymbol{\Sigma}_\alpha^{(k)})^{-1} \mathbf{C}^{(k)}, \quad (4.11c)$$

where the mean vector $\tilde{\boldsymbol{\mu}}_{-\alpha}^{(k)}$ and the covariance matrix $\tilde{\boldsymbol{\Sigma}}_{-\alpha}^{(k)}$ pertain to $\phi_{-\alpha}$ conditioned on a known ϕ_α value. Matrix $\tilde{\boldsymbol{\Sigma}}_{-\alpha}^{(k)}$ is also known as the schur complement of matrix $\boldsymbol{\Sigma}_{-\alpha}^{(k)}$ defined in Eq. (4.10).

Substituting the expansion from Eq. (4.11a) in Eq. (4.9) leads to

$$\hat{p}(\mathcal{D}|\boldsymbol{\alpha}) = \sum_{k=1}^K a^{(k)} \int \mathcal{N}(\phi_{-\alpha}|\tilde{\boldsymbol{\mu}}_{-\alpha}^{(k)}, \tilde{\boldsymbol{\Sigma}}_{-\alpha}^{(k)}) \underbrace{\mathcal{N}(\phi_\alpha|\boldsymbol{\mu}_\alpha^{(k)}, \boldsymbol{\Sigma}_\alpha^{(k)})\mathcal{N}(\phi_\alpha|\mathbf{0}, \mathbf{A}^{-1})}_{\text{Product of two Gaussian pdfs}} d\phi. \quad (4.12)$$

The product of two Gaussian pdfs in Eq. (4.12) can be evaluated analytically using the identity

listed in Appendix F.2 as

$$\mathcal{N}(\phi_\alpha | \boldsymbol{\mu}_\alpha^{(k)}, \boldsymbol{\Sigma}_\alpha^{(k)}) \mathcal{N}(\phi_\alpha | \mathbf{0}, \mathbf{A}^{-1}) = \mathcal{N}(\boldsymbol{\mu}_\alpha^{(k)} | \mathbf{0}, \mathbf{B}_\alpha^{(k)}), \mathcal{N}(\phi_\alpha | \mathbf{m}_\alpha^{(k)}, \mathbf{P}_\alpha^{(k)}), \quad (4.13a)$$

$$\mathbf{B}_\alpha^{(k)} = \boldsymbol{\Sigma}_\alpha^{(k)} + \mathbf{A}^{-1}, \quad (4.13b)$$

$$\mathbf{P}_\alpha^{(k)} = ((\boldsymbol{\Sigma}_\alpha^{(k)})^{-1} + \mathbf{A})^{-1}, \quad (4.13c)$$

$$\mathbf{m}_\alpha^{(k)} = \mathbf{P}_\alpha^{(k)} (\boldsymbol{\Sigma}_\alpha^{(k)})^{-1} \boldsymbol{\mu}_\alpha^{(k)}, \quad (4.13d)$$

where $\mathbf{m}_\alpha^{(k)}$ is the posterior mean and $\mathbf{P}_\alpha^{(k)}$ is the posterior covariance of questionable parameters ϕ_α , pertaining to the k^{th} kernel. Notice that $\mathcal{N}(\boldsymbol{\mu}_\alpha^{(k)} | \mathbf{0}, \mathbf{B}_\alpha^{(k)})$ is independent of ϕ and acts as a normalizing factor in Eq. (4.13a). Substituting Eq. (4.13a) in Eq. (4.12) and taking $\mathcal{N}(\boldsymbol{\mu}_\alpha^{(k)} | \mathbf{0}, \mathbf{B}_\alpha^{(k)})$ out of the integral leads to

$$\begin{aligned} \hat{p}(\mathcal{D} | \boldsymbol{\alpha}) &= \sum_{k=1}^K a^{(k)} \int \mathcal{N}(\phi_{-\alpha} | \tilde{\boldsymbol{\mu}}_{-\alpha}^{(k)}, \tilde{\boldsymbol{\Sigma}}_{-\alpha}^{(k)}) \mathcal{N}(\boldsymbol{\mu}_\alpha^{(k)} | \mathbf{0}, \mathbf{B}_\alpha^{(k)}) \mathcal{N}(\phi_\alpha | \mathbf{m}_\alpha^{(k)}, \mathbf{P}_\alpha^{(k)}) d\phi \\ &= \sum_{k=1}^K a^{(k)} \mathcal{N}(\boldsymbol{\mu}_\alpha^{(k)} | \mathbf{0}, \mathbf{B}_\alpha^{(k)}) \int \left\{ \int \mathcal{N}(\phi_{-\alpha} | \tilde{\boldsymbol{\mu}}_{-\alpha}^{(k)}, \tilde{\boldsymbol{\Sigma}}_{-\alpha}^{(k)}) d\phi_{-\alpha} \right\} \mathcal{N}(\phi_\alpha | \mathbf{m}_\alpha^{(k)}, \mathbf{P}_\alpha^{(k)}) d\phi_\alpha \\ &= \sum_{k=1}^K a^{(k)} \mathcal{N}(\boldsymbol{\mu}_\alpha^{(k)} | \mathbf{0}, \mathbf{B}_\alpha^{(k)}) \int \mathcal{N}(\phi_\alpha | \mathbf{m}_\alpha^{(k)}, \mathbf{P}_\alpha^{(k)}) d\phi_\alpha \\ &= \sum_{k=1}^K a^{(k)} \mathcal{N}(\boldsymbol{\mu}_\alpha^{(k)} | \mathbf{0}, \mathbf{B}_\alpha^{(k)}) \end{aligned} \quad (4.14)$$

where $a^{(k)}$ is known from Eq. (4.3), $\boldsymbol{\mu}_\alpha^{(k)}$ is known from Eq. (4.10), and $\mathbf{B}_\alpha^{(k)}$ is known from Eq. (4.13b). As noted from Eq. (4.14), the dependence of model evidence on $\boldsymbol{\alpha}$ is through matrix $\mathbf{B}_\alpha^{(k)} = \boldsymbol{\Sigma}_\alpha^{(k)} + \mathbf{A}^{-1}$ as $\mathbf{A} = \text{Diag}(\boldsymbol{\alpha})$. The entities $a^{(k)}$, $\boldsymbol{\mu}_\alpha^{(k)}$ and $\boldsymbol{\Sigma}_\alpha^{(k)}$ are independent of $\boldsymbol{\alpha}$ and are solely dependent on $p(\mathcal{D} | \phi)p(\phi_{-\alpha})$ based on Eq. (4.3).

In summary, Eq. (4.14) provides a Gaussian kernel-based estimate of model evidence for a given $\boldsymbol{\alpha}$ value. This computation of model evidence for varying $\boldsymbol{\alpha}$ without accessing the likelihood function or the model is a powerful tool in itself. This tool offers a significant computational relief for high-dimensional models (large N_ϕ or large N_α) as it eliminates the need for

performing time-consuming sampling of posterior parameter pdf (like MCMC) for estimating evidence for varying α values.

4.2.4.2 Posterior parameter pdf

The probabilistic predictions of a given quantity-of-interest require the posterior pdf of sparse ϕ identified through sparse learning. An estimate $\hat{p}(\phi|\mathcal{D}, \alpha)$ of the posterior parameter pdf is obtained by substituting $p(\mathcal{D}|\phi)p(\phi_{-\alpha})$ approximation from Eq. (4.3) in Eq. (4.2) to obtain

$$\hat{p}(\phi|\mathcal{D}, \alpha) = \sum_{k=1}^K \frac{a^{(k)} \mathcal{N}(\phi|\boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}) \mathcal{N}(\phi_{-\alpha}|\mathbf{0}, \mathbf{A}^{-1})}{\hat{p}(\mathcal{D}|\alpha)}. \quad (4.15)$$

Substituting $\mathcal{N}(\phi|\boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)})$ from Eq. (4.11a) reduces Eq. (4.15) to

$$\begin{aligned} \hat{p}(\phi|\mathcal{D}, \alpha) &= \sum_{k=1}^K \frac{a^{(k)} \mathcal{N}(\phi_{-\alpha}|\tilde{\boldsymbol{\mu}}_{-\alpha}^{(k)}, \tilde{\boldsymbol{\Sigma}}_{-\alpha}^{(k)}) \mathcal{N}(\phi_{\alpha}|\boldsymbol{\mu}_{\alpha}^{(k)}, \boldsymbol{\Sigma}_{\alpha}^{(k)}) \mathcal{N}(\phi_{\alpha}|\mathbf{0}, \mathbf{A}^{-1})}{\hat{p}(\mathcal{D}|\alpha)} \\ &= \sum_{k=1}^K \underbrace{\left(\frac{a^{(k)} \mathcal{N}(\boldsymbol{\mu}_{\alpha}^{(k)}|\mathbf{0}, \mathbf{B}_{\alpha}^{(k)})}{\sum_{r=1}^K a^{(r)} \mathcal{N}(\boldsymbol{\mu}_{\alpha}^{(r)}|\mathbf{0}, \mathbf{B}_{\alpha}^{(r)})} \right)}_{w^{(k)}} \underbrace{\mathcal{N}(\phi_{-\alpha}|\tilde{\boldsymbol{\mu}}_{-\alpha}^{(k)}, \tilde{\boldsymbol{\Sigma}}_{-\alpha}^{(k)}) \mathcal{N}(\phi_{\alpha}|\mathbf{m}_{\alpha}^{(k)}, \mathbf{P}_{\alpha}^{(k)})}_{\mathcal{N}(\phi|\mathbf{m}^{(k)}, \mathbf{P}^{(k)})} \\ &= \sum_{k=1}^K w^{(k)} \mathcal{N}(\phi|\mathbf{m}^{(k)}, \mathbf{P}^{(k)}), \end{aligned} \quad (4.16)$$

where $1 \geq w^{(k)} \geq 0$ is the weight coefficient, $\mathbf{m}^{(k)}$ is the posterior mean of ϕ , and $\mathbf{P}^{(k)}$ is the posterior covariance of ϕ ; all pertaining to the k^{th} kernel. Notice that the sum of all weight coefficients is always one, i.e. $\sum_k w^{(k)} = 1$.

The posterior pdf from Eq. (4.16) is rewritten in an expanded form as

$$\hat{p}(\phi|\mathcal{D}, \alpha) = \sum_{k=1}^K w^{(k)} \mathcal{N} \left(\left\{ \begin{array}{c} \phi_{\alpha} \\ \phi_{-\alpha} \end{array} \right\} \middle| \left\{ \begin{array}{c} \mathbf{m}_{\alpha}^{(k)} \\ \mathbf{m}_{-\alpha}^{(k)} \end{array} \right\}, \left[\begin{array}{cc} \mathbf{P}_{\alpha}^{(k)} & \mathbf{D}^{(k)} \\ (\mathbf{D}^{(k)})^T & \mathbf{P}_{-\alpha}^{(k)} \end{array} \right] \right), \quad (4.17)$$

where $\mathbf{m}_{\alpha}^{(k)}$ and $\mathbf{P}_{\alpha}^{(k)}$ are the posterior entities pertaining to ϕ_{α} ; $\mathbf{m}_{-\alpha}^{(k)}$ and $\mathbf{P}_{-\alpha}^{(k)}$ pertain to $\phi_{-\alpha}$;

and $\mathbf{D}^{(k)}$ is the posterior cross-covariance of ϕ_α and $\phi_{-\alpha}$.

Using the Woodbury identity (Eq. 156 in [137]), the posterior covariance $\mathbf{P}_\alpha^{(k)}$ of ϕ_α in Eq. (4.13c) can be rewritten as

$$\mathbf{P}_\alpha^{(k)} = \Sigma_\alpha^{(k)} - \Sigma_\alpha^{(k)} (\mathbf{B}_\alpha^{(k)})^{-1} \Sigma_\alpha^{(k)} \quad (4.18)$$

Using this expansion, the posterior mean $\mathbf{m}_\alpha^{(k)}$ of ϕ_α in Eq. (4.13d) can be rewritten as

$$\begin{aligned} \mathbf{m}_\alpha^{(k)} &= \mathbf{P}_\alpha^{(k)} (\Sigma_\alpha^{(k)})^{-1} \boldsymbol{\mu}_\alpha^{(k)} = (\Sigma_\alpha^{(k)} - \Sigma_\alpha^{(k)} (\mathbf{B}_\alpha^{(k)})^{-1} \Sigma_\alpha^{(k)}) (\Sigma_\alpha^{(k)})^{-1} \boldsymbol{\mu}_\alpha^{(k)} \\ &= \boldsymbol{\mu}_\alpha^{(k)} - \Sigma_\alpha^{(k)} (\mathbf{B}_\alpha^{(k)})^{-1} \boldsymbol{\mu}_\alpha^{(k)}, \end{aligned} \quad (4.19)$$

where $\mathbf{B}_\alpha^{(k)} = \Sigma_\alpha^{(k)} + \mathbf{A}^{-1}$ is known from Eq. (4.13b). Notice that matrix $\Sigma_\alpha^{(k)} (\mathbf{B}_\alpha^{(k)})^{-1}$ or $\Sigma_\alpha^{(k)} (\Sigma_\alpha^{(k)} + \mathbf{A}^{-1})^{-1}$ holds a special significance in Eq. (4.19) since it computes the ratio of contribution by measurements (through $\Sigma_\alpha^{(k)}$) in the posterior pdf (as in $\Sigma_\alpha^{(k)} + \mathbf{A}^{-1}$). This matrix $\Sigma_\alpha^{(k)} (\mathbf{B}_\alpha^{(k)})^{-1}$ plays a central role in determining the relevance of questionable parameters following the estimation of hyperparameters. More details regarding this in Sec. 4.2.5.

Next, using the law of total expectation [62], the posterior mean $\mathbf{m}_\alpha^{(k)}$ of ϕ_α is computed by taking the expectation of conditional mean $\tilde{\boldsymbol{\mu}}_{-\alpha}^{(k)}$ in Eq. (4.11b) as

$$\mathbf{m}_\alpha^{(k)} = \mathbb{E}_k [\mathbb{E}_k [\phi_{-\alpha} | \phi_\alpha]] = \mathbb{E}_k [\tilde{\boldsymbol{\mu}}_{-\alpha}^{(k)}] = \mathbb{E}_k [\boldsymbol{\mu}_{-\alpha}^{(k)} + (\mathbf{C}^{(k)})^T (\Sigma_\alpha^{(k)})^{-1} (\phi_\alpha - \boldsymbol{\mu}_\alpha^{(k)})] \quad (4.20a)$$

$$= \boldsymbol{\mu}_{-\alpha}^{(k)} + (\mathbf{C}^{(k)})^T (\Sigma_\alpha^{(k)})^{-1} (\mathbf{m}_\alpha^{(k)} - \boldsymbol{\mu}_\alpha^{(k)}), \quad (4.20b)$$

where the inner expectation is with respect to $\phi_{-\alpha}$ and the outer expectation is with respect to ϕ_α . Substituting $\mathbf{m}_\alpha^{(k)}$ from Eq. (4.19) in Eq. (4.20b) leads to

$$\begin{aligned} \mathbf{m}_\alpha^{(k)} &= \boldsymbol{\mu}_{-\alpha}^{(k)} + (\mathbf{C}^{(k)})^T (\Sigma_\alpha^{(k)})^{-1} (\cancel{\boldsymbol{\mu}_\alpha^{(k)}} - \Sigma_\alpha^{(k)} (\mathbf{B}_\alpha^{(k)})^{-1} \boldsymbol{\mu}_\alpha^{(k)} - \cancel{\boldsymbol{\mu}_\alpha^{(k)}}) \\ &= \boldsymbol{\mu}_{-\alpha}^{(k)} - \mathbf{C}^{(k)} (\mathbf{B}_\alpha^{(k)})^{-1} \boldsymbol{\mu}_\alpha^{(k)} \end{aligned} \quad (4.21)$$

Rewriting Eq. (4.19) and Eq. (4.21) in a matrix form produces the solution to posterior mean vector $\mathbf{m}^{(k)}$ as

$$\mathbf{m}^{(k)} = \begin{Bmatrix} \mathbf{m}_\alpha^{(k)} \\ \mathbf{m}_{-\alpha}^{(k)} \end{Bmatrix} = \begin{Bmatrix} \boldsymbol{\mu}_\alpha^{(k)} \\ \boldsymbol{\mu}_{-\alpha}^{(k)} \end{Bmatrix} - \begin{bmatrix} \boldsymbol{\Sigma}_\alpha^{(k)} (\mathbf{B}_\alpha^{(k)})^{-1} & \mathbf{0} \\ (\mathbf{C}^{(k)})^T (\mathbf{B}_\alpha^{(k)})^{-1} & \mathbf{0} \end{bmatrix} \begin{Bmatrix} \boldsymbol{\mu}_\alpha^{(k)} \\ \boldsymbol{\mu}_{-\alpha}^{(k)} \end{Bmatrix}. \quad (4.22)$$

Next, the posterior covariance $\mathbf{P}_{-\alpha}^{(k)}$ of $\phi_{-\alpha}$ in Eq. (4.17) is computed using the law of total variance [62] as

$$\begin{aligned} \mathbf{P}_{-\alpha}^{(k)} &= \text{Var}_k(\phi_{-\alpha}) = \text{Var}_k(\mathbb{E}_k[\phi_{-\alpha}|\phi_\alpha]) + \mathbb{E}_k[\text{Var}_k(\phi_{-\alpha}|\phi_\alpha)] = \text{Var}_k(\tilde{\boldsymbol{\mu}}_{-\alpha}^{(k)}) + \mathbb{E}_k[\tilde{\boldsymbol{\Sigma}}_{-\alpha}^{(k)}] \\ &= \text{Var}_k(\boldsymbol{\mu}_{-\alpha}^{(k)} + (\mathbf{C}^{(k)})^T (\boldsymbol{\Sigma}_\alpha^{(k)})^{-1} (\phi_\alpha - \boldsymbol{\mu}_\alpha^{(k)})) + \mathbb{E}_k[\boldsymbol{\Sigma}_{-\alpha}^{(k)} - (\mathbf{C}^{(k)})^T (\boldsymbol{\Sigma}_\alpha^{(k)})^{-1} \mathbf{C}^{(k)}] \\ &= (\mathbf{C}^{(k)})^T (\boldsymbol{\Sigma}_\alpha^{(k)})^{-1} \text{Var}_k(\phi_\alpha) (\boldsymbol{\Sigma}_\alpha^{(k)})^{-1} \mathbf{C}^{(k)} + \boldsymbol{\Sigma}_{-\alpha}^{(k)} - (\mathbf{C}^{(k)})^T (\boldsymbol{\Sigma}_\alpha^{(k)})^{-1} \mathbf{C}^{(k)} \\ &= \boldsymbol{\Sigma}_{-\alpha}^{(k)} + (\mathbf{C}^{(k)})^T (\boldsymbol{\Sigma}_\alpha^{(k)})^{-1} (\mathbf{P}_\alpha^{(k)} (\boldsymbol{\Sigma}_\alpha^{(k)})^{-1} - \mathbf{I}_{N_\alpha}) \mathbf{C}^{(k)}, \end{aligned} \quad (4.23)$$

where $\tilde{\boldsymbol{\mu}}_{-\alpha}^{(k)}$ and $\tilde{\boldsymbol{\Sigma}}_{-\alpha}^{(k)}$ are substituted from Eq. (4.11). Substituting $\mathbf{P}_\alpha^{(k)}$ from Eq. (4.18) in Eq. (4.23) results in

$$\mathbf{P}_{-\alpha}^{(k)} = \boldsymbol{\Sigma}_{-\alpha}^{(k)} + (\mathbf{C}^{(k)})^T (\boldsymbol{\Sigma}_\alpha^{(k)})^{-1} (-\boldsymbol{\Sigma}_\alpha^{(k)} (\mathbf{B}_\alpha^{(k)})^{-1}) \mathbf{C}^{(k)} = \boldsymbol{\Sigma}_{-\alpha}^{(k)} - (\mathbf{C}^{(k)})^T (\mathbf{B}_\alpha^{(k)})^{-1} \mathbf{C}^{(k)} \quad (4.24)$$

Next, the posterior cross-covariance matrix $\mathbf{D}^{(k)}$ in Eq. (4.17) is evaluated as

$$\begin{aligned} \mathbf{D}^{(k)} &= \mathbb{E}_k \left[(\phi_\alpha - \mathbf{m}_\alpha^{(k)}) (\phi_{-\alpha} - \mathbf{m}_{-\alpha}^{(k)})^T \right] \\ &= \mathbb{E}_k \left[(\phi_\alpha - \boldsymbol{\mu}_\alpha^{(k)} + \boldsymbol{\Sigma}_\alpha^{(k)} (\mathbf{B}_\alpha^{(k)})^{-1} \boldsymbol{\mu}_\alpha^{(k)}) (\phi_{-\alpha} - \boldsymbol{\mu}_{-\alpha}^{(k)} + \mathbf{C}^{(k)} (\mathbf{B}_\alpha^{(k)})^{-1} \boldsymbol{\mu}_\alpha^{(k)})^T \right] \\ &= \mathbf{C}^{(k)} + (\mathbf{m}_\alpha^{(k)} - \boldsymbol{\mu}_\alpha^{(k)}) (\boldsymbol{\mu}_\alpha^{(k)})^T (\mathbf{B}_\alpha^{(k)})^{-1} \mathbf{C}^{(k)} + \boldsymbol{\Sigma}_\alpha^{(k)} (\mathbf{B}_\alpha^{(k)})^{-1} \boldsymbol{\mu}_\alpha^{(k)} (\mathbf{m}_{-\alpha}^{(k)} - \boldsymbol{\mu}_{-\alpha}^{(k)})^T \\ &\quad + \boldsymbol{\Sigma}_\alpha^{(k)} (\mathbf{B}_\alpha^{(k)})^{-1} \boldsymbol{\mu}_\alpha^{(k)} (\boldsymbol{\mu}_\alpha^{(k)})^T (\mathbf{B}_\alpha^{(k)})^{-1} \mathbf{C}^{(k)}. \end{aligned} \quad (4.25)$$

Substituting $\mathbf{m}_\alpha^{(k)}$ from Eq. (4.19) and $\mathbf{m}_{-\alpha}^{(k)}$ from Eq. (4.21) reduces Eq. (4.25) to

$$\begin{aligned} \mathbf{D}^{(k)} &= \mathbf{C}^{(k)} - \Sigma_\alpha^{(k)} (\mathbf{B}_\alpha^{(k)})^{-1} \boldsymbol{\mu}_\alpha^{(k)} (\boldsymbol{\mu}_\alpha^{(k)})^T (\mathbf{B}_\alpha^{(k)})^{-1} \mathbf{C}^{(k)} - \Sigma_\alpha^{(k)} (\mathbf{B}_\alpha^{(k)})^{-1} \boldsymbol{\mu}_\alpha^{(k)} (\boldsymbol{\mu}_\alpha^{(k)})^T (\mathbf{B}_\alpha^{(k)})^{-1} \mathbf{C}^{(k)} \\ &\quad + \Sigma_\alpha^{(k)} (\mathbf{B}_\alpha^{(k)})^{-1} \boldsymbol{\mu}_\alpha^{(k)} (\boldsymbol{\mu}_\alpha^{(k)})^T (\mathbf{B}_\alpha^{(k)})^{-1} \mathbf{C}^{(k)} \\ &= \mathbf{C}^{(k)} - \Sigma_\alpha^{(k)} (\mathbf{B}_\alpha^{(k)})^{-1} \boldsymbol{\mu}_\alpha^{(k)} ((\mathbf{C}^{(k)})^T (\mathbf{B}_\alpha^{(k)})^{-1} \boldsymbol{\mu}_\alpha^{(k)})^T \end{aligned} \quad (4.26)$$

This completes the calculation of posterior pdf $\hat{p}(\phi|\mathcal{D}, \alpha)$ in Eq. (4.17), wherein $\mathbf{m}^{(k)}$ is known from Eq. (4.22), $\mathbf{P}_\alpha^{(k)}$ is known from Eq. (4.18), $\mathbf{P}_{-\alpha}^{(k)}$ is known from Eq. (4.23), and $\mathbf{D}^{(k)}$ is known from Eq. (4.26). Notice that when using flat prior for questionable parameters (i.e. prior precision $\alpha \approx \mathbf{0}$), we get $\mathbf{m}_\alpha^{(k)} \approx \boldsymbol{\mu}_\alpha^{(k)}$, $\mathbf{m}_{-\alpha}^{(k)} \approx \boldsymbol{\mu}_{-\alpha}^{(k)}$, $\mathbf{P}_\alpha^{(k)} \approx \Sigma_\alpha^{(k)}$, $\mathbf{P}_{-\alpha}^{(k)} \approx \Sigma_{-\alpha}^{(k)}$ and $\mathbf{D}^{(k)} \approx \mathbf{C}^{(k)}$.

4.2.4.3 Gradient vector

For the sake of gradient and Hessian calculations, the optimization problem in Eq. (4.5) is reposed in terms of the logarithm of the posterior pdf $p(\alpha|\mathcal{D})$ in Eq. (4.4). In addition, the optimization is performed with respect to $\log \alpha$ instead of α . This change of variable facilitates straightforward analytical differentiation, while ensuring a well-behaved numerical computation of the gradient and Hessian information. Also, optimizing $\log \alpha$ over α automatically enforces the positivity constraint of the precision parameter α during the optimization in Eq. (4.5). Following these modifications, the sparse learning optimization problem in Eq. (4.4) is rewritten as

$$\begin{aligned} \log \alpha^{\text{MAP}} &= \arg \max_{\log \alpha} \{\mathcal{L}(\log \alpha)\} = \arg \max_{\log \alpha} \{\log p(\log \alpha|\mathcal{D})\} \\ &= \arg \max_{\log \alpha} \{\log \hat{p}(\mathcal{D}|\log \alpha) + \log p(\log \alpha)\}, \end{aligned} \quad (4.27)$$

where the model evidence estimate $\hat{p}(\mathcal{D}|\log \alpha)$ is available from Eq. (4.14), and $\mathcal{L}(\log \alpha)$ is the objective function that needs to be optimized for identifying the sparse structure of ϕ . Given

$p(\boldsymbol{\alpha})$ in Eq. (4.6), hyperprior $p(\log \boldsymbol{\alpha})$ required in Eq. (4.27) is obtained using the univariate transformation of random variables as [62]

$$p(\log \boldsymbol{\alpha}) = \prod_{i=1}^{N_\alpha} p(\log \alpha_i) = \prod_{i=1}^{N_\alpha} \frac{s_i^{r_i}}{\Gamma(r_i)} \alpha_i^{r_i} e^{-s_i \alpha_i}. \quad (4.28)$$

Subsequently, the objective function $\mathcal{L}(\log \boldsymbol{\alpha})$ from Eq. (4.27) is rewritten using Eq. (4.28) as (ignoring constant terms)

$$\mathcal{L}(\log \boldsymbol{\alpha}) = \log \hat{p}(\mathcal{D} | \log \boldsymbol{\alpha}) + \sum_{i=1}^{N_\alpha} (r_i \log \alpha_i - s_i \alpha_i). \quad (4.29)$$

Notice that when using Jeffrey's prior ($p(\log \alpha_i) \propto 1$) for α_i , hyperprior parameters r_i and s_i are close to zero, and the objective function in Eq. (4.29) is solely dictated by the log-evidence estimator $\log \hat{p}(\mathcal{D} | \log \boldsymbol{\alpha})$.

The i^{th} element of the gradient vector $\mathbf{J}(\log \boldsymbol{\alpha})$, denoted as $J_i(\log \boldsymbol{\alpha})$, is obtained by differentiating Eq. (4.29) with respect to $\log \alpha_i$ as

$$\begin{aligned} J_i(\log \boldsymbol{\alpha}) &= \frac{\partial \mathcal{L}(\log \boldsymbol{\alpha})}{\partial \log \alpha_i} = \frac{\partial}{\partial \log \alpha_i} \left\{ \log \hat{p}(\mathcal{D} | \log \boldsymbol{\alpha}) + \sum_{i=1}^{N_\alpha} (r_i \log \alpha_i - s_i \alpha_i) \right\} \\ &= \frac{\partial \log \hat{p}(\mathcal{D} | \log \boldsymbol{\alpha})}{\partial \log \alpha_i} + r_i - s_i \alpha_i. \end{aligned} \quad (4.30)$$

The log-evidence $\log \hat{p}(\mathcal{D} | \log \boldsymbol{\alpha})$ from Eq. (4.14) is differentiated with respect to $\log \alpha_i$ to obtain

$$\begin{aligned} \frac{\partial \log \hat{p}(\mathcal{D} | \log \boldsymbol{\alpha})}{\partial \log \alpha_i} &= \frac{1}{\hat{p}(\mathcal{D} | \log \boldsymbol{\alpha})} \frac{\partial}{\partial \log \alpha_i} \left\{ \sum_{k=1}^K a^{(k)} \mathcal{N}(\boldsymbol{\mu}_\alpha^{(k)} | \mathbf{0}, \mathbf{B}_\alpha^{(k)}) \right\} \\ &= \frac{1}{\hat{p}(\mathcal{D} | \log \boldsymbol{\alpha})} \sum_{k=1}^K a^{(k)} \frac{\partial}{\partial \log \alpha_i} \{ \exp(\log \mathcal{N}(\boldsymbol{\mu}_\alpha^{(k)} | \mathbf{0}, \mathbf{B}_\alpha^{(k)})) \} \\ &= \sum_{k=1}^K \underbrace{\left(\frac{a^{(k)} \mathcal{N}(\boldsymbol{\mu}_\alpha^{(k)} | \mathbf{0}, \mathbf{B}_\alpha^{(k)})}{\hat{p}(\mathcal{D} | \log \boldsymbol{\alpha})} \right)}_{w^{(k)}} \underbrace{\left(\frac{\partial \log \mathcal{N}(\boldsymbol{\mu}_\alpha^{(k)} | \mathbf{0}, \mathbf{B}_\alpha^{(k)})}{\partial \log \alpha_i} \right)}_{v_i^{(k)}}, \end{aligned} \quad (4.31)$$

where the weight coefficient $w^{(k)}$ has been previously defined in Eq. (4.16) and factor $v_i^{(k)}$ is defined here for simplicity in the ensuing mathematical exposition. $v_i^{(k)}$ in Eq. (4.31) is further evaluated as

$$v_i^{(k)} = -\frac{1}{2} \left\{ \frac{\partial \log |\mathbf{B}_\alpha^{(k)}|}{\partial \log \alpha_i} + (\boldsymbol{\mu}_\alpha^{(k)})^T \frac{\partial (\mathbf{B}_\alpha^{(k)})^{-1}}{\partial \log \alpha_i} \boldsymbol{\mu}_\alpha^{(k)} \right\}. \quad (4.32)$$

An identity involving the derivative of a matrix determinant is available as (Eq. (46) in [137])

$$\frac{\partial \log |\mathbf{Z}|}{\partial x} = \text{Trace} \left(\mathbf{Z}^{-1} \frac{\partial \mathbf{Z}}{\partial x} \right), \quad (4.33)$$

where $\mathbf{Z}$ is any square matrix and x is a scalar. The derivative of term $\log |\mathbf{B}_\alpha^{(k)}|$ in Eq. (4.32) is evaluated using this identity as

$$\frac{\partial \log |\mathbf{B}_\alpha^{(k)}|}{\partial \log \alpha_i} = \alpha_i \frac{\partial \log |\mathbf{B}_\alpha^{(k)}|}{\partial \alpha_i} = \alpha_i \text{Trace} \left((\mathbf{B}_\alpha^{(k)})^{-1} \frac{\partial (\boldsymbol{\Sigma}_\alpha^{(k)} + \mathbf{A}^{-1})}{\partial \alpha_i} \right), \quad (4.34)$$

where $\mathbf{B}_\alpha^{(k)} = \boldsymbol{\Sigma}_\alpha^{(k)} + \mathbf{A}^{-1}$ is substituted from Eq. (4.13b). As observed from Eq. (4.3), the matrix $\boldsymbol{\Sigma}_\alpha^{(k)}$ is solely dictated by the measurements $\mathcal{D}$ and the known prior $p(\phi_{-\alpha})$, and is not a function of α_i . On the other hand, matrix $\mathbf{A} = \text{Diag}(\boldsymbol{\alpha})$ is a diagonal matrix with $\boldsymbol{\alpha}$ as its diagonal. Hence, differentiating $\mathbf{A}^{-1} = \text{Diag}(1/\boldsymbol{\alpha})$ with respect to α_i produces $(-1/\alpha_i^2) \boldsymbol{\Delta}_{ii}$, where the matrix $\boldsymbol{\Delta}_{ii} \in \mathbb{R}^{N_\alpha \times N_\alpha}$ has only (i, i) element equal to one and the rest of the elements are all zeros. Consequently, Eq. (4.34) can be evaluated as

$$\frac{\partial \log |\mathbf{B}_\alpha^{(k)}|}{\partial \log \alpha_i} = \alpha_i \text{Trace} \left((\mathbf{B}_\alpha^{(k)})^{-1} \frac{\partial \mathbf{A}^{-1}}{\partial \alpha_i} \right) = -\frac{1}{\alpha_i} \text{Trace} \left((\mathbf{B}_\alpha^{(k)})^{-1} \boldsymbol{\Delta}_{ii} \right). \quad (4.35)$$

Further, the matrix $(\mathbf{B}_\alpha^{(k)})^{-1}$ from Eq. (4.13d) is expanded using the Woodbury identity (Eq. 156 in [137]) to produce

$$(\mathbf{B}_\alpha^{(k)})^{-1} = (\boldsymbol{\Sigma}_\alpha^{(k)} + \mathbf{A}^{-1})^{-1} = \mathbf{A} - \mathbf{A} \mathbf{P}_\alpha^{(k)} \mathbf{A}, \quad (4.36)$$

where $\mathbf{P}_\alpha^{(k)}$ is known from Eq. (4.13b). Substituting Eq. (4.36) in Eq. (4.35) leads to

$$\begin{aligned} \frac{\partial \log |\mathbf{B}_\alpha^{(k)}|}{\partial \log \alpha_i} &= -\frac{1}{\alpha_i} \text{Trace} (\mathbf{A} \Delta_{ii} - \mathbf{A} \mathbf{P}_\alpha^{(k)} \mathbf{A} \Delta_{ii}) \\ &= -\frac{1}{\alpha_i} \left(\alpha_i - \alpha_i^2 P_{ii}^{(k)} \right) = -1 + \alpha_i P_{ii}^{(k)}, \end{aligned} \quad (4.37)$$

where the variance $P_{ii}^{(k)}$ is the (i, i) element of matrix $\mathbf{P}_\alpha^{(k)}$ or $\mathbf{P}^{(k)}$ (see Eq. (4.17)).

Next, the derivative of $(\mathbf{B}_\alpha^{(k)})^{-1}$ in Eq. (4.32) is evaluated by expanding $(\mathbf{B}_\alpha^{(k)})^{-1}$ using the Woodbury identity [137] as

$$(\mathbf{B}_\alpha^{(k)})^{-1} = (\Sigma_\alpha^{(k)} + \mathbf{A}^{-1})^{-1} = (\Sigma_\alpha^{(k)})^{-1} - (\Sigma_\alpha^{(k)})^{-1} \mathbf{P}_\alpha^{(k)} (\Sigma_\alpha^{(k)})^{-1}, \quad (4.38)$$

where only the posterior covariance $\mathbf{P}_\alpha^{(k)}$ depends on α_i . As a result,

$$\begin{aligned} (\boldsymbol{\mu}_\alpha^{(k)})^T \frac{\partial (\mathbf{B}_\alpha^{(k)})^{-1}}{\partial \log \alpha_i} \boldsymbol{\mu}_\alpha^{(k)} &= \alpha_i (\boldsymbol{\mu}_\alpha^{(k)})^T \left\{ \frac{\partial}{\partial \alpha_i} \left((\Sigma_\alpha^{(k)})^{-1} - (\Sigma_\alpha^{(k)})^{-1} \mathbf{P}_\alpha^{(k)} (\Sigma_\alpha^{(k)})^{-1} \right) \right\} \boldsymbol{\mu}_\alpha^{(k)} \\ &= -\alpha_i (\boldsymbol{\mu}_\alpha^{(k)})^T (\Sigma_\alpha^{(k)})^{-1} \frac{\partial \mathbf{P}_\alpha^{(k)}}{\partial \alpha_i} (\Sigma_\alpha^{(k)})^{-1} \boldsymbol{\mu}_\alpha^{(k)}. \end{aligned} \quad (4.39)$$

The derivative of $\mathbf{P}_\alpha^{(k)}$ is evaluated by using an identity involving the derivative of matrix inverse: (Eq. (59) in [137]),

$$\frac{\partial \mathbf{Z}^{-1}}{\partial x} = -\mathbf{Z}^{-1} \frac{\partial \mathbf{Z}}{\partial x} \mathbf{Z}^{-1}, \quad (4.40)$$

where $\mathbf{Z}$ is any square matrix and x is a scalar. Using this identity in Eq. (4.39) leads to

$$\begin{aligned} (\boldsymbol{\mu}_\alpha^{(k)})^T \frac{\partial (\mathbf{B}_\alpha^{(k)})^{-1}}{\partial \log \alpha_i} \boldsymbol{\mu}_\alpha^{(k)} &= -\alpha_i (\boldsymbol{\mu}_\alpha^{(k)})^T (\Sigma_\alpha^{(k)})^{-1} \left\{ -\mathbf{P}_\alpha^{(k)} \frac{\partial (\mathbf{P}_\alpha^{(k)})^{-1}}{\partial \alpha_i} \mathbf{P}_\alpha^{(k)} \right\} (\Sigma_\alpha^{(k)})^{-1} \boldsymbol{\mu}_\alpha^{(k)} \\ &= \alpha_i \left\{ \mathbf{P}_\alpha^{(k)} (\Sigma_\alpha^{(k)})^{-1} \boldsymbol{\mu}_\alpha^{(k)} \right\}^T \left\{ \frac{\partial}{\partial \alpha_i} \left((\Sigma_\alpha^{(k)})^{-1} + \mathbf{A} \right) \right\} \left\{ \mathbf{P}_\alpha^{(k)} (\Sigma_\alpha^{(k)})^{-1} \boldsymbol{\mu}_\alpha^{(k)} \right\} \\ &= \alpha_i (\mathbf{m}_\alpha^{(k)})^T \Delta_{ii} \mathbf{m}_\alpha^{(k)} = \alpha_i (m_i^{(k)})^2, \end{aligned} \quad (4.41)$$

where $(\mathbf{P}_\alpha^{(k)})^{-1} = (\Sigma_\alpha^{(k)})^{-1} + \mathbf{A}$ is substituted from Eq. (4.13b), $\mathbf{m}_\alpha^{(k)} = \mathbf{P}_\alpha^{(k)} (\Sigma_\alpha^{(k)})^{-1} \boldsymbol{\mu}_\alpha^{(k)}$ is sub-

stituted from Eq. (4.13d), and $m_i^{(k)}$ is the i^{th} element of posterior mean vector $\mathbf{m}_\alpha^{(k)}$. Substituting Eq. (4.37) and Eq. (4.41) in Eq. (4.32) leads to

$$v_i^{(k)} = -\frac{1}{2} \left\{ -1 + \alpha_i P_{ii}^{(k)} + \alpha_i (m_i^{(k)})^2 \right\} = \frac{\gamma_i^{(k)} - \alpha_i (m_i^{(k)})^2}{2}, \quad (4.42)$$

where $\gamma_i^{(k)} = 1 - \alpha_i P_{ii}^{(k)}$ is defined as the relevance indicator for parameter ϕ_i corresponding to the k^{th} kernel. As demonstrated later in Sec. 4.2.5, the relevance indicator $\gamma_i^{(k)}$ provides a quantitative measure for determining the relevancy of ϕ_i for a given value of α_i . Nevertheless, the gradient $J_i(\log \alpha)$ in Eq. (4.30) is computed using Eq. (4.31) and Eq. (4.42) as

$$J_i(\log \alpha) = \left\{ \sum_{k=1}^K \left(\frac{a^{(k)} \mathcal{N}(\boldsymbol{\mu}_\alpha^{(k)} | \mathbf{0}, \mathbf{B}_\alpha^{(k)})}{\hat{\mathbf{p}}(\mathcal{D} | \log \alpha)} \right) \left(\frac{\gamma_i^{(k)} - \alpha_i (m_i^{(k)})^2}{2} \right) \right\} + r_i - s_i \alpha_i \quad (4.43a)$$

$$= \left\{ \sum_{k=1}^K w^{(k)} v_i^{(k)} \right\} + r_i - s_i \alpha_i = \bar{v}_i + r_i - s_i \alpha_i, \quad (4.43b)$$

where $\bar{v}_i$ is defined as the weighted average of $v_i^{(k)}$ across all kernels, weighted according to weights $w^{(k)}$ defined in Eq. (4.31). In summary, Eq. (4.43) provides an analytical solution to the gradient vector $\mathbf{J}(\log \alpha)$ of the objective function $\mathcal{L}(\log \alpha)$ in Eq. (4.29) with respect to $\log \alpha$.

4.2.4.4 Hessian matrix

The (i, j) element of the Hessian matrix $\mathbf{H}(\log \alpha)$, denoted as $H_{ij}(\log \alpha)$, is evaluated by differentiating Eq. (4.43b) as

$$\begin{aligned} H_{ij}(\log \alpha) &= \frac{\partial^2 \mathcal{L}(\log \alpha)}{\partial \log \alpha_i \partial \log \alpha_j} = \frac{\partial J_j(\log \alpha)}{\partial \log \alpha_i} = \frac{\partial}{\partial \log \alpha_i} \left\{ \sum_{k=1}^K w^{(k)} v_j^{(k)} + r_j - s_j \alpha_j \right\} \\ &= \sum_{k=1}^K \left\{ w^{(k)} \frac{\partial v_j^{(k)}}{\partial \log \alpha_i} + v_j^{(k)} \frac{\partial w^{(k)}}{\partial \log \alpha_i} \right\} - \delta_{ij} s_i \alpha_i \end{aligned} \quad (4.44)$$

where $\delta_{ij} = 1$ when $i = j$ and $\delta_{ij} = 0$ when $i \neq j$. The derivative of the factor $v_j^{(k)}$ in Eq. (4.44) is obtained using Eq. (4.42) as

$$\frac{\partial v_j^{(k)}}{\partial \log \alpha_i} = \frac{\alpha_i}{2} \left\{ -\alpha_j \left(\frac{\partial P_{jj}^{(k)}}{\partial \alpha_i} + \frac{\partial (m_j^{(k)})^2}{\partial \alpha_i} \right) - \delta_{ij} \left(P_{ii}^{(k)} + (m_i^{(k)})^2 \right) \right\}. \quad (4.45)$$

Note that Δ_{ii} matrix can also be written as $\delta_i \delta_i^T$ where $\delta_i \in \mathbb{R}^{N_\alpha}$ has the i^{th} element as one and rest are all zeros. Using this definition, the derivative of the posterior variance $P_{jj}^{(k)}$ in Eq. (4.45) is computed as

$$\begin{aligned} \frac{\partial P_{jj}^{(k)}}{\partial \alpha_i} &= \frac{\partial}{\partial \alpha_i} (\delta_j^T \mathbf{P}_\alpha^{(k)} \delta_j) = \delta_j^T \left(\frac{\partial}{\partial \alpha_i} \mathbf{P}_\alpha^{(k)} \right) \delta_j = \delta_j^T (-\mathbf{P}_\alpha^{(k)} \Delta_{ii} \mathbf{P}_\alpha^{(k)}) \delta_j \\ &= -(\delta_j^T \mathbf{P}_\alpha^{(k)} \delta_i) (\delta_i^T \mathbf{P}_\alpha^{(k)} \delta_j) = -P_{ji}^{(k)} P_{ij}^{(k)} = -(P_{ij}^{(k)})^2, \end{aligned} \quad (4.46)$$

where $P_{ij}^{(k)}$ is the (i, j) element of matrix $\mathbf{P}_\alpha^{(k)}$ and $\partial \mathbf{P}_\alpha^{(k)} / \partial \alpha_i = -\mathbf{P}_\alpha^{(k)} \Delta_{ii} \mathbf{P}_\alpha^{(k)}$ has been previously evaluated in Eq. (4.41). Next, the derivative of $(m_j^{(k)})^2$ in Eq. (4.45) is evaluated using Eq. (4.13c) as

$$\begin{aligned} \frac{\partial (m_j^{(k)})^2}{\partial \alpha_i} &= \frac{\partial}{\partial \alpha_i} ((\mathbf{m}_\alpha^{(k)})^T \Delta_{jj} \mathbf{m}_\alpha^{(k)}) = 2(\mathbf{m}_\alpha^{(k)})^T \Delta_{jj} \frac{\partial \mathbf{m}_\alpha^{(k)}}{\partial \alpha_i} \\ &= 2(\mathbf{m}_\alpha^{(k)})^T \Delta_{jj} \frac{\partial \mathbf{P}_\alpha^{(k)}}{\partial \alpha_i} (\Sigma_\alpha^{(k)})^{-1} \boldsymbol{\mu}_\alpha^{(k)} = 2(\mathbf{m}_\alpha^{(k)})^T \Delta_{jj} (-\mathbf{P}_\alpha^{(k)} \Delta_{ii} \mathbf{P}_\alpha^{(k)}) (\Sigma_\alpha^{(k)})^{-1} \boldsymbol{\mu}_\alpha^{(k)} \\ &= -2 \underbrace{(\mathbf{m}_\alpha^{(k)})^T \delta_j}_{m_j^{(k)}} \underbrace{\delta_j^T \mathbf{P}_\alpha^{(k)} \delta_i}_{P_{ij}^{(k)}} \underbrace{\delta_i^T \mathbf{P}_\alpha^{(k)} (\Sigma_\alpha^{(k)})^{-1} \boldsymbol{\mu}_\alpha^{(k)}}_{m_i^{(k)}} = -2m_i^{(k)} m_j^{(k)} P_{ij}^{(k)} \end{aligned} \quad (4.47)$$

Eq. (4.45) is then re-written using Eq. (4.46) and Eq. (4.47) as

$$\begin{aligned} \frac{\partial v_j^{(k)}}{\partial \log \alpha_i} &= \frac{\alpha_i}{2} \left\{ -\alpha_j \left(-(P_{ij}^{(k)})^2 - 2m_i^{(k)} m_j^{(k)} P_{ij}^{(k)} \right) - \delta_{ij} \left(P_{ii}^{(k)} + (m_i^{(k)})^2 \right) \right\} \\ &= \alpha_i \alpha_j \left(\frac{(P_{ij}^{(k)})^2}{2} + m_i^{(k)} m_j^{(k)} P_{ij}^{(k)} \right) + \delta_{ij} \left(\frac{-\alpha_i P_{ii}^{(k)} - \alpha_i (m_i^{(k)})^2}{2} \right) \\ &= \alpha_i \alpha_j \left(\frac{(P_{ij}^{(k)})^2}{2} + m_i^{(k)} m_j^{(k)} P_{ij}^{(k)} \right) + \delta_{ij} \left(v_i^{(k)} - \frac{1}{2} \right) \end{aligned} \quad (4.48)$$

Next, derivative of weight coefficient $w^{(k)}$ in Eq. (4.44) is evaluated by expanding $w^{(k)}$ according to Eq. (4.16) as

$$\begin{aligned}
\frac{\partial w^{(k)}}{\partial \log \alpha_i} &= \frac{\partial}{\partial \log \alpha_i} \left\{ \frac{a^{(k)} \mathcal{N}(\boldsymbol{\mu}_\alpha^{(k)} | \mathbf{0}, \mathbf{B}_\alpha^{(k)})}{\hat{\mathbf{p}}(\mathcal{D} | \log \boldsymbol{\alpha})} \right\} \\
&= \frac{a^{(k)}}{\hat{\mathbf{p}}(\mathcal{D} | \log \boldsymbol{\alpha})} \frac{\partial \mathcal{N}(\boldsymbol{\mu}_\alpha^{(k)} | \mathbf{0}, \mathbf{B}_\alpha^{(k)})}{\partial \log \alpha_i} - \frac{a^{(k)} \mathcal{N}(\boldsymbol{\mu}_\alpha^{(k)} | \mathbf{0}, \mathbf{B}_\alpha^{(k)})}{(\hat{\mathbf{p}}(\mathcal{D} | \log \boldsymbol{\alpha}))^2} \frac{\partial \hat{\mathbf{p}}(\mathcal{D} | \log \boldsymbol{\alpha})}{\partial \log \alpha_i} \\
&= \frac{a^{(k)} \mathcal{N}(\boldsymbol{\mu}_\alpha^{(k)} | \mathbf{0}, \mathbf{B}_\alpha^{(k)})}{\hat{\mathbf{p}}(\mathcal{D} | \log \boldsymbol{\alpha})} \left(\frac{\partial \log \mathcal{N}(\boldsymbol{\mu}_\alpha^{(k)} | \mathbf{0}, \mathbf{B}_\alpha^{(k)})}{\partial \log \alpha_i} - \frac{\partial \log \hat{\mathbf{p}}(\mathcal{D} | \log \boldsymbol{\alpha})}{\partial \log \alpha_i} \right) \\
&= w^{(k)} \left(v_i^{(k)} - \bar{v}_i \right)
\end{aligned} \tag{4.49}$$

where $v_i^{(k)}$ is substitute from Eq. (4.42), and $\bar{v}_i$ is substituted from Eq. (4.43b). Using Eq. (4.48) and Eq. (4.49), the Hessian $H_{ij}(\log \boldsymbol{\alpha})$ from Eq. (4.44) is further evaluated as

$$\begin{aligned}
H_{ij}(\log \boldsymbol{\alpha}) &= \sum_{k=1}^K \left[w^{(k)} \left\{ \alpha_i \alpha_j \left(\frac{(P_{ij}^{(k)})^2}{2} + m_i^{(k)} m_j^{(k)} P_{ij}^{(k)} \right) + \delta_{ij} \left(v_i^{(k)} - \frac{1}{2} \right) \right\} \right. \\
&\quad \left. + v_j^{(k)} \left\{ w^{(k)} \left(v_i^{(k)} - \bar{v}_i \right) \right\} \right] - \delta_{ij} s_i \alpha_i
\end{aligned} \tag{4.50}$$

Re-arranging terms and substituting $\bar{v}_i$ from Eq. (4.43) leads to

$$\begin{aligned}
H_{ij}(\log \boldsymbol{\alpha}) &= \sum_{k=1}^K w^{(k)} \alpha_i \alpha_j \left(\frac{(P_{ij}^{(k)})^2}{2} + m_i^{(k)} m_j^{(k)} P_{ij}^{(k)} \right) + \sum_{k=1}^K w^{(k)} v_i^{(k)} v_j^{(k)} \\
&\quad - \bar{v}_i \sum_{k=1}^K w^{(k)} v_j^{(k)} + \delta_{ij} \left\{ \sum_{k=1}^K w^{(k)} v_i^{(k)} - \frac{1}{2} \sum_{k=1}^K w^{(k)} - s_i \alpha_i \right\} \\
&= \sum_{k=1}^K w^{(k)} \left\{ \alpha_i \alpha_j \left(\frac{(P_{ij}^{(k)})^2}{2} + m_i^{(k)} m_j^{(k)} P_{ij}^{(k)} \right) + v_i^{(k)} v_j^{(k)} - \bar{v}_i \bar{v}_j \right\} \\
&\quad + \delta_{ij} \left\{ \bar{v}_i - \frac{1}{2} - s_i \alpha_i \right\},
\end{aligned} \tag{4.51}$$

where $\bar{v}_i$ or $\bar{v}_j$ is known from Eq. (4.43); $w^{(k)}$ is known from Eq. (4.16); $v_i^{(k)}$ is known from Eq. (4.42); $m_i^{(k)}$ or $m_j^{(k)}$ (elements of $\mathbf{m}_\alpha^{(k)}$) is known from Eq. (4.13d); and $P_{ij}^{(k)}$ (elements of $\mathbf{P}_\alpha^{(k)}$) is known from Eq. (4.13c). This completes the analytical computation of Hessian matrix

$\mathbf{H}(\log \alpha)$ for the objective function $\mathcal{L}(\alpha)$ in Eq. (4.29). Table 4.2 summarizes the kernel-based analytical solution for all Bayesian entities involved in NSBL.

Entity	Solution
Parameter decomposition	$\phi = \{\phi_\alpha, \phi_{-\alpha}\}$
GMM approximation	$p(\mathcal{D} \phi)p(\phi_{-\alpha}) \approx \sum_{k=1}^K a^{(k)} \mathcal{N}(\phi \mu^{(k)}, \Sigma^{(k)})$
Hybrid prior pdf	$p(\phi \alpha) = p(\phi_{-\alpha})p(\phi_\alpha \alpha) = p(\phi_{-\alpha})\mathcal{N}(\phi_\alpha \mathbf{0}, \mathbf{A}^{-1})$
Hyperprior	$p(\alpha) = \prod_{i=1}^{N_\alpha} \mathcal{G}(\alpha_i r_i, s_i)$
Model evidence	$\hat{p}(\mathcal{D} \alpha) = \sum_{k=1}^K a^{(k)} \mathcal{N}(\mu_\alpha^{(k)} \mathbf{0}, \mathbf{B}_\alpha^{(k)}) ; \mathbf{B}_\alpha^{(k)} = \Sigma_\alpha^{(k)} + \mathbf{A}^{-1}$
Parameter posterior pdf	$\hat{p}(\phi \mathcal{D}, \alpha) = \sum_{k=1}^K w^{(k)} \mathcal{N}(\phi \mathbf{m}^{(k)}, \mathbf{P}^{(k)}) ; w^{(k)} = \frac{a^{(k)} \mathcal{N}(\mu_\alpha^{(k)} \mathbf{0}, \mathbf{B}_\alpha^{(k)})}{\hat{p}(\mathcal{D} \log \alpha)}$ $\mathbf{m}^{(k)}$ and $\mathbf{P}^{(k)}$ solution in eqs. (4.13c), (4.13d), (4.17), (4.20b), (4.23) and (4.26)
Objective function	$\mathcal{L}(\log \alpha) = \log \hat{p}(\mathcal{D} \log \alpha) + \sum_{i=1}^{N_\alpha} (r_i \log \alpha_i - s_i \alpha_i)$
Gradient of $\mathcal{L}(\log \alpha)$	$J_i(\log \alpha) = \sum_{k=1}^K w^{(k)} v_i^{(k)} + r_i - s_i \alpha_i = \bar{v}_i + r_i - s_i \alpha_i$ $v_i^{(k)} = (\gamma_i^{(k)} - \alpha_i (m_i^{(k)})^2)/2 ; \gamma_i^{(k)} = 1 - \alpha_i P_{ii}^{(k)}$
Hessian of $\mathcal{L}(\log \alpha)$	$H_{ij}(\log \alpha) = \sum_{k=1}^K w^{(k)} \left\{ \alpha_i \alpha_j \left(\frac{(P_{ij}^{(k)})^2}{2} + m_i^{(k)} m_j^{(k)} P_{ij}^{(k)} \right) + v_i^{(k)} v_j^{(k)} - \bar{v}_i \bar{v}_j \right\}$ $+ \delta_{ij} \left\{ \bar{v}_i - \frac{1}{2} - s_i \alpha_i \right\}$

Table 4.2: Summary of the analytical Bayesian apparatus of NSBL.

4.2.5 Numerical implementation details

NSBL involves following two numerical tasks: 1) Building the GMM approximation for the unnormalized pdf $p(\mathcal{D}|\phi)p(\phi_{-\alpha})$ in Eq. (4.3), and 2) solving the unconstrained, non-convex optimization of $\mathcal{L}(\log \alpha)$ in Eq. (4.27). Lets first focus on the task of building a GMM for $p(\mathcal{D}|\phi)p(\phi_{-\alpha})$. The unknown entities $a^{(k)}$, $\mu_\alpha^{(k)}$, $\Sigma^{(k)}$ and K of the GMM in Eq. (4.3) can be estimated using a set of stationary samples generated from the unnormalized pdf $p(\mathcal{D}|\phi)p(\phi_{-\alpha})$. These samples can be easily generated using an MCMC sampler [46]. For example, RWM [62] and its variants are best suited to sample from unimodal pdfs while TMCMC [48] can sample from multimodal pdfs. Note that this MCMC sampling needs to be executed only once for the

purpose of sparse learning.

A vast literature exists in the machine learning practice on ways to estimate the GMM parameters following the availability of training data (in this case, stationary MCMC samples). KDE is a rudimentary but quick way to construct a GMM for $p(\mathcal{D}|\phi)p(\phi_{-\alpha})$ as its parameters $a^{(k)}, \mu_{\alpha}^{(k)}, \Sigma^{(k)}$ and K are exactly known given the stationary samples [138]. EM algorithm is the preferred algorithm in machine learning for building GMMs with a minimum number (K) of kernels [40]. Pettit and Wilson [139] demonstrated the applicability of variational Bayesian inference as an alternative to the EM algorithm for the case of limited training data. Also, a Monte Carlo based approximation of $p(\mathcal{D}|\phi)p(\phi_{-\alpha})$ can be considered as a special type of GMM where the Gaussian kernels are reduced to Dirac-delta functions centered at individual ϕ values. In this case, evidence from Eq. (4.8) reduces to the average of ARD prior $\mathcal{N}(\phi_{\alpha}|\mathbf{0}, \mathbf{A}^{-1})$ values computed at stationary ϕ samples from $p(\mathcal{D}|\phi)p(\phi_{-\alpha})$. Although computationally efficient, such an approach needs to be carefully adjusted for cases when samples from $p(\mathcal{D}|\phi)p(\phi_{-\alpha})$ are located far from the prior space $\mathcal{N}(\phi_{\alpha}|\mathbf{0}, \mathbf{A}^{-1})$, which in turn varies with changing α value.

In this work, we employ the KDE approach to build the GMM in Eq. (4.3) since the model dimensionality considered in the numerical investigations is manageable. However, we recommend using EM or the variational Bayesian inference approach for large model dimensionality (N_{ϕ}) cases to ensure the computational efficiency of the GMM construction task and the subsequent calculation of NSBL entities from Table 4.2. Note that the sparse learning apparatus is independent of the algorithm used to build the GMM for $p(\mathcal{D}|\phi)p(\phi_{-\alpha})$, and therefore we leave it to the end-user to choose an appropriate algorithm for GMM construction.

The second numerical task involved in NSBL is the optimization of the objective function $\mathcal{L}(\log \alpha)$ in Eq. (4.27). Due to the unconstrained, non-convex nature of the optimization [136], we pursue a multistart Newton's method to estimate α^{MAP} in Eq. (4.27). Newton's method is powered by the analytically available gradient vector $\mathbf{J}(\log \alpha_i)$ and the Hessian matrix $\mathbf{H}(\log \alpha_i)$ from Eq. (4.43) and Eq. (4.51), respectively. Newton's method provides a quadratic rate of convergence and is a faster alternative to other gradient-based algorithms such

as the steepest-descent and the Quasi-Newton methods [140]. Newton's method operates by generating a sequence of iterates $\{\log \alpha_i\}$ which are guaranteed (in local sense) to converge to $\log \alpha^{\text{MAP}}$. For brevity, we denote $\mathcal{L}(\log \alpha_j) = \mathcal{L}_j$, $\mathbf{J}(\log \alpha_j) = \mathbf{J}_j$ and $\mathbf{H}(\log \alpha_j) = \mathbf{H}_j$. A Newton's iteration to obtain the new iterate α_{j+1} is written as

$$\log \alpha_{j+1} = \log \alpha_j + \beta_j \mathbf{p}_j, \text{ where } \mathbf{H}_j \mathbf{p}_j = -\mathbf{J}_j \quad (4.52)$$

and β_j is the step-length determined by satisfying Wolfe, Goldstein, or Armijo backtracking conditions [140]. Depending on the size of α vector (or N_α), the linear system of equations $\mathbf{H}_j \mathbf{p}_j = -\mathbf{J}_j$ in Eq. (4.52) can be solved using direct (i.e. Gauss elimination, Cramer's rule) or iterative (i.e. Jacobi, Gauss-Seidal, Conjugate-gradient) solvers [141, 142]. The choice of the solver is dictated by the dimensionality (in this case N_α) and the kind of matrix $\mathbf{H}_j$. However, for non-convex optimization, the Hessian $\mathbf{H}_j$ in Eq. (4.52) is not guaranteed to be a positive definite matrix or an invertible matrix. As a result, several of the commonly used direct and iterative solvers are inapplicable to solve $\mathbf{H}_j \mathbf{p}_j = -\mathbf{J}_j$ in Eq. (4.52). Nocedal and Wright [140] provide a detailed discussion on many variants of Newton's algorithm designed for solving non-convex optimization problems where the Hessian is not a positive definite matrix. Most of these variants of Newton's method fall under two categories:

1. Modified Newton method [140]: Instead of solving $\mathbf{H}_j \mathbf{p}_j = -\mathbf{J}_j$, the modified Newton method solves $(\mathbf{H}_j + \mathbf{E}_j) \mathbf{p}_j = -\mathbf{J}_j$ where $\mathbf{E}_j$ is an appropriate matrix added to the Hessian matrix $\mathbf{H}_j$ to make it positive definite.
2. Trust-region Newton method: The trust-region approach relies on building a quadratic approximation of $\mathcal{L}(\log \alpha)$ around the current iterate $\log \alpha_i$, and determining an appropriate search direction $\mathbf{p}_j$ by solving a constrained optimization subproblem

$$\min \left(\mathcal{L}_j + \mathbf{J}_j^T \mathbf{p} + \frac{1}{2} \mathbf{p}^T \mathbf{H}_j \mathbf{p} \right) \quad \text{such that} \quad \|\mathbf{p}\| \leq \Delta_j \quad (4.53)$$

where Δ_j is the trust-region radius [143]. The trust-region Newton method does not require the Hessian matrix $\mathbf{H}_j$ to be positive definite, thus making it suitable for non-convex optimization such as Eq. (4.27). See Conn *et al.* [143] for a detailed review of trust-region methods.

In this work, we employ trust-region Newton methods to optimize $\mathcal{L}(\log \alpha)$ for estimating α^{MAP} . In particular, we employ the nearly-exact algorithm [140, 143] which solves the trust-region subproblem of constrained optimization in Eq. (4.53) almost exactly within the trust-region (more details in [143]). This algorithm is suitable when the dimensionality N_α of the optimization is not large and the storage/processing of Hessian is not creating a bottleneck in the optimization of $\mathcal{L}(\log \alpha)$. For large dimensionality N_α cases, algorithms such as trust-Region Newton-Conjugate-Gradient Algorithm and trust-region Truncated Generalized Lanczos are a more efficient alternatives [140]. We use the Scipy library [144] implementation of the trust-region algorithms to execute the Newton iteration. Furthermore, we initiate multiple instances of the optimizer from different starting α values to locate the global optimum α^{MAP} . This multistart of Newton iteration can be easily implemented using parallel computing. In addition, the kernel-based computation of NSBL entities in Table 4.2 can be executed in parallel where each core computes NSBL entities pertaining to a single kernel (not pursued here).

Once the MAP estimate $\log \alpha^{\text{MAP}}$ is computed, the relevance of each questionable parameter $\phi_i \in \phi_\alpha$ needs to be determined using the corresponding $\log \alpha_i^{\text{MAP}}$ values. In SBL [78], the relevance of each questionable parameter ϕ_i was determined using the relevance indicator $\gamma_i = 1 - \alpha_i P_{ii}$ where P_{ii} is the posterior variance of ϕ_i . However, in NSBL, there exist a relevance indicator $\gamma_i^{(k)}$ pertaining to each Gaussian kernel $a^{(k)} \mathcal{N}(\phi | \mu^{(k)}, \Sigma^{(k)})$ in Eq. (4.3). The true nature of $\gamma_i^{(k)}$ is revealed by substituting $P_{ii}^{(k)}$ from Eq. (4.13c) to write

$$\gamma_i^{(k)} = 1 - \frac{\alpha_i}{(P_{ii}^{(k)})^{-1}} = 1 - \frac{\alpha_i}{\alpha_i + \{(\Sigma_\alpha^{(k)})^{-1}\}_{(i,i)}} \in [0, 1], \quad (4.54)$$

where $\{(\Sigma_\alpha^{(k)})^{-1}\}_{(i,i)}$ is the (i, i) element of precision matrix $(\Sigma_\alpha^{(k)})^{-1}$ in Eq. (4.3). For an

irrelevant parameter ϕ_i , the prior precision α_i is large (low prior variance) and the posterior precision is dictated by the prior, i.e. $(P_{ii}^{(k)})^{-1} \approx \alpha_i$ or $\gamma_i^{(k)} \approx 0$. Alternatively, for a relevant parameter ϕ_i , the posterior precision $(P_{ii}^{(k)})^{-1}$ is dictated by the measurements, thereby forcing $P_{ii}^{-1} \approx (\Sigma_i^{(k)})^{-1}$ or $\gamma_i^{(k)} \approx 1$. In other words, $\gamma_i^{(k)}$ indicates the percentage of posterior precision that is due to measurements $\mathcal{D}$ in the k^{th} kernel. Also, $\gamma_i^{(k)}$ varies between zero and one and therefore provides a consistent quantitative measure of relevance for $\phi_i \in \phi_\alpha$ according to the k^{th} kernel.

We extend this idea of relevance indicator for a multi-kernel setting by taking root-mean-square (RMS) of K relevance indicators $\gamma_i^{(k)}$ as

$$\gamma_i^{\text{rms}} = \left(\frac{1}{K} \sum_{k=1}^K (\gamma_i^{(k)})^2 \right)^{1/2} = \left(\frac{1}{K} \sum_{k=1}^K \left(1 - \alpha_i P_{ii}^{(k)} \right)^2 \right)^{1/2}. \quad (4.55)$$

This summarized measure γ_i^{rms} in Eq. (4.55) also varies in range $[0, 1]$ since each $\gamma_i^{(k)} \in [0, 1]$ from Eq. (4.54). A γ_i^{rms} value close to zero implies irrelevance and a γ_i^{rms} value close to one implies relevance. Notice that γ_i^{rms} is a scale-invariant measure since it involves the ratio of prior (α_i) and posterior $((P_{ii}^{(k)})^{-1})$ precision. This is the reason γ_i^{rms} is preferred over α_i values (which are scale-dependent) for monitoring parameter relevance during NSBL. We propose a tolerance γ^{tol} for γ_i^{rms} where a γ_i^{rms} value greater than γ^{tol} implies relevance. The impact of the choice of γ^{tol} on sparsity levels produced by NSBL is investigated in Sec. 4.3.1. Further guidance on selecting an appropriate γ^{tol} value is also provided in Sec. 4.3.4.2. The resulting algorithm of NSBL is summarized in Algorithm 2.

4.2.6 Relation to SBL

In this section, we demonstrate the relationship between NSBL and SBL (or RVM) for the case of linear regression models and Gaussian errors. Consider the model $\mathbf{d} = \Psi\phi + \epsilon$, where $\phi \in \mathbb{R}^{N_\phi}$ is the unknown coefficient vector, $\mathcal{D} \equiv \mathbf{d} \in \mathbb{R}^{N_d \times 1}$ is the measurement vector, $\Psi \in \mathbb{R}^{N_d \times N_\phi}$ is the design matrix, and $\epsilon \sim \mathcal{N}(\mathbf{0}, \rho^{-1} \mathbf{I}_{N_d})$ is the model error where ρ is the error

Algorithm 2 NSBL algorithm

- 1: Decompose ϕ as $\{\phi_\alpha, \phi_{-\alpha}\}$ and assign a known prior pdf $p(\phi_{-\alpha})$;
 - 2: Build a GMM for $p(\mathcal{D}|\phi)p(\phi_{-\alpha})$ by estimating $a^{(k)}, \mu_\alpha^{(k)}, \Sigma^{(k)}$ in Eq. (4.3);
 - 3: Choose a hyperprior $\mathcal{G}(\phi_i|r_i, s_i)$ for each $\phi_i \in \phi_\alpha$ using Table 4.1;
 - 4: Choose a starting hyperparameter value $\log \alpha_0$; set $j=0$;
 - 5: **while** not converged **do**
 - 6: Given α_j , compute $\mathbf{m}^{(k)}, \mathbf{P}^{(k)}, \mathbf{B}_\alpha^{(k)}$ using Eq. (4.13);
 - 7: Compute weight $w^{(k)}$ using Eq. (4.16) and factor $v_i^{(k)}$ using Eq. (4.42);
 - 8: Compute gradient vector $\mathbf{J}(\log \alpha_j)$ using Eq. (4.43);
 - 9: Compute Hessian matrix $\mathbf{H}(\log \alpha_j)$ using Eq. (4.51);
 - 10: Compute the new iterate $\log \alpha_{j+1}$ using Newton's iteration as per Eq. (4.52);
 - 11: Compute relevance indicator γ_i^{rms} using Eq. (4.55); Set $j=j+1$;
 - 12: **end while**
-

precision. SBL operates by conditioning the sparse learning apparatus on ρ , and then estimating it iteratively following each SBL update in α . This explicit conditioning on ρ is not shown here for brevity. SBL treats all parameters in ϕ as questionable, and so $N_\phi=N_\alpha$ and $\phi=\phi_\alpha$ in the NSBL setup outlined in Sec. 4.2.1. Consequently, $\phi_{-\alpha}$ is a null vector and the prior pdf $p(\phi|\alpha)$ in Eq. (4.1) is just the ARD prior $\mathcal{N}(\phi|\mathbf{0}, \mathbf{A}^{-1})$. Under these circumstances, the approximation for $p(\mathcal{D}|\phi)$ in Eq. (4.3) is known analytically from MLE as [40]

$$p(\mathcal{D}|\phi) = \mathcal{N}(\mathbf{d}|\Psi\phi, \rho^{-1}\mathbf{I}_{N_d}) = \hat{p}(\mathcal{D}|\phi) = a^{(1)}\mathcal{N}(\phi|\mu^{(1)}, \Sigma^{(1)}) \quad (4.56)$$

where $\mu^{(1)} = (\Psi^T\Psi)^{-1}\Psi^T\mathbf{d}$ is the MLE estimate, and $\Sigma^{(1)} = (\rho\Psi^T\Psi)^{-1}$ is the covariance matrix of the MLE estimate $\mu^{(1)}$. Note that the sampling distribution $\mathcal{N}(\phi|\mu^{(1)}, \Sigma^{(1)})$ produced from MLE is same as the posterior pdf obtained using Bayesian linear regression with no or flat priors [84]. Given $\mu^{(1)}$ and $\Sigma^{(1)}$, the coefficient $a^{(1)}$ is also available analytically from Eq. (4.56).

The model evidence from Eq. (4.14) is obtained as $p(\mathcal{D}|\alpha)=\mathcal{N}(\mathbf{d}|\mathbf{0}, \Psi\mathbf{A}^{-1}\Psi^T + \rho\mathbf{I}_{N_\phi})$. Given the single kernel ($K=1$) representation of the likelihood function in Eq. (4.56), the only weight coefficient $w^{(1)}$ in Eq. (4.16) is equal to one. The posterior pdf from Eq. (4.17) is available as $p(\phi|\mathcal{D}, \alpha)=\mathcal{N}(\phi|^{(1)}, \mathbf{P}^{(1)})$, where $\mathbf{P}^{(1)}=(\rho\Psi^T\Psi + \mathbf{A})^{-1}$ and $^{(1)}=\rho\mathbf{P}^{(1)}\Psi^T\mathbf{d}$ from

Eq. (4.13). The gradient $J_i(\log \alpha)$ of $\mathcal{L}(\log \alpha)$ in Eq. (4.43a) is obtained as

$$J_i(\log \alpha) = \frac{\gamma_i^{(1)} - \alpha_i(m_i^{(1)})^2}{2} + r_i - s_i\alpha_i \quad (4.57)$$

Setting this gradient $J_i(\log \alpha)$ to zero leads to

$$\alpha_i = \frac{\gamma_i^{(1)} + 2r_i}{(m_i^{(1)})^2 + 2s_i} \quad (4.58)$$

This solution to α_i is exactly the expression exploited in SBL to perform iterative re-estimation of α_i [78]. The posterior entities $^{(1)}$ and $\mathbf{P}^{(1)}$ are updated after each update in α_i . This approach of setting the gradient to zero is motivated by the convex nature of $\mathcal{L}(\log \alpha)$ with respect to α_i [136]. Tipping and Faul [87] proposed a more efficient implementation of SBL, called as fastSBL, which involved iterative addition-deletion of basis. Later, Babacan *et al.* [79] pointed out that SBL is a special case of BCS for a certain prior setting. See Appendix A for a detailed mathematical and numerical contrast between SBL and BCS.

In summary, NSBL and SBL possess the same sparsity-inducing Bayesian apparatus for linear regression models; the only difference being the way the model evidence or $\mathcal{L}(\log \alpha)$ is optimized. SBL (or fastSBL) operates by setting the gradient with respect to α_i (Eq. (4.57)) to zero, resulting in an iterative re-estimation procedure. On the contrary, NSBL exploits the analytically available gradient and Hessian information with respect to α vector to execute a multistart Newton iteration. Notice that while SBL updates only a single α_i per iteration, NSBL updates the entire hyperparameter vector α at each Newton iteration. In Sec. 4.3.1 we explore how this difference affects the numerical performance of the two algorithms. Nevertheless, NSBL is applicable to a broad range of linear/nonlinear mathematical models and non-Gaussian likelihood functions encountered in physics-based modelling practice. On the contrary, SBL is designed for data-based modelling which is only applicable to linear-in-parameter models (regression or classification) or Gaussian posterior pdf scenarios (Laplace approximation).

4.3 Numerical investigations

4.3.1 Linear regression: Sparse PCE expansion of Ishigami function

In this section, we consider a linear regression test bed for investigating the performance of NSBL in comparison with SBL and BCS. Since the sparse learning apparatus of NSBL is similar to SBL for a linear regression setting (as illustrated in Sec. 4.2.6), we expect similar sparsity levels from both the algorithms. We acknowledge that the conclusions from this exercise may not be transferable to physics-based sparse learning where the models are typically in the form of stochastic nonlinear differential equations. This exercise is also intended to contrast the efficiency of these algorithms in terms of the number of evidence computations required to reach the optimal sparse solution.

The linear-in-parameter model considered here is a PCE surrogate. PCE surrogates have become omnipresent in computational physics applications as a cheap replacement for high-fidelity, time-consuming numerical solvers for forward propagation of uncertainties. In addition, the orthogonality of PCE bases with respect to a probability measure helps power many desirable properties needed for performing uncertainty quantification and sensitivity analysis tasks [145, 146]. One such property of PCE surrogates is the analytical availability of Sobol sensitivity indices following the estimation of PCE coefficients [147]. See Appendix E for a brief introduction to PCE surrogates and their relation to Sobol sensitivity indices.

Given a d -dimensional input $\mathbf{X}$, the PCE surrogate of a scalar output Y is written as

$$Y = \sum_{k=0}^{P-1} y_k \Psi_k(\boldsymbol{\xi}) + \epsilon \quad (4.59)$$

where y_k are the unknown PCE coefficients, $\Psi_k(\boldsymbol{\xi})$ is the k^{th} PCE basis, $\boldsymbol{\xi}$ is the set of d standardized random variables (called germs), P is the total number of PCE basis, and ϵ is the model discrepancy error. The PCE basis $\Psi(\boldsymbol{\xi})$ are obtained through the tensorization of univariate PCE polynomials chosen from the Askey family of polynomials based on the probability

distribution of germs ξ [148]. A total-order PCE truncation with input dimension d and maximum order of p produces $P = (p + d)!/(p!d!)$ number of PCE basis in Eq. (4.59). A slight increase in PCE order p or the dimension d creates a tremendous increase in the number of PCE terms. Estimation of these large numbers of PCE coefficients demands a proportionally large number of model evaluations from time-consuming computer codes. This creates a computational bottleneck commonly known as the *curse of dimensionality*. Since the PCE coefficients are inherently sparse (only a few PCE terms are consequential), this computational bottleneck is remedied by seeking a sparse PCE representation. See Ghanem [146] for a comprehensive text on PCEs.

Most of the recent work aimed at constructing sparse PCE surrogates involves using regularized least-square techniques such as LASSO [149–152]. Sargsyan *et al.* [81] demonstrated the usefulness of BCS for constructing sparse PCE representation using a manageable number of runs from a high-fidelity earth system model. Shao *et al.* [153] approached the sparse PCE construction by maximizing the Kashyap information criterion, which is based on Gaussian approximation of posterior parameter pdf. On the contrary, NSBL applies to any general non-Gaussian posterior pdf. In this numerical exercise, we consider the probabilistic algorithms of NSBL, SBL and BCS for seeking a sparse PCE surrogate.

Ishigami function is a popular test bed for benchmarking sparse PCE construction algorithms due to its nonlinear and non-monotonic behaviour with respect to input variables [154]. Ishigami function is written as [154]

$$Y = \sin X_1 + a \sin^2 X_2 + b X_3^4 \sin X_1 \quad (4.60)$$

where the input vector is $\mathbf{X} = \{X_1, X_2, X_3\}$ and the output quantity-of-interest is Y . Each input X_i is uniformly distributed within $[-\pi, \pi]$ and parameters $a = 7.0$ and $b = 0.1$ are known. Given the uniformly distributed germs ξ_i , Legendre polynomials are employed for constructing a PCE surrogate of the Ishigami function.

An inverse problem is posed by generating 250 samples of input $\mathbf{X} = \{X_1, X_2, X_3\}$ using Latin Hypercube Sampling (LHS) [154]. These samples are then pushed forward through the Ishigami function in Eq. (4.60) to generate the corresponding samples for output Y . Next, multiple PCE surrogates with varying orders are proposed to fit this data. The order of PCE surrogates is varied as $p = 1, 2, 3, 4, 5, 6, 7$. Notice that the number of PCE terms for a 7th-order PCE surrogate are $P = (7 + 3)!/(7!3!) = 120$. Next, NSBL, SBL and BCS are employed to seek a sparse representation of PCE coefficients. The numerical implementation of SBL and BCS algorithms considered here involves iterative addition-deletion of basis, also known as fastSBL [87] and fastLAPLACE [79], respectively. More details regarding SBL and BCS are provided in Appendix A. For both SBL and BCS, we employ flat hyperpriors, and α_i at each iteration of fastSBL or fastLAPLACE is chosen in a deterministic fashion by iterating index i from zero to N_α .

The NSBL algorithm employed for this exercise is summarized in Algorithm 2. Since PCE surrogates are linear-in-parameter and all the unknown parameters (PCE coefficients) are assigned ARD priors, the GMM approximation in Eq. (4.3) is known analytically, as detailed in Sec. 4.2.6. Hence, the number of Gaussian kernels in Eq. (4.3) is one, and the mean and covariance of this kernel are known using MLE theory. Furthermore, the sparsity levels produced from NSBL are determined using the relevance indicator γ_i^{rms} defined in Sec. 4.2.5. Three different values are chosen for tolerance γ^{tol} for implementing NSBL, wherein $\gamma_i < \gamma^{\text{tol}}$ implies the corresponding PCE basis is irrelevant. Using Table 4.1, hyperprior for α_i in NSBL is chosen to be flat in log-space, i.e. $p(\log \alpha_i) \propto 1$ by assigning $r_i = s_i = 1\text{e-}05$. The model error variance ρ^{-1} is estimated in similar fashion as in SBL and BCS as $(N_d - \sum \gamma_i + 2a)/(\|\mathbf{d} - \Psi\|^2 + 2b)$ where $\mathbf{d}$ is the posterior mean.

Fig. 4.2a shows the index-of-sparsity identified by SBL, BCS and NSBL. The index of sparsity is defined as the ratio of relevant coefficients with the total number of coefficients (N_ϕ or N_α). For example, SBL identified 45 relevant PCE basis for the sixth-order PCE surrogate having a total of 84 terms, resulting in an index-of-sparsity of 0.536. As evident from Fig. 4.2a,

BCS produces sparser solution than SBL. This increased sparsity in BCS is at the expense of lower model evidence at the optimum (as seen in Fig. 4.2c). This fact about SBL vs BCS has been previously proven through analytical means [79]. The index-of-sparsity for NSBL is reported for the relevance indicator tolerance γ^{tol} value of 0.25, 0.50 and 0.75. For instance, when $\gamma^{\text{tol}} = 0.50$, any questionable parameter having the relevance indicator γ_i^{rms} value less than 0.5 is deemed irrelevant.

As evident from Fig. 4.2a, the index-of-sparsity from NSBL with $\gamma^{\text{tol}} = 0.25$ resembles closely to those obtained using SBL. This observation provides reassurance in NSBL methodology as both NSBL and SBL are optimizing the same cost function (as per Sec. 4.2.6). Also, as γ^{tol} increases, NSBL tends to produce a sparser solution. NSBL with $\gamma^{\text{tol}} = 0.75$ results in similar sparsity levels as BCS. Also, notice that the index-of-sparsity decreases with increasing PCE order. This observation implies that the addition of a higher-order PCE basis does not contribute much to the understanding of the Ishigami function, and most of these PCE bases end up being irrelevant.

Fig. 4.2b shows the iteration count to reach the optimum. NSBL involves a significantly less number of evidence computations since each iteration updates the entire vector α using analytically available gradient and Hessian information. On the contrary, SBL and BCS iterate through individual α_i 's and exploit only the gradient information, leading to a higher iteration count to reach the optimum. Fig. 4.2d shows the model error variance pertaining to the optimum. The model error variance decreases continuously with increasing PCE order as the models are getting better, even if they are getting sparser (as per Fig. 4.2a). The model error variance stagnates beyond sixth-order PCE as no significant knowledge remains to be gained regarding the Ishigami function. Note that the results reported in Fig. 4.2 were generated for multiple instances of 250 $\mathbf{X} - Y$ samples generated using LHS. Although the index-of-sparsity identified by SBL, BCS and NSBL were slightly different, the relative trend in index-of-sparsity, optimal model evidence and model error variance were observed to be similar to those reported in Fig. 4.2. Also, the kernel-based computation of gradient and Hessian of model evidence from

NSBL was validated using finite-difference for varying PCE order and $\log \alpha$ values.

The Sobol sensitivity indices pertaining to the sparse representation of the seventh-order PCE surrogate were within 1% error of the analytical values (results not reported here). Even when using only 50 LHS samples, the sensitivity indices obtained using sparse PCE were within 5% error of the analytical values. This is possible since only 42 PCE terms in seventh-order PCE were relevant. In summary, the similarity in results validates the proposed algorithm of NSBL against SBL for a linear regression setting. Also, it is shown that NSBL requires a lesser number of evidence computations than SBL since it exploits the Hessian information of model evidence to expedite the evidence maximization.

Figure 4.2: Variation of sparse learning entities with increasing PCE order. NSBL is executed with γ_i tolerance set at 0.25, 0.50 and 0.75.

4.3.2 Polynomial regression with a multimodal prior pdf

We consider a polynomial regression example wherein a multimodal prior pdf is assigned to an *a priori* relevant parameter, which then leads to multimodality in posterior parameter pdf and model evidence. We investigate the applicability of NSBL under such adverse circumstances where SBL and BCS are inapplicable. This example demonstrates the performance of NSBL for physics-based inverse problems where multimodality exists in either the parameter space of the hyperparameter space.

Fig. 4.3 shows the noisy measurement data generated using the polynomial $y_i = 1 + x_i^2 + \epsilon_i$, where ϵ_i is a Gaussian white noise process with pdf $\mathcal{N}(\epsilon_i|0, 0.02)$. The measurements consist of 50 noisy samples equally spaced in $x \in [0.75, 1.25]$. An inverse problem is posed to understand the truth ($y = 1 + x^2$) by using these measurements. To mimic the practical circumstances, an over-parameterized polynomial $y = a_0 + a_1x + a_2x^2 + \epsilon$ is proposed to model the measurements. Subsequently, sparsity in $\phi = \{a_0, a_1, a_2\}$ needs to be identified. The measurement noise ϵ is assumed as known ($\mathcal{N}(\epsilon|0, 0.02)$) for the sake of simplicity. Although the noise strength could also be estimated within the proposed NSBL setup, this is not considered necessary to highlight the usefulness of NSBL for this numerical exercise.

Figure 4.3: Noisy measurements versus the truth.

To mimic the physics-based modelling circumstances, parameter a_0 is treated as *a priori*

relevant. As a result, ϕ is decomposed into $\phi_\alpha = \{a_1, a_2\}$ and $\phi_{-\alpha} = \{a_0\}$. The known prior pdf $p(a_0)$ is chosen by realizing (using Fig. 4.4a) that the following three nested models demonstrate a reasonable fit to the measurements: 1) $y = 2x$, 2) $y = 1 + x^2$, and 3) $y = -1 + 4x - x^2$. Using this information, the prior pdf for a_0 is chosen as a mixture of equally-weighted Gaussian kernels centered at minus one, zero, and one, as shown in Fig. 4.4b. The location of these kernels demonstrate our prior belief about the potential fit of the three nested models. Note that one of these modes ($a_0 = 1$) also represent the truth ($y = 1 + x^2$).

Figure 4.4: Inverse problem setup.

Next, we employ NSBL to investigate the sparse structure of the questionable parameter vector $\phi_\alpha = \{a_1, a_2\}$ of the quadratic polynomial $y = a_0 + a_1x + a_2x^2 + \epsilon$. This problem set-up is summarized in Table 4.3. ARD priors are assigned to a_1 and a_2 with precision α_1 and α_2 , respectively. Notice that the likelihood function and the posterior pdf are a three-dimensional function of ϕ , while the model evidence is a two-dimensional function of $\alpha = \{\alpha_1, \alpha_2\}$. Also, the hyperprior for each $\alpha_i \in \alpha$ is chosen to be uninformative by assigning $\log r_i = \log s_i = -10$ in Eq. (4.6). In this case, the optimum of objective function $\mathcal{L}(\log \alpha)$ is entirely dictated by the log-evidence estimator $\log \hat{p}(\mathcal{D}|\alpha)$ from Eq. (4.14).

Initiating NSBL, TMCMC algorithm [48] from Appendix B.1 is employed to generate 2500 stationary samples from the partial posterior pdf $\propto p(\mathcal{D}|\phi)p(\phi_{-\alpha})$. The TMCMC algorithm is well-suited to generate iid samples from multimodal and/or high-dimensional pdfs with mini-

Proposed model	$y = a_0 + a_1x + a_2x^2 + \epsilon ; \quad \epsilon \sim \mathcal{N}(0, 0.02)$
ϕ decomposition	$\phi_\alpha = \{a_1, a_2\} , \quad \phi_{-\alpha} = \{a_0\}$
Known prior, $p(\phi_{-\alpha})$	$\{\mathcal{N}(a_0 -1, 0.2^2) + \mathcal{N}(a_0 0, 0.2^2) + \mathcal{N}(a_0 1, 0.2^2)\} / 3$
ARD prior, $p(\phi_\alpha \alpha)$	$\mathcal{N}(a_1 0, \alpha_1^{-1}) \mathcal{N}(a_2 0, \alpha_2^{-1})$
Hyperprior, $p(\alpha)$	$\prod_{i=1}^2 \mathcal{G}(\alpha_i r, s) ; \log r = \log s = -10$

Table 4.3: NSBL setup for the polynomial regression model with a multimodal prior pdf for a_0 .

mum manual intervention [48]. Given these iid samples, a multivariate KDE approximation of the partial posterior is constructed. This KDE-based GMM of partial posterior involves 2500 equally-weighted Gaussian kernels centered at individual TMCMC sample locations and having the same covariance. The covariance matrix of the KDE kernels is equal to the sample covariance of 2500 TMCMC samples, scaled by a factor computed using Scott’s rule [89]. The multivariate KDE approximation is constructed using the Scipy library’s *gaussian_kde* function [144].

Fig. 4.5a show the marginal parameter pdfs, and Fig. 4.5b shows the joint parameter pdfs, both pertaining to the KDE approximation of $p(\mathcal{D}|\phi)p(\phi_{-\alpha})$ using Gaussian kernels. As noted from Fig. 4.5a, the partial posterior pdf is peaked highest at the true value of $\phi = \{1, 0, 1\}$ (or model $y = 1 + x^2$). The other two possibilities of $\phi = \{0, 2, 0\}$ (or model $y = 2x$) and $\phi = \{-1, 4, -1\}$ (or model $y = -1 + 4x - x^2$) possess a lower yet non-zero posterior probability. This behavior of $p(\mathcal{D}|\phi)p(\phi_{-\alpha})$ is indicative of the irrelevance of a_1 (value of zero) and the relevance of a_2 (value of one). However, determining parameter relevance by examining posterior pdfs is an impractical and arbitrary process for high-dimensional physics-based models. As we demonstrate next, the evidence-based NSBL framework provides an efficient quantitative alternative to identify sparsity while dealing with non-Gaussian posterior pdfs.

The next step of NSBL involves executing the Newton iteration to compute the optimal hyperparameter vector α^{MAP} by maximizing the objective function $\mathcal{L}(\log \alpha)$ from Eq. (4.29). Fig. 4.6 show the NSBL results when initiating the Newton iteration from $\log \alpha = \{6, 8\}$. In

Figure 4.5: GMM representation for $p(\mathcal{D}|\phi)p(\phi_{-\alpha})$ using multivariate KDE approximation.

Fig. 4.6a, the optimal hyperparameter vector is computed as $\log \alpha^{\text{MAP}} = \{-1.42, 6.76\}$ in only ten Newton iterations. In Fig. 4.6b, the relevance indicator for a_1 approaches zero while that of a_2 approaches a value of one, thereby indicating the irrelevance of a_2 and the relevance of a_1 . This optimum pertains to the nested model of $y = 2x$. The resulting posterior pdf $p(\phi|\mathcal{D}, \alpha^{\text{MAP}})$ is also available analytically and is shown in Fig. 4.6c.

Fig. 4.7 shows the NSBL results when initiating from $\log \alpha = \{-3, -3\}$. The optimal hyperparameter vector is computed as $\log \alpha^{\text{MAP}} = \{5.24, 0.00\}$. In Fig. 4.7b, the relevance indicator for a_1 converges to one, while that for a_2 converges to zero. This demonstrates the irrelevance of a_1 and the relevance of a_2 . Based on Fig. 4.7a, this optimum is the global optimum, and the sparse model thus identified is the true model ($y = 1 + x^2$). In Fig. 4.7c, the posterior pdf for a_1 following the sparse learning approaches a Dirac-delta function centered at zero, indicating the removal of a_1 from the proposed model. On the other hand, the posterior parameter pdf of relevant parameters a_0 and a_2 is a uniquely-peaked Gaussian pdf centered around the true values.

Figure 4.6: NSBL results when initiating from $\alpha = \{6, 8\}$.

The results reported in Fig. 4.6 and Fig. 4.7 brings out two key points. First, multimodality in log-evidence or $\mathcal{L}(\log \alpha)$ can be induced by multimodality in the prior pdf. To our best knowledge, multimodality in model evidence has not been studied or reported previously in the scientific literature. This example demonstrates the need for cautious use of evidence-based model selection tasks in the presence of multimodal prior pdfs. Second, a multistart of the Newton iteration is necessary for the convergence of NSBL to the global optimum.

The benefit of sparse learning can be realized by contrasting the extrapolated response from the sparse model with that of the proposed model. Fig. 4.8a shows prediction made using the posterior parameter pdf ($\propto p(\mathcal{D}|\phi)p(\phi_{-\alpha})$) while using flat priors for questionable parameters, Fig. 4.8b shows the prediction made using the sparse model based on the local evidence optima of $\log \alpha = \{-1.42, 6.76\}$ (Fig. 4.6), and Fig. 4.8c shows the prediction made using the sparse model based on the global optimum of $\log \alpha = \{6.24, -4.00\}$ (Fig. 4.7). In Fig. 4.8a, using flat prior pdfs for questionable parameters lead to multimodality and a large variance in the

Figure 4.7: NSBL results when initiating from $\alpha = \{-3, -3\}$.

predictions. The use of flat priors has been the engineering practice under the lack of prior knowledge. This approach is problematic in this case. NSBL thus offers a much more principled alternative to assigning flat priors using ARD. As shown in Fig. 4.8c, the extrapolated response from the sparse model is a drastic improvement over the use of an over-parameterized quadratic model with flat priors.

This numerical example demonstrates the applicability of NSBL as a robust sparse learning tool while dealing with multimodality in the posterior parameter space or the hyperparameter space. Also, NSBL allows for the inclusion of prior knowledge in the form of non-Gaussian prior pdfs, making it well-suited for physics-based applications. Reiterating, the inclusion of prior knowledge and the handling of multimodal posterior pdfs are two key benefits of NSBL over SBL and BCS.

Figure 4.8: Model predictions using the log α value of a) $\{-\infty, -\infty\}$ (flat priors), b) $\{-1.42, 6.76\}$ (local optimum), and c) $\{6.24, -4.00\}$ (global optimum).

4.3.3 Three-dimensional mass-spring-damper system

In this section we consider a multi-storey shear building frame with rigid floors being modeled as a three-dof mass-spring-damper system, as shown in Fig. 4.9a. Given the displacement vector $\mathbf{u} = \{u_1, u_2, u_3\}$, the equation-of-motion for the mass-spring-damper system is obtained as

$$\mathbf{M}\ddot{\mathbf{u}} + \mathbf{C}\dot{\mathbf{u}} + \mathbf{K}\mathbf{u} = \mathbf{f}(t), \quad (4.61)$$

where $\mathbf{M}$ is the mass matrix, $\mathbf{C}$ is the damping matrix, $\mathbf{K}$ is the stiffness matrix, and $\mathbf{f}(t)$ is the external forcing. For the current numerical study we consider the case of free-vibration, i.e. $\mathbf{f}(t) = 0$. Given the state vector $\mathbf{x} = \{\mathbf{u}, \dot{\mathbf{u}}\}$, Eq. (4.61) is represented in the state space form as $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$, where matrix $\mathbf{A} = [\mathbf{0}, \mathbf{I}; -\mathbf{M}^{-1}\mathbf{K}, -\mathbf{M}^{-1}\mathbf{C}]$. Given the initial state $\mathbf{x}_0$, the solution

to the linear first-order differential equation $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ is obtained as $\mathbf{x}(t) = e^{\mathbf{A}t}\mathbf{x}_0$. Following values are used for simulating the three-dof system for this study: $\mathbf{x}_0 = \{0, 1, 0, 0, 0, 0\}$, $m_1 = m_2 = m_3 = 1.0$, $k_1 = k_2 = k_3 = 1000.0$, $c_1 = 10$, $c_2 = 0$, $c_3 = 0$. Notice that the damping is only present between the first and the second storey. Fig. 4.9b shows a noisy time-history of the third-floor (u_3) displacement obtained by corrupting the simulated (true) response with Gaussian white noise process that has a zero mean and a variance of 0.1. The measurements consist of 100 points spread over four seconds. The measurements are deemed incomplete since the displacement at the first and second floor is not observed.

Figure 4.9: Inverse problem setup.

Given the measurements in Fig. 4.9b, we pose an inverse problem to estimate the interstorey damping and stiffness coefficients. For simplicity, the mass ($m_i = 1.0$) and the measurement noise strength are assumed to be known. To imitate the real-life circumstances, we assume that we have no prior knowledge of the underlying damping representation. In other words, it is not known *a priori* that the damping is only active between the first and second storey. Under such conditions, the damping is assumed to be present in-between all floors. The unknown parameter vector for this model becomes $\phi = \{c_1, c_2, c_3, k_1, k_2, k_3\}$. Given that stiffness coefficients are always positive, we assign a uniform pdf $\mathcal{U}(k_i|0, 5000)$ to all the three stiffness coefficients. Also, since we have no prior knowledge regarding damping coefficients, they are deemed questionable and are assigned ARD priors. In other words, ϕ is decomposed into the

questionable parameter vector $\phi_\alpha = \{c_1, c_2, c_3\}$, and the *a priori* relevant parameter vector $\phi_{-\alpha} = \{k_1, k_2, k_3\}$. Jeffrey's prior is used for ARD hyperparameters as per Table 4.1 such that the optima are solely dictated by model evidence. This NSBL setup is summarized in Table 4.4.

Proposed model	$\mathbf{M}\ddot{\mathbf{u}} + \mathbf{C}\dot{\mathbf{u}} + \mathbf{K}\mathbf{u} = \mathbf{0}$
ϕ decomposition	$\phi_\alpha = \{c_1, c_2, c_3\}$, $\phi_{-\alpha} = \{k_1, k_2, k_3\}$
Known prior, $p(\phi_{-\alpha})$	$\mathcal{U}(k_1 0, 5000) \mathcal{U}(k_2 0, 5000) \mathcal{U}(k_3 0, 5000)$
ARD prior, $p(\phi_\alpha \alpha)$	$\mathcal{N}(c_1 0, \alpha_1^{-1}) \mathcal{N}(c_2 0, \alpha_2^{-1}) \mathcal{N}(c_3 0, \alpha_3^{-1})$
Hyperprior, $p(\alpha)$	$\prod_{i=1}^3 \mathcal{G}(\alpha_i r, s) ; \log r = \log s = -10$

Table 4.4: NSBL setup for the three-dof mass-spring-damper model.

Initiating NSBL, the TMCMC algorithm is exploited to generate 2500 iid samples from the partial posterior pdf ($p(\mathcal{D}|\phi)p(\phi_{-\alpha})$). Subsequently, a multivariate KDE approximation using Gaussian kernels is constructed for $p(\mathcal{D}|\phi)p(\phi_{-\alpha})$ that consists of 2500 kernels, as per Eq. (4.3). The marginal pdfs pertaining to this KDE approximation are shown in Fig. 4.10. Notice that although we are dealing with a linear structural dynamics model, the non-Gaussian nature of the partial posterior pdf is due to the nonlinear relation between the unknown parameters (damping and stiffness coefficients) and the observable entity (displacement at the third storey).

Although the GMM approximation of $p(\mathcal{D}|\phi)p(\phi_{-\alpha})$ is centered around the true value (in Fig. 4.10), the fitted model suffers from overfitting. One possible indication of overfitting is the non-zero posterior probability for negative values of the damping coefficients c_2 and c_3 in Fig. 4.10. Since the structural damping is always positive, this behavior of posterior pdf indicates the non-physical aspect of the posterior pdf. Further, Fig. 4.11 shows the predicted velocity response at the second storey, i.e. $\dot{u}_2$. The large variation in predictions is the result of overfitting produced by the over-parameterized model consisting of non-zero damping at all floors. NSBL is designed to handle such scenarios by sparsifying the parameter space to produce an optimally-parameterized model (in the sense of model evidence) that is free from overfitting.

Figure 4.10: Marginal posterior pdf pertaining to the multivariate KDE approximation of $p(\mathcal{D}|\phi)p(\phi_{-\alpha})$.

Figure 4.11: Predicted versus true time-history of velocity $\dot{u}_2$.

Next, a multistart Newton iteration is initiated as per Algorithm 2 to compute the sparse representation of damping coefficients. Fig. 4.12 show the NSBL results for three different choices of starting $\log \alpha$ value. Notice that log-evidence and objective function are the same for the range of $\log \alpha$ values attempted during Newton iteration; hence their plots are indistinguishable. A unique solution of $\alpha^{\text{MAP}} = \{-5.02, 4.03, 4.42\}$ is obtained from multistart Newton

iterations. At this optimum, the relevance indicator for the damping coefficients c_2 and c_3 converges to a value of zero, indicating their irrelevance towards the system physics. The relevance indicator for c_1 approaches the value of one, indicating its relevance. Also note that the log-evidence and objective function values are identical for the optimum, demonstrating that the sparse learning process is solely dictated by model evidence. In summary, the sparse relevant parameter vector is identified as $\phi = \{c_1, k_1, k_2, k_3\}$, which is identical to the data-generating model. Notice that this sparse structure was identified while accounting for the non-Gaussian behavior of posterior pdf and the prior knowledge regarding stiffness coefficients. This aspect distinguishes NSBL from SBL and BCS, that are incapable of handling non-Gaussian parameter pdfs and prior knowledge.

Figure 4.12: NSBL results for identifying sparsity among damping coefficients of the three-dof mass-spring-damper model.

As derived in Sec. 4.2.4.2, the posterior pdf of the relevant parameters is also available analytically as a sum of 2500 Gaussian kernels with an updated mean and covariance. This posterior pdf is computed using the user-supplied prior pdf for stiffness coefficients k_i (listed in Table 4.4), and a data-informed ARD prior $\mathcal{N}(c_1|0, e^{5.02})$ using the optimal hyperparameter value. Fig. 4.13 shows the marginal pdf of the relevant parameters. Fig. 4.14 shows the predicted time-history of velocity $\dot{u}_1$. Notice the reduction in variance in predictions before (Fig. 4.11) and after (Fig. 4.14) sparse learning. This reduction in variance is the result of removal of redundant parameters c_2 and c_3 that were causing overfitting. Notice that even with the removal of two damping coefficients, significant improvement is obtained in terms of reduction in the uncertainty in predictions. This low level of sparsity in mechanistic models is in contrast to data-driven models (such as PCE surrogate in Sec. 4.3.1) wherein such improvements are only realized for high level of sparsity.

Figure 4.13: Marginal posterior pdf for the relevant parameters of the three-dof mass-spring-damper model.

4.3.4 Nonlinear aeroelastic oscillator

In this section, we demonstrate the performance of NSBL for identifying sparsity in the unknown parameters of nonlinear stochastic differential equations. The synthetically generated LCO response from Sec. 3.2 is considered for the application of NSBL. This synthetic LCO resembles the nonlinear aeroelastic oscillator captured in the wind-tunnel. See Sec. 2.2.1 for more details regarding the wind-tunnel oscillator. The synthetic LCO is generated using the

Figure 4.14: Predicted vs true time-history of velocity $\dot{u}_2$ using the optimal nested model.

coupled fluid and structural system described by the following ODEs: (from Eq. (3.6))

$$\ddot{\theta} = c_1 \text{sign}(\dot{\theta}) + c_2 \theta + c_3 C_M + c_4 \dot{\theta} + c_5 \theta^3, \quad (4.62a)$$

$$\frac{\dot{C}_M}{B} + C_M = e_1 \theta + e_2 \dot{\theta} + e_3 \theta^3 + e_4 \theta^2 \dot{\theta} + \frac{c_6}{B} \ddot{\theta} + \sigma \xi(\tau). \quad (4.62b)$$

Here, Eq. (4.62a) is the structural equation of motion whose parameters are listed in Table 2.1. The data-generating model for the aerodynamic loading (C_M) is shown in Eq. (4.62b). The pitch measurements $\mathcal{D}$ needed for posing the inverse problem are shown in Fig. 3.2. See Sec. 2.3.2.1 for parameter values and other settings used for generating the measurements $\mathcal{D}$ shown in Fig. 3.2. Next, we exploit the NSBL algorithm from Sec. 4.2 to explore sparsity among over-parameterized models of C_M in Eq. (4.62a).

4.3.4.1 Unidimensional sparse learning

We first consider a unidimensional sparse learning problem, meaning the questionable parameter vector ϕ_α and the associated hyperparameter vector α are both scalar entities. Given the measurements $\mathcal{D}$, the proposed model for C_M is chosen to be the same as the data-generating

model shown in Eq. (4.62b). The structural equation-of-motion in Eq. (4.62a) is assumed to be known for the sake of inverse modelling. The unknown parameter vector for the coupled ODE system of Eq. (4.62a) and Eq. (4.62b) is $\phi = \{B, e_1, e_2, e_3, e_4, e_5, \sigma\}$. The inverse problems considered here are identical to those considered in Sec. 3.2.1. Hence, only the salient details are reported in this section, and the reader is referred to Sec. 3.2.1 for additional details.

Next, the cubic aerodynamic stiffness parameter e_3 (or term $e_3\theta^3$) in Eq. (4.62b) is treated as a questionable parameter, whose relevance to the LCO aerodynamics needs to be determined using the NSBL algorithm from Sec. 4.2. As per Sec. 4.2.1, ϕ is decomposed into the questionable parameter $\phi_\alpha = \{e_3\}$ and the *a priori* relevant parameter vector $\phi_{-\alpha} = \{B, e_1, e_2, e_4, \sigma\}$. The prior pdf $p(\phi_{-\alpha})$ is known, while e_3 is assigned an ARD prior $\mathcal{N}(e_3|0, \alpha^{-1})$. Table 4.5 summarizes this unidimensional sparse learning setup, including the known prior pdf of $\phi_{-\alpha}$. Note that $\mathcal{L}_n(\cdot|r, s)$ represents a log-normal distribution with median at r and coefficient of variation of s ; and $\mathcal{U}(\cdot|e, f)$ represents a uniform distribution with bounds (e, f) .

Aerodynamic model	$\frac{\dot{C}_M}{B} + C_M = e_1\theta + e_2\dot{\theta} + e_3\theta^3 + e_4\theta^2\dot{\theta} + \frac{c_6}{B}\ddot{\theta} + \sigma\xi(\tau)$
ϕ decomposition	$\phi_\alpha = \{e_3\}, \quad \phi_{-\alpha} = \{B, e_1, e_2, e_4, \sigma\}$
ARD prior, $p(\phi_\alpha \alpha)$	$\mathcal{N}(e_3 0, \alpha^{-1})$
Known prior, $p(\phi_{-\alpha})$	$\mathcal{L}_n(B 0.2, 0.5)\mathcal{U}(e_1 -2, 0)\mathcal{U}(e_2 -2, 0)\mathcal{U}(e_4 -600, 0)\mathcal{L}_n(\sigma 0.002, 0.5)$

Table 4.5: Unidimensional sparse learning setup where parameter e_3 is treated as questionable.

The likelihood function $p(\mathcal{D}|\phi)$ is computed using EKF to propagate the joint state pdf through the nonlinear model. Given the known prior pdf $p(\phi_{-\alpha})$ from Table 4.5, the unnormalized pdf $p(\mathcal{D}|\phi)p(\phi_{-\alpha})$ is sampled using the MCMC sampler detailed in Appendix B.2. A total of 5000 stationary (i.e. post burn-in) MCMC samples are generated from the six-dimensional pdf $p(\mathcal{D}|\phi)p(\phi_{-\alpha})$. These 5000 samples possess some correlation due to the Markovian nature of MCMC sampling. In an effort to alleviate the affect of correlation, every tenth sample is extracted to produce 500 iid samples from $p(\mathcal{D}|\phi)p(\phi_{-\alpha})$. These 500 iid samples are then decomposed to into 10 sets of 50 iid samples each. These iid sets are used to investigate the finite

sample properties of NSBL algorithm. Note that this sampling of $p(\mathcal{D}|\phi)p(\phi_{-\alpha})$ is only required to be performed once since both $p(\mathcal{D}|\phi)$ and $p(\phi_{-\alpha})$ remains unchanged during sparse learning. Moreover, generating 5000 stationary samples from a six-dimensional pdf is computationally cheap using MCMC algorithms such as DRAM (Appendix B.2) and TMCMC (Appendix B.1).

Next, a multivariate KDE approximation is built for $p(\mathcal{D}|\phi)p(\phi_{-\alpha})$ using each of the 10 sets of 50 iid samples. When using a Gaussian kernel, the KDE approximation resembles the kernel-based approximation in Eq. (4.3). Henceforth, the number of kernels K is the same as the number of samples, which is 50 for the current case. The coefficient $a^{(k)}$ are all equal to one since the Gaussian kernels in KDE are all equally weighted. The mean vector $\mu^{(k)}$ for each kernel is equal to the corresponding iid sample values (vector). The covariance matrix $\Sigma^{(k)}$ is computed automatically using Scott's rule for estimating KDE bandwidth [89] and is the same for all 50 kernels. Fig. 4.15 shows the marginal parameter pdf pertaining to each of the 10 instances of KDE approximation of $p(\mathcal{D}|\phi)p(\phi_{-\alpha})$. The marginal pdfs show a reasonable variation across multiple instances considering only 50 iid samples were used.

Figure 4.15: Marginal parameter pdf obtained using 10 instances of 50 iid samples from $p(\mathcal{D}|\phi)p(\phi_{-\alpha})$. The baseline (converged) pdf is obtained by using all of the 500 iid samples.

Fig. 4.16a shows the objective function $\mathcal{L}(\log \alpha)$ and the log-evidence $\log \hat{p}(\mathcal{D}|\log \alpha)$ for

varying $\log \alpha$ value, pertaining to each instance of the KDE approximation. The objective function is computed using hyperprior parameters of $\log r_i = \log s_i = -6.0$ in Eq. (4.29). The behaviour of $\mathcal{L}(\log \alpha)$ (although not shown) is observed to be almost identical to log-evidence in Fig. 4.16a. This implies the negligible effect of hyperprior in $\mathcal{L}(\log \alpha)$ for the range of $\log \alpha$ values shown in Fig. 4.16a. Also, the variation in $\mathcal{L}(\log \alpha)$ across different KDE instances is much less than that for the marginal pdfs shown in Fig. 4.15. This behaviour of $\mathcal{L}(\log \alpha)$ indicates that the relevance property of a parameter is more robust to sampling variability than the KDE approximation of $p(\mathcal{D}|\phi)p(\phi_{-\alpha})$.

Figure 4.16: NSBL results for the relevance determination of an *a priori* relevant parameter e_3 .

Next, the NSBL algorithm from Algorithm 2 is exploited to optimize $\mathcal{L}(\log \alpha)$ shown in Fig. 4.16a. Fig. 4.16b plots the $\log \alpha$ iterates during the multistart Newton iteration, and Fig. 4.16c plots the corresponding value of the relevance indicator γ_i^{rms} from Eq. (4.55). The $\log \alpha$ iterates and γ_i^{rms} are shown for each of the 10 KDE instances of $p(\mathcal{D}|\phi)p(\phi_{-\alpha})$. The Newton iteration is initiated from three $\log \alpha$ values of -20.0, -10.0 and 0.0. The analytical solution for the gradient and Hessian of $\mathcal{L}(\log \alpha)$ employed in the Newton iteration was first validated using a finite-difference scheme (results not reported here for brevity). As noted from Fig. 4.16b, all the multistart Newton iterations converge to a unique optimum of $\alpha^{\text{MAP}} = -9.1$. This optimum was also obtained in Fig. 3.3 in Sec. 3.2.1, further validating NSBL against the sampling-based ARD scheme. Notice that the Newton iteration converges in just 2 iterations when starting from $\log \alpha = -10.0$. This computational efficiency of NSBL is a significant

improvement over the sampling-based model reduction approach from Chapter 3.

The relevance indicator γ^{rms} in Fig. 4.16c converges to one, implying the strong relevance of parameter e_3 to the LCO aerodynamics. This relevance of e_3 is insensitive to the choice of tolerance γ^{tol} as the optimal relevance indicator is very close to one. The posterior pdf of ϕ can be obtained as per Sec. 4.2.4.2 using the optimal ARD prior $\mathcal{N}(e_3|0, e^{9.1})$. Since this optimal prior has large variance ($e^{9.1}$), the posterior parameter pdf obtained from NSBL is the same as that obtained by sampling $p(\mathcal{D}|\phi)p(\phi_{-\alpha})$ (the exact pdf shown in Fig. 4.15). In other words, the optimal ARD prior did not bias the posterior pdf of e_3 , an observation also made in the numerical investigations reported in Sec. 3.2.1. This behaviour is crucial since we would not want a data-informed prior to influence the posterior parameter pdf. Also, we acknowledge that the results reported in this exercise are conditioned on the prior pdf $p(\phi_{-\alpha})$ chosen for the *a priori* relevant parameters (shown in Table 4.5). This subjective nature of the Bayesian framework is crucial for physics-based modelling as it facilitates the update in one's knowledge of system physics with the incoming experimental data.

Next, consider the unidimensional sparse learning setup in Table 4.6 where parameter e_5 is treated as questionable. The aerodynamic model in Table 4.6 has an additional fifth-order aerodynamic stiffness term $e_5\theta^5$, in comparison to the data-generating model in Eq. (4.62b). Therefore, we expect the e_5 parameter will be deemed irrelevant by NSBL. We initiate NSBL following the same process as was used in the previous numerical exercise reported in this section. Fig. 4.17 shows the marginal parameter pdf for each of the 10 instances of the KDE approximation of $p(\mathcal{D}|\phi)p(\phi_{-\alpha})$. Again, the marginal pdfs show a reasonable sampling variation considering only 50 iid samples were used.

Fig. 4.18a shows the log-evidence $\log \hat{p}(\mathcal{D}|\log \alpha)$ and Fig. 4.18b shows the objective function $\mathcal{L}(\log \alpha)$ using the hyperprior parameters of $\log r_i = \log s_i = -6$. Both these entities possess large sampling variability across multiple KDE instances of $p(\mathcal{D}|\phi)p(\phi_{-\alpha})$. However, the relevance of e_5 is determined by the optimum of $\mathcal{L}(\log \alpha)$, which remains same across multiple KDE instances. This finite sample property of $\mathcal{L}(\log \alpha)$ reiterates the notion that relevance is

Aerodynamic model	$\frac{\dot{C}_M}{B} + C_M = e_1\theta + e_2\dot{\theta} + e_3\theta^3 + e_4\theta^2\dot{\theta} + e_5\theta^5 + \frac{c_6}{B}\ddot{\theta} + \sigma\xi(\tau)$
ϕ decomposition	$\phi_\alpha = \{e_5\}$, $\phi_{-\alpha} = \{B, e_1, e_2, e_3, e_4, \sigma\}$
ARD prior, $p(\phi_\alpha \alpha)$	$\mathcal{N}(e_5 0, \alpha^{-1})$; $\alpha = \{\alpha\}$
Known prior, $p(\phi_{-\alpha})$	$\mathcal{L}_n(B 0.2, 0.5) \mathcal{U}(e_1 -2, 0) \mathcal{U}(e_2 -2, 0) \mathcal{U}(e_3 -250, 250)$ $\mathcal{U}(e_4 -600, 0) \mathcal{L}_n(\sigma 0.002, 0.5)$

Table 4.6: Unidimensional sparse learning setup where parameter e_5 is treated as questionable.

Figure 4.17: Marginal parameter pdf obtained using 10 instances of 50 iid samples from $p(\mathcal{D}|\phi)p(\phi_{-\alpha})$. The baseline (converged) pdf is obtained by using all of the 500 iid samples.

well-determined using limited iid samples from $p(\mathcal{D}|\phi)p(\phi_{-\alpha})$, even when the KDE instances possess large sampling variability. Also notice that, unlike log-evidence, the objective function $\mathcal{L}(\log \alpha)$ in Fig. 4.18b is free from flat regions for large $\log \alpha$ values. The Hessian matrix is singular in flat regions. Therefore, the presence of flat regions in log-evidence makes it unsuitable for the application of Newton's method. The absence of flat regions in $\mathcal{L}(\log \alpha)$ is due to the influence of hyperprior $p(\alpha)$ for large $\log \alpha$ values. This ensures $\mathcal{L}(\log \alpha)$ is amenable to optimization by Newton's method wherein the Hessian is non-singular at all times. This regularization of the sparse learning optimization problem by the hyperprior $p(\log \alpha)$ is the main reason behind its inclusion into NSBL.

Fig. 4.18c shows the $\log \alpha$ iterates and Fig. 4.18d shows the RMS relevance indicator γ^{rms} ; when initiating at $\log \alpha$ values of -15.0, -5.0 and 5.0. The multistart Newton iteration converges

to a unique optimum of $\alpha^{\text{MAP}} = 0$ and $\gamma^{\text{rms}} = 0$ for all multistart iterations and for all KDE instances of $p(\mathcal{D}|\phi)p(\phi_{-\alpha})$. Note that since the objective function is relatively flat close to the optimum, the estimated optimum can vary with a varying function tolerance set for terminating the Newton iteration. However, the key entity to monitor is the relevance indicator. Even with a small variation in the optimum hyperparameter value, the relevance indicator converges to the value of zero. According to Eq. (4.55), a γ^{rms} value of zero implies irrelevance since the posterior pdf is entirely dictated by the Dirac-delta ARD prior centered at zero. This change in the posterior pdf of e_5 following NSBL is shown in Fig. 4.18e.

Figure 4.18: NSBL results for the relevance determination of an *a priori* irrelevant parameter e_5 .

Fig. 4.18f shows the posterior pdf of parameter e_3 before and after the inclusion of optimal ARD prior $\mathcal{N}(e_5|0, e^{0.0})$. Notice that the posterior pdf of e_3 following the removal of e_5 has a lower uncertainty and is in close agreement with the true value. This behaviour can be explained by realizing that e_3 and e_5 both quantify the nonlinear aerodynamic stiffness. When both the parameters are present in the model, overfitting occurs. When e_5 is removed, overfitting is remedied, and e_3 is estimated with an increased accuracy using $\mathcal{D}$. This numerical exercise

demonstrates the benefit of sparse learning in preventing overfitting and enabling robust probabilistic predictions outside the regime of measurements. The two unidimensional cases studied here enforces our confidence in NSBL as a practical tool to perform sparse learning among physics-based models. Next, we study two-dimensional sparse learning cases to gain a more in-depth understanding of NSBL.

4.3.4.2 Two-dimensional sparse learning

In this section, we consider two-dimensional sparse learning cases where the number of questionable parameters is two, i.e. $N_\alpha = 2$. The hyperparameter vector is $\alpha = \{\alpha_1, \alpha_2\}$. NSBL is executed with the same settings as Sec. 4.3.4.1. Hence, the likelihood function $p(\mathcal{D}|\phi)$ is computed using EKF, followed by the generation of 500 iid samples from $p(\mathcal{D}|\phi)p(\phi_\alpha)$ using the DRAM MCMC algorithm detailed in Appendix B.2. NSBL is executed for each of the 10 KDE instances of $p(\mathcal{D}|\phi)p(\phi_\alpha)$ containing 50 iid samples each. Also, the hyperprior parameters are chosen as $\log r_i = \log s_i = -6.0$ for all the cases reported in this section.

We first consider the Bayesian setup in Table 4.7 where the proposed model is same as the data-generating model from Eq. (4.62b), and parameters e_3 and e_4 are treated as questionable. Fig. 4.19a shows the $\log \alpha$ iterates during Newton iteration initiated from $\{-20, -20\}$, $\{-5, -20\}$, $\{-20, -5\}$ and $\{-5, -5\}$, for each of the 10 KDE instances. Notice that all of the multistart Newton iterations converges to a unique optimum of $\log \alpha^{\text{MAP}} = \{-9.1, -12.4\}$. Also, convergence to a unique optimum for varying KDE instances of $p(\mathcal{D}|\phi)p(\phi_\alpha)$ demonstrates that the curvature of objective function $\mathcal{L}(\log \alpha)$ remains fairly insensitive to sampling variability. Fig. 4.19b and Fig. 4.19c shows the variation in RMS relevance indicator γ^{rms} for parameter e_3 and e_4 , respectively. Both the relevance indicators converge to one, implying the relevance of both the parameters to the LCO physics. NSBL converges in less than 10 Newton iterations, demonstrating the power of the Hessian informed optimizer.

Next, consider the Bayesian setup in Table 4.8 where parameters e_3 and e_5 are treated as questionable. The data-generating model in Eq. (4.62b) contains the cubic aerodynamic stiff-

Aerodynamic model	$\frac{\hat{C}_M}{B} + C_M = e_1\theta + e_2\dot{\theta} + e_3\theta^3 + e_4\theta^2\dot{\theta} + \frac{c_6}{B}\ddot{\theta} + \sigma\xi(\tau)$
ϕ decomposition	$\phi_\alpha = \{e_3, e_4\}$, $\phi_{-\alpha} = \{B, e_1, e_2, \sigma\}$
ARD prior, $p(\phi_\alpha \alpha)$	$\mathcal{N}(e_3 0, \alpha_1^{-1}) \mathcal{N}(e_4 0, \alpha_2^{-1})$
Known prior, $p(\phi_{-\alpha})$	$\mathcal{L}_n(B 0.2, 0.5)\mathcal{U}(e_1 -2, 0)\mathcal{U}(e_2 -2, 0)\mathcal{L}_n(\sigma 0.002, 0.5)$

Table 4.7: Two-dimensional sparse learning setup where parameters e_3 and e_4 are treated as questionable.

Figure 4.19: NSBL results for the relevance determination of parameter e_3 and e_4 .

ness coefficient e_3 , while the fifth-order aerodynamic stiffness coefficient e_5 is absent. Fig. 4.20a shows the $\log \alpha$ iterates pertaining to starting $\log \alpha$ values of $\{-20, -20\}$, $\{-5, -20\}$, $\{-20, -5\}$ and $\{-5, -5\}$. The multistart Newton iteration converges to a unique optimum of $\log \alpha^{\text{MAP}} = \{-9.1, 0.0\}$. Fig. 4.20b and Fig. 4.20c shows the RMS relevance indicators for parameter e_3 and e_5 , respectively. Parameter e_3 is determined to be relevant since γ_1^{rms} converges to exactly one. On the contrary, parameter e_5 is determined to be irrelevant as γ_2^{rms} converges to zero. These conclusion are true for all multistart locations and for all KDE instances of $p(\mathcal{D}|\phi)p(\phi_{-\alpha})$.

Aerodynamic model	$\frac{\hat{C}_M}{B} + C_M = e_1\theta + e_2\dot{\theta} + e_3\theta^3 + e_4\theta^2\dot{\theta} + e_5\theta^5 + \frac{c_6}{B}\ddot{\theta} + \sigma\xi(\tau)$
ϕ decomposition	$\phi_\alpha = \{e_3, e_5\}$, $\phi_{-\alpha} = \{B, e_1, e_2, e_4, \sigma\}$
ARD prior, $p(\phi_\alpha \alpha)$	$\mathcal{N}(e_3 0, \alpha_1^{-1}) \mathcal{N}(e_5 0, \alpha_2^{-1})$
Known prior, $p(\phi_{-\alpha})$	$\mathcal{L}_n(B 0.2, 0.5)\mathcal{U}(e_1 -2, 0)\mathcal{U}(e_2 -2, 0)\mathcal{U}(e_4 -600, 0)\mathcal{L}_n(\sigma 0.002, 0.5)$

Table 4.8: Two-dimensional sparse learning setup where parameters e_3 and e_5 are treated as questionable.

Next, consider the Bayesian setup in Table 4.9, where parameters e_5 and e_6 are treated as

Figure 4.20: NSBL results for the relevance determination of parameter e_3 and e_5 .

questionable. Both these parameters were absent from the data-generating model in Eq. (4.62b). Fig. 4.21a shows the $\log \alpha$ iterates for the multistart Newton iteration beginning from $\{-20, -20\}$, $\{-5, -20\}$, $\{-20, -5\}$ and $\{-5, -5\}$. The Newton iterations converge to two different optima of $\{0.0, -20.5\}$ and $\{0.0, 0.0\}$. Based on the $\mathcal{L}(\log \alpha)$ values at these optima, $\{0.0, -20.5\}$ was found to be the global optimum (for the range of α_1 and α_2 considered here). This ability of the multistart Newton iteration to capture multiple optima was further validated using the case of bimodal likelihood functions. The findings of this study are reported and discussed in Appendix D.

Aerodynamic model	$\frac{\dot{C}_M}{B} + C_M = e_1\theta + e_2\dot{\theta} + e_3\theta^3 + e_4\theta^2\dot{\theta} + e_5\theta^5 + e_6\theta^4\dot{\theta} + \frac{c_6}{B}\ddot{\theta} + \sigma\xi(\tau)$
ϕ decomposition	$\phi_\alpha = \{e_5, e_6\}$, $\phi_{-\alpha} = \{B, e_1, e_2, e_3, e_4, \sigma\}$
ARD prior, $p(\phi_\alpha \alpha)$	$\mathcal{N}(e_5 0, \alpha_1^{-1}) \mathcal{N}(e_6 0, \alpha_2^{-1})$
Known prior, $p(\phi_{-\alpha})$	$\mathcal{L}_n(B 0.2, 50)\mathcal{U}(e_1 -2, 0)\mathcal{U}(e_2 -2, 0)\mathcal{U}(e_3 -250, 250)$ $\mathcal{U}(e_4 -1e+4, 0)\mathcal{L}_n(\sigma 0.002, 0.5)$

Table 4.9: Two-dimensional sparse learning setup where parameters e_5 and e_6 are treated as questionable.

Fig. 4.21b and Fig. 4.21c show the RMS relevance indicator for parameter e_5 and e_6 during multistart Newton iterations. Parameter e_5 is rendered irrelevant irrespective of the optimum as γ_1^{rms} converges to zero for both the optima. The parameter e_6 requires a special attention as γ_2^{rms} in Fig. 4.21c converges to a value in the range $[0.75, 0.90]$ for the case of the global optimum. As

demonstrated previously in the unidimensional setting, γ_i^{rms} value converges to exactly one for relevant parameters. Also, based on our experience, the convergence of relevance indicator γ_i^{rms} to one is a necessary condition for a parameter to be deemed relevant. This condition implies that the posterior pdf of a relevant parameter should be dictated by the likelihood function and not the data-informed ARD prior. Based on this principle, the parameter e_6 should be deemed irrelevant. Alternatively, one could define a tolerance γ^{tol} for determining the relevance of such parameters. A γ^{tol} value of 0.5 implies parameter e_6 as relevant, whereas a γ^{tol} value of 0.9 implies an irrelevant e_6 . This choice of an appropriate γ^{tol} imparts greater flexibility to the modeller, thereby allowing the sparse learning process to align with the modelling goals. The effect of γ^{tol} on sparsity levels was previously studied in Sec. 4.3.1 for a linear regression setting.

Figure 4.21: NSBL results for the relevance determination of parameter e_5 and e_6 .

Fig. 4.22 shows the marginal parameter pdf before and after the inclusion of the optimal ARD prior for questionable parameters e_5 and e_6 . The pre-NSBL pdf is obtained from the KDE representation of $p(\mathcal{D}|\phi)p(\phi_{-\alpha})$ using 500 iid samples. The post-NSBL pdf is obtained using the analytical solution derived in Sec. 4.2.4.2. In Fig. 4.22, the posterior pdf of e_5 and e_6 following NSBL reduces to a Dirac-delta function centered at zero. This removal of parameters e_5 and e_6 results in a decrease in posterior uncertainty in other parameters, except the model error strength σ . This behaviour of the marginal posterior pdf of σ indicates that the removal of redundant parameters during NSBL does not affect the data-fit property of the model. Notice the decrease in posterior uncertainty of parameters e_3 and e_4 following the removal of e_5 and

e_6 . This decrease is attributed to the remediation of overfitting by the sparse learning process. Fig. 4.23 shows the corresponding marginal posterior pdfs for the case when e_5 is deemed irrelevant and e_6 is deemed relevant.

Figure 4.22: Marginal posterior pdf of model parameters before and after sparse learning, for the case when e_5 and e_6 are deemed irrelevant.

Figure 4.23: Marginal posterior pdf of model parameters before and after sparse learning, for the case when e_5 is deemed irrelevant and e_6 is deemed relevant.

The objective function $\mathcal{L}(\log \alpha)$ was observed to be dictated by log-evidence for the range of α values shown in Fig. 4.21a, meaning both log-evidence and $\mathcal{L}(\log \alpha)$ are bimodal. To our best knowledge, this is the first reporting of a multimodal log-evidence for a physics-based Bayesian inverse modelling. This multimodality also demonstrates the issue of *global identifiability* in the model (or hyperparameter) space, while the inverse problem may or may not be globally

identifiable in the likelihood (or model parameter) space. This multimodality also warrants careful consideration of evidence-based Bayesian model comparison while dealing with nested models with closely-related nonlinear terms.

NSBL was also executed for cases with 100, 500, and 1000 iid samples per KDE instance of $p(\mathcal{D}|\phi)p(\phi_{-\alpha})$. These results are not reported here since they were identical to those reported in this section but with even less variation in $\log \alpha$ iterates and relevance indicator γ^{rms} across multiple instances. This conclusion was true for the three numerical cases studied in this section.

4.3.4.3 Four-dimensional sparse learning

Consider the four-dimensional sparse learning problem detailed in Table 4.10. NSBL is executed to determine the relevance of questionable parameters e_3 , e_4 , e_5 and e_6 . The NSBL setup is the same as Sec. 4.3.4.1. Fig. 4.24 shows the $\log \alpha$ iterates and relevance indicator γ_i^{rms} for multistart Newton iterations beginning from $\{-20, -20, -20, -20\}$, $\{-20, -20, -5, -5\}$, $\{-5, -5, -20, -20\}$ and $\{-5, -5, -5, -5\}$. Notice that all but $\log \alpha_4$ converges to a unique optimum. The $\log \alpha_4$ converges to two different optimums depending on the starting point. This type of behaviour has been previously reported in Fig. 4.21c. Based on the $\mathcal{L}(\log \alpha)$ values at these optimums, $\log \alpha_4^{\text{MAP}} = 0.0$ was found to be the global optimum. This produces the solution $\log \alpha^{\text{MAP}} = \{-9.1, -12.4, 0.0, 0.0\}$, and the corresponding value of relevance indicator as $\{1.0, 1.0, 0.0, 0.0\}$. This optimal value of γ_i^{rms} indicates that parameter e_3 and e_4 are important, while parameter e_5 and e_6 are redundant. This conclusion is in agreement with the data-generating model in Eq. (4.62b) that was used to generate $\mathcal{D}$. Fig. 4.25 shows the marginal posterior pdf before and after the inclusion of the optimal ARD prior computed using NSBL. The pre-NSBL pdf is obtained using 500 iid samples from $p(\mathcal{D}|\phi)p(\phi_{-\alpha})$, while the post-NSBL pdf is obtained using the kernel-based analytical solution derived in Sec. 4.2.4.2. Notice the decrease in uncertainty in relevant parameters following the removal of parameter e_5 and e_6 . Note that the level of sparsity in physical models such as the LCO models is significantly lower than data-based models ((such as PCE surrogate in Sec. 4.3.1)) due to the presence of physics-based parameters. Sig-

nificant benefits are realized in terms of decrease in uncertainty in the posterior pdf even with the removal of two parameters.

Aerodynamic model	$\frac{\dot{C}_M}{B} + C_M = e_1\theta + e_2\dot{\theta} + e_3\theta^3 + e_4\theta^2\dot{\theta} + e_5\theta^5 + a_6\theta^4\dot{\theta} + \frac{c_6}{B}\ddot{\theta} + \sigma\xi(\tau)$
ϕ decomposition	$\phi_\alpha = \{e_3, e_4, e_5, e_6\}, \quad \phi_{-\alpha} = \{B, e_1, e_2, \sigma\}$
ARD prior, $p(\phi_\alpha \alpha)$	$\mathcal{N}(e_3 0, \alpha_1^{-1})\mathcal{N}(e_4 0, \alpha_2^{-1})\mathcal{N}(e_5 0, \alpha_3^{-1})\mathcal{N}(e_6 0, \alpha_4^{-1}),$ $\alpha = \{\alpha_1, \alpha_2, \alpha_3, \alpha_4\}$
Known prior, $p(\phi_{-\alpha})$	$\mathcal{L}_n(B 0.2, 0.5)\mathcal{U}(e_1 -2, 0)\mathcal{U}(e_2 -2, 0)\mathcal{L}_n(\sigma 0.002, 0.5)$

Table 4.10: Four-dimensional sparse learning setup.

Figure 4.24: NSBL results for the four-dimensional sparse learning.

Figure 4.25: Marginal posterior pdf of model parameters before the after the inclusion of optimal ARD prior.

Note that as the size of MCMC sample set increases, the number of kernels K of the KDE approximation also increases in Eq. (4.3). This increases the computational cost of evaluating the gradient vector $\mathbf{J}(\log \boldsymbol{\alpha})$ and the Hessian matrix $\mathbf{H}(\log \boldsymbol{\alpha})$ in Newton's method. This computational issue can be resolved using alternate approaches such as constructing a GMM where an optimal number of kernels can be computed by standard machine learning tools such as EM.

In addition, when dealing with a large number of hyperparameters, the size of $\mathbf{J}$ and $\mathbf{H}$ also increases, which then creates computational and memory bottleneck. This bottleneck can be alleviated in a number of ways. First, a diagonal representation of the Hessian matrix could be constructed for performing Newton's iteration in an approximate manner. Alternatively, the GMM based computation of Hessian matrix could be replaced by algorithms such as limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm that employs the gradient information to compute an approximation of Hessian by efficiently managing the memory requirements [7].

Chapter 5

Conclusion and future research direction

5.1 Conclusion

In this thesis, the issue of overfitting among nonlinear physics-based models is tackled through efficient Bayesian algorithms. The Bayesian algorithms presented in this thesis were motivated by the need to model the nonlinear aeroelastic system observed in the wind-tunnel. Table 5.1 provides a comparative review of the Bayesian algorithms studied in this thesis. This thesis dealt with three different Bayesian techniques to tackle the issue of overfitting.

First, we presented a Bayesian model comparison algorithm powered by a parallel and adaptive MCMC algorithm for computing model evidence for models in the form of nonlinear stochastic ODEs. The Bayesian algorithm is validated successfully on a synthetically generated LCO data similar to the wind-tunnel LCO. The model selection algorithm performs well in selecting the aerodynamic moment model that generated the synthetic data, while the MAP estimates and the marginal posterior pdfs are in good agreement with the true parameter values. The Bayesian model selection results for the wind-tunnel LCO data are presented for four transitional Reynolds numbers. The candidate model set consists of both quasi-steady and unsteady aerodynamic models with varying degrees of nonlinearities. The unsteady models are inspired by the linear unsteady aerodynamic theory wherein Wagner function quantifies the ef-

fects of trailing edge vortices on the aerodynamic loading. The Bayesian algorithm proved to be a robust tool for characterizing and predicting the nonlinear and unsteady aerodynamics.

Next, we presented an ARD-based model reduction algorithm to identify the relevance of questionable parameters in an over-parameterized model. In contrast to SBL and BCS, a hybrid approach for choosing prior pdfs is presented wherein model parameters with significant prior knowledge are assigned a known pdf, and the rest are assigned ARD priors. This approach permits a prior-informed sparse learning process. The optimal hyperparameters are estimated by maximizing model evidence. This model reduction methodology is applied to synthetic measurements generated from the nonlinear aeroelastic oscillator. Three intentionally over-parameterized models are proposed with varying capabilities in modelling the unsteady and nonlinear aerodynamics of the LCO. The proposed models have different state-space vectors and cannot be represented as nested under a common encompassing model. Subsequently, an optimal nested model is obtained for each proposed model using ARD. The three optimal nested models are compared among themselves using their optimal evidence values using Bayesian model selection to obtain posterior model probabilities. The optimal nested model with the highest posterior probability is the same as the data-generating model. The posterior parameter samples are obtained as a by-product of the evidence computation algorithm. The posterior parameter pdfs exhibit close resemblance with the true parameter values used in generating the synthetic data. It is also shown that the zero-mean ARD priors introduce negligible bias in the posterior parameter statistics. Hence, the data-generating model is obtained as the best (maximum evidence) model among all the models nested under the three over-parameterized models. It is concluded that using ARD with Bayesian model selection enlarges the model domain and increases the chances of obtaining a better model than the discrete model comparison.

Next, NSBL is offered as a computationally efficient alternative to sampling-based sparse learning of nonlinear physics-based models. SBL serves as a motivation for designing the semi-analytical apparatus of NSBL. The analytical tractability of the Bayesian analysis is enabled by a GMM approximation of the unnormalized posterior pdf with flat priors assigned

to questionable parameters, and the subsequent assignment of Gaussian ARD priors to questionable parameters. A detailed mathematical derivation into the analytical solution for model evidence, posterior parameter pdf, and the gradient and Hessian information of model evidence is presented in this thesis. SBL and NSBL are shown (numerically and analytically) to be the same for a linear regression setting where all parameters are assigned a Gaussian ARD prior.

NSBL is first validated using a linear regression example, wherein a sparse PCE surrogate of the Ishigami function is constructed for computing the Sobol' sensitivity indices. This numerical exercise further illustrated that the sparsity levels identified from NSBL are close to those from SBL and BCS. This numerical exercise also helped validate the analytical apparatus of NSBL against SBL. Next, a polynomial regression example with a multimodal prior pdf is considered to investigate the performance of NSBL for a multimodal posterior pdf, and a multimodal model evidence cases. The multistart Newton's method is exploited to solve the non-convex optimization problem (having multiple optima) involving the maximization of evidence with respect to hyperparameters of ARD prior. Finally, NSBL is applied to identify a sparse representation of damping coefficients for a three-dof mass-spring-damper model of a shear building frame. The true sparse structure of damping coefficients is identified by NSBL while estimating stiffness coefficients and including their prior knowledge into the sparse learning process. In summary, NSBL is shown to identify the data-generating model in all the numerical cases, demonstrating its effectiveness as a sparse learning tool in the presence of multimodal and non-Gaussian posterior/prior pdfs. This automatic dimensionality reduction property of NSBL is being offered as an efficient alternative to sampling-based Bayesian inverse modelling of complex physical systems.

5.2 Future research direction

Following possible avenues could be pursued based on the findings reported in this thesis:

- In this thesis, we dealt with a rigid wing undergoing an LCO, which is essentially a single

dof system. Application of sparse learning algorithms such as NSBL to continuous spatio-temporal models require a thorough investigation. Previous work by Matachniouk [155], Rocha Da Costa [156] and Robinson [157] was aimed at developing continuous aeroelastic models described by a coupled set of nonlinear PDEs and ODEs. These models were proposed to model the wind-tunnel response of a flexible wing undergoing self-sustained aeroelastic oscillation. Since NSBL is model-free approach that deals with only the likelihood function, we expect NSBL can execute a computationally efficient calibration of over-parameterized continuous models.

- Murphy [7] demonstrated that the variational Bayes approach of sparse learning provides similar results to SBL for a linear regression setting and a large number of measurements. A thorough comparative review of variational Bayes and the proposed ARD-based approach is needed when dealing with nonlinear dynamical systems. Also, an extension of variational Bayes in a multi-kernel setting such as the NSBL is worth exploring.
- Since the introduction of physics-informed neural networks [83], the use of machine learning tools in engineering practice has received a considerable attention. Generally speaking, a complex network configuration (in terms of number of hidden layers and the number of neurons per layer) of a neural network is inherently sparse and nonlinear. Clearly, the challenges in applying NSBL for sparse learning will include dealing with high-dimensionality of the weight matrix and the bias vector. An associated challenge includes the large dimensionality of evidence and its gradient and Hessian since all weights will be assigned ARD priors. It would be interesting to see how NSBL performs for sparsifying the unknown weight matrix and bias vector for an overfitted neural network, especially for physics-informed neural network. Also, it would be interesting to see how the size of data effects the sparsity in neural networks.
- In NSBL, the GMM construction of the unnormalized pdf in Eq. (4.3) is implemented using KDE with Gaussian kernels. In KDE, the number of kernels is equal to the number of station-

ary MCMC samples from the unnormalized pdf. Hence, for models with a large number of unknown parameters, KDE based GMM approximation can produce a computational bottleneck for the semi-analytical computation of Bayesian entities in NSBL. Alternatively, algorithms such as expectation-maximization and variational inference provide an efficient way to construct GMM using a low number of kernels. These algorithms are particularly appealing at representing multimodal probability distributions using a minimum number of kernels. We expect such algorithms to further expedite sparse learning among high-dimensional and high-fidelity numerical models of physical systems.

- In this thesis, we have employed ARD to identify sparsity in the model parameter space. It will be interesting to see how ARD and NSBL performs when seeking sparsity in the state space or the measurement space. For instance, the quasi-steady models studied in this thesis has a state vector that is nested under the state space of unsteady models. In this case, handling sparsity in state space could enable an all encompassing model that can encapsulate all the quasi-steady and unsteady models of the LCO aerodynamics, thereby further expediting the sparse learning process.

Method	Capabilities			Methodology			Scalability	More details provided in
	Nonlinear models?	Prior knowledge?	Sparse learning?	Prior pdf type	Evidence estimation	Evidence optimization	# of likelihood evaluations	
SBL/BCS	×	×	✓	Gaussian ARD (Conjugate prior)	Analytical	Gradient set to zero	N/A	Appendix A
Bayesian model comparison	✓	✓	×	Known pdf	Numerical (MCMC-based)	N/A	$N_{\text{Models}} \times N_{\text{MCMC}}$	Chapter 2
ARD (Sampling-based)	✓	✓	✓	Known + Gaussian or Laplace ARD	Numerical (MCMC-based)	Gradient-free	$N_{\text{Iter}} \times N_{\text{MCMC}}$	Chapter 3
ARD (NSBL)	✓	✓	✓	Known + Gaussian ARD	Semianalytical (GMM-based)	Gradient and Hessian based	N_{MCMC}	Chapter 4

N_{MCMC} : # of MCMC samples needed for computing evidence (for an average model).

N_{Models} : # of candidate models.

N_{Iter} : # of iterations required for optimizing evidence.

Table 5.1: Comparative review of Bayesian algorithms studied in this thesis.

Appendices

Appendix A

Sparse learning in linear regression

In this section, we investigate the performance of LASSO, SBL, and BCS in a linear regression setting. These algorithms have received significant attention from the engineering community as a way to induce sparsity among over-parameterized physical and statistical models [74, 76, 80–82]. Hastie *et al.* [25] presented a detailed comparison of frequentist sparse learning techniques for linear regression models. In this study, we focus on investigating the numerical and mathematical underpinnings of the Bayesian techniques of SBL and BCS and how they perform in comparison to LASSO.

A linear regression model of a scalar output is defined as

$$y^{(i)} = \sum_{j=0}^{N-1} \phi_j(\mathbf{u}^{(i)})w_j + \epsilon^{(i)} \Rightarrow \mathbf{y} = \Phi\mathbf{w} + \epsilon \quad (\text{A.1})$$

where $\phi_j(\cdot)$ is a proposed basis function, $\mathbf{u} \in \mathbb{R}^{N_d}$ is the model input (independent variable), $\mathbf{y} \in \mathbb{R}^M$ is the output measurement vector, $\mathbf{w} \in \mathbb{R}^N$ is the unknown model parameter vector, $\epsilon \in \mathbb{R}^M$ is the model discrepancy error vector and $\Phi \in \mathbb{R}^{M \times N}$ is the design matrix. The training data consists of M measurements of $\{y^{(i)}, \mathbf{u}^{(i)}\}$. Also, N denotes the number of model parameters or the number of proposed basis functions or the model dimensionality. The model discrepancy error ϵ quantifies the mismatch between the model and measurements. We assume that the model error ϵ is zero-mean and has a precision ρ .

The most commonly used basis functions in statistical modelling are radial basis function, SVM, PCE, among many others. These basis functions transform the independent variable $\mathbf{u}$ space into a new space $\phi_j(\mathbf{u})$ whose scale and variation aligns well with $\mathbf{y}$. This transformed space of $\phi_j(\mathbf{u})$ is also known as a *feature* of the model. In practice, the statistical models often contain more features than it is required to obtain a reasonably good fit to the measurements. This unrestricted complexity leads to overfitting during parameter estimation, thereby hindering the predictive capability of the learned model for unobserved data. Sparse learning algorithms are aimed at identifying the subset of proposed features that are sufficient to provide a reasonable data-fit and a good extrapolating capability. The removal of redundant features also leads to a reduced model dimensionality (N), thereby expediting the training and prediction phase of supervised learning.

The remaining chapter is organized as follows. In Sec. A.1, we detail the frequentist techniques to parameter estimation, including LASSO. In Sec. A.2, we detail the analytically tractable Bayesian inference framework using the concept of conjugate priors, which is later used in deriving SBL and BCS. In Sec. A.3, we provide a detailed derivation of SBL. In Sec. A.4, we detail the derivation of BCS, along with its relation to SBL. In Sec. A.6, we report and discuss the numerical performance of LASSO, SBL, and BCS for a polynomial regression example.

A.1 Regularized least-square estimation

The most popular frequentist approach of estimating $\mathbf{w}$ in linear regression setup of Eq. (A.1) is the ordinary least-square (OLS) estimation. OLS is based on the assumptions of 1) linearity of the model ($\mathbf{y} = \Phi\mathbf{w} + \epsilon$), 2) zero-mean model error ($E[\epsilon] = \mathbf{0}$), 3) iid model errors ($E[\epsilon\epsilon^T] = \mathbf{I}_M\rho^{-1}$), and 4) no multicollinearity within basis functions ($\text{Rank}(\Phi) = N$). Following these assumptions, the OLS solution $\mathbf{w}_{\text{ols}}$ is obtained as [6]

$$\mathbf{w}_{\text{ols}} = \arg \min_{\mathbf{w}} \{E_D(\mathbf{w})\} = \arg \min_{\mathbf{w}} \left\{ \frac{1}{2} \|\mathbf{y} - \Phi\mathbf{w}\|_2^2 \right\} = (\Phi^T\Phi)^{-1}\Phi^T\mathbf{y} \quad (\text{A.2})$$

where $\|\cdot\|_2$ denotes the L_2 -norm and $E_D(\mathbf{w})$ is the OLS cost function. The convex optimization task in Eq. (A.2) is solved by differentiating $E_D(\mathbf{w})$ analytically with respect to $\mathbf{w}$ and then substituting it to zero to obtain the OLS estimate $\mathbf{w}_{\text{ols}} = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{y}$. The assumption of no multicollinearity among basis functions implies that no two columns of the design matrix Φ are linearly dependent, thereby ensuring that the matrix $\Phi^T \Phi$ in Eq. (A.2) is invertible. Also, the probability distribution of model error ϵ is not required to be known for computing $\mathbf{w}_{\text{ols}}$ using Eq. (A.2). In OLS, ϵ only need to satisfy the property of $E[\epsilon] = \mathbf{0}$ and $E[\epsilon \epsilon^T] = \mathbf{I}_M \rho^{-1}$. Given $\mathbf{w}_{\text{ols}}$, an unbiased estimate for error precision ρ can also be obtained analytically as $\rho_{\text{ols}} = (M - N) / \|\mathbf{y} - \Phi \mathbf{w}_{\text{ols}}\|_2^2$ [6]. The popularity of OLS is attributed to the following special properties that $\mathbf{w}_{\text{ols}}$ holds: 1) linearity ($\mathbf{w}_{\text{ols}}$ is a linear function of $\mathbf{y}$), 2) unbiasedness ($E[\mathbf{w}_{\text{ols}}] = \mathbf{w}$), 3) efficiency (Variance of $\mathbf{w}_{\text{ols}}$ is minimum for all unbiased estimators of $\mathbf{w}$) [25]. As a result, $\mathbf{w}_{\text{ols}}$ is also known as the best (efficient) linear unbiased estimator of $\mathbf{w}$.

MLE is an another popular frequentist technique of estimating $\mathbf{w}$ by maximizing the log-likelihood function $\log p(\mathbf{y}|\mathbf{w})$. Unlike OLS, MLE is applicable only when the probability distribution of model error ϵ is known. Gaussian model errors are generally used in practice since a Gaussian distribution contains maximum entropy (least information) for an unbounded random variable with a finite mean and a finite variance [158]. The likelihood function for a Gaussian and iid ϵ_i in Eq. (A.1) is written as $p(\mathbf{y}|\mathbf{w}, \rho) = \mathcal{N}(\mathbf{y}|\Phi \mathbf{w}, \mathbf{I}_M \rho^{-1})$. The MLE solution $\mathbf{w}_{\text{mle}}$ of $\mathbf{w}$ is then obtained as [21]

$$\begin{aligned}
 \mathbf{w}_{\text{mle}} &= \arg \max_{\mathbf{w}} \{ \log \mathcal{N}(\mathbf{y}|\Phi \mathbf{w}, \mathbf{I}_M \rho^{-1}) \} \\
 &= \arg \max_{\mathbf{w}} \left\{ -\frac{N}{2} \log 2\pi + \frac{1}{2} \log \rho - \rho E_D(\mathbf{w}) \right\} \\
 &= \arg \min_{\mathbf{w}} \{ E_D(\mathbf{w}) \} = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{y} = \mathbf{w}_{\text{ols}}
 \end{aligned} \tag{A.3}$$

Hence, MLE provides the same solution as OLS for $\mathbf{w}$ for Gaussian and iid model error. The MLE solution for ρ is different from the OLS solution and is obtained analytically as $\rho_{\text{mle}} = M / \|\mathbf{y} - \Phi \mathbf{w}_{\text{mle}}\|_2^2$ [21].

Ideally speaking, the training data consists of multiple realizations of $\mathbf{y}$, where each realization leads to a unique point estimate $\mathbf{w}_{\text{ols}}$ or $\mathbf{w}_{\text{mle}}$ of $\mathbf{w}$. The distribution of these multiple estimates is known as the sampling distribution of $\mathbf{w}$. However, in practice, the training data usually consists of a single realization of input-output measurements due to budgetary or safety constraints. The sampling distribution cannot be computed for such cases, except for 1) Gaussian model errors or 2) ‘large’ number of measurements (large M). For these special circumstances, the sampling distribution of $\mathbf{w}$ is analytically available as $\mathcal{N}(\mathbf{w}|\mathbf{w}_{\text{mle}}, (\rho\Phi^T\Phi)^{-1})$ for both OLS and MLE techniques [40]. This sampling distribution can be subsequently used in computing confidence intervals on model predictions.

The OLS/MLE estimation techniques suffer from the following disadvantages while dealing with over-parameterized models (large N):

- **Under-determined system:** When the number of features exceeds the number of measurements ($N > M$), the OLS/MLE estimation is ill-posed as there exist infinite solutions for $\mathbf{w}$.
- **Multicollinearity:** Even for an over-determined system ($N < M$), the OLS/MLE estimation can become ill-posed if some features are linearly related to each other. This overlap among features is called multicollinearity, causing an increase in the condition number of matrix $\Phi^T\Phi$. The ill-conditioned matrix $\Phi^T\Phi$ causes the OLS/MLE solution $(\Phi^T\Phi)^{-1}\Phi^T\mathbf{y}$ to be highly sensitive to the noise in measurements $\mathbf{y}$, making the OLS/MLE estimation ill-posed. Note that matrix $\Phi^T\Phi$ is singular for the case of perfect collinearity among features since matrix Φ is no longer a full-rank matrix.
- **Overfitting:** For a well-posed estimation involving an over-parameterized model, the OLS/MLE solution is heavily dependent on a single realization of the measurement. An over-parameterized model tends to capture the noise in measurements leading to overfitting and poor generalization to unseen measurements.

RLS is a family of methods designed to remedy these issues with OLS/MLE estimation by

means of regularization. A general form of RLS estimation of $\mathbf{w}$ is defined as

$$\mathbf{w}_{\text{rls}} = \arg \min_{\mathbf{w}} \left\{ \frac{1}{2} \|\mathbf{y} - \Phi \mathbf{w}\|_2^2 + \frac{\lambda}{2} \sum_{i=1}^N |w_i|^p \right\} = \arg \min_{\mathbf{w}} \{E_D(\mathbf{w}) + \lambda E_w(\mathbf{w})\} \quad (\text{A.4})$$

where $E_w(\mathbf{w})$ is the regularization term, $E_D(\mathbf{w})$ is the OLS cost function, and $\lambda > 0$ is the regularization coefficient. The mean term w_0 is excluded from the term $E_w(\mathbf{w})$ to eliminate the dependence of the regularization on the origin (or scale) of the measurements. The following two types of regularization is popular in practice [25]: 1) LASSO, 2) Ridge regression. LASSO uses L_1 -norm regularization ($p = 1$ in Eq. (A.4)) while Ridge regression uses L_2 -norm regularization ($p = 2$ in Eq. (A.4)), also known as Tikhonov regularization. When using a combination of L_1 and L_2 penalty in Eq. (A.4), the RLS estimation is known as *elastic net* [159].

Ridge regression cost function ($p=2$ in Eq. (A.4)) can be differentiated analytically and substituted to zero to obtain the estimate $\mathbf{w}_{\text{ridge}} = (\Phi^T \Phi + \lambda \mathbf{I}_M)^{-1} \Phi^T \mathbf{y}$. On the contrary, LASSO cost function ($p=1$ in Eq. (A.4)) cannot be differentiated analytically and therefore the LASSO solution $\mathbf{w}_{\text{lasso}}$ is computed through numerical optimization [25]. The LARS algorithm provides an efficient alternative to the numerical optimization for providing the LASSO solution for a range of λ values though a stepwise regression approach [75]. Table A.1 summarizes the three RLS techniques for estimating $\mathbf{w}$ in linear regression. Notice that when the regularization coefficient λ is zero, both LASSO and Ridge regression reduces to OLS estimation.

Method	Cost function	$\mathbf{w}$ estimate	Precision of $\mathbf{w}$ estimate
OLS	$\frac{1}{2} \ \mathbf{y} - \Phi \mathbf{w}\ _2^2$	$(\Phi^T \Phi)^{-1} \Phi^T \mathbf{y}$	$\rho \Phi^T \Phi$
LASSO	$\frac{1}{2} \ \mathbf{y} - \Phi \mathbf{w}\ _2^2 + \frac{\lambda}{2} \sum_{i=1}^N w_i $	Numerically (ex. LARS)	Numerically
Ridge	$\frac{1}{2} \ \mathbf{y} - \Phi \mathbf{w}\ _2^2 + \frac{\lambda}{2} \sum_{i=1}^N w_i ^2$	$(\Phi^T \Phi + \lambda \mathbf{I})^{-1} \Phi^T \mathbf{y}$	$\rho(\Phi^T \Phi + \lambda \mathbf{I})(\Phi^T \Phi)^{-1}(\Phi^T \Phi + \lambda \mathbf{I})$

Table A.1: Summary of frequentist estimation techniques for $\mathbf{w}$.

The regularization term $E_w(\mathbf{w})$ with non-zero λ in Eq. (A.4) helps alleviate the aforementioned issues with the OLS/MLE techniques. For example, from Table A.1, computing $\mathbf{w}_{\text{ridge}}$

requires the inverse of matrix $\Phi^T \Phi + \lambda \mathbf{I}_M$, which is always invertible even when matrix $\Phi^T \Phi$ is singular. Also, LARS implementation of LASSO alleviates the need to deal with a potentially ill-conditioned matrix $\Phi^T \Phi$. In addition, the regularization term $E_D(\mathbf{w})$ in Eq. (A.4) act as a penalty on large values of $\mathbf{w}$. This helps restrict the model complexity, resulting in a parsimonious model. The amount of penalty is dictated by the regularization coefficient λ . A large λ over-penalizes the model complexity, thereby producing a model that fails to capture even the salient trends in the measurements. On the contrary, a small λ value is unable to limit model complexity, thereby leading to overfitting by allowing large $\mathbf{w}$ values. Therefore, there exist an optimal value of λ that provides the best balance between model parsimony and data-fit.

In practice, the optimal λ is chosen using the CV technique of minimizing the model prediction error. The CV prediction error is computed for various values of λ and the optimal λ is chosen corresponding to the minimum CV prediction error. The calculation of CV prediction error operates by randomly grouping the M number of measurements into K groups as $\{\mathbf{y}_{(1)}, \mathbf{y}_{(2)}, \dots, \mathbf{y}_{(K)}\}$ where each group contains $M' = M/K$ measurements. For each k^{th} group, LASSO or Ridge regression is used to estimate $\mathbf{w}$ using measurements from all but k^{th} group, denoted as $\mathbf{y}_{(-k)}$. Subsequently, the CV prediction error for a given λ is computed as

$$E_K(\lambda) = \frac{1}{K} \sum_{k=1}^K \sum_{j=1}^{M'} \left(y_{(-k)}^{(j)} - \hat{y}_{(-k)}^{(j)} \right)^2 \quad (\text{A.5})$$

where $\hat{y}_{(-k)}^{(j)}$ is the model prediction pertaining to j^{th} measurement. The optimal λ is chosen corresponding to the minimum CV error $E_K(\lambda)$. This form of cross-validation where the measurements are randomly segmented into K groups is called K-fold cross-validation (K-fold CV). Multiple estimates of K-fold CV error $E_K(\lambda)$ can be obtained for each λ through multiple instances of random grouping of measurements. This repeated estimation of K-fold CV error provides the variation in CV error, thereby ensuring prudent selection of optimal λ . Studies have shown that choosing K between 5 to 10 ensures minimum variation of CV error for any given λ [25]. When the number of measurements and the number of CV data groups are equal

($K = M$), the technique becomes leave-one-out CV. The CV error for this case can be computed without performing the grouping task and just using the $\mathbf{w}$ estimate computed using all the measurements $\mathbf{y}$ as [25]

$$E_M(\lambda) = \frac{1}{M} \sum_{j=1}^M \left(\frac{y^{(j)} - \hat{y}^{(j)}}{1 - h_{jj}} \right)^2 \quad (\text{A.6})$$

where h_{jj} is the j^{th} diagonal term of matrix $\Phi(\Phi^T \Phi)^{-1} \Phi^T$. This estimation of λ by minimizing CV error ensures an optimal amount of penalty in the LASSO or Ridge regression cost function in Eq. (A.4).

The regularization has different effect on $\mathbf{w}$ values for LASSO and Ridge regression. In Ridge regression, increasing λ value forces many parameters towards zero, but never exactly zero [74]. On the contrary, increasing λ in LASSO forces increasingly number of parameters to be exactly zero, thereby providing a sparse solution to $\mathbf{w}$. This sparsity inducing property of LASSO is the reason behind its widespread usage as a sparse learning tool. We will revisit these differences between LASSO and Ridge regression and their applicability as a sparse learning tool during numerical investigations.

A.2 Bayesian linear regression

This section details the Bayesian alternative for estimating $\mathbf{w}$ in the linear regression setup of Eq. (A.1). This Bayesian estimation framework is commonly known as Bayesian linear regression (BLR) [160]. BLR is later exploited in deriving Bayesian sparse learning algorithms of SBL and BCS. In BLR, the unknown model parameters are treated as random variables whose posterior pdf $p(\mathbf{w}|\mathbf{y}, \rho)$ is obtained using Bayes' theorem as

$$p(\mathbf{w}|\mathbf{y}, \rho) = \frac{p(\mathbf{y}|\mathbf{w}, \rho)p(\mathbf{w})}{p(\mathbf{y}|\rho)} \propto p(\mathbf{y}|\mathbf{w}, \rho)p(\mathbf{w}) \quad (\text{A.7})$$

where $p(\mathbf{y}|\mathbf{w}, \rho)$ is the likelihood function, $p(\mathbf{w})$ is the prior pdf and $p(\mathbf{y}|\rho)$ is the model evidence or marginal likelihood or type-II likelihood. We assume that the model errors ϵ_i are Gaussian whose precision ρ is known. As a result, the likelihood function is known as $p(\mathbf{y}|\mathbf{w}, \rho) = \mathcal{N}(\mathbf{y}|\Phi\mathbf{w}, \Gamma)$, where $\Gamma = \rho^{-1}\mathbf{I}_M$.

In most practical scenarios, an analytical evaluation of posterior $p(\mathbf{w}|\mathbf{y}, \rho)$ using Eq. (A.7) is not feasible and stochastic simulation techniques such as MCMC are used to generate stationary samples from the unnormalized posterior pdf in Eq. (A.7) [46]. The stochastic simulation techniques are often computationally expensive and thus create a bottleneck for an efficient application of Bayesian methods to inverse problems. However, the posterior pdf $p(\mathbf{w}|\mathbf{y}, \rho)$ becomes analytically tractable in BLR by choosing the prior pdf $p(\mathbf{w})$ to be a conjugate prior to the given likelihood function. A conjugate prior is defined as the prior that produces posterior in the same probability distribution family as the prior [161]. The popularity of conjugate priors in BLR is attributed to its ability to obtain closed-form solution to the posterior pdf for large parameter dimensionality with minimal computational effort.

For a Gaussian likelihood function $p(\mathbf{y}|\mathbf{w}, \rho) = \mathcal{N}(\mathbf{y}|\Phi\mathbf{w}, \Gamma)$, the conjugate prior for $\mathbf{w}$ is also Gaussian and is assigned as $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{P}_0)$ where $\mathbf{m}_0 \in \mathbb{R}^N$ is the prior mean vector and $\mathbf{P}_0 \in \mathbb{R}^N$ is the prior covariance matrix [161]. The analytical evaluation of posterior $p(\mathbf{w}|\mathbf{y}, \rho)$ requires computation of following relations:

$$\mathbf{E}[\mathbf{y}] = \Phi\mathbf{E}[\mathbf{w}] + \mathbf{E}[\boldsymbol{\epsilon}] = \Phi\mathbf{m}_0, \quad (\text{A.8a})$$

$$\text{Var}(\mathbf{y}) = \Phi\text{Var}(\mathbf{w})\Phi^T + \text{Var}(\boldsymbol{\epsilon}) = \Phi\mathbf{P}_0\Phi^T + \Gamma,$$

$$\mathbf{E}[(\mathbf{w} - \mathbf{m}_0)(\mathbf{y} - \Phi\mathbf{m}_0)^T] = \mathbf{E}[(\mathbf{w} - \mathbf{m}_0)(\Phi\mathbf{w} + \boldsymbol{\epsilon} - \Phi\mathbf{m}_0)^T] \quad (\text{A.8b})$$

$$\begin{aligned} &= \mathbf{E}[(\mathbf{w} - \mathbf{m}_0)(\mathbf{w} - \mathbf{m}_0)^T]\Phi^T + \mathbf{E}[(\mathbf{w} - \mathbf{m}_0)\boldsymbol{\epsilon}^T] \\ &= \mathbf{P}_0\Phi^T, \end{aligned} \quad (\text{A.8c})$$

where $\text{Var}(\cdot)$ represents the covariance. Using the relations from Eq. (A.8), the joint pdf

$p(\mathbf{w}|\mathbf{y}, \rho)$ is obtained as

$$p(\mathbf{w}, \mathbf{y}|\rho) = \mathcal{N} \left(\begin{Bmatrix} \mathbf{w} \\ \mathbf{y} \end{Bmatrix} \middle| \begin{Bmatrix} \mathbf{m}_0 \\ \Phi \mathbf{m}_0 \end{Bmatrix}, \begin{bmatrix} \mathbf{P}_0 & \mathbf{P}_0 \Phi^T \\ \Phi \mathbf{P}_0^T & \Phi \mathbf{P}_0 \Phi^T + \mathbf{\Gamma} \end{bmatrix} \right) \quad (\text{A.9})$$

Using the Gaussian conditional distribution formula listed in Appendix F, the posterior pdf of $\mathbf{w}$ is obtained as

$$p(\mathbf{w}|\mathbf{y}, \rho) = \mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{P}) \quad (\text{A.10a})$$

$$\mathbf{m} = \mathbf{m}_0 + \mathbf{K}(\mathbf{y} - \Phi \mathbf{m}_0) \quad (\text{A.10b})$$

$$\mathbf{P} = (\mathbf{I} - \mathbf{K}\Phi)\mathbf{P}_0 \quad (\text{A.10c})$$

$$\mathbf{K} = \mathbf{P}_0 \Phi^T (\Phi \mathbf{P}_0 \Phi^T + \mathbf{\Gamma})^{-1} \quad (\text{A.10d})$$

where $\mathbf{m}$ is the posterior mean (MAP), $\mathbf{P}$ is the posterior covariance, and $\mathbf{K}$ is the gain matrix. Eq. (A.10) represents the Bayesian updating formula that also laid the foundation for data assimilation techniques such as the Kalman filter [13]. For an unknown error precision ρ , the joint posterior pdf $p(\mathbf{w}, \rho|\mathbf{y})$ can also be computed analytically using a Normal-Gamma conjugate prior for $\mathbf{w}$ and ρ (not discussed here) [161].

The model evidence $p(\mathbf{y}|\rho)$ (denominator in Eq. (A.7)) can also be evaluated analytically using expressions for $E[\mathbf{y}]$ and $\text{Var}(\mathbf{y})$ in Eq. (A.8) leading to $p(\mathbf{y}|\rho) = \mathcal{N}(\mathbf{y}|\Phi \mathbf{m}_0, \Phi \mathbf{P}_0 \Phi^T + \mathbf{\Gamma})$. The evidence can also written in the form (using Eq. (A.7)) [54]

$$\begin{aligned} \log p(\mathbf{y}|\rho) &= \log p(\mathbf{y}|\rho) \int p(\mathbf{w}|\mathbf{y}, \rho) d\mathbf{w} \\ &= \int \log \left(\frac{p(\mathbf{y}|\mathbf{w}, \rho)p(\mathbf{w})}{p(\mathbf{w}|\mathbf{y}, \rho)} \right) p(\mathbf{w}|\mathbf{y}, \rho) d\mathbf{w} \\ &= \int \log p(\mathbf{y}|\mathbf{w}, \rho) p(\mathbf{w}|\mathbf{y}, \rho) d\mathbf{w} - \int \log \left(\frac{p(\mathbf{w}|\mathbf{y}, \rho)}{p(\mathbf{w})} \right) p(\mathbf{w}|\mathbf{y}, \rho) d\mathbf{w} \\ &= \underbrace{E[\log p(\mathbf{y}|\mathbf{w}, \rho)]}_{\text{Goodness-of-fit (GOF)}} - \underbrace{E \left[\log \left(\frac{p(\mathbf{w}|\mathbf{y}, \rho)}{p(\mathbf{w})} \right) \right]}_{\text{Expected Information Gain (EIG)}} \end{aligned} \quad (\text{A.11})$$

where the expectation $E[\cdot]$ is with respect to posterior $p(\mathbf{w}|\mathbf{y}, \rho)$. The GOF term quantifies the data-fit capability of the model as it relates to the likelihood function. The EIG term captures the KL divergence between the prior and posterior pdf, which in turn quantifies model complexity. For complex models the posterior is vastly different from the prior due to the overfitting caused by the unrestricted complexity of the model. In turn, the KL divergence or EIG is higher for complex models which then penalizes the model by reducing the log-evidence according to Eq. (A.11). This competing nature of GOF and EIG imparts model evidence the quantitative Occam's razor property that promotes the compromise between data-fit and model complexity.

Given the current BLR setup, the GOF term in Eq. (A.11) can be calculated analytically by writing the likelihood function as $\mathcal{N}(\mathbf{y}|\Phi\mathbf{w}, \mathbf{\Gamma}) = c\mathcal{N}(\mathbf{w}|\mathbf{w}_{\text{mle}}, \mathbf{P}_{\text{mle}})$, where c is a constant of proportionality, $\mathbf{w}_{\text{mle}} = (\Phi^T\Phi)^{-1}\Phi^T\mathbf{y}$ is the MLE estimate, and $\mathbf{P}_{\text{mle}}^{-1} = \Phi^T\mathbf{\Gamma}^{-1}\Phi$ is the MLE precision matrix. As a result, the GOF term in Eq. (A.11) can be written as

$$\begin{aligned} \text{GOF} &= E[\log p(\mathbf{y}|\mathbf{w}, \rho)] = E[\log \mathcal{N}(\mathbf{y}|\Phi\mathbf{w}, \mathbf{\Gamma})] = E[\log(c\mathcal{N}(\mathbf{w}|\mathbf{w}_{\text{mle}}, \mathbf{P}_{\text{mle}}))] \\ &= \log c + \int \mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{P}) \log \mathcal{N}(\mathbf{w}|\mathbf{w}_{\text{mle}}, \mathbf{P}_{\text{mle}}) d\mathbf{w} . \end{aligned} \quad (\text{A.12})$$

The integral in Eq. (A.12) is the cross-entropy between the posterior pdf $\mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{P})$ from BLR. Using the cross-entropy relation for multivariate Gaussian pdfs derived in Eq. (F.2) in Appendix F, we get.

$$\begin{aligned} \text{GOF} &= \log c + \log \mathcal{N}(\mathbf{m}|\mathbf{w}_{\text{mle}}, \mathbf{P}_{\text{mle}}) - \frac{1}{2} \text{Trace}(\mathbf{P}_{\text{mle}}^{-1} \mathbf{P}) \\ &= \log \mathcal{N}(\mathbf{y}|\Phi\mathbf{m}, \mathbf{\Gamma}) - \frac{1}{2} \text{Trace}(\mathbf{P}_{\text{mle}}^{-1} \mathbf{P}) . \end{aligned} \quad (\text{A.13})$$

Next, the EIG term in Eq. (A.11) can also be evaluated analytically using Eq. (F.2) from Ap-

pendix **F** as

$$\begin{aligned}
 \text{EIG} &= \mathbb{E} \left[\log \left(\frac{p(\mathbf{w}|\mathbf{y}, \rho)}{p(\mathbf{w})} \right) \right] = \mathbb{E}[\log \mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{P})] - \mathbb{E}[\log \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{P}_0)] \\
 &= \log \mathcal{N}(\mathbf{m}|\mathbf{m}, \mathbf{P}) - \frac{1}{2} \text{Trace}(\mathbf{P}^{-1} \mathbf{P}) - \log \mathcal{N}(\mathbf{m}|\mathbf{m}_0, \mathbf{P}_0) + \frac{1}{2} \text{Trace}(\mathbf{P}_0^{-1} \mathbf{P}) \\
 &= \log \left(\frac{\mathcal{N}(\mathbf{m}|\mathbf{m}, \mathbf{P})}{\mathcal{N}(\mathbf{m}|\mathbf{m}_0, \mathbf{P}_0)} \right) - \frac{N}{2} + \frac{1}{2} \text{Trace}(\mathbf{P}_0^{-1} \mathbf{P}) .
 \end{aligned} \tag{A.14}$$

Note that when the posterior $p(\mathbf{w}|\mathbf{y}, \rho)$ is entirely dictated by measurements $\mathbf{y}$, the posterior precision $\mathbf{P}^{-1}$ is same as the OLS/MLE precision $\mathbf{P}_{\text{mle}}^{-1}$. In this case, $\text{Trace}(\mathbf{P}_{\text{mle}}^{-1} \mathbf{P}) \approx N$ in Eq. (A.13) and $\text{Trace}(\mathbf{P}_0^{-1} \mathbf{P}) \approx 0$ in Eq. (A.14).

Next, the posterior predictive distribution of unseen measurements $\tilde{\mathbf{y}}$ can be computed using the new design matrix $\tilde{\Phi}$ (computed at the new input $\tilde{\mathbf{u}}$ of our choice) as

$$\begin{aligned}
 p(\tilde{\mathbf{y}}|\mathbf{y}, \rho) &= \int p(\tilde{\mathbf{y}}|\mathbf{w}, \rho) p(\mathbf{w}|\mathbf{y}, \rho) d\mathbf{w} \\
 &= \int \mathcal{N}(\tilde{\mathbf{y}}|\tilde{\Phi}\mathbf{w}, \mathbf{\Gamma}) \mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{P}) d\mathbf{w} \\
 &= \mathcal{N}(\tilde{\mathbf{y}}|\tilde{\Phi}\mathbf{m}, \tilde{\Phi}\mathbf{P}\tilde{\Phi}^T + \mathbf{\Gamma}) ,
 \end{aligned} \tag{A.15}$$

where $\mathbf{\Gamma} = \rho^{-1} \mathbf{I}_M$ is known. Table A.2 provides the summary of BLR estimation of $\mathbf{w}$ using conjugate priors.

Next, we provide an interesting contrast between the posterior pdf $\mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{P})$ from BLR in Eq. (A.10) and the sampling distribution $\mathcal{N}(\mathbf{w}|\mathbf{w}_{\text{mle}}, \mathbf{P}_{\text{mle}})$ from the frequentist approaches of OLS/MLE [84]. Rewriting the posterior covariance $\mathbf{P}$ from Eq. (A.9) using the Woodbury formula (Eq. 156 in [137]) leads to

$$\begin{aligned}
 \mathbf{P} &= \mathbf{P}_0 - \mathbf{P}_0 \Phi^T (\Phi \mathbf{P}_0 \Phi^T + \mathbf{\Gamma})^{-1} \Phi \mathbf{P}_0 = (\mathbf{P}_0^{-1} + \Phi^T \mathbf{\Gamma}^{-1} \Phi)^{-1} \\
 &= \mathbf{P}_0^{-1} + \Phi^T \mathbf{\Gamma}^{-1} \Phi = \mathbf{P}_0^{-1} + \mathbf{P}_{\text{mle}}^{-1} .
 \end{aligned} \tag{A.16}$$

Hence, the posterior precision $\mathbf{P}^{-1}$ of $\mathbf{w}$ is the sum of prior precision $\mathbf{P}_0^{-1}$ and the precision

Model	$\mathbf{y} = \Phi \mathbf{w} + \boldsymbol{\epsilon}$
measurements	$\mathbf{y} = \{y_1, \dots, y_M\}^T$
Model error	$p(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{\epsilon}, \mathbf{0}, \mathbf{I}_M \rho^{-1}) = \mathcal{N}(\boldsymbol{\epsilon} \mathbf{0}, \mathbf{\Gamma})$
Prior pdf	$p(\mathbf{w}) = \mathcal{N}(\mathbf{w} \mathbf{m}_0, \mathbf{P}_0)$
Likelihood function	$p(\mathbf{y} \mathbf{w}, \rho) = \mathcal{N}(\mathbf{y} \Phi \mathbf{w}, \mathbf{\Gamma})$
Posterior pdf	$p(\mathbf{w} \mathbf{y}, \rho) = \mathcal{N}(\mathbf{w} \mathbf{m}, \mathbf{P})$ $\mathbf{m} = \mathbf{m}_0 + \mathbf{K}(\mathbf{y} - \Phi \mathbf{m}_0)$ $\mathbf{P} = (\mathbf{I} - \mathbf{K}\Phi) \mathbf{P}_0$ $\mathbf{K} = \mathbf{P}_0 \Phi^T (\Phi \mathbf{P}_0 \Phi^T + \mathbf{\Gamma})^{-1}$
Evidence	$p(\mathbf{y} \rho) = \mathcal{N}(\mathbf{y} \Phi \mathbf{m}_0, \Phi \mathbf{P}_0 \Phi^T + \mathbf{\Gamma})$
Predictive pdf	$p(\tilde{\mathbf{y}} \mathbf{y}, \rho) = \mathcal{N}(\tilde{\mathbf{y}} \tilde{\Phi} \mathbf{m}, \tilde{\Phi} \mathbf{P} \tilde{\Phi}^T + \mathbf{\Gamma})$

Table A.2: Summary of BLR.

$\Phi^T \mathbf{\Gamma}^{-1} \Phi$ of OLS/MLE estimate. Given that precision is analogous to knowledge, this inference re-iterates the fact that the posterior knowledge obtained through Bayesian inference is the sum total of prior knowledge and the information provided by the measurements. It is also realized that precision adds up when solving inverse problem, while variance adds up when solving a forward UQ problem.

Next, posterior mean in Eq. (A.9) is re-written in the form

$$\begin{aligned}
\mathbf{m} &= \mathbf{m}_0 + \mathbf{K}(\mathbf{y} - \Phi \mathbf{m}_0) = (\mathbf{I} - \mathbf{K}\Phi) \mathbf{m}_0 + \mathbf{P}_0 \Phi^T (\Phi \mathbf{P}_0 \Phi^T + \mathbf{\Gamma})^{-1} \mathbf{y} \\
&= \mathbf{P} \mathbf{P}_0^{-1} \mathbf{m}_0 + (\mathbf{P}_0^{-1} + \Phi^T \mathbf{\Gamma}^{-1} \Phi)^{-1} \Phi^T \mathbf{\Gamma}^{-1} \mathbf{y} \\
&= \mathbf{P} \mathbf{P}_0^{-1} \mathbf{m}_0 + \mathbf{P} (\Phi^T \mathbf{\Gamma}^{-1} \Phi) (\Phi^T \mathbf{\Gamma}^{-1} \Phi)^{-1} \Phi^T \mathbf{\Gamma}^{-1} \mathbf{y} \\
&= (\mathbf{P} \mathbf{P}_0^{-1}) \mathbf{m}_0 + (\mathbf{P} \mathbf{P}_{\text{mle}}^{-1}) \mathbf{w}_{\text{mle}} .
\end{aligned} \tag{A.17}$$

Hence, the posterior mean $\mathbf{m}$ is the weighted sum of prior mean $\mathbf{m}_0$ and the OLS/MLE estimate $\mathbf{w}_{\text{ols}}$. Following observations can be made regarding the Bayesian apparatus by analyzing Eq. (A.16) and Eq. (A.17):

- When using flat priors (i.e. $\mathbf{P}_0^{-1} \approx 0$), the posterior pdf from BLR is same as the sampling distribution from OLS/MLE since $\mathbf{P} \approx \mathbf{P}_{\text{mle}}$ in Eq. (A.16) and $\mathbf{m} \approx \mathbf{w}_{\text{mle}}$ in Eq. (A.17). This

resemblance between BLR and OLS/MLE is attributed to the purely data-driven Bayesian estimation of $\mathbf{w}$ due to flat priors.

- In the case of high measurement noise ($\Gamma^{-1} = \rho \approx 0$) or large prior knowledge ($\mathbf{P}_0^{-1} \gg \Phi^T \Gamma^{-1} \Phi$), the posterior pdf is same as the prior pdf $\mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{P}_0)$, meaning no information is gained from the data due to the lack of information in measurements or abundance of knowledge in prior, respectively.

This subjectivity of the Bayesian framework in conditioning the inference using prior knowledge is the primary reason behind its widespread usage in modelling engineering systems. Next, we detail the sparse learning algorithm of SBL as a Bayesian alternative to remedy overfitting in linear regression.

A.3 Sparse Bayesian learning

SBL is a sparse learning algorithm first proposed by Tipping [78] to obtain sparse solution to linear-in-parameter models as an alternative to SVMs [7]. SBL provides a Bayesian alternative to sparse SVM construction, and the resulting algorithm is termed as RVM. Nevertheless, SBL is not exclusive to SVM models and can be applied to any linear-in-parameter model of form Eq. (A.1) where the model error is additive, iid and Gaussian. The mathematical exposition of SBL presented in this section is derived from [78, 87, 136].

Given the linear-in-parameter model in Eq. (A.1) and a Gaussian likelihood function, SBL operates by assigning a zero-mean conjugate prior to $\mathbf{w}$ as

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{i=0}^{N-1} \mathcal{N}(w_i|0, \alpha_i^{-1}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{A}^{-1}) , \quad (\text{A.18})$$

where $\mathbf{A} = \text{Diag}(\boldsymbol{\alpha})$ is the prior precision matrix, and $\boldsymbol{\alpha}$ is the unknown hyperparameter vector. This zero-mean Gaussian prior facilitates the addition/removal of each parameter w_i by varying the value of the corresponding hyperparameter α_i . For example, when α_i is infinity, the marginal

prior pdf $\mathcal{N}(w_i|0, \alpha^{-1})$ is close to a Dirac-delta function $\mathcal{N}(w_i|0, 0)$. According to Eq. (A.16) and Eq. (A.17), this prior pdf forces the posterior to be a Dirac-delta function centered at zero, thereby eliminating the contribution of basis $\phi_i(\mathbf{u})$ from the model. On the contrary, a finite value of α_i entails a non-zero contribution of basis $\phi_i(\mathbf{u})$ in the model. This property of the prior $p(\mathbf{w}|\boldsymbol{\alpha})$ to control the contribution of each basis by varying α_i is called ARD and the prior $\mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{A}^{-1})$ in Eq. (A.18) is termed as ARD prior.

Given the ARD prior $\mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{A}^{-1})$, the posterior pdf $p(\mathbf{w}|\mathbf{y}, \boldsymbol{\alpha}, \rho)$ is analytically available using BLR from Sec. A.2 as

$$p(\mathbf{w}|\mathbf{y}, \boldsymbol{\alpha}, \rho) = \frac{p(\mathbf{y}|\mathbf{w}, \boldsymbol{\alpha}, \rho)p(\mathbf{w}|\boldsymbol{\alpha})}{p(\mathbf{y}|\boldsymbol{\alpha}, \rho)} \propto \mathcal{N}(\mathbf{y}|\Phi\mathbf{w}, \rho^{-1}\mathbf{I}_M)\mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{A}^{-1}) = \mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{P}) \quad (\text{A.19})$$

where the posterior mean $\mathbf{m}$ and the posterior covariance $\mathbf{P}$ are available from Eq. (A.10) as

$$\mathbf{B} = \Phi\mathbf{A}^{-1}\Phi^T + \mathbf{I}_M\rho^{-1} \quad (\text{A.20a})$$

$$\mathbf{K} = \mathbf{A}^{-1}\Phi^T(\Phi\mathbf{A}^{-1}\Phi^T + \mathbf{I}_M\rho^{-1})^{-1} = \mathbf{A}^{-1}\Phi^T\mathbf{B}^{-1} \quad (\text{A.20b})$$

$$\mathbf{m} = (\mathbf{A} + \rho\Phi^T\Phi)^{-1}(\rho\Phi^T\mathbf{y}) = \mathbf{K}\mathbf{y} \quad (\text{A.20c})$$

$$\mathbf{P} = (\mathbf{A} + \rho\Phi^T\Phi)^{-1} = (\mathbf{I} - \mathbf{K}\Phi)\mathbf{A}^{-1} \quad (\text{A.20d})$$

Consequently, the model evidence $p(\mathbf{y}|\boldsymbol{\alpha}, \rho)$ (denominator in Eq. (A.19)) is also available analytically from BLR as $\mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{B})$, where $\mathbf{B}$ is known from Eq. (A.20a).

Next, the prior pdfs for the error precision ρ and the hyperparameter vector $\boldsymbol{\alpha}$ is chosen as [78]

$$p(\boldsymbol{\alpha}) = \prod_{i=0}^{N-1} \mathcal{G}(\alpha_i|c, d), \quad p(\rho) = \mathcal{G}(\rho|a, b), \quad (\text{A.21})$$

where $a, b, c, d > 0$ are some known parameters. The prior $p(\boldsymbol{\alpha})$ is known as the hyper-prior (prior for hyper-parameter). Note that $\mathcal{G}(x|r, s)$ denotes a gamma distribution with pdf [137]

$$\mathcal{G}(x|r, s) = \frac{s^r}{\Gamma(r)} x^{r-1} e^{-sx}; \quad \mathbb{E}[x] = \frac{s}{r}; \quad \text{Var}(x) = \frac{s}{r^2}. \quad (\text{A.22})$$

Using the hierarchical Bayes approach, the posterior pdf $p(\boldsymbol{\alpha}, \rho|\mathbf{y})$ is written as [78]

$$\begin{aligned} p(\boldsymbol{\alpha}, \rho|\mathbf{y}) &= \frac{p(\mathbf{y}|\boldsymbol{\alpha}, \rho)p(\boldsymbol{\alpha}, \rho)}{p(\mathbf{y})} \propto p(\mathbf{y}|\boldsymbol{\alpha}, \rho)p(\rho)p(\boldsymbol{\alpha}) \\ &\propto \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{B})\mathcal{G}(\rho|a, b) \prod_{i=0}^{N-1} \mathcal{G}(\alpha_i|c, d) \end{aligned} \quad (\text{A.23})$$

For the sake of seeking sparsity in $\mathbf{w}$, our interest lies in finding the mode or the MAP estimate $(\boldsymbol{\alpha}^{\text{MAP}}, \rho^{\text{MAP}})$ obtained by maximizing the posterior $p(\boldsymbol{\alpha}, \rho|\mathbf{y})$ in Eq. (A.23). The MAP estimate $\boldsymbol{\alpha}^{\text{MAP}}$ produces a sparse construction of $\mathbf{w}$ as many α_i^{MAP} approach infinity during evidence maximization, thereby forcing both prior and posterior of w_i to be a Dirac-delta function centered at zero. Therefore, the problem of finding a sparse solution to $\mathbf{w}$ is transformed into finding $\boldsymbol{\alpha}^{\text{MAP}}$ that maximizes the posterior pdf $p(\boldsymbol{\alpha}, \rho|\mathbf{y})$. This sparse learning optimization is reposed in terms of $\log p(\boldsymbol{\alpha}, \rho|\mathbf{y})$ as

$$(\boldsymbol{\alpha}^{\text{MAP}}, \rho^{\text{MAP}}) = \arg \max \{\log p(\boldsymbol{\alpha}, \rho|\mathbf{y})\} = \arg \max \{\mathcal{L}(\boldsymbol{\alpha}, \rho)\} , \quad (\text{A.24})$$

where $\mathcal{L}(\boldsymbol{\alpha}, \rho)$ is the objective function to be maximized. $\mathcal{L}(\boldsymbol{\alpha}, \rho)$ is obtained using Eq. (A.23) as (ignoring constant terms)

$$\mathcal{L}(\boldsymbol{\alpha}, \rho) = -\frac{1}{2} \{\log |\mathbf{B}| + \mathbf{y}^T \mathbf{B}^{-1} \mathbf{y}\} + (a-1) \log \rho - b\rho + \sum_{i=0}^{N-1} \{(c-1) \log \alpha_i - d\alpha_i\} . \quad (\text{A.25})$$

The optimization problem posed in Eq. (A.24) is a convex optimization problem (proved later in this section), and can be solved semi-analytically through analytical differentiation of $\mathcal{L}(\boldsymbol{\alpha}, \rho)$. Based on the approach undertaken to differentiate $\mathcal{L}(\mathbf{w}, \rho)$, there exist two types of SBL methods: 1) the original SBL method by Tipping [78], and 2) the accelerated SBL method by Tipping [87] and Faul [136].

A.3.1 Original algorithm

The idea behind this algorithm is to isolate the effect of hyperparameter α_i and error precision ρ by expanding terms $\log |\mathbf{B}|$ and $\mathbf{y}^T \mathbf{B}^{-1} \mathbf{y}$ present in the objective function $\mathcal{L}(\alpha, \rho)$ in Eq. (A.25). To begin with, the $\log |\mathbf{B}|$ term in Eq. (A.25) is re-written using the matrix determinant identity [137]

$$|\mathbf{A}| |\Phi \mathbf{A}^{-1} \Phi^T + \mathbf{I}_M \rho^{-1}| = |\mathbf{I}_M \rho^{-1}| |\mathbf{A} + \rho \Phi^T \Phi|, \quad (\text{A.26})$$

leading to

$$\begin{aligned} \log |\mathbf{B}| &= \log |\Phi \mathbf{A}^{-1} \Phi^T + \mathbf{I}_M \rho^{-1}| = \log |\mathbf{I}_M \rho^{-1}| + \log |\mathbf{A} + \rho \Phi^T \Phi| - \log |\mathbf{A}| \\ &= -M \log \rho + \log |\mathbf{P}^{-1}| - \sum_{i=1}^{N-1} \log \alpha_i. \end{aligned} \quad (\text{A.27})$$

Furthermore, the term $\mathbf{y}^T \mathbf{B}^{-1} \mathbf{y}$ in Eq. (A.25) is evaluated using the Woodbury identity (Eq. 156 in [137])

$$\mathbf{B}^{-1} = (\Phi \mathbf{A}^{-1} \Phi^T + \mathbf{I}_M \rho^{-1})^{-1} = \mathbf{I}_M \rho - \rho \Phi (\mathbf{A} + \rho \Phi^T \Phi)^{-1} \Phi^T \rho = \mathbf{I}_M \rho - \rho^2 \Phi \mathbf{P} \Phi^T \quad (\text{A.28})$$

to obtain

$$\begin{aligned} \mathbf{y}^T \mathbf{B}^{-1} \mathbf{y} &= \mathbf{y}^T (\mathbf{I}_M \rho - \rho^2 \Phi \mathbf{P} \Phi^T) \mathbf{y} = \rho \mathbf{y}^T (\mathbf{y} - \Phi (\rho \mathbf{P} \Phi^T \mathbf{y})) \\ &= \rho (\mathbf{y} - \Phi \mathbf{m} + \Phi \mathbf{m})^T (\mathbf{y} - \Phi \mathbf{m}) \\ &= \rho \|\mathbf{y} - \Phi \mathbf{m}\|_2^2 + \mathbf{m}^T \mathbf{P}^{-1} (\rho \mathbf{P} \Phi^T \mathbf{y}) - \rho \mathbf{m}^T \Phi^T \Phi \mathbf{m} \\ &= \rho \|\mathbf{y} - \Phi \mathbf{m}\|_2^2 + \mathbf{m}^T (\mathbf{P}^{-1} - \rho \Phi^T \Phi) \mathbf{m} \\ &= \rho \|\mathbf{y} - \Phi \mathbf{m}\|_2^2 + \mathbf{m}^T \mathbf{A} \mathbf{m} \\ &= \rho \|\mathbf{y} - \Phi \mathbf{m}\|_2^2 + \sum_{i=0}^{N-1} \alpha_i m_i^2, \end{aligned} \quad (\text{A.29})$$

where m_i is the posterior mean of w_i .

Substituting $\log |\mathbf{B}|$ from Eq. (A.27) and $\mathbf{y}^T \mathbf{B}^{-1} \mathbf{y}$ from Eq. (A.29) in the objective function $\mathcal{L}(\boldsymbol{\alpha}, \rho)$ in Eq. (A.25) produces

$$\begin{aligned} \mathcal{L}(\boldsymbol{\alpha}, \rho) = & -\frac{1}{2} \left\{ -M \log \rho + \log |\mathbf{P}^{-1}| + \rho \|\mathbf{y} - \Phi \mathbf{m}\|_2^2 + \sum_{i=0}^{N-1} (\alpha_i m_i^2 - \log \alpha_i) \right\} \\ & + (a-1) \log \rho - b\rho + \sum_{i=0}^{N-1} \{(c-1) \log \alpha_i - d\alpha_i\} . \end{aligned} \quad (\text{A.30})$$

Differentiating $\mathcal{L}(\mathbf{w}, \rho)$ from Eq. (A.30) with respect to α_i leads to

$$\frac{\partial \mathcal{L}(\boldsymbol{\alpha}, \rho)}{\partial \alpha_i} = -\frac{1}{2} \left\{ \frac{\partial(\log |\mathbf{P}^{-1}|)}{\partial \alpha_i} - \frac{1}{\alpha_i} + m_i^2 \right\} + \frac{(c-1)}{\alpha_i} - d . \quad (\text{A.31})$$

The differentiation of $\log |\mathbf{P}^{-1}|$ is computed using the relation $\mathbf{P}^{-1} = \mathbf{A} + \rho \Phi^T \Phi$ and an identity involving derivative of a matrix determinant (Eq. 46 in [137]) as

$$\begin{aligned} \frac{\partial(\log |\mathbf{P}^{-1}|)}{\partial \alpha_i} &= \frac{1}{|\mathbf{P}^{-1}|} \frac{\partial |\mathbf{P}^{-1}|}{\partial \alpha_i} = \text{Trace} \left(\mathbf{P} \frac{\partial(\mathbf{A} + \rho \Phi^T \Phi)}{\partial \alpha_i} \right) \\ &= \text{Trace} \left(\mathbf{P} \frac{\partial(\text{Diag}(\boldsymbol{\alpha}))}{\partial \alpha_i} \right) = P_{ii}, \end{aligned} \quad (\text{A.32})$$

where P_{ii} is the $(i, i)^{th}$ element of the posterior covariance matrix $\mathbf{P}$ in Eq. (A.20). Equating Eq. (A.31) to zero we get

$$\begin{aligned} \frac{\partial \mathcal{L}(\boldsymbol{\alpha}, \rho)}{\partial \alpha_i} &= -\frac{1}{2} \left\{ P_{ii} - \frac{1}{\alpha_i} + m_i^2 \right\} + \frac{(c-1)}{\alpha_i} - d = 0 \\ \frac{2(c-1) + 1}{\alpha_i} &= P_{ii} + m_i^2 + 2d \\ \alpha_i &= \frac{1 + 2(c-1)}{P_{ii} + m_i^2 + 2d} \end{aligned} \quad (\text{A.33})$$

This update formula for α_i is similar to the one obtained using EM (More details in Appendix A in [78]). Following Mackay [53], Tipping [87] suggested the use of a new entity $r_i = 1 - \alpha_i P_{ii}$ while writing the iterative re-estimation equation in Eq. (A.33). As a result, Eq. (A.33) is re-

written as

$$\alpha_i = \frac{r_i + 2(c - 1)}{m_i^2 + 2d} \quad (\text{A.34})$$

Similarly, differentiating $\mathcal{L}(\mathbf{w}, \rho)$ in Eq. (A.30) with respect to ρ and equating it to zero leads to

$$\begin{aligned} \frac{\partial \mathcal{L}(\boldsymbol{\alpha}, \rho)}{\partial \rho} &= -\frac{M}{\rho} + \frac{1}{|\mathbf{P}^{-1}|} \frac{\partial(|\mathbf{P}^{-1}|)}{\partial \rho} + \|\mathbf{y} - \Phi \mathbf{m}\|_2^2 - \frac{2(a-1)}{\rho} + 2b = 0 \\ \frac{M + 2a - 2}{\rho} &= \frac{1}{|\mathbf{P}^{-1}|} |\mathbf{P}^{-1}| \text{Trace} \left(\mathbf{P} \frac{\partial(\mathbf{A} + \rho \Phi^T \Phi)}{\partial \rho} \right) + \|\mathbf{y} - \Phi \mathbf{m}\|_2^2 + 2b \\ M + 2(a-1) &= \rho \left(\text{Trace}(\mathbf{P} \Phi^T \Phi) + \|\mathbf{y} - \Phi \mathbf{m}\|_2^2 + 2b \right) \\ M + 2(a-1) &= \rho \left(\rho^{-1} \sum r_i + \|\mathbf{y} - \Phi \mathbf{m}\|_2^2 + 2b \right) \\ \rho &= \frac{M - \sum r_i + 2(a-1)}{\|\mathbf{y} - \Phi \mathbf{m}\|_2^2 + 2b} \end{aligned} \quad (\text{A.35})$$

In summary, Eq. (A.34) and Eq. (A.35) provide the iterative re-estimation equations for computing the MAP estimates $\boldsymbol{\alpha}^{\text{MAP}}$ and ρ^{MAP} , wherein the posterior statistics ($\mathbf{m}$ and $\mathbf{P}$) is updated using Eq. (A.20) after each iteration.

The introduction of entity $r_i = 1 - \alpha_i P_{ii}$ in SBL and in Eq. (A.34) and Eq. (A.35) can be explained by rewriting r_i using Eq. (A.20) as

$$r_i = 1 - \frac{\alpha_i}{P_{ii}^{-1}} = 1 - \frac{\alpha_i}{\alpha_i + \rho \Phi_i^T \Phi_i} \in [0, 1], \quad (\text{A.36})$$

where $\rho \Phi_i^T \Phi_i$ is the OLS/MLE contribution to the posterior precision P_{ii}^{-1} of w_i . For an irrelevant parameter w_i , the prior pdf approaches a Dirac-delta function during SBL, meaning the prior precision α_i is large (low prior variance) and the posterior is dictated by prior so that $P_{ii}^{-1} \approx \alpha_i$ or $r_i \approx 0$. Alternatively, for a relevant parameter w_i , the posterior precision P_{ii}^{-1} is dictated by measurements, thereby forcing $P_{ii}^{-1} \approx \rho^{-1} \Phi_i^T \Phi_i$ or $r_i \approx 1$. Hence, r_i can be thought of as a quantitative indicator of relevance for parameter w_i or basis ϕ_i . An r_i value of zero imply irrelevance, while an r_i value of one implies relevance. We hereby refer to r_i as the

relevance indicator.

Algorithm 3 summarizes the steps needed for implementing the SBL algorithm. The convergence of the original SBL algorithm is determined by monitoring the change in the objective function $\mathcal{L}(\alpha, \rho)$ value. Following the convergence, the relevance of each w_i is determined by examining relevance indicator values pertaining to the optima. Unlike α_i values, the relevance indicator provides a scale-independent measure of relevance that varies between zero and one.

Algorithm 3 Original SBL method by Tipping [78]

- 1: Training data: $\{(y^{(i)}, \mathbf{u}^{(i)}) ; i = 1, 2, \dots, M\}$;
 - 2: Propose N basis functions, and construct the design matrix Φ according to Eq. (A.1);
 - 3: Choose hyper-prior parameter a, b, c and d in Eq. (A.21);
 - 4: Choose a starting value of α and ρ ;
 - 5: **while** no convergence **do**
 - 6: Compute posterior entities $\mathbf{m}, \mathbf{P}$ and $\mathbf{B}$ using Eq. (A.20);
 - 7: Compute the relevance indicator $r_i = 1 - \alpha_i P_{ii}$ for all basis;
 - 8: Compute the objective function $\mathcal{L}(\alpha, \rho)$ using Eq. (A.30);
 - 9: Choose an index $1 \leq i \leq N$ randomly or in a predefined fashion;
 - 10: Update α_i using Eq. (A.34);
 - 11: Update ρ using Eq. (A.35)
 - 12: **end while**
-

A.3.2 Accelerated algorithm

The accelerated SBL algorithm by Faul and Tipping [136] maximizes model evidence through an explicit addition and deletion of basis functions from the model in Eq. (A.1). The addition of the parameter w_i is performed by assigning α_i a finite value, while a basis is deleted by setting α_i to infinity. This approach is contrary to the original SBL algorithm discussed in Sec. A.3.1 that always has a finite hyper-parameter value which in-turn controls the contribution of each basis function. The mathematical exposition presented here is derived from Faul and Tipping [136].

The accelerated SBL algorithm assumes flat hyperpriors for α and ρ using $a = b = 1$ and $c = d = 0$ in Eq. (A.18). In that case, the SBL objective function in Eq. (A.25) is same as

log-evidence and is obtained as

$$\mathcal{L}(\boldsymbol{\alpha}, \rho) = -\frac{1}{2} \left\{ \log |\mathbf{B}| + \mathbf{y}^T \mathbf{B}^{-1} \mathbf{y} \right\} \quad (\text{A.37})$$

where $\mathbf{B} = \Phi \mathbf{A}^{-1} \Phi^T + \rho^{-1} \mathbf{I}_M$. The underlying structure of matrix $\mathbf{B}$ is revealed when expanding it as

$$\begin{aligned} \mathbf{B} &= \{\Phi_0 \ \Phi_1 \ \dots \ \Phi_{N-1}\} \begin{bmatrix} \frac{1}{\alpha_0} & & & \\ & \frac{1}{\alpha_1} & & \\ & & \ddots & \\ & & & \frac{1}{\alpha_{N-1}} \end{bmatrix} \begin{Bmatrix} \Phi_0^T \\ \Phi_1^T \\ \vdots \\ \Phi_{N-1}^T \end{Bmatrix} + \rho^{-1} \mathbf{I}_M \\ &= \rho^{-1} \mathbf{I}_M + \sum_{j=0}^{N-1} \frac{\Phi_j \Phi_j^T}{\alpha_j} \end{aligned} \quad (\text{A.38})$$

where Φ_j is the j^{th} column of the design matrix Φ . The dependence of matrix $\mathbf{B}$ on each hyper-parameter α_i can be isolated by rewriting Eq. (A.38) as

$$\mathbf{B} = \left\{ \rho^{-1} \mathbf{I}_M + \sum_{\substack{j=0 \\ j \neq i}}^{N-1} \frac{\Phi_j \Phi_j^T}{\alpha_j} \right\} + \frac{\Phi_i \Phi_i^T}{\alpha_i} = \mathbf{B}_{-i} + \frac{\Phi_i \Phi_i^T}{\alpha_i} \quad (\text{A.39})$$

where $\mathbf{B}_{-i}$ quantifies the effect of all hyper-parameters but α_i , while the term $\Phi_i \Phi_i^T / \alpha_i$ is a sole function of α_i .

Next, matrix determinant lemma of $|X + vv^T| = |X|(1 + v^T X^{-1}v)$ is invoked where X is any invertible matrix [162]. Using this lemma and the decomposition of $\mathbf{B}$ in Eq. (A.39), the term $\log |\mathbf{B}|$ in Eq. (A.37) is expanded as

$$\begin{aligned} \log |\mathbf{B}| &= \log \left| \mathbf{B}_{-i} + \frac{\Phi_i \Phi_i^T}{\alpha_i} \right| = \log \left(|\mathbf{B}_{-i}| \left(1 + \frac{\Phi_i^T \mathbf{B}_{-i} \Phi_i}{\alpha_i} \right) \right) \\ &= \log |\mathbf{B}_{-i}| + \log(\alpha_i + \Phi_i^T \mathbf{B}_{-i} \Phi_i) - \log \alpha_i. \end{aligned} \quad (\text{A.40})$$

Using Sherman-Morrison identity (Eq. 160 in [137]), the term $\mathbf{y}^T \mathbf{B} \mathbf{y}$ in Eq. (A.37) is expanded as

$$\begin{aligned} \mathbf{y}^T \mathbf{B}^{-1} \mathbf{y} &= \mathbf{y}^T \left(\mathbf{B}_{-i} + \frac{\Phi_i \Phi_i^T}{\alpha_i} \right)^{-1} \mathbf{y} \\ &= \mathbf{y}^T \left(\mathbf{B}_{-i}^{-1} - \frac{\mathbf{B}_{-i}^{-1} \Phi_i \Phi_i^T \mathbf{B}_{-i}^{-1}}{\alpha_i + \Phi_i^T \mathbf{B}_{-i}^{-1} \Phi_i} \right) \mathbf{y} \\ &= \mathbf{y}^T \mathbf{B}_{-i}^{-1} \mathbf{y} - \frac{\mathbf{y}^T \mathbf{B}_{-i}^{-1} \Phi_i \Phi_i^T \mathbf{B}_{-i}^{-1} \mathbf{y}}{\alpha_i + \Phi_i^T \mathbf{B}_{-i}^{-1} \Phi_i}. \end{aligned} \quad (\text{A.41})$$

Using Eq. (A.40) and Eq. (A.41), the SBL objective function $\mathcal{L}(\boldsymbol{\alpha}, \rho)$ in Eq. (A.37) is obtained as

$$\mathcal{L}(\boldsymbol{\alpha}, \rho) = \mathcal{L}(\boldsymbol{\alpha}_{-i}, \rho) + l(\alpha_i) \quad (\text{A.42a})$$

$$\mathcal{L}(\boldsymbol{\alpha}_{-i}, \rho) = -\frac{1}{2} \{ \log |\mathbf{B}_{-i}| + \mathbf{y}^T \mathbf{B}_{-i}^{-1} \mathbf{y} \} \quad (\text{A.42b})$$

$$l(\alpha_i) = \frac{1}{2} \left\{ -\log(\alpha_i + s_i) + \log \alpha_i + \frac{q_i^2}{\alpha_i + s_i} \right\}, \quad (\text{A.42c})$$

where $\boldsymbol{\alpha}_{-i}$ is a vector of size $N - 1$ that contains all hyperparameters except α_i . In addition, $q_i = \Phi_i^T \mathbf{B}_{-i}^{-1} \mathbf{y}$ and $s_i = \Phi_i^T \mathbf{B}_{-i}^{-1} \Phi_i$ are entities that are independent of α_i , and are defined to simplify the subsequent mathematical exposition. This decomposition of $\mathcal{L}(\boldsymbol{\alpha}, \rho)$ in Eq. (A.42) therefore decouples the effect of each α_i on log-evidence in the form of function $l(\alpha_i)$ while function $\mathcal{L}(\boldsymbol{\alpha}_{-i}, \rho)$ quantifies the effect of rest of the hyperparameters (i.e. $\boldsymbol{\alpha}_{-i}$).

To compute the MAP estimate $\boldsymbol{\alpha}^{\text{MAP}}$, the objective function $\mathcal{L}(\boldsymbol{\alpha}, \rho)$ in Eq. (A.42) is differ-

entiated with respect to α_i to obtain

$$\begin{aligned}
 \frac{\partial \mathcal{L}(\boldsymbol{\alpha}, \rho)}{\partial \alpha_i} &= \frac{\partial \mathcal{L}(\boldsymbol{\alpha}_{\cancel{i}}, \rho)}{\partial \alpha_i} + \frac{\partial l(\alpha_i)}{\partial \alpha_i} \\
 &= \frac{1}{2} \frac{\partial}{\partial \alpha_i} \left\{ -\log(\alpha_i + s_i) + \log \alpha_i + \frac{q_i^2}{\alpha_i + s_i} \right\} \\
 &= \frac{1}{2} \left\{ -\frac{1}{\alpha_i + s_i} + \frac{1}{\alpha_i} - \frac{q_i^2}{(\alpha_i + s_i)^2} \right\} \\
 &= \frac{1}{2} \left\{ \frac{\alpha_i^{-1} s_i^2 - (q_i^2 - s_i)}{(\alpha_i + s_i)^2} \right\}. \tag{A.43}
 \end{aligned}$$

The solution for α_i^{MAP} is then obtained by equating the derivative of log-evidence in Eq. (A.43) to zero, leading to an iterative re-estimation relation

$$\alpha_i = \begin{cases} \frac{s_i^2}{q_i^2 - s_i}, & q_i^2 > s_i \\ \infty & q_i^2 \leq s_i \end{cases} \tag{A.44}$$

In other words, if $q_i^2 \leq s_i$, the basis ϕ_i is deemed irrelevant by setting $\alpha_i^{\text{MAP}} = \infty$. Else, if $q_i^2 > s_i$, basis ϕ_i is deemed relevant by setting $\alpha_i^{\text{MAP}} = s_i^2/(q_i^2 - s_i)$ according to Eq. (A.44). This step is repeated for each basis ϕ_i in an iterative manner. The MAP estimate ρ^{MAP} of model error precision is computed using Eq. (A.35) by assuming flat prior for ρ (i.e. set $a = 1$ and $b = 0$). This combination of Eq. (A.44) and Eq. (A.35) powers the iterative addition/deletion scheme of the accelerated SBL algorithm.

The variation of log-evidence with varying α_i can be further explored by computing the curvature of $\mathcal{L}(\boldsymbol{\alpha}, \rho)$ (or $l(\alpha_i)$) using Eq. (A.43) as

$$\begin{aligned}
 \frac{\partial^2 \mathcal{L}(\boldsymbol{\alpha}, \rho)}{\partial \alpha_i^2} &= \frac{\partial^2 l(\alpha_i)}{\partial \alpha_i^2} = \frac{-\alpha_i^{-2} s_i^2}{2(\alpha_i + s_i)^2} - \frac{(\alpha_i^{-1} s_i^2 - (q_i^2 - s_i))}{(\alpha_i + s_i)^3} \\
 &= -\frac{1}{2} \left\{ \frac{s_i}{\alpha_i(\alpha_i + s_i)} \right\}^2 - \frac{2}{(\alpha_i + s_i)} \frac{\partial \mathcal{L}(\boldsymbol{\alpha}, \rho)}{\partial \alpha_i}. \tag{A.45}
 \end{aligned}$$

At the optimal $\boldsymbol{\alpha}$ the slope $\partial l(\alpha_i)/\partial \alpha_i = 0$ and therefore the curvature of log-evidence in

Eq. (A.45) is always negative (concave down) at the optimum computed using Eq. (A.44). This proves the convex nature of optimization of log-evidence with respect to each α_i (but not α). Also, when $q_i^2 \leq s_i$, the slope of log-evidence in Eq. (A.43) is always positive. Note that $s_i = \mathbf{y}^T \mathbf{B}_{-i} \mathbf{y} \geq 0$ since $\mathbf{B}_{-i}$ is a covariance matrix (a positive semi-definite matrix). As a result, log-evidence has negative curvature for all values of α , except at $\alpha_i = \infty$ when it is zero. Hence, α_i^{MAP} keeps on increasing during the iterative re-estimation procedure of the original SBL algorithm. These large α_i values produce ill-conditioned matrices during the Bayesian updating of covariance/precision matrices. This numerical hurdle is the main reason behind the addition/deletion idea of the accelerated SBL algorithm where α_i^{MAP} is set to infinity, and the corresponding parameter is explicitly removed from the model.

The accelerated SBL algorithm operates by selecting a candidate basis function among the proposed basis functions, and then determining whether the basis needs to added or removed from the model based on $q_i^2 - s_i$ values in Eq. (A.44). This selection of a basis function can be performed in a random or deterministic fashion. Once an appropriate basis is selected, factors $q_i = \Phi_i^T \mathbf{B}_{-i}^{-1} \mathbf{y}$ and $s_i = \Phi_i^T \mathbf{B}_{-i}^{-1} \Phi_i$ are computed using the updated $\mathbf{B}_{-i}^{-1}$ from Eq. (A.38). For large number of basis (large N) or large data size (large M), calculating inverse of a $M \times M$ matrix $\mathbf{B}_{-i}$ for all N basis at every iteration degrades the computational efficiency of the algorithm. As a result, entities $\bar{s}_m = \Phi_m^T \mathbf{B}^{-1} \Phi_m$ and $\bar{q}_m = \Phi_m^T \mathbf{B}^{-1} \mathbf{y}$ are introduced that depend on matrix $\mathbf{B}$ which is fixed for all N basis. The relation between $\bar{s}_i$ and s_i is obtained using Eq. (A.39) as

$$\begin{aligned}
 \bar{s}_i &= \Phi_i^T \mathbf{B}^{-1} \Phi_i = \Phi_i^T \left(\mathbf{B}_{-i} + \frac{\Phi_i \Phi_i^T}{\alpha_i} \right)^{-1} \Phi_i \\
 &= \Phi_i^T \mathbf{B}_{-i}^{-1} \Phi_i - \frac{\Phi_i^T \mathbf{B}_{-i}^{-1} \Phi_i \Phi_i^T \mathbf{B}_{-i}^{-1} \Phi_i}{\alpha_i + \Phi_i^T \mathbf{B}_{-i}^{-1} \Phi_i} \\
 &= s_i - \frac{s_i^2}{\alpha_i + s_i} \\
 s_i &= \frac{\alpha_i \bar{s}_i}{\alpha_i - \bar{s}_i}
 \end{aligned} \tag{A.46}$$

Note that $s_i = \bar{s}_i$ when $\alpha_i = \infty$, while $s_i > \bar{s}_i$ for finite α_i . Similarly, the relation between $\bar{q}_i$ and q_i is obtained using Eq. (A.39) as

$$\begin{aligned}\bar{q}_i &= \Phi_i^T \mathbf{B}_{-i}^{-1} \mathbf{y} - \frac{\Phi_i^T \mathbf{B}_{-i}^{-1} \Phi_i \Phi_i^T \mathbf{B}_{-i}^{-1} \mathbf{y}}{\alpha_i + \Phi_i^T \mathbf{B}_{-i}^{-1} \mathbf{y}} \\ &= q_i - \frac{q_i^2}{\alpha_i + s_i} \\ q_i &= \frac{\alpha_i \bar{q}_i}{\alpha_i - \bar{s}_i}\end{aligned}\tag{A.47}$$

Furthermore, the calculation of inverse $\mathbf{B}^{-1}$ is avoided by rewriting the relationship for $\bar{q}_i$ and $\bar{s}_i$ using the Woodbury identity (similar to Eq. (A.28)) as

$$\bar{s}_m = \Phi_m^T \mathbf{B}^{-1} \Phi_m = \rho \Phi_m^T \Phi_m - \rho^2 \Phi_m^T (\Phi \mathbf{P} \Phi^T) \Phi_m \tag{A.48a}$$

$$\bar{q}_m = \Phi_m^T \mathbf{B}^{-1} \mathbf{y} = \rho \Phi_m^T \mathbf{y} - \rho^2 \Phi_m^T (\Phi \mathbf{P} \Phi^T) \mathbf{y} \tag{A.48b}$$

where Φ and $\mathbf{P}$ only contain those basis functions that are currently present in the model. This approach of altering the design matrix Φ further enhances the computational efficiency of the addition/deletion algorithm, and helps ensure the matrix Φ is free of multicollinearity as it only contains the basis functions that are absolutely necessary for fitting the training data. Algorithm 4 summarizes the steps needed to implement the accelerated SBL algorithm. Further details about the practical aspects of the algorithm are discussed during numerical investigations reported in Sec. A.6.

A.4 Bayesian compressive sensing

BCS originated in the signal processing literature as a way to enable compact representation of discrete signals [163]. Ji *et al.* [163] proposed BCS as a fully Bayesian alternative to LASSO to solve ill-posed inverse problems encountered in signal reconstruction, albeit analytical evaluation of the sparse solution (like SBL) was not explored. Later, Babacan *et al.* [79] provided a

Algorithm 4 Accelerated SBL algorithm by Faul and Tipping [136]

- 1: Training data: $\{(y^{(i)}, \mathbf{u}^{(i)}) ; i = 1, 2, \dots, M\}$;
 - 2: Propose N basis functions, and construct the design matrix Φ according to Eq. (A.1);
 - 3: Choose a starting value of α and ρ ;
 - 4: **while** no convergence **do**
 - 5: Compute posterior entities $\mathbf{m}$, $\mathbf{P}$ and $\mathbf{B}$ using Eq. (A.20);
 - 6: Compute the log-evidence using Eq. (A.37);
 - 7: Choose an index $1 \leq i \leq N$ randomly or in a predefined fashion;
 - 8: Compute $\bar{s}_i = \Phi_i^T \mathbf{B}^{-1} \Phi_i$ and $\bar{q}_i = \Phi_i^T \mathbf{B}^{-1} \mathbf{y}$ for the chosen basis;
 - 9: Compute s_i using Eq. (A.46), and q_i using Eq. (A.47);
 - 10: **if** $q_i^2 > s_i$ **then**
 - 11: Relevance: $\alpha_i = s_i^2 / (q_i^2 - s_i)$;
 - 12: **else if** $q_i^2 < s_i$ **then**
 - 13: Irrelevance: $\alpha_i = \infty$;
 - 14: **end if**
 - 15: Update ρ using Eq. (A.35)
 - 16: **end while**
-

semi-analytical solution to BCS, and formalized the algorithm for linear-in-parameter models. BCS involves assigning a parameterized Laplace prior as a way to induce sparsity among model parameters. The use of Laplace prior over Gaussian prior (as in SBL) is being promoted due to the sparsity inducing property of L1-regularization (LASSO) over L2-regularization (Ridge regression) in least-square estimation. The mathematical exposition presented in this section is derived from Babacan *et al.* [79].

BCS operates by assigning a parametrized Laplace prior to each model parameter w_i . However, a Laplace prior on $\mathbf{w}$ is not conjugate to a Gaussian likelihood function in the Bayesian linear regression setting, as detailed in Eq. (A.1). This loss of conjugacy hinders the analytical tractability of Bayesian analysis. In order to preserve prior-posterior conjugacy, a Laplace prior

is constructed through a two-level hierarchical prior assignment of

$$\mathbf{p}(\mathbf{w}|\boldsymbol{\gamma}) = \prod_{i=0}^{N-1} \mathcal{N}(w_i|0, \gamma_i) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{A}^{-1}) \quad (\text{A.49a})$$

$$\mathbf{p}(\boldsymbol{\gamma}|\lambda) = \prod_{i=0}^{N-1} \mathcal{G}(\gamma_i|1, \lambda/2) = \prod_{i=0}^{N-1} \frac{\lambda}{2} \exp \left\{ -\frac{\lambda \gamma_i}{2} \right\} \quad (\text{A.49b})$$

$$\mathbf{p}(\lambda) = \mathcal{G}(\lambda|c, d), \quad (\text{A.49c})$$

where $\boldsymbol{\gamma}$ is the unknown hyperparameter vector (prior variance of $\mathbf{w}$), λ is the unknown parameter that dictates the prior pdf of $\boldsymbol{\gamma}$, and $c, d > 0$ are some known constants. The prior $\mathcal{G}(\gamma_i|1, \lambda/2)$ is a special type of Gamma distribution, and is also called as the exponential distribution with rate parameter $\lambda/2$.

A Laplace prior on w_i is obtained by integrating out hyperparameter γ_i from the joint distribution of w_i and γ_i as [136]

$$\begin{aligned} \mathbf{p}(w_i|\lambda) &= \int \mathbf{p}(w_i, \gamma_i|\lambda) d\gamma_i = \int \mathbf{p}(w_i|\gamma_i) \mathbf{p}(\gamma_i|\lambda) d\gamma_i \\ &= \int \mathcal{N}(w_i|0, \gamma_i) \mathcal{G}(\gamma_i|1, \lambda/2) d\gamma_i \\ &= \mathcal{L}_p(w_i|0, \lambda^{-0.5}) \end{aligned} \quad (\text{A.50})$$

where $\mathcal{L}_p(x|r, s)$ denotes a laplace distribution with properties:

$$\mathcal{L}_p(x|r, s) = \frac{1}{2s} \exp \left\{ -\frac{|x - r|}{s} \right\}; \quad \mathbf{E}[x] = r; \quad \mathbf{Var}(x) = 2s^2 \quad (\text{A.51})$$

where s is the scale parameter. Therefore, the prior for w_i is a laplace pdf when conditioned on λ , while it is a Gaussian pdf when conditioned on the hyperparameter γ_i . The prior for model error precision is assigned similar to SBL as $\mathbf{p}(\rho) = \mathcal{G}(\rho|a, b)$.

Following the SBL formulation in Sec. A.3, the posterior parameter pdf conditioned on $\boldsymbol{\gamma}, \lambda$

and ρ can be evaluated analytically as

$$\begin{aligned}
 p(\mathbf{w}|\mathbf{y}, \gamma, \lambda, \rho) &= \frac{p(\mathbf{y}|\mathbf{w}, \gamma, \lambda, \rho)p(\mathbf{w}|\gamma, \lambda, \rho)}{p(\mathbf{y}|\gamma, \lambda, \rho)} \\
 &\propto p(\mathbf{y}|\mathbf{w}, \rho)p(\mathbf{w}|\gamma) \\
 &\propto \mathcal{N}(\mathbf{y}|\Phi\mathbf{w}, \mathbf{I}_M\rho^{-1})\mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{A}^{-1}) \\
 &= \mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{P})
 \end{aligned} \tag{A.52}$$

where the expressions for $\mathbf{m}$ and $\mathbf{P}$ are listed in Eq. (A.20) with $\mathbf{A}^{-1} = \text{Diag}(\gamma)$. Consequently, the model evidence $p(\mathbf{y}|\gamma, \lambda, \rho)$ (denominator in Eq. (A.52)) is also available analytically as $\mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{B})$ where matrix $\mathbf{B}$ is known from Eq. (A.20). The remaining pdf $p(\gamma, \lambda, \rho|\mathbf{y})$ in Eq. (A.52) is evaluated using Bayes' theorem as

$$p(\gamma, \lambda, \rho|\mathbf{y}) = \frac{p(\mathbf{y}|\gamma, \lambda, \rho)p(\gamma, \lambda, \rho)}{p(\mathbf{y})} \tag{A.53a}$$

$$\propto p(\mathbf{y}|\gamma, \lambda, \rho)p(\rho)p(\lambda)p(\gamma|\lambda) \tag{A.53b}$$

$$\propto \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{B})\mathcal{G}(\rho|a, b)\mathcal{G}(\lambda|a, b) \prod_{i=0}^{N-1} \mathcal{G}(\gamma_i|1, \lambda/2) . \tag{A.53c}$$

The sparse solution for $\mathbf{w}$ pertains to the MAP estimate γ^{MAP} , λ^{MAP} and ρ^{MAP} obtained by maximizing the posterior $p(\gamma, \lambda, \rho|\mathbf{y})$ in Eq. (A.53). The optimization is posed in terms of the objective function $\mathcal{L}(\gamma, \lambda, \rho)$ obtained by taking logarithm of $p(\gamma, \lambda, \rho|\mathbf{y})$ and ignoring the constant terms, leading to [79]

$$\begin{aligned}
 \mathcal{L}(\gamma, \lambda, \rho) &= -\frac{1}{2} \{ \log |\mathbf{B}| + \mathbf{y}^T \mathbf{B}^{-1} \mathbf{y} \} + N \log \frac{\lambda}{2} - \frac{\lambda}{2} \sum_{i=0}^{N-1} \gamma_i \\
 &\quad + (a-1) \log \rho - b\rho + (c-1) \log \lambda - d\lambda .
 \end{aligned} \tag{A.54}$$

Using the expansions of term $\log |\mathbf{B}|$ in Eq. (A.27) and $\mathbf{y}^T \mathbf{B}^{-1} \mathbf{y}$ in Eq. (A.29), the objective

function $\mathcal{L}(\cdot)$ in Eq. (A.54) is rewritten as [79]

$$\begin{aligned} \mathcal{L}(\boldsymbol{\alpha}, \lambda, \rho) = & -\frac{1}{2} \left\{ -M \log \rho + \log |\mathbf{P}^{-1}| + \rho \|\mathbf{y} - \Phi \mathbf{m}\|_2^2 + \sum_{i=0}^{N-1} \left(\frac{m_i^2}{\gamma_i} + \log \gamma_i \right) \right\} \\ & + N \log \frac{\lambda}{2} - \frac{\lambda}{2} \sum_{i=0}^{N-1} \gamma_i + (a-1) \log \rho - b\rho + (c-1) \log \lambda - d\lambda. \end{aligned} \quad (\text{A.55})$$

BCS incorporates a fast sub-optimal approach similar to the accelerated SBL algorithm from Sec. A.3.2 that involves an iterative addition/deletion of bases. Using the decomposition of matrix $\mathbf{B}$ in Eq. (A.39) and the subsequent expansion of $\log |\mathbf{B}|$ in Eq. (A.40) and $\mathbf{y}^T \mathbf{B}^{-1} \mathbf{y}$ in Eq. (A.41), the objective function is re-written as [79]

$$\mathcal{L}(\boldsymbol{\gamma}, \lambda, \rho) = \mathcal{L}(\boldsymbol{\gamma}_{-i}, \lambda, \rho) + l(\gamma_i) \quad (\text{A.56a})$$

$$\mathcal{L}(\boldsymbol{\gamma}_{-i}, \lambda, \rho) = -\frac{1}{2} \left\{ \log |\mathbf{B}_{-i}| + \mathbf{y}^T \mathbf{B}_{-i}^{-1} \mathbf{y} \right\} + N \log \frac{\lambda}{2} - \frac{\lambda}{2} \sum_{\substack{j=0 \\ j \neq i}}^{N-1} \gamma_j \quad (\text{A.56b})$$

$$+ (a-1) \log \rho - b\rho + (c-1) \log \lambda - d\lambda \quad (\text{A.56c})$$

$$l(\gamma_i) = \frac{1}{2} \left\{ \log \left(\frac{1}{1 + \gamma_i s_i} \right) + \frac{\gamma_i q_i^2}{1 + \gamma_i s_i} - \lambda \gamma_i \right\} \quad (\text{A.56d})$$

where $q_i = \Phi_i^T \mathbf{B}_{-i}^{-1} \mathbf{y}$ and $s_i = \Phi_i^T \mathbf{B}_{-i}^{-1} \Phi_i$. The calculation of $\bar{q}_i$ and $\bar{s}_i$ is performed by defining entities $\bar{s}_m = \Phi_m^T \mathbf{B}^{-1} \Phi_m$ and $\bar{q}_m = \Phi_m^T \mathbf{B}^{-1} \mathbf{y}$ and using relations $s_i = \bar{s}_i / (1 - \gamma_i \bar{s}_i)$ and $q_i = \bar{q}_i / (1 - \gamma_i \bar{s}_i)$ from Eq. (A.46) and Eq. (A.47), respectively. Subsequently, differentiating

$\mathcal{L}(\gamma, \lambda, \rho)$ with respect to γ_i leads to

$$\begin{aligned}
 \frac{\partial \mathcal{L}(\gamma, \lambda, \rho)}{\partial \gamma_i} &= \frac{\partial \mathcal{L}(\gamma_{-i}, \lambda, \rho)}{\partial \gamma_i} + \frac{\partial l(\gamma_i)}{\partial \gamma_i} \\
 &= \frac{1}{2} \left\{ \frac{-1}{(1 + \gamma_i s_i)^2} s_i + \frac{q_i^2}{1 + \gamma_i s_i} - \frac{q_i^2 \gamma_i}{(1 + \gamma_i s_i)^2} s_i - \lambda \right\} \\
 &= \frac{1}{2} \left\{ \frac{-s_i - \gamma_i s_i^2 + q_i^2 + q_i^2 \gamma_i s_i - q_i^2 \gamma_i s_i - \lambda - \lambda \gamma_i^2 s_i^2 - 2\lambda \gamma_i s_i}{(1 + \gamma_i s_i)^2} \right\} \\
 &= -\frac{1}{2} \left\{ \frac{\gamma_i^2 (\lambda s_i^2) + \gamma_i (s_i^2 + 2\lambda s_i) + (s_i - q_i^2 + \lambda)}{(1 + \gamma_i s_i)^2} \right\} \tag{A.57}
 \end{aligned}$$

Equating the derivative of the objective function in the above equation to zero leads to a quadratic equation

$$\gamma_i^2 (\lambda s_i^2) + \gamma_i (s_i^2 + 2\lambda s_i) + (s_i - q_i^2 + \lambda) = 0 \tag{A.58}$$

for which the solution is obtained as

$$\begin{aligned}
 \gamma_i^{\text{MAP}} &= \frac{-s_i(s_i + 2\lambda) \pm \sqrt{(s_i^2 + 2\lambda s_i)^2 - 4\lambda s_i^2(s_i - q_i^2 + \lambda)}}{2\lambda s_i^2} \\
 &= \frac{-(s_i + 2\lambda) \pm \sqrt{(s_i + 2\lambda)^2 - 4\lambda(s_i - q_i^2 + \lambda)}}{2\lambda s_i}. \tag{A.59}
 \end{aligned}$$

Given that λ, γ_i, s_i are all greater than zero, following two scenarios exist regarding the MAP estimate γ_i^{MAP} obtained in Eq. (A.59):

1. $s_i - q_i^2 + \lambda \geq 0$ or $q_i^2 - s_i \leq \lambda$: For this case, both the roots of γ_i in Eq. (A.59) is negative. Also, the slope of the objective function in Eq. (A.57) is negative for all $\gamma_i \geq 0$. Given that $\gamma_i \geq 0$, the only possible solution that can satisfy these conditions is $\gamma_i^{\text{MAP}} = 0$.
2. $s_i - q_i^2 + \lambda < 0$ or $q_i^2 - s_i > \lambda$: For this case, we have one positive and one negative solution for γ_i^{MAP} in Eq. (A.59). From Eq. (A.57) it can be deduced that the slope of the objective function $\mathcal{L}$ is positive at $\gamma_i = 0$ and is negative at $\gamma_i = \infty$. Given that $\gamma_i \geq 0$, the positive solution from Eq. (A.59) maximizes $\mathcal{L}$ and is taken to be the MAP estimate.

In summary, the MAP estimate of γ_i obtained by maximizing the posterior $p(\gamma, \lambda, \rho | \mathbf{y})$ in Eq. (A.52) or the objective function $\mathcal{L}(\gamma, \lambda, \rho)$ in Eq. (A.56) is obtained as

$$\gamma_i^{\text{MAP}} = \begin{cases} \frac{-(s_i + 2\lambda) \pm \sqrt{(s_i + 2\lambda)^2 - 4\lambda(s_i - q_i^2 + \lambda)}}{2\lambda s_i}, & q_i^2 - s_i \geq \lambda \\ 0 & q_i^2 - s_i < \lambda \end{cases} \quad (\text{A.60})$$

Further, differentiating $\mathcal{L}(\gamma, \lambda, \rho)$ in Eq. (A.54) or Eq. (A.55) with respect to λ and equating it to zero leads to (note that matrices $\mathbf{B}$ and $\mathbf{P}$ are independent of λ)

$$\begin{aligned} \frac{\partial \mathcal{L}(\gamma, \lambda, \rho)}{\partial \lambda} &= \frac{\partial}{\partial \lambda} \left\{ N \log \frac{\lambda}{2} - \frac{\lambda}{2} \sum_{i=0}^{N-1} \gamma_i + (c-1) \log \lambda - d\lambda \right\} \\ 0 &= \frac{N}{\lambda/2} \cdot \frac{1}{2} - \frac{1}{2} \sum_{i=0}^{N-1} \gamma_i + \frac{c-1}{\lambda} - d \\ \lambda^{\text{MAP}} &= \frac{N + (c-1)}{0.5 \sum \gamma_i + d} \end{aligned} \quad (\text{A.61})$$

Similarly, differentiating $\mathcal{L}(\gamma, \lambda, \rho)$ in Eq. (A.55) with respect to ρ and equating it to zero leads to Eq. (A.35). Hence, the MAP estimate for ρ is the same for BCS and SBL. In summary, Eq. (A.60), Eq. (A.61), and Eq. (A.35) completes the derivation of iterative re-estimation equations for computing γ_i^{MAP} , λ^{MAP} AND ρ^{MAP} . Algorithm 5 summarizes the steps involved in the BCS algorithm.

A.5 Relationship between LASSO, SBL, and BCS

We first attempt to establish the Bayesian interpretation of the frequentist techniques of Ridge regression and LASSO. Consider the MAP estimate calculation for $\mathbf{w}$ using the BLR setup outlined in Sec. A.2, but with following special circumstances:

- Consider a zero-mean Gaussian prior pdf $\mathcal{N}(w_i | 0, (\rho\lambda)^{-1})$ with precision $\rho\lambda$ for each model

Algorithm 5 BCS algorithm by Babacan *et al.* [136]

```

1: Training data:  $\{(y^{(i)}, \mathbf{u}^{(i)}) ; i = 1, 2, \dots, M\}$ ;
2: Propose  $N$  basis functions, and construct the design matrix  $\Phi$  according to Eq. (A.1);
3: Initialize  $\gamma$ ,  $\lambda$  and  $\rho$ ;
4: while no convergence do
5:   Compute posterior entities  $\mathbf{m}$ ,  $\mathbf{P}$  and  $\mathbf{B}$  using Eq. (A.20);
6:   Compute the log-evidence using Eq. (A.37);
7:   Choose an index  $1 \leq i \leq N$  randomly or in a predefined fashion;
8:   Compute  $\bar{s}_i = \Phi_i^T \mathbf{B}^{-1} \Phi_i$  and  $\bar{q}_i = \Phi_i^T \mathbf{B}^{-1} \mathbf{y}$  for the chosen basis;
9:   Compute  $s_i$  using Eq. (A.46), and  $q_i$  using Eq. (A.47);
10:  if  $q_i^2 - s_i \geq \lambda$  then
11:    Relevance: Update  $\gamma_i$  using Eq. (A.60);
12:  else if  $q_i^2 - s_i < \lambda$  then
13:    Irrelevance: Set  $\gamma_i = \infty$ ;
14:  end if
15:  Update  $\gamma$  using Eq. (A.59);
16:  Update  $\rho$  using Eq. (A.35);
17: end while

```

parameter w_i . The MAP estimate calculation using Eq. (A.7) leads to

$$\begin{aligned}
\mathbf{w}_{\text{map}} &= \arg \max_{\mathbf{w}} \{p(\mathbf{y}|\mathbf{w}, \rho)p(\mathbf{w})\} = \arg \max_{\mathbf{w}} \{\log p(\mathbf{y}|\mathbf{w}, \rho) + \log p(\mathbf{w})\} \\
&= \arg \max_{\mathbf{w}} \left\{ \log \mathcal{N}(\mathbf{y}|\Phi\mathbf{w}, \rho^{-1}\mathbf{I}_M) + \log \left(\prod_{i=1}^{N-1} \mathcal{N}(w_i|0, (\rho\lambda)^{-1}) \right) \right\} \\
&= \arg \min_{\mathbf{w}} \left\{ \frac{1}{2} \|\mathbf{y} - \Phi\mathbf{w}\|_2^2 + \frac{\lambda}{2} \sum_{i=1}^{N-1} w_i^2 \right\} \tag{A.62}
\end{aligned}$$

The cost function in the above equation is the same as Ridge regression cost function in Eq. (A.4) with $p = 2$. Therefore, the MAP estimate $\mathbf{w}_{\text{map}}$ in BLR while choosing prior $\mathcal{N}(w_i|0, (\rho\lambda)^{-1})$ is same as the Ridge regression estimate $\mathbf{w}_{\text{ridge}}$.

- Next, consider a zero-mean Laplace prior pdf with the scale parameter $s = 2(\rho\lambda)^{-1}$ (see Eq. (A.51) for the definition of Laplace pdf). The MAP estimate calculation using Eq. (A.7)

leads to

$$\begin{aligned}
\mathbf{w}_{\text{map}} &= \arg \max_{\mathbf{w}} \{p(\mathbf{y}|\mathbf{w}, \rho)p(\mathbf{w})\} = \arg \max_{\mathbf{w}} \{\log p(\mathbf{y}|\mathbf{w}, \rho) + \log p(\mathbf{w})\} \\
&= \arg \max_{\mathbf{w}} \left\{ \log \mathcal{N}(\mathbf{y}|\Phi\mathbf{w}, \rho^{-1}\mathbf{I}_M) + \log \left(\prod_{i=1}^{N-1} \mathcal{L}_p(w_i|0, 2(\rho\lambda)^{-1}) \right) \right\} \\
&= \arg \min_{\mathbf{w}} \left\{ \frac{1}{2} \|\mathbf{y} - \Phi\mathbf{w}\|_2^2 + \frac{\lambda}{2} \sum_{i=1}^{N-1} |w_i| \right\} \tag{A.63}
\end{aligned}$$

The cost function in the above equation is the same as LASSO cost function in Eq. (A.4) with $p = 1$. Therefore, the MAP estimate $\mathbf{w}_{\text{map}}$ in BLR with prior $\mathcal{L}_p(w_i|0, 2(\rho\lambda)^{-1})$ is same as the the LASSO regression estimate $\mathbf{w}_{\text{lasso}}$.

Next, we explore the relationship between the Bayesian sparse learning techniques of SBL and BCS. The most salient difference between SBL and BCS is the type of sparsity inducing prior pdf. The parametrized prior pdf assigned to the unknown parameter $\mathbf{w}$ in both SBL and BCS is a zero-mean Gaussian prior $\mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{A}^{-1})$. The sparsity inducing property of this prior is realized while obtaining the marginal prior pdf of $\mathbf{w}$ by integrating out the hyperparameters (α in SBL and γ in BCS), as discussed by Babacan *et al.* [79]. The marginal pdf of model parameters w_i for SBL and BCS is obtained as follows:

- In SBL, the marginal prior pdf for an unknown w_i is obtained as

$$\begin{aligned}
p(w_i) &= \int p(w_i|\alpha_i)p(\alpha_i)d\alpha_i \\
&= \int \mathcal{N}(w_i|0, 1/\alpha_i)\mathcal{G}(\alpha_i|c, d)d\alpha_i \\
&= \mathcal{T}(w_i|0, 2c, d/c) \tag{A.64}
\end{aligned}$$

where $\mathcal{T}(x|\mu, \nu, V)$ is a student's t-distribution with mean μ , dof ν and shape parameter V (section 7.7 in [137]).

- In BCS, the marginal prior pdf $p(w_i|\lambda)$ for BCS is obtained as

$$\begin{aligned}
 p(w_i|\lambda) &= \int p(w_i|\gamma_i)p(\gamma_i|\lambda)d\gamma_i \\
 &= \int \mathcal{N}(w_i|0, \gamma_i)\mathcal{G}(\gamma_i|1, \lambda/2)d\gamma_i \\
 &= \mathcal{L}_p(w_i|0, \lambda^{-0.5})
 \end{aligned} \tag{A.65}$$

where $\mathcal{L}_p(x|m, b)$ denotes a Laplace distribution whose properties are outlined in Eq. (A.51).

Hence, SBL uses a student's t-distribution as the marginal prior while BCS uses a Laplace distribution. This choice of Laplace prior results in a more sparser solution for BCS relative to SBL [79]. This fact is realized in posing the maximization problem by rewriting the objective function for SBL (accelerated) using Eq. (A.37) and for BCS using Eq. (A.56) in terms of prior precision $\gamma_i = 1/\alpha_i$ as

$$\text{SBL: } \gamma_i^{SBL} = \arg \max \{l^{SBL}(\gamma_i)\} = \arg \max \left\{ \frac{1}{2} \left(\log \left(\frac{1}{1 + \gamma_i s_i} \right) + \frac{q_i^2 \gamma_i}{1 + \gamma_i s_i} \right) \right\} \tag{A.66}$$

$$\text{BCS: } \gamma_i^{BCS} = \arg \max \{l^{BCS}(\gamma_i)\} = \arg \max \left\{ \frac{1}{2} \left(\log \left(\frac{1}{1 + \gamma_i s_i} \right) + \frac{q_i^2 \gamma_i}{1 + \gamma_i s_i} - \lambda \gamma_i \right) \right\}. \tag{A.67}$$

Notice the additional term of $-\lambda \gamma_i$ in the BCS objective function when compared to the SBL objective function. This term results in additional penalty to large prior variance γ_i values. Further understanding is gained by differentiating the objective functions for BCS and SBL, leading to

$$\left. \frac{\partial \mathcal{L}^{BCS}(\gamma_i)}{\partial \gamma_i} \right|_{\gamma_i = \gamma_i^{SBL}} < 0 \quad \& \quad \left. \frac{\partial \mathcal{L}^{BCS}(\gamma_i)}{\partial \gamma_i} \right|_{\gamma_i = 0} > 0 \Rightarrow \gamma_i^{SBL} \geq \gamma_i^{BCS}, \tag{A.68}$$

implying that BCS always produces lower γ_i^{MAP} value than the SBL solution at the optimum (i.e. for maximum model evidence). A lower γ_i^{MAP} value means a lower prior variance for BCS solution. This means BCS always produces equal or more number of parameters than SBL

whose prior and posterior variance are reduced to zero. In other words, BCS produces same or sparser solution than SBL. Notice that SBL provides the same solution as BCS when $\lambda = 0$. Therefore, it is fair to say that SBL is a special case of BCS when λ is zero. A λ value of zero also implies a hyperprior $\mathcal{G}(\alpha_i|1, 0)$, which is also known as Jeffrey's prior [161].

A.6 Numerical study: Polynomial regression

In this section, we detail the performance of frequentist (LASSO) and Bayesian (SBL, BCS) sparse learning algorithms by taking an example of polynomial regression. A realization of noisy measurements is generated using the polynomial $y_i = 1 + x_i^2 + \epsilon_i$, where $\epsilon \sim \mathcal{N}(0, 1/100)$. Eq. (A.1) shows the training data consisting of 25 noisy samples of y_i . Subsequently, an inverse problem is posed wherein a N^{th} -order polynomial of the form $y_j = \sum_{i=0}^N w_i x_j^i + \epsilon_j$ is proposed to model the training data. The unknown parameter vector is denoted as $\mathbf{w} = \{w_0, w_1, \dots, w_N\}$. By choosing a higher-order polynomial, most of the coefficients in $\mathbf{w}$ is redundant, leading to a sparse learning problem. We first consider the frequentist estimation approaches outlined in Sec. A.1 for solving this sparse learning problem.

Figure A.1: Training data generated using $y_i = 1 + x_i^2 + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, 1/100)$.

A.6.1 Frequentist approach

Consider a 5^{th} order polynomial for modelling the training data shown in Fig. A.1. Table A.3 lists the point estimates for w using OLS, LASSO and Ridge regression from Sec. A.1. The mean parameter w_0 is excluded from the regularization term in order to eliminate the dependence on the scale of measurements. In other words, the measurements are considered to be of zero mean when quantifying the penalty imposed by RLS techniques.

The results reported in Table A.3 are computed for the three choices of regularization coefficient λ : a low value of 0.001, the optimal value computed by minimizing the CV prediction error, and a high value of 0.5. Fig. A.2 shows the variation of CV prediction error with varying λ for the case of five-fold CV and leave-one-out CV. The optimal λ value pertaining to the minimum CV error is same for both five-fold CV and leave-one-out CV, which is $\lambda = 0.15$ for LASSO and $\lambda = 0.4$ for Ridge regression. Note that the five-fold CV error is computed using Eq. (A.5) and the leave-one-out CV error is computed using Eq. (A.6).

When computing five-fold CV, each of the five groups of data contain five randomly chosen measurement points (since $M = 25$). The CV error shown in Fig. A.2 is the average of 20 CV calculations wherein each calculation pertain to a different instance of measurements in each of the five folds. Following key inferences can be made by examining Table A.3 and Fig. A.2:

1. The OLS estimates for w shown in Table A.3 possess large magnitude for higher-order coefficients. Note that the OLS solution is same as LASSO or Ridge regression solution with no regularization (i.e. $\lambda = 0$). This behaviour of OLS estimates is the result of unrestricted complexity of the model leading to overfitting and poor generalization of the learned model to future data. This issue of overfitting hinders the use of OLS/MLE technique in dealing with over-parameterized models and noisy measurements.
2. The w estimates in Table A.3 for LASSO and Ridge regression have a lower magnitude than their OLS counterparts. This behaviour is due to the regularization imposed by the penalty term in RLS cost function from Eq. (A.4). Also, in Fig. A.2, the CV error for

RLS (LASSO or Ridge) is always less than the corresponding error for OLS estimation. This implies that the regularization is helping to alleviate overfitting without degrading the data-fit of the learned model.

3. Increasing λ increases the amount of penalty imposed by the regularization. This behaviour can be seen in Table A.1 as the magnitude of RLS estimates decreases with increasing λ value. For a low λ value of 0.001, the RLS estimates are dictated by data-fit while for a large value of 0.5 the RLS estimates are dictated by the penalty term. There exists an optimal λ that results in an optimal trade-off between data-fit and model complexity. This optimal λ pertains to minimum CV error. Note that the sudden increase in CV error for large λ value is due to the removal of relevant model coefficients due to the higher penalty.
4. Unlike OLS/MLE, Ridge regression and LASSO estimation is well-posed even for the case of multicollinearity. For example, the condition number of matrix $\Phi^T \Phi$ for the case of a 10th-order polynomial model is of the order 10^{14} while that for matrix $\Phi^T \Phi + 0.001 \mathbf{I}_M$ encountered in Ridge regression is of order 10^4 . This ill-conditioning of matrix $\Phi^T \Phi$ is due to the fact that the input x varies in range $[0, 1]$, and the higher-order basis, say x^9 and x^{10} , become almost colinear. OLS/MLE solution in this case is not possible since the matrix $\Phi^T \Phi$ is no longer a full-rank matrix. On the contrary, Ridge regression and LASSO estimation is well-posed due to the regularization.
5. The penalty imposed by the regularization have different affects on $\mathbf{w}$ values for LASSO and Ridge regression. In case of LASSO, the penalty forces the redundant parameters to exactly zero, where the number of parameters driven to zero depends on the value of λ (Table A.3). This property of LASSO is used in variable selection to produce sparse representation for $\mathbf{w}$. On the contrary, Ridge regression forces the parameter values towards zero but never exactly zero (Table A.3). Even though the CV prediction error pertaining to optimal λ is comparable for LASSO and Ridge regression, LASSO is preferred in prac-

tice due to its sparsity inducing property. It can be observed from Table A.3 that LASSO solution pertaining to optimal $\lambda = 0.15$ produces the same model as the data-generating model where the estimates for $\mathbf{w}$ are close to the true values.

w_i	True	OLS/MLE	LASSO			Ridge Regression		
		$\lambda = 0$	0.001	0.15*	5.0	0.001	0.4*	5.0
w_0	1.0	1.041	0.948	1.017	1.353	0.947	1.003	1.123
w_1	0.0	-3.350	0.588	0	0	0.616	0.297	0.163
w_2	1.0	29.78	-0.080	0.950	0	0.175	0.249	0.167
w_3	0.0	-81.73	0	0	0	0.067	0.207	0.156
w_4	0.0	93.26	0.735	0	0	0.805	0.179	0.144
w_5	0.0	-36.98	0	0	0	-0.153	0.155	0.133

Table A.3: Frequentist estimation results for $\mathbf{w}$. (*optimal λ value computed by minimizing five-fold CV error)

Figure A.2: Cross-validation prediction error for varying regularization coefficient λ .

A.6.2 Bayesian linear regression

In this study, we investigate the applicability of BLR (detailed in Sec. A.2) to fitting a fifth-order polynomial to the noisy measurements shown in Fig. A.1. The joint prior pdf of $\mathbf{w}$ is chosen as a zero-mean Gaussian pdf $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, 10^3\mathbf{I}_N)$, and the model error precision is assumed to be known as $\rho = 100$ (true value). Fig. A.3 shows the marginal posterior pdf

of $\mathbf{w} = \{w_0, w_1, w_2, w_3, w_4, w_5\}$ obtained using BLR. Fig. A.4 shows the probabilistic predictions computed using this posterior pdf, along with the point predictions made using the MLE estimate and the true parameter values. Next, we discuss some interesting measurements made from the BLR parameter estimation results reported in Fig. A.3 and Fig. A.4:

1. In Table A.4, the posterior precision of all model parameters is solely dictated by the MLE precision computed using Table A.1. This BLR behaviour is indicative of the non-informativeness or flatness of the prior pdf $\mathcal{N}(\mathbf{w}|\mathbf{0}, 10^3\mathbf{I}_N)$ due to its large variance of 10^3 . In Fig. A.4, this type of Bayesian estimation results in an exceedingly good data-fit for training data ($0 \leq x \leq 1$) and a large bias and a large variance during extrapolation (i.e. for $x \geq 1$ and $x \leq 0$). This poor extrapolation performance is attributed to the large variance in the posterior parameter pdf, as shown in Fig. A.3. This overfitting of over-parameterized models demonstrates the drawback of using flat priors in Bayesian inference.
2. The large, negative correlations among adjacent parameters shown in Table A.4 allude to redundancy in the model parameter (or basis) space. For example, parameters w_4 and w_5 have a correlation of -0.98 , which basically means that w_4 captures the same behaviour as w_5 . Hence, removing w_5 has no affect on the data-fit capability of the model as w_4 will compensate for the removed basis. The parameters w_1, w_2, w_3, w_4 , and w_5 are all correlated to each other with large, negative correlation coefficients. This behaviour is expected since $0 \leq x \leq 1$ and basis x, x^2, x^3, x^4, x^5 exhibit some collinearity among each other. This collinearity implies that eliminating some of these basis from the model has little effect on the data-fit capability of the model. Sparse learning algorithms are aimed to achieve exactly this.

Figure A.3: Marginal posterior pdf of $\mathbf{w}$ for a fifth-order polynomial with prior $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, 10^3\mathbf{I}_N)$.

Figure A.4: Model predictions using the posterior pdf from BLR, using MLE estimates, and using the true model.

w_i	Precision of w_i			Correlation coefficient					
	prior	MLE	posterior	w_0	w_1	w_2	w_3	w_4	w_5
w_0	1e-03	2.5e+03	2.5e+03	1.00					
w_1	1e-03	8.5e+02	8.5e+02	-0.76	1.00				
w_2	1e-03	5.3e+02	5.3e+02	0.52	-0.92	1.00			
w_3	1e-03	3.9e+02	3.9e+02	-0.34	0.76	-0.95	1.00		
w_4	1e-03	3.2e+02	3.2e+02	0.22	-0.60	0.85	-0.97	1.00	
w_5	1e-03	2.7e+02	2.7e+02	-0.13	0.47	-0.74	0.91	-0.98	1.00

Table A.4: Bayesian linear regression results for precision and correlation coefficient for a 5th order polynomial model.

A.6.3 Bayesian model comparison

We realize from the BLR results that the proposed fifth-order polynomial model can be reduced to a simpler model in order to achieve good extrapolation capability. As a result, we consider several polynomial models with increasing order and investigate the applicability of Bayesian model selection methodology in ranking these models. Table A.5 reports log-evidence and its decomposition (according to Eq. (A.11)) in terms of goodness-of-fit (GOF) and expected information-gain (EIG) for varying order of the polynomial model. Note that GOF quantifies data-fit while EIG quantifies model complexity, and log-evidence equals to GOF minus EIG. Table A.5 report results for three choices of prior pdf: 1) $\mathcal{N}(\mathbf{w}|\mathbf{0}, 10^0\mathbf{I}_N)$, 2) $\mathcal{N}(\mathbf{w}|\mathbf{0}, 10^3\mathbf{I}_N)$, and 3) $\mathcal{N}(\mathbf{w}|\mathbf{0}, 10^6\mathbf{I}_N)$. Remember that the BLR results reported previously pertained to the prior $\mathcal{N}(\mathbf{w}|\mathbf{0}, 10^3\mathbf{I}_N)$. Here are some observation made from the Bayesian model comparison results reported in Table A.5:

1. We first focus on the base case of prior $\mathcal{N}(\mathbf{w}|\mathbf{0}, 10^3\mathbf{I}_N)$. Increasing complexity (order) of the model increases GOF only until the second-order polynomial, and the increase is marginal thereafter. On the other hand, EIG increases with increasing order of the polynomial for all models. This stagnation in GOF along with the an increasing EIG results in complex models being penalized heavily. This penalization of complex models leads to the selection of second-order model as the optimal model with the highest log-evidence and having a posterior model probability of 89%. This selection of optimal model makes sense since a second-order polynomial is the closest to the data-generating model $y_i = 1 + x_i^2$.
2. Notice that the true model has the largest log-evidence among all models for all prior choices. Hence, if the true model is included in the candidate model set, it would have been selected with 100% posterior model probability for all prior choices. However, this is rarely the case in engineering practice since the true model is never known. As Box [38] famously said: "All models are wrong but some are useful". Hence, the goal is to find the

best useful model, not the true model. The best useful model for the prior $\mathcal{N}(\mathbf{w}|\mathbf{0}, 10^3\mathbf{I}_N)$ is the second-order model.

3. Next, we examine the trends in posterior model probability with increasing prior variance as shown in Table A.5. We realize that Bayesian model selection increasingly favors simpler models with increasing prior variance. Notice the increase in posterior model probability of simpler models with increasing prior variance. This behaviour can be explained by noticing the corresponding trend in EIG for a given model. For example, for the second-order model the increase in prior variance has no influence on GOF (=20.0 for all prior choices) while EIG increases proportional to the increase in prior variance. This increase is higher for models with more parameters, which results in complex models being penalized more than simpler models. This influence of prior variance on log-evidence is well documented in literature, and has been a source of concern in practical Bayesian community as it leads to conclusions that are sensitive to the choice of prior width [21, 164].
4. This issue of prior sensitivity is explored further. Fig. A.5 shows the variation in GOF, EIG and log-evidence with increasing prior variance of w_0 , while the remaining parameter have a fixed prior $\mathcal{N}(w_i|0, 10^3)$. The reduction of prior variance beyond a certain point results in a drastic decrease in GOF and hence the log-evidence. This reduction in GOF is caused by the fact that the posterior pdf of w_0 is being pushed towards zero by the zero-mean prior with a very low variance. Since the data-generating model has a non-zero value of w_0 , the removal of w_0 results in a drastic loss of data-fit. In fact, there exist an optimal prior variance for which the log-evidence is maximum, providing an optimal trade-off between data-fit (GOF) and model complexity (EIG).
5. Next, we vary the prior variance for w_1 while the prior for all other parameters is fixed at $\mathcal{N}(w_i|0, 10^3)$. Fig. A.5 shows that the decrease in prior variance of w_1 has no effect on GOF. This behaviour of GOF is expected since the data-generating model does

not contain the w_1x term. Further, the removal of w_1 through reduction of prior variance decreases model complexity or EIG which then increases the log-evidence. In fact, Fig. A.5 suggests that the prior variance of w_1 should be set to zero in order to achieve the maximum log-evidence. This implies to removal of parameter w_1 from the model. This behaviour of log-evidence with varying prior variance laid the foundation for sparse learning algorithms of SBL and BCS. The notion that the model complexity is controllable through prior variance is exploited in these algorithms to obtain the optimal model nested under an over-parameterized model.

Prior:	$\mathcal{N}(\mathbf{w} \mathbf{0}, 10^0\mathbf{I}_N)$			$\mathcal{N}(\mathbf{w} \mathbf{0}, 10^3\mathbf{I}_N)$			$\mathcal{N}(\mathbf{w} \mathbf{0}, 10^6\mathbf{I}_N)$		
order	GOF	EIG	Log-evid	GOF	EIG	Log-evid	GOF	EIG	Log-evid
0	-112.5	4.3	-116.8	-112.5	6.9	-119.4	-112.5	10.3	-122.8
1	11.8	6.5	5.3	11.8	12.5	(2) -0.7	11.8	19.4	(40) -7.6
2	20.0	7.6	(26) 12.4	20.0	16.9	(89) 3.1	20.0	27.2	(59) -7.2
3	20.6	8.0	(32) 12.6	20.4	19.9	(7) 0.5	20.4	33.7	-13.3
4	20.7	8.4	(24) 12.3	20.0	21.6	-1.6	20.0	38.8	-18.8
5	20.6	8.7	(16) 11.9	21.4	23.5	-2.1	22.8	42.6	-19.8
True	20.5	6.7	13.8	20.5	12.6	7.9	20.5	19.5	1.0

Table A.5: Bayesian model selection results for varying polynomial order. The data in parenthesis is the posterior model probability in percentage (rounded), without considering the true model.

Figure A.5: Variation in GOF, EIG and log-evidence with varying prior variance of parameter w_0 and w_1 for a fifth-order polynomial model. All other model parameters are assigned a fixed prior $\mathcal{N}(w_i|0, 10^3)$.

A.6.4 Sparse Bayesian learning

In this section we detail the performance of SBL in obtaining a sparse representation of $\mathbf{w}$ for a fifth-order polynomial model. SBL is initiated from the prior pdf $\mathcal{N}(\mathbf{w}|\mathbf{0}, 10^3\mathbf{I}_N)$, which amounts to choosing a starting value of $\mathbf{A} = 10^{-3}\mathbf{I}_N$ or $\alpha_i = 10^{-3}$ (for all w_i) in the ARD prior definition from Eq. (A.18). This choice of model and prior pdf has been previously studied in this section using BLR and RLS estimation techniques. The starting value of ρ is chosen as 100.0 (the true value). In addition, flat hyperprior is assumed for all α_i and ρ . We first focus on the original SBL algorithm as detailed in Sec. A.3.1 and summarized in Algorithm 3. The key findings are as follows:

1. Fig. A.6a shows the variation in log-evidence, GOF and EIG during SBL iterations. The increase in log-evidence is primarily due to the decrease in EIG (quantifies model complexity) while GOF (quantifies data-fit) remains largely unchanged. This decrease in EIG is driven by the increase in α_i values of irrelevant parameters during SBL, as shown in Fig. A.6d. This increase in prior precision forces the posterior pdf to be a Dirac-delta function, thereby eliminating the corresponding model parameter from the model and leading to a reduced model complexity or EIG.
2. The relevance indicator r_i shown in Fig. A.6c quantifies the importance of each w_i at each SBL iteration. As SBL operates, the r_i value for parameters w_0 and w_2 approach the value of one, while the r_i value for other parameters approach the value of zero. Based on Eq. (A.36), w_0 and w_2 are deemed as relevant, while rest of the parameters are pruned off from the model. Hence, the data-generating model of $y_i = 1 + x_i^2$ is recovered by SBL using mere 25 noisy measurements.
3. Once the SBL algorithm has converged, the optimal hyper-parameter value is obtained as $\boldsymbol{\alpha}^{\text{MAP}} = \{1.00, 2406, 0.915, 1102, 1676, 2016\}$ and $\rho^{\text{MAP}} = 87.8$. The prior and posterior pdf obtained using these optimal values is shown in Fig. A.7. As expected, both the prior and posterior pdfs for redundant parameters w_1, w_3, w_4 , and w_5 are Dirac-delta functions

centered at zero. On the other hand, the posterior pdfs of relevant parameters w_0 and w_2 are centered around the true values and have low variances. The MAP estimates for relevant parameters are computed as $m_0 = 1.00$ and $m_2 = 1.04$.

4. Fig. A.8 contrasts the model predictions computed using the posterior parameter pdf pertaining to the initial prior $\mathcal{N}(\mathbf{w}|\mathbf{0}, 10^3\mathbf{I}_N)$ (pre-SBL) and the optimal ARD prior pdf with precision α^{MAP} computed using SBL (post-SBL). Notice that the predictions from the sparse model (post-SBL) have lower bias and lower variance while predicting future data (extrapolation). This performance of the sparse model is at the expense of having a marginally higher bias than the over-parameterized model for predicting training data. This higher bias is however negligible in the comparison to the reduction in bias and variance during extrapolation. This fact is analogous to comparing the marginal decrease in GOF with the significant decrease in EIG during SBL, as shown in Fig. A.6a. The sparse model thus balances model parsimony and data-fit as facilitated by the Bayesian model selection framework through model evidence maximization.

Next, we apply the accelerated SBL algorithm from Sec. A.3.2 using the same settings as the original SBL algorithm. Fig. A.9 shows the variation of SBL entities during accelerated SBL iteration. Notice that the sparse solution obtained from accelerated algorithm is same as that obtained previously using the original SBL algorithm. The only difference arises from the faster convergence of the accelerated algorithm. This increase in efficiency can be attributed to the fact that when a parameter w_i is deemed irrelevant the accelerated algorithm proceeds by removing the corresponding basis from the model by explicitly assigning $\alpha_i = \infty$ and $r_i = 0$. This explicit removal of basis expedites the convergence of SBL as α for redundant parameter is no longer being estimated (like original SBL) and is fixed at infinity. Since the resulting sparse solution is identical, the resulting posterior pdf of relevant parameters w_0 and w_2 and the model predictions are also identical for both the SBL algorithms, and have already been reported in Fig. A.7 and Fig. A.8, respectively.

Figure A.6: Original SBL algorithm results for a fifth-order polynomial model.

Figure A.7: Post-SBL marginal posterior pdf of w for a fifth-order polynomial.

Figure A.8: Model predictions using joint posterior pdf of $\mathbf{w}$ before (pre-SBL) and after (post-SBL) the application of SBL.

Figure A.9: Accelerated SBL algorithm results.

A.6.5 Bayesian compressive sensing

Next, we report the performance of BCS algorithm (as detailed in Sec. A.4) in identifying the sparsity in $\mathbf{w}$ for the fifth-order polynomial model with the initial values of $\gamma_i = 1/\alpha_i = 10^3$ (prior pdf $\mathcal{N}(\mathbf{w}|\mathbf{0}, 10^3\mathbf{I}_N)$), $\lambda = 1.0$ and $\rho = 100.0$. The hyper-priors are chosen as $\mathcal{G}(\lambda|10, 10)$ and $\mathcal{G}(\rho|1, 10^5)$. Fig. A.10 shows the variation in BCS entities during BCS. The sparse solution obtained from BCS is identical to that obtained using SBL.

This numerical experiment demonstrates the usefulness of sparse learning algorithms in identifying a sparse parameter structure of an over-parameterized model. The sparse model obtained using evidence-based Bayesian apparatus is the result of optimal trade-off between model complexity and data-fit. This trade-off ensures the sparse model is free from overfitting and can be adequately used to predict unseen data.

Figure A.10: BCS results.

Appendix B

Bayesian algorithms

B.1 Transitional MCMC

In this section we provide mathematical details behind the TMCMC algorithm employed in Chapter 4 for sampling from non-Gaussian posterior pdfs. This section closely follows the work by Ching and Chen [48]. The mathematical notations used in this section follow Sec. 2.1. The explicit conditioning of Bayesian entities on model $\mathcal{M}$ is dropped here for brevity in mathematical exposition.

TMCMC is primarily aimed at addressing the limitations of RWM algorithms while sampling from the multimodal or non-Gaussian distributions encountered during Bayesian estimation of complex engineering systems [48]. TMCMC algorithm uses the principle of *simulated annealing* to proceed from prior pdf to posterior pdf through a series of intermediate pdfs. Given the observations $\mathcal{D}$, the intermediate pdf $p_j(\phi|\mathcal{D})$ at the j^{th} stage is written as [48]

$$p_j(\phi|\mathcal{D}) \propto p(\mathcal{D}|\phi)^{q_j} p(\phi) \quad (\text{B.1})$$

where $p(\phi)$ is the prior parameter pdf, $p(\mathcal{D}|\phi)$ is the likelihood function and $q_j \in [0, 1]$ is the annealing parameter. Note that each stage q_j is chosen such that $q_0 = 0 < q_1 < q_2 < \dots, q_m = 1$ where m is the total number of stages. Note that since the variation between pdfs

$p_j(\phi|\mathcal{D})$ and $p_{j+1}(\phi|\mathcal{D})$ is small, it is convenient to perform MCMC sampling from $p_{j+1}(\phi|\mathcal{D})$ using samples from $p_j(\phi|\mathcal{D})$ to build the proposal pdf for the MH algorithm. TMCMC uses this principle to execute the multistage sampling beginning from prior pdf ($q_0 = 0$) and reaching the posterior pdf ($q_m = 1$) in certain number of stages.

To begin with, samples from the known prior pdf $p_0(\phi|\mathcal{D}) = p(\phi)$ are generated using Monte Carlo sampling. Next, we assume that M_{j-1} number of stationary samples are available from $p_{j-1}(\phi|\mathcal{D})$ before initiating the j^{th} stage sampling. These samples are denoted as $\{\phi_{j-1,k} : k = 1, \dots, M_{j-1}\}$. The goal is to obtain stationary samples $\{\phi_{j,k} : k = 1, \dots, M_j\}$ from the pdf $p_j(\phi|\mathcal{D})$. The plausibility weights of $\phi_{j-1,k}$ samples with respect to the pdf $p_j(\phi|\mathcal{D})$ is computed as

$$w_{j,k} = \frac{p(\phi_{j-1,k})p(\mathcal{D}|\phi_{j-1,k})^{q_j}}{p(\phi_{j-1,k})p(\mathcal{D}|\phi_{j-1,k})^{q_{j-1}}} = p(\mathcal{D}|\phi_{j-1,k})^{q_j - q_{j-1}}. \quad (\text{B.2})$$

These plausibility weights quantify the change in intermediate pdfs with each stage. Given the weights $w_{j,k}$ in Eq. (B.2), a Gaussian RWM proposal pdf is constructed with covariance [48]

$$\Sigma_j = \beta^2 \sum_{k=1}^{M_j} \frac{w_{j,k}}{\sum_{l=1}^{M_j} w_{j,l}} \{\phi_{j-1,k} - \mu_j\} \{\phi_{j-1,k} - \mu_j\}^T, \quad (\text{B.3})$$

where

$$\mu_j = \frac{\sum_{l=1}^{M_j} w_{j,l} \phi_{j-1,k}}{\sum_{l=1}^{M_j} w_{j,l}} \quad (\text{B.4})$$

and β is the proposal scaling factor that is chosen such that the rejection rate among proposed RWM moves is optimal. The β value of $2.4/\sqrt{N_\phi}$ is often recommended for Gaussian RWM proposals [165]. Here, N_ϕ is the model dimension (size of ϕ).

The samples $\{\phi_{j,k}; j = 1, \dots, M_j\}$ from the intermediate pdf $p_j(\phi|\mathcal{D})$ are generated as following. First, M_{j-1} Markov chains are initiated with $\phi_{j-1,k}$ as their starting location. For each sampling step, k^{th} Markov chain is selected with the probability $w_{j,k} / \sum_{l=1}^{M_j} w_{j,l}$. The Markov chains located at $\phi_{j-1,k}$ samples with higher weights will be selected more often than those with lower weights. The chosen Markov chain is then moved forward using a RWM step while the

remaining Markov chains remains fixed at their current locations. Given the current state $\phi_{j,l}^c$ of l^{th} chain, a new sample ϕ' is proposed from a Gaussian proposal $\mathcal{N}(\phi_{j,l}^c, \Sigma_j)$. The proposed move ϕ' is accepted with the probability

$$\alpha(\phi', \phi_{j,l}^c) = \min \left(1, \frac{p_j(\phi'|\mathcal{D})}{p_j(\phi_{j,l}^c|\mathcal{D})} \right). \quad (\text{B.5})$$

If the proposed move is accepted, set $\phi_{j,k} = \phi'$ and $\phi_{j,l}^c = \phi'$. This step is repeated N_j times. Note that it is likely that some Markov chains will never get chosen since they are located in low posterior probability regions (lower weights). On the other hand, the Markov chains with higher weights will be chosen frequently as a starting point for the RWM step. Also, when a Markov chain is selected again, the RWM step is initiated from the current state of the chain, and not the initial value. At the end, $\phi_{j,k}$ samples are obtained which are distributed according to $p_j(\phi|\mathcal{D})$.

TMCMC also produces an estimate for model evidence $p(\mathcal{D})$ needed for executing Bayesian model comparison. Given that $q_0 = 0 < q_1 < q_2 < \dots, q_m = 1$, the model evidence integral in Eq. (2.6) can be written as [48]

$$\begin{aligned} p(\mathcal{D}) &= \int p(\phi) p(\mathcal{D}|\phi)^{q_m} d\phi \\ &= \frac{\int p(\phi) p(\mathcal{D}|\phi)^{q_m} d\phi}{\int p(\phi) p(\mathcal{D}|\phi)^{q_{m-1}} d\phi} \frac{\int p(\phi) p(\mathcal{D}|\phi)^{q_{m-1}} d\phi}{\int p(\phi) p(\mathcal{D}|\phi)^{q_{m-2}} d\phi} \cdots \\ &\quad \cdots \frac{\int p(\phi) p(\mathcal{D}|\phi)^{q_j} d\phi}{\int p(\phi) p(\mathcal{D}|\phi)^{q_{j-1}} d\phi} \cdots \frac{\int p(\phi) p(\mathcal{D}|\phi)^{q_1} d\phi}{\int p(\phi) p(\mathcal{D}|\phi)^{q_0} d\phi}. \end{aligned} \quad (\text{B.6})$$

Note that the denominator of the last term is unity, i.e. $\int p(\phi) p(\mathcal{D}|\phi)^{q_0} d\phi = \int p(\phi) d\phi = 1$.

Hence, we can rewrite Eq. (B.6) as

$$\begin{aligned}
 p(\mathcal{D}) &= \prod_{j=1}^m \frac{\int p(\phi) p(\mathcal{D}|\phi)^{q_j} d\phi}{\int p(\phi) p(\mathcal{D}|\phi)^{q_{j-1}} d\phi} = \prod_{j=1}^m \int \frac{p(\phi) p(\mathcal{D}|\phi)^{q_j} d\phi}{\int p(\phi) p(\mathcal{D}|\phi)^{q_{j-1}} d\phi} \\
 &= \prod_{j=1}^m \int \frac{p(\mathcal{D}|\phi)^{q_j}}{p(\mathcal{D}|\phi)^{q_{j-1}}} \left\{ \frac{p(\phi) p(\mathcal{D}|\phi)^{q_{j-1}}}{\int p(\phi) p(\mathcal{D}|\phi)^{q_{j-1}} d\phi} \right\} d\phi \\
 &= \prod_{j=1}^m \int p(\mathcal{D}|\phi)^{q_j - q_{j-1}} p_{j-1}(\phi|\mathcal{D}) d\phi.
 \end{aligned} \tag{B.7}$$

The integral in the above equation computes the expectation of weights for j^{th} stage of TMCMC sampling. An estimate of evidence is obtained using Eq. (B.2) as [48]

$$\hat{p}(\mathcal{D}) = \prod_{j=1}^m \frac{1}{M_j} \sum_{k=1}^{M_j} w_{j,k} \tag{B.8}$$

B.2 Delayed rejection adaptive Metropolis MCMC

In this section we detail the DRAM algorithm that was employed in Chapter 2 and Chapter 3 for generating stationary samples from the unnormalized posterior parameter pdf $p(\phi|\mathcal{D}) \propto p(\mathcal{D}|\phi)p(\phi)$. The mathematical notations used here were defined in Sec. 2.1.

The DRAM algorithm implemented in this work operates as following. An RWM algorithm with a Gaussian proposal pdf is initiated in parallel on N_c number of cores from dispersed ϕ values. These multiple RWM chains ensure that the multiple optima of posterior pdf are discovered during MCMC sampling. For each parallel chain, given the current state ϕ^c , a future Markov move ϕ' is proposed using the proposal pdf $q_1(\phi^c, \phi') = \mathcal{N}(\phi|\phi^c, \mathbf{C})$, where $\mathbf{C}$ is the proposal covariance matrix. The probability that the proposal move ϕ' is accepted is [60]

$$\alpha_1(\phi^c, \phi') = \min \left(1, \frac{p(\phi'|\mathcal{D})}{p(\phi^c|\mathcal{D})} \right). \tag{B.9}$$

In this work, we initiate $\mathbf{C} = \text{Diag}(s_d \text{abs}(\phi^{\text{MAP}}))$ where $s_d = 2.38^2/N_\phi$ is a constant well

known to provide an optimal RWM proposal covariance for MCMC sampling of Gaussian distributions using a Gaussian proposal [62, 166]. Here, N_ϕ is the size of ϕ vector. This choice of a diagonal $\mathbf{C}$ matrix is poor since it ignores the correlation among model parameters. The MCMC samples generated by the RWM algorithm will exhibit large autocorrelation and poor convergence properties. In this work, we employ AM [64], DR [63, 94] and INCA [93] to tune this poor choice of RWM proposal covariance.

Consider that the proposal covariance $\mathbf{C}$ is tuned using AM after every M_w samples, using the entire sample history upto that point. This tuning is implemented using 1) the sample sum $\mathbf{v}^k$ and 2) the sample sum-of-squares $\mathbf{S}^k$, both pertaining to the k^{th} Markov chain. In this work, we employ the *running_stat_vec* class of the Armadillo C++ library [127] to implement an recursive computation of $\mathbf{v}^k$ and $\mathbf{S}^k$. This approach is computationally advantageous since it does not require the storing of entire sample history, especially when dealing with high-dimensional ϕ .

Instead of updating the proposal covariance locally on each chain, we employ the INCA principle, where samples from all chains are used to tune a global proposal covariance [93]. This idea is implemented by broadcasting the entities $\mathbf{v}^k$ and $\mathbf{S}^k$ from each core to the master core using the global communication commands provided by the message passing library (MPI) [128]. Note that the broadcasting and storage of $\mathbf{v}^k$ and $\mathbf{S}^k$ is much more efficient than that of the entire sample history.

The updated covariance $\mathbf{C}$ is computed using the pairwise algorithm provided by Chan *et al.* [167]. Consider a pair $\{\mathbf{v}^a, \mathbf{S}^a\}$ and $\{\mathbf{v}^b, \mathbf{S}^b\}$ from the chain a and chain b , respectively. Further, assume that the pair $\{\mathbf{v}^a, \mathbf{S}^a\}$ is computed using M_s^a samples while the pair $\{\mathbf{v}^b, \mathbf{S}^b\}$ is computed using M_s^b samples. The resulting pair $\{\mathbf{v}, \mathbf{S}\}$ that incorporates the sample history for both the chains is computed as [167]

$$\mathbf{v} = \mathbf{v}^a + \mathbf{v}^b ; \quad \mathbf{S} = \mathbf{S}^a + \mathbf{S}^b + \frac{M_s^a M_s^b}{M_s^a + M_s^b} \delta \delta^T ; \quad \delta = \frac{1}{M_s^a} \mathbf{v}^a - \frac{1}{M_s^b} \mathbf{v}^{(b)}. \quad (\text{B.10})$$

This pairwise formula can be applied iteratively to incorporate the results from all parallel chains to obtain the global sum-of-squares matrix $\mathbf{S}^G$. In this work, the parallel chains are synchronized and hence M_s is same for all cores.

The updated proposal covariance C can be computed using the global sum-of-squares matrix $\mathbf{S}^G$ as

$$\mathbf{C} = s_d \times \left\{ \frac{1}{M_s - 1} \mathbf{S}^G + \epsilon \mathbf{I}_p \right\} \quad (\text{B.11})$$

where $\mathbf{I}_p$ is an identity matrix and ϵ is a small constant assigned to prevent the singularity of the resulting covariance matrix. This updated covariance $\mathbf{C}$ is then broadcasted to all the cores, and each core starts sampling using this $\mathbf{C}$. Due to the diminishing nature of the adaptation for AM, M_w can be increased after burn-in to save CPU time while M_w is assigned lower during the burn-in to allow frequent tuning of the proposal. For the range of ϕ dimensions ($5 < p < 10$) considered in this work, the following choice of M_w is a reasonable one: $M_w = 500$ during burn-in and $M_w = 1000$ after burn-in.

When using INCA, the exchange of information across parallel chains decreases as the sampling progresses. The amount of information exchanged can be monitored using the Brooks-Gelman-Rubin (BGR) diagnostic [61, 168]. The diagnostic relies on a normalized ratio of the within-chain and between-chain sample variances and can also be considered as a convergence indicator. Brooks and Gelman [168] stated this criterion of monitoring the convergence through the first two moments (mean and variance) works well even for non-Gaussian target distributions. In this work, we use the BGR diagnostic to determine the burn-in period.

During the initial stages of RWM sampling, the proposal adaptation is slow due to a large sample rejection ratio as a result of a poor proposal covariance [49]. This issue is partially remedied by the DR technique [63, 94]. Here, DR implies a multi-level sampling procedure for the case of a rejected Markov chain move. In other words, if the proposed move ϕ' is rejected, instead of duplicating the current position, a new move ϕ'' is proposed using a second stage

proposal pdf $q_2(\phi, \phi', \phi'')$ that has a lower covariance. A lower covariance implies a higher chance of acceptance for ϕ'' and, therefore, a decrease in the rejection ratio. Multiple DR iterations can be performed on subsequent rejections, albeit at an additional computational cost of computing the posterior $p(\phi|\mathcal{D})$ at each proposed move. Although a second stage proposal pdf is allowed to depend on the current (ϕ) and the rejected (ϕ') position, in this work we choose $q_2(\phi, \phi', \phi'') = \mathcal{N}(\phi, s_{dr}\mathbf{C})$, where s_{dr} is the scale factor and $\mathbf{C}$ is the covariance of the first stage proposal pdf $q_1(\phi, \phi')$. Note that the scaling factor has to be less than one to ensure a higher acceptance probability of the proposed Markov chain during the second stage. Note that since $\mathbf{C}$ gets adapted, the proposal for DR gets adapted as well.

The probability of accepting the second DR move ϕ'' is [63]

$$\alpha_2(\phi^c, \phi', \phi'') = \min \left(1, \frac{p(\phi''|\mathcal{D})\{1 - \alpha_1(\phi'', \phi')\}}{p(\phi^c|\mathcal{D})\{1 - \alpha_1(\phi^c, \phi')\}} \right), \quad (\text{B.12})$$

where the first stage acceptance probability $\alpha_1(\phi, \phi')$ is given in Eq. (B.11). The acceptance probability $\alpha_2(\phi, \phi', \phi'')$ is calculated by satisfying the detailed balance equation and ensures the reversibility of the resulting Markov chain [63]. For Gaussian-like target distributions, the benefit of using DR diminishes with time due to the improvement in the rejection ratio as a result of an improved $\mathbf{C}$. However, for non-Gaussian targets, DR provides a significant improvement in the rejection ratio throughout the sampling. In summary, the INCA coupled DRAM algorithm provides a robust and efficient solution for the issue of a poor choice of the starting ϕ value and the RWM proposal covariance.

B.3 The Chib-Jeliazkov method

In this section, we provide mathematical details regarding the Chib-Jeliazkov method of computing model evidence needed for Bayesian model comparison. This method relies on the stationary samples generated from the posterior parameter pdf $p(\phi|\mathcal{D}, \mathcal{M})$ using an MH based MCMC algorithm. The mathematical notations followed in this section are given in Sec. 2.1.

Consider that a stationary sample set $\{\phi^{(g)}, g = 1, 2, \dots, M\}$ has been generated from the posterior pdf $p(\phi|\mathcal{D}, \mathcal{M})$ for each plausible model $\mathcal{M}$ using a MH algorithm. The proposed move ϕ' of the Markov chain is generated by sampling from a proposal distribution $q(\phi'|\phi)$, where ϕ is the current position of the chain. In an MH algorithm, the proposed move ϕ' is accepted with the probability $r(\phi, \phi')$, where [67]

$$r(\phi, \phi') = \min \left(1, \frac{p(\phi'|\mathcal{D}, \mathcal{M})q(\phi|\phi')}{p(\phi|\mathcal{D}, \mathcal{M})q(\phi'|\phi)} \right). \quad (\text{B.13})$$

Invoking the reversibility property of the MH transition kernel, the Chib-Jeliazkov method estimates the posterior ordinate $p(\phi^*|\mathcal{D}, \alpha)$ at any preselected point ϕ^* in the posterior space [67] as

$$\hat{p}(\phi^*|\mathcal{D}, \mathcal{M}) = \frac{\frac{1}{M} \sum_{g=1}^M r(\phi^{(g)}, \phi^*)q(\phi^{(g)}|\phi^*)}{\frac{1}{J} \sum_{j=1}^J r(\phi^*, \phi^{(j)})}. \quad (\text{B.14})$$

where samples $\phi^{(j)}, j = 1, 2, \dots, J$ are generated from the fixed proposal pdf $q(\phi'|\phi^*)$ using Monte Carlo sampling. The fixed point ϕ^* should be chosen in the high-probability region of the posterior pdf for best performance of the Chib-Jeliazkov method. Using $\hat{p}(\phi^*|\mathcal{D}, \mathcal{M})$, the model evidence is estimated as [67]

$$\hat{p}(\mathcal{D}|\mathcal{M}) = \frac{p(\mathcal{D}|\phi^*)p(\phi^*|\mathcal{M})}{\hat{p}(\phi^*|\mathcal{D}, \mathcal{M})}. \quad (\text{B.15})$$

The accuracy of the evidence estimate $\hat{p}(\mathcal{D}|\mathcal{M})$ obtained using Eq. (B.14) and Eq. (B.15) relies on the quality (auto-correlation) of MCMC samples $\phi^{(g)}$ and the preselected point ϕ^* .

In this work, we choose $\phi^* = \phi^{\text{MAP}}$ to be the MAP estimate, and $\tilde{q}(\phi, \phi') = \mathcal{N}(\phi'|\phi, \mathbf{C})$ where $\mathbf{C}$ is the proposal covariance at the end of the burn-in of the MCMC sampling. GOF and EIG in Eq. (2.7) can be reasonably estimated for nearly Gaussian posterior pdfs as [45]

$$\text{Goodness-of-fit: } E[\ln p(\mathcal{D}|\phi, \mathcal{M})] \approx \ln p(\mathcal{D}|\phi^{\text{MAP}}, \mathcal{M}) \quad (\text{B.16})$$

$$\text{EIG: } E \left[\ln \frac{p(\phi|\mathcal{D}, \mathcal{M})}{p(\phi|\mathcal{M})} \right] \approx \ln p(\mathcal{D}|\phi^{\text{MAP}}, \mathcal{M}) - \ln \hat{p}(\mathcal{D}|\mathcal{M}) \quad (\text{B.17})$$

B.4 Extended Kalman filter

In this section, we list the expressions of EKF for obtaining the update and forecast mean and covariance of the state vector given the nonlinear model and measurement equations in Eq. (2.1) and Eq. (2.2), respectively. This section follows the references [3, 14]. See Sec. 2.1.1 for details about the mathematical notations used in this section.

B.4.1 Update step

$$\text{Updated mean, } \mathbf{u}_{j(k)}^a = \mathbf{u}_{j(k)}^f + \mathbf{K}_k(\mathbf{d}_k - \mathbf{h}_k(\mathbf{u}_{j(k)}^f, \phi, \bar{\epsilon}_k)) \quad (\text{B.18})$$

$$\text{Updated covariance, } \mathbf{P}_{j(k)}^a = (\mathbf{I} - \mathbf{K}_k \mathbf{C}_k) \mathbf{P}_{j(k)}^f \quad (\text{B.19})$$

where the Kalman gain matrix is

$$\mathbf{K}_k = \mathbf{P}_{j(k)}^f \mathbf{C}_k^T (\mathbf{C}_k \mathbf{P}_{j(k)}^f \mathbf{C}_k^T + \mathbf{D}_k \mathbf{\Gamma}_k \mathbf{D}_k^T)^{-1} \quad (\text{B.20})$$

and the Jacobian matrices are

$$\mathbf{C}_k = \left. \frac{\partial \mathbf{h}_k(\cdot)}{\partial \mathbf{u}_{j(k)}} \right|_{\mathbf{u}_{j(k)} = \mathbf{u}_{j(k)}^f, \epsilon_k = \bar{\epsilon}_k}, \quad \mathbf{D}_k = \left. \frac{\partial \mathbf{h}_k(\cdot)}{\partial \epsilon_k} \right|_{\mathbf{u}_{j(k)} = \mathbf{u}_{j(k)}^f, \epsilon_k = \bar{\epsilon}_k} \quad (\text{B.21})$$

B.4.2 Forecast step

$$\text{Forecast mean, } \mathbf{u}_{j+1}^f = \mathbf{g}_j(\mathbf{u}_j^a, \phi, \mathbf{f}_j, \bar{\mathbf{q}}_j) \quad (\text{B.22})$$

$$\text{Forecast covariance, } \mathbf{P}_{j+1}^f = \mathbf{A}_j \mathbf{P}_j^a \mathbf{A}_j^T + \mathbf{B}_j \mathbf{Q}_j \mathbf{B}_j^T \quad (\text{B.23})$$

where the Jacobian matrices are

$$\mathbf{A}_j = \left. \frac{\partial \mathbf{g}_j(\cdot)}{\partial \mathbf{u}_j} \right|_{\mathbf{u}_j = \mathbf{u}_j^a, \mathbf{q}_j = \bar{\mathbf{q}}_j}, \quad \mathbf{B}_j = \left. \frac{\partial \mathbf{g}_j(\cdot)}{\partial \mathbf{q}_j} \right|_{\mathbf{u}_j = \mathbf{u}_j^a, \mathbf{q}_j = \bar{\mathbf{q}}_j}. \quad (\text{B.24})$$

B.4.3 Expression for likelihood function

Given the mean and covariance of the forecast state pdf computed using EKF, the expression for the likelihood function in Eq. (2.4) can be reduced to

$$p(\mathcal{D}|\phi) = \prod_{k=1}^{N_d} \mathcal{N}(\mathbf{d}_k | \mathbf{h}_k(\mathbf{u}_k^f, \bar{\epsilon}_k), \mathbf{C}_k \mathbf{P}_k^f \mathbf{C}_k^T + \mathbf{D}_k \mathbf{\Gamma}_k \mathbf{D}_k^T). \quad (\text{B.25})$$

B.5 Laplace approximation and BIC

Laplace approximation [96] provides an estimate of model evidence by approximating the posterior parameter distribution as a Gaussian distribution centered at the MAP estimate ϕ^{MAP} . This approximation is only valid for the case of *large* measurements, meaning the number of data points in $\mathcal{D}$ is large. The mathematical exposition presented here closely follows Murphy [7].

The logarithm of model evidence $p(\mathcal{D}|\alpha)$ can be written as

$$\ln p(\mathcal{D}|\alpha) = \ln \left[\int \exp\{\mathcal{L}(\phi, \alpha)\} d\phi \right], \quad (\text{B.26})$$

where $\mathcal{L}(\phi, \alpha) = \ln\{p(\mathcal{D}|\phi)p(\phi|\alpha)\}$. When the posterior $p(\phi|\mathcal{D}, \alpha)$ is close to Gaussian, $\mathcal{L}(\phi, \alpha)$ is expanded around ϕ^{MAP} using the Taylor series expansion as

$$\begin{aligned} \mathcal{L}(\phi, \alpha) &= \mathcal{L}(\phi^{\text{MAP}}, \alpha) + \frac{1}{2}(\phi - \phi^{\text{MAP}})^T \frac{\partial^2 \mathcal{L}(\phi, \alpha)}{\partial \phi \partial \phi^T} \Big|_{\phi = \phi^{\text{MAP}}} (\phi - \phi^{\text{MAP}}) + \dots \\ &\approx \mathcal{L}(\phi^{\text{MAP}}, \alpha) - \frac{1}{2}(\phi - \phi^{\text{MAP}})^T \Sigma(\alpha)^{-1} (\phi - \phi^{\text{MAP}}) \end{aligned} \quad (\text{B.27})$$

where the Hessian matrix $\mathbf{H}(\alpha)$ of $\mathcal{L}(\phi, \alpha)$ and the covariance matrix $\Sigma(\alpha)$ of the posterior $p(\phi|\mathcal{D}, \alpha)$ are defined as

$$\mathbf{H}(\alpha) = \Sigma(\alpha)^{-1} = - \frac{\partial^2 \mathcal{L}(\phi, \alpha)}{\partial \phi \partial \phi^T} \Big|_{\phi = \phi^{\text{MAP}}}. \quad (\text{B.28})$$

Substituting $\mathcal{L}(\phi, \alpha)$ from Eq. (B.27) to Eq. (B.26), we get

$$\begin{aligned}
\ln p(\mathcal{D}|\alpha) &\approx \ln \left[\int \exp \left\{ \mathcal{L}(\phi^{\text{MAP}}, \alpha) - \frac{1}{2}(\phi - \phi^{\text{MAP}})^T \Sigma(\alpha)^{-1}(\phi - \phi^{\text{MAP}}) \right\} d\phi \right] \\
&\approx \ln \left[\int \exp \{ \mathcal{L}(\phi^{\text{MAP}}, \alpha) \} \exp \left\{ -\frac{1}{2}(\phi - \phi^{\text{MAP}})^T \Sigma(\alpha)^{-1}(\phi - \phi^{\text{MAP}}) \right\} d\phi \right] \\
&\approx \mathcal{L}(\phi^{\text{MAP}}, \alpha) + \ln \left[\int \exp \left\{ -\frac{1}{2}(\phi - \phi^{\text{MAP}})^T \Sigma(\alpha)^{-1}(\phi - \phi^{\text{MAP}}) \right\} d\phi \right].
\end{aligned} \tag{B.29}$$

Notice that the term inside the integral is an unnormalized Gaussian pdf $\mathcal{N}(\phi|\phi^{\text{MAP}}, \Sigma)$. Hence, Eq. (B.29) can be reduced to

$$\begin{aligned}
\ln p(\mathcal{D}|\alpha) &\approx \mathcal{L}(\phi^{\text{MAP}}, \alpha) + \ln \sqrt{\frac{(2\pi)^{N_\phi}}{|\Sigma(\alpha)^{-1}|}} \\
&\approx \mathcal{L}(\phi^{\text{MAP}}, \alpha) + \frac{N_\phi}{2} \ln 2\pi + \frac{1}{2} \ln |\Sigma(\alpha)|.
\end{aligned} \tag{B.30}$$

Eq. (B.30) provides an estimate of the model evidence using the MAP estimate ϕ^{MAP} and the covariance matrix $\Sigma(\alpha)$ or the Hessian $\mathbf{H}(\alpha)$ computed at ϕ^{MAP} .

Given $\mathcal{L}(\phi, \alpha) = \ln\{p(\mathcal{D}|\phi)p(\phi|\alpha)\}$ and $\mathbf{H}(\alpha) = \Sigma(\alpha)^{-1}$, Eq. (B.30) can also be written as

$$\ln p(\mathcal{D}|\alpha) \approx \ln p(\mathcal{D}|\phi^{\text{MAP}}) + \ln p(\phi^{\text{MAP}}|\alpha) + \frac{N_\phi}{2} \ln 2\pi + \frac{1}{2} \ln |\Sigma(\alpha)| \tag{B.31}$$

$$\approx \underbrace{\ln p(\mathcal{D}|\phi^{\text{MAP}})}_{\text{GOF}} - \underbrace{\ln \left\{ \frac{\mathcal{N}(\phi^{\text{MAP}}|\phi^{\text{MAP}}, \Sigma(\alpha))}{p(\phi^{\text{MAP}}|\alpha)} \right\}}_{\text{EIG or Occam factor}} \tag{B.32}$$

Here, GOF is the maximum log-likelihood and quantifies the data-fit capability of the model. EIG is the KL divergence between the prior and the Gaussian posterior. Notice that Eq. (B.32) is a simplified version of Eq. (2.7) when dealing with a Gaussian posterior pdf produced by large measurement data $\mathcal{D}$.

Next, we describe the derivation of BIC by assuming a large sample size N_d of iid measure-

ments in $\mathcal{D}$. In this case, the following two simplifications can be made in Eq. (B.31):

- Given large $\mathcal{D}$, the prior is considered as inconsequential and the posterior is considered dictated by the likelihood function. In other words, the prior is considered flat in comparison to the likelihood function and can be ignored in Eq. (B.31). Also, the conditioning of prior and posterior on hyperparameters α can be removed, and the MLE estimate ϕ^{MLE} can be substituted in for the MAP estimate ϕ^{MAP} .
- Given iid $\mathcal{D}$, the observed Hessian matrix $\mathbf{H}$ can be written as $\mathbf{H} = N_d \hat{\mathbf{H}}$, where $\hat{\mathbf{H}}$ is the Fisher information matrix. The Fisher matrix represents the amount of information conveyed by each measurement data point, which is the same for all data points in case of iid $\mathcal{D}$. In other words, the information conveyed by N_d number of data points is equal to N_d times the information conveyed by each data point, i.e. $\hat{\mathbf{H}}$.

Based on these two arguments, Eq. (B.31) can be written as

$$\begin{aligned} \ln p(\mathcal{D}|\alpha) &\approx \ln p(\mathcal{D}|\phi^{\text{MLE}}) + \frac{N_\phi}{2} \ln 2\pi - \frac{1}{2} \ln \{N_d^{N_\phi} |\hat{\mathbf{H}}|\} \\ &\approx \ln p(\mathcal{D}|\phi^{\text{MLE}}) + \frac{N_\phi}{2} \ln 2\pi - \frac{N_\phi}{2} \ln N_d - \frac{1}{2} \ln |\hat{\mathbf{H}}| \end{aligned} \quad (\text{B.33})$$

For large N_d , the term $N_\phi \ln 2\pi$ will be small in comparison to the term $N_\phi \ln N_d$, and, therefore, can be ignored. Also, the term $\ln |\hat{\mathbf{H}}|$ is a constant for all models. Hence, the BIC model comparison measure is obtained by further simplifying Eq. (B.33) as

$$\text{BIC} = \underbrace{2 \ln p(\mathcal{D}|\phi^{\text{MLE}})}_{\text{Data-fit}} - \underbrace{N_\phi \ln N_d}_{\text{Model complexity}}. \quad (\text{B.34})$$

Notice that the BIC promotes a trade-off between data-fit and model complexity, wherein the data-fit is quantified by the maximum log-likelihood and the complexity (the penalty term) is quantified in terms of the number of unknown model parameters (N_ϕ). In contrast, AIC operates

as [6]

$$\text{AIC} = \underbrace{2 \ln p(\mathcal{D}|\phi^{\text{MLE}})}_{\text{Data-fit}} - \underbrace{2N_\phi}_{\text{Model complexity}} . \quad (\text{B.35})$$

Notice that the penalty term is higher for BIC than AIC. See Konishi and Kitagawa [6] for a detailed comparison of AIC and BIC.

Appendix C

Derivation of higher-order Wagner unsteady models

We present a detailed mathematical derivation for converting the integro-differential equation in Eq. (2.13) into an ODE for a second-order and third-order Wagner function approximation. This section closely follows the references: [106, 169].

C.1 Second-order Wagner function

A second-order Wagner function of the exponential form is known as

$$\varphi(\tau) = 1 - A_1 e^{-B_1 \tau} - A_2 e^{-B_2 \tau}. \quad (\text{C.1})$$

Taking the Laplace transform of the Wagner function, we get

$$\begin{aligned}
 \mathcal{L}\{\varphi(\tau)\} &= \mathcal{L}\{1 - A_1 e^{-B_1 \tau} - A_2 e^{-B_2 \tau}\} \\
 &= \frac{1}{r} - \frac{A_1}{r + B_1} - \frac{A_2}{r + B_2} \\
 &= \frac{(r + B_1)(r + B_2) - A_1 r(r + B_2) - A_2 r(r + B_1)}{s.(r + B_1).(r + B_2)} \\
 &= \frac{0.5r^2 + r(B_1(1 - A_2) + B_2(1 - A_1)) + B_1 B_2}{r(r^2 + r(B_1 + B_2) + B_1 B_2)}.
 \end{aligned}$$

where $A_1 + A_2 = 0.5$ was imposed based on the constraint that $\varphi(\tau = 0) = 0.5$. Consider that $\mathcal{L}\{\theta(\tau)\} = \Theta(r)$ and $\mathcal{L}\{\dot{\theta}(\tau)\} = \mathcal{X}(r)$. Next, we substitute $\mathcal{L}\{\varphi(\tau)\}$ from the above equation into Eq. (2.15b) to obtain

$$\begin{aligned}
 \mathcal{L}\{C_M\} &= \frac{2\pi e}{c\mathcal{U}} \mathcal{U}\{\Theta(r) + (0.5 - a_h)\mathcal{X}(r)\} \left\{ \frac{0.5r^2 + r(B_1(1 - A_2) + B_2(1 - A_1)) + B_1 B_2}{r(r^2 + r(B_1 + B_2) + B_1 B_2)} \right\} \\
 r^2 \mathcal{L}\{C_M\} &+ (B_1 + B_2)r \mathcal{L}\{C_M\} + B_1 B_2 \mathcal{L}\{C_M\} \\
 &= \frac{2\pi e}{c} \{B_1 B_2 \Theta(r) + (B_1(1 - A_2) + B_2(1 - A_1))r \Theta(r) + (0.5 - a_h)B_1 B_2 \mathcal{X}(r) + \\
 &\quad 0.5r^2 \Theta(r) + (0.5 - a_h)(B_1(1 - A_2) + B_2(1 - A_1))r \mathcal{X}(r) + 0.5(0.5 - a_h)r^2 \mathcal{X}(r)\}.
 \end{aligned}$$

The inverse Laplace transform of the above equation leads to

$$\begin{aligned}
 \ddot{C}_M &+ (B_1 + B_2)\dot{C}_M + B_1 B_2 C_M \\
 &= \frac{2\pi e}{c} \{B_1 B_2 \dot{\theta} + (B_1(1 - A_2) + B_2(1 - A_1))\dot{\theta} + (0.5 - a_h)B_1 B_2 \ddot{\theta} + \\
 &\quad 0.5\ddot{\theta} + (0.5 - a_h)(B_1(1 - A_2) + B_2(1 - A_1))\ddot{\theta} + 0.5(0.5 - a_h)\ddot{\theta}\}.
 \end{aligned}$$

Rearranging terms in the above equation, we get

$$\left(\frac{1}{B_1 B_2}\right) \ddot{C}_M + \left(\frac{1}{B_1} + \frac{1}{B_2}\right) \dot{C}_M + C_M = a_1 \theta + a_2 \dot{\theta} + a_3 \ddot{\theta} + a_4 \dddot{\theta} \quad (\text{C.2a})$$

$$\text{where } a_1 = \frac{2\pi e}{c} \quad (\text{C.2b})$$

$$a_2 = \frac{2\pi e}{c} \left\{ \left(\frac{1}{2} - a_h\right) + \left(\frac{1 - A_1}{B_1} + \frac{1 - A_2}{B_2}\right) \right\} \quad (\text{C.2c})$$

$$a_3 = \frac{2\pi e}{c} \left\{ \left(\frac{1}{2} - a_h\right) \times \left(\frac{1 - A_1}{B_1} + \frac{1 - A_2}{B_2}\right) + \left(\frac{1}{2B_1 B_2}\right) \right\} \quad (\text{C.2d})$$

$$a_4 = \frac{2\pi e}{c} \left\{ \left(\frac{1}{2} - a_h\right) \times \left(\frac{1}{2B_1 B_2}\right) \right\}, \quad (\text{C.2e})$$

where $A_1 + A_2 = 0.5$ and $B_1, B_2 > 0$. Notice that when setting B_1 and B_2 to infinity or $\varphi(\tau)$ as one, Eq. (C.2) reduces to the quasi-steady ODE of $C_M = a_1 \theta + a_2 \dot{\theta}$ as the terms involving $\ddot{C}_M$, $\dot{C}_M$, $\ddot{\theta}$ and $\dddot{\theta}$ are eliminated.

C.2 Third-order Wagner function

A third-order Wagner function is

$$\varphi^{(3)}(t) = 1 - A_1 e^{-B_1 \tau} - A_2 e^{-B_2 \tau} - A_3 e^{-B_3 \tau} \quad (\text{C.3})$$

Following the same procedure as described in the case of the second-order Wagner function, we obtain

$$\left(\frac{1}{B_1 B_2 B_3}\right) \ddot{C}_M + \left(\frac{1}{B_2 B_3} + \frac{1}{B_1 B_3} + \frac{1}{B_1 B_2}\right) \dot{C}_M + \left(\frac{1}{B_1} + \frac{1}{B_2} + \frac{1}{B_3}\right) C_M + C_M \quad (\text{C.4a})$$

$$= a_1 \theta + a_2 \dot{\theta} + a_3 \ddot{\theta} + a_4 \ddot{\theta} + a_5 \ddot{\theta} \quad (\text{C.4b})$$

$$\text{where } a_1 = \frac{2\pi e}{c} \left\{ \left(\frac{1}{2} - a_h \right) \times 0.0 + 1.0 \right\} \quad (\text{C.4c})$$

$$a_2 = \frac{2\pi e}{c} \left\{ \left(\frac{1}{2} - a_h \right) \times 1.0 + S_1 \right\} \quad (\text{C.4d})$$

$$a_3 = \frac{2\pi e}{c} \left\{ \left(\frac{1}{2} - a_h \right) \times S_1 + S_2 \right\} \quad (\text{C.4e})$$

$$a_4 = \frac{2\pi e}{c} \left\{ \left(\frac{1}{2} - a_h \right) \times S_2 + S_3 \right\} \quad (\text{C.4f})$$

$$a_5 = \frac{2\pi e}{c} \left\{ \left(\frac{1}{2} - a_h \right) \times S_3 + 0.0 \right\} \quad (\text{C.4g})$$

$$S_1 = \frac{1 - A_1}{B_1} + \frac{1 - A_2}{B_2} + \frac{1 - A_3}{B_3} \quad (\text{C.4h})$$

$$S_2 = \frac{1 - (A_2 + A_3)}{B_2 B_3} + \frac{1 - (A_1 + A_3)}{B_1 B_3} + \frac{1 - (A_1 + A_2)}{B_1 B_2} \quad (\text{C.4i})$$

$$S_3 = \frac{1}{2B_1 B_2 B_3}, \quad (\text{C.4j})$$

where $A_1 + A_2 + A_3 = 0.5$ and $B_1, B_2, B_3 > 0$. Again, setting B_i 's to infinity would reduce Eq. (C.4) to the quasi-steady ODE, i.e. $C_M = a_1 \theta + a_2 \dot{\theta}$.

Appendix D

Application of NSBL to bimodal likelihood functions

In this section, we demonstrate the applicability of NSBL to two-dimensional models where both the parameters are treated as questionable ($\phi=\phi_\alpha$) and $\phi_{-\alpha}$ is a null vector. In addition, we assume that the kernel-based approximation for the likelihood function $p(\mathcal{D}|\phi)$ in Eq. (4.3) is known in terms of two Gaussian kernels ($K=2$). The model parameter vector for this case is denoted as $\phi=\{\phi_1, \phi_2\}$. Following the NSBL algorithm from Section 4.2, an ARD prior is assigned to ϕ in terms of hyperparameter vector $\alpha = \{\alpha_1, \alpha_2\}$ as

$$p(\phi|\alpha) = \mathcal{N}(\phi_1|0, \alpha_1^{-1})\mathcal{N}(\phi_2|0, \alpha_2^{-1}). \quad (\text{D.1})$$

Table D.1 shows the two synthetic choices of $p(\mathcal{D}|\phi)$ considered in this numerical exercise. The NSBL algorithm outlined in Algorithm 2 is applied to identify sparsity in ϕ for each case listed in Table D.1.

Figure D.1 shows the likelihood function $p(\mathcal{D}|\phi)$, log-evidence $\log p(\mathcal{D}|\log \alpha)$ and objective function $\mathcal{L}(\log \alpha)$ pertaining to Case A in Table D.1. The objective function is obtained using hyperprior parameters of $\log r_i=\log s_i=-12$ in Eq. (4.29). The likelihood function in Figure D.1a is centered at $[1,0]$ and $[0,1]$, thereby introducing bimodality in ϕ space. This multimodality

	Likelihood function $p(\mathcal{D} \phi)$ approximation in Eq. (4.3)
Case A	$\mathcal{N}\left(\begin{Bmatrix}\phi_1 \\ \phi_2\end{Bmatrix} \middle \begin{Bmatrix}1.0 \\ 0.0\end{Bmatrix}, \begin{bmatrix}0.05 & 0.00 \\ 0.00 & 0.04\end{bmatrix}\right) + \mathcal{N}\left(\begin{Bmatrix}\phi_1 \\ \phi_2\end{Bmatrix} \middle \begin{Bmatrix}0.0 \\ 1.0\end{Bmatrix}, \begin{bmatrix}0.05 & -0.03 \\ -0.03 & 0.04\end{bmatrix}\right)$
Case B	$\mathcal{N}\left(\begin{Bmatrix}\phi_1 \\ \phi_2\end{Bmatrix} \middle \begin{Bmatrix}-1.0 \\ -1.0\end{Bmatrix}, \begin{bmatrix}0.05 & 0.00 \\ 0.00 & 0.04\end{bmatrix}\right) + \mathcal{N}\left(\begin{Bmatrix}\phi_1 \\ \phi_2\end{Bmatrix} \middle \begin{Bmatrix}1.0 \\ 1.0\end{Bmatrix}, \begin{bmatrix}0.05 & -0.03 \\ -0.03 & 0.04\end{bmatrix}\right)$

Table D.1: Two cases of likelihood function considered for the application of NSBL.

in likelihood function mimics the nonlinear relation between model parameters ϕ and observations $\mathcal{D}$ encountered in engineering mechanics. This choice of likelihood kernels induces multimodality in log-evidence (Figure D.1b) and in objective function (Figure D.1c). The objective function $\mathcal{L}(\alpha)$ in Figure D.1c was computed using flat hyperprior for $\log \alpha_i$ by selecting $\log r_i = \log s_i = -12.0$ in Eq. (4.29). This choice of hyperprior results in objective function being dictated by log-evidence for all but large $\log \alpha_i$ values. As a result, the objective function possess unique local peaks, unlike log-evidence where the two peaks extend to infinity. In other words, hyperprior $p(\log \alpha)$ regularizes the log-evidence, resulting in an objective function that is free from flat regions and, therefore, amicable to gradient-based optimization of Newton's method. We will revisit the influence of hyperprior parameters r_i and s_i on the sparse solution later in this section.

Figure D.1: Bayesian entities pertaining to Case A from Table D.1.

Next, NSBL is applied to optimize the objective function shown in Figure D.1c. Note that the data-based algorithms of SBL and BCS are inapplicable for such cases. NSBL is initiated from

$\log \alpha = \{10, 10\}$. A function tolerance for $\mathcal{L}(\log \alpha)$ is set at $1e-08$ for terminating the Newton iteration. Figure D.2a shows α iterates; Figure D.2b shows log-evidence $\log p(\mathcal{D} | \log \alpha)$ and objective function $\mathcal{L}(\log \alpha)$; Figure D.2c shows relevance indicator γ_i^{rms} ; and Figure D.2d shows the resulting posterior pdf computed using α^{MAP} . The final solution provided by NSBL is $\log \alpha^{\text{MAP}} = \{0.0, 7.0\}$. Notice that α iterates in Figure D.2a move in the orthogonal direction to $\mathcal{L}(\log \alpha)$ curvature. This property of Newton iteration enables efficient calculation of α^{MAP} using minimum iterations. From Figure D.2b, log-evidence and objective function are same when using a flat hyperprior for $\log \alpha_i$ ($\log r_i = \log s_i = -12.0$). This observation reiterates the notion that objective function is same as log-evidence when using flat hyperprior for α_i , i.e. $p(\log \alpha_i) \propto 1$.

Figure D.2: NSBL results for Case A in Table D.1 when initiated from $\log \alpha = \{10, 10\}$.

The relevance indicator in Figure D.2c is zero in the beginning due to a large starting value of $\log \alpha = \{10, 10\}$. During Newton iteration, $\log \alpha_1$ value decreases while γ_1^{rms} approaches a

value of one. At the optimum, $\gamma_1^{\text{rms}} \approx 1.0$ implying ϕ_1 is deemed relevant by NSBL. On the other hand, γ_2^{rms} stays zero as $\log \alpha_2$ is unchanged during Newton iteration. $\gamma_2^{\text{rms}} \approx 0.0$ implies ϕ_2 is redundant. This irrelevance of ϕ_2 is also evident from Figure D.2d as the posterior pdf for ϕ_2 resembles a Dirac-delta function centered at zero. Henceforth, ϕ_2 could be removed from the model, leading to dimensionality reduction. On the other hand, the relevant parameter ϕ_1 has a posterior pdf centered at one. This centering of posterior pdf of ϕ_2 resembles that of the likelihood function shown in Figure D.1a. This posterior pdf of ϕ_2 has no contribution from the ARD prior $\mathcal{N}(\phi_i|0, 1/\alpha_1^{\text{MAP}})$.

Figure D.3 shows NSBL results for Case A when initiating from $\alpha = \{-5, -5\}$. Rest of the NSBL settings are kept same as before. Notice that NSBL produces a different solution for α^{MAP} than previously obtained in Figure D.2. This difference in sparse solution reinforces the need for multistart Newton iteration for a general physics-based sparse learning problem. Notice that relevance indicator γ_i^{rms} starts from a value of one due to a low starting value of $\log \alpha_i = -5$. Parameter ϕ_1 is deemed irrelevant as γ_1^{rms} approaches a value of zero, while ϕ_2 is deemed relevant as γ_2^{rms} approaches a value of one. When using multistart, the global optimum pertains to solution reported in Figure D.2. Results in Figure D.2 and Figure D.3 pertain to using hyperprior parameters of $\log r_i = \log s_i = -12.0$.

Next, we perform a sensitivity study to investigate the effect of hyperprior parameters $\log r_i$ ($= \log s_i$) on NSBL. Since we are assigning Jeffery's prior ($p(\log \alpha_i) \propto 1$) to hyperparameters α_1 and α_2 , hyperprior parameters are kept same ($r_i = s_i$). NSBL is executed for Case A in Table D.1 with varying values of $\log r_i$. Figure D.4a shows optimal log-evidence and optimal objective function; Figure D.4b shows optimal relevance indicator γ_i^{rms} ; and Figure D.4c shows the optimal hyperparameter value α^{MAP} for an increasing value of $\log r_i$. As evident from Figure D.4b, the optimal relevance indicator stays unchanged for all $\log r_i < -10$. For all $\log r_i > -10$, hyperprior $p(\log \alpha)$ starts to overcome the contribution of log-evidence in the objective function, as per Eq. (4.29). A hyperprior dominated objective function forces α^{MAP} towards zero, thereby making both parameters relevant as γ_i^{rms} approaches one for all ϕ_i . This

Figure D.3: NSBL results for Case A in Table D.1 when initiating at $\alpha = \{-5, -5\}$.

observation demonstrates the need for a cautious selection of hyperprior parameter $\log r_i$ or $\log s_i$ in NSBL. We recommend performing a sensitivity study to investigate whether the sparse solution varies with changing $\log r_i$ values. Note this study can be executed with minimal computational effort due to the analytical apparatus of NSBL. When in doubt, lower values of r_i should be preferred to ensure the objective function remains dictated by log-evidence.

Figure D.4: Variation in optimal Bayesian entities for varying hyperprior parameters $\log r_i$ or $\log s_i$.

Next, we execute NSBL for the Case B in Table D.1. Figure D.5 shows the likelihood function, log-evidence and objective function pertaining to Case B in Table D.1. As seen from Table D.1, the only difference between Case A and Case B is the centering of Gaussian kernels of the likelihood function approximation. For case B, the kernels are centered at $[-1, -1]$ and $[1, 1]$. As seen in Figure D.5, this centering of likelihood function results in a uniquely peaked log-evidence and objective function. This observation demonstrates that multimodality in likelihood function can produce unimodal or multimodal log-evidence depending on the location of local modes of the likelihood function.

Figure D.5: Bayesian entities pertaining to Case B from Table D.1.

Figure D.6 shows the NSBL results when initiating from $\log \alpha = \{10, 10\}$ and using $\log r_i = \log s_i = -12.0$. As seen from Figure D.6a, NSBL reaches the unique optimum of $\log \alpha^{\text{MAP}} = \{0.0, 0.5\}$ in 12 Newton iterations. The optimal relevance indicator value of one in Figure D.6c indicates the relevance of both parameters ϕ_1 and ϕ_2 . The posterior pdf is obtained as a bimodal pdf centered at -1.0 and 1.0, resembling that of the likelihood function shown in Figure D.5a. In summary, NSBL provides a robust way to determine sparsity in multimodal log-evidence or multimodal likelihood function scenarios.

Figure D.6: NSBL results for Case B in Table D.1 when initiating from $\log \alpha = \{10, 10\}$.

Appendix E

PCE surrogate and Sobol sensitivity indices

Consider a PCE surrogate with dimension $d = 2$ and order $p = 3$. The standardized input (germ) vector is denoted as $\xi = \{\xi_1, \xi_2\}$. The univariate PCE polynomials of j^{th} input ξ_j are denoted as $\psi_i(\xi_j)$ where i is the polynomial degree. The number of PCE terms for a total-order truncation will be $(2 + 3)!/(2!3!) = 10$. Each PCE basis $\Psi_k(\xi)$ is obtained through tensorization of these univariate PCE polynomials $\psi_i(\xi_j)$ obtained from Askey family of polynomials [145]. The choice of these PCE polynomials is dictated by the probability distribution of germs ξ [146]. The coefficient pertaining to PCE basis $\Psi_k(\xi)$ is denoted as y_k . The variance of each PCE basis $\Psi_k(\xi)$ is known analytically (product of variance of univariate polynomials) and is denoted as $||\Psi_k||^2$. Table E.1 illustrates the variance decomposition principle of computing Sobol indices through the PCE surrogate for this simple two-dimensional input case.

Given the PCE coefficients are available, the Sobol indices are obtained using Table E.1

Order	k	Multi-index α_k	PCE basis $\Psi_k(\xi)$	Variance contribution solely due to		
				ξ_1	ξ_2	interaction
0	0	[0,0]	1.0			
1	1	[1,0]	$\psi_1(\xi_1)$	$y_1^2 \Psi_1 ^2$		
	2	[0,1]	$\psi_1(\xi_2)$		$y_2^2 \Psi_2 ^2$	
2	3	[2,0]	$\psi_2(\xi_1)$	$y_3^2 \Psi_3 ^2$		
	4	[1,1]	$\psi_1(\xi_1)\psi_1(\xi_2)$			$y_4^2 \Psi_4 ^2$
	5	[0,2]	$\psi_2(\xi_2)$		$y_5^2 \Psi_5 ^2$	
3	6	[3,0]	$\psi_3(\xi_1)$	$y_6^2 \Psi_6 ^2$		
	7	[2,1]	$\psi_2(\xi_1)\psi_1(\xi_2)$			$y_7^2 \Psi_7 ^2$
	8	[1,2]	$\psi_1(\xi_1)\psi_2(\xi_2)$			$y_8^2 \Psi_8 ^2$
	9	[0,3]	$\psi_3(\xi_2)$		$y_9^2 \Psi_9 ^2$	
				V_1	V_2	V_{12}

Table E.1: Sobol indices calculation from PCE surrogates through variance decomposition.

as [147]

$$\text{Total output variance: } V_{\text{total}} = V_1 + V_2 + V_{12}$$

$$\text{First-order Sobol indices: } S_1 = \frac{V_1}{V_{\text{total}}} \quad , \quad S_2 = \frac{V_2}{V_{\text{total}}}$$

$$\text{Second-order Sobol index: } S_{12} = \frac{V_{12}}{V_{\text{total}}}$$

$$\text{Total-order Sobol index: } S_1^T = S_1 + S_{12} \quad , \quad S_2^T = S_2 + S_{12}$$

Appendix F

Gaussian distribution identities

F.1 Conditional distribution relations for Gaussian pdfs

Given a jointly Gaussian random vectors $\mathbf{X}_1$ and $\mathbf{X}_2$ with marginal distributions $p(\mathbf{X}_1) = \mathcal{N}(\mathbf{X}_1|\boldsymbol{\mu}_1, \Sigma_{11})$ and $p(\mathbf{X}_2) = \mathcal{N}(\mathbf{X}_2|\boldsymbol{\mu}_2, \Sigma_{22})$ and the joint distribution as [40]

$$p(\mathbf{X}_1, \mathbf{X}_2) = \mathcal{N} \left(\left\{ \begin{matrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{matrix} \right\} \middle| \left\{ \begin{matrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{matrix} \right\}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{bmatrix} \right),$$

then the conditional distribution can be computed as

$$p(\mathbf{X}_1|\mathbf{X}_2) = \mathcal{N}(\mathbf{X}_1|\mathbf{X}_2|\boldsymbol{\mu}_{1|2}, \Sigma_{1|2}) \quad (\text{F.1a})$$

$$\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{X}_2 - \boldsymbol{\mu}_2) \quad (\text{F.1b})$$

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T \quad (\text{F.1c})$$

where $\mathbf{K}_{1|2} = \Sigma_{12}\Sigma_{22}^{-1}$ is called the gain matrix and $\Sigma_{1|2}$ is also called the schur complement of matrix Σ_{22} .

F.2 Cross-entropy between two multivariate Gaussian pdfs

Consider that we have a random vector $\mathbf{w}$ of size N that has two possible probability distributions denoted as $r_1 = \mathcal{N}(\mathbf{w}|\mathbf{m}_1, \mathbf{P}_1)$ and $r_2 = \mathcal{N}(\mathbf{w}|\mathbf{m}_2, \mathbf{P}_2)$. The cross-entropy $H(r_1, r_2)$ between pdfs r_1 and r_2 is defined as the expectation of logarithm of the pdf r_2 where the expectation is taken with respect to the pdf r_1 . The cross-entropy $H(r_1, r_2)$ for Gaussian pdfs can be computed analytically as follows [6]:

$$\begin{aligned} H(r_1, r_2) &= \mathbb{E}[\log \mathcal{N}(\mathbf{w}|\mathbf{m}_2, \mathbf{P}_2)] = \int \mathcal{N}(\mathbf{w}|\mathbf{m}_1, \mathbf{P}_1) \log \mathcal{N}(\mathbf{w}|\mathbf{m}_2, \mathbf{P}_2) d\mathbf{w} \\ &= \mathbb{E} \left[-\frac{N}{2} \log 2\pi - \frac{1}{2} \log |\mathbf{P}_2| - \frac{1}{2} (\mathbf{w} - \mathbf{m}_2)^T \mathbf{P}_2^{-1} (\mathbf{w} - \mathbf{m}_2) \right] \\ &= -\frac{N}{2} \log 2\pi - \frac{1}{2} \log |\mathbf{P}_2| - \frac{1}{2} \mathbb{E} [(\mathbf{w} - \mathbf{m}_2)^T \mathbf{P}_2^{-1} (\mathbf{w} - \mathbf{m}_2)] \end{aligned}$$

Using the property $\mathbf{z}^T \mathbf{z} = \text{Trace}(\mathbf{z} \mathbf{z}^T)$, we get

$$\begin{aligned} H(r_1, r_2) &= -\frac{N}{2} \log 2\pi - \frac{1}{2} \log |\mathbf{P}_2| - \frac{1}{2} \mathbb{E} [\text{Trace}(\mathbf{P}_2^{-1} (\mathbf{w} - \mathbf{m}_2) (\mathbf{w} - \mathbf{m}_2)^T)] \\ &= -\frac{N}{2} \log 2\pi - \frac{1}{2} \log |\mathbf{P}_2| - \frac{1}{2} \text{Trace}(\mathbb{E}[\mathbf{P}_2^{-1} (\mathbf{w} - \mathbf{m}_2) (\mathbf{w} - \mathbf{m}_2)^T]) \\ &= -\frac{N}{2} \log 2\pi - \frac{1}{2} \log |\mathbf{P}_2| - \frac{1}{2} \text{Trace}(\mathbf{P}_2^{-1} \mathbb{E}[\mathbf{w} \mathbf{w}^T - 2\mathbf{m}_2 \mathbf{w}^T + \mathbf{m}_2 \mathbf{m}_2^T]) \\ &= -\frac{N}{2} \log 2\pi - \frac{1}{2} \log |\mathbf{P}_2| - \frac{1}{2} \text{Trace}(\mathbf{P}_2^{-1} (\mathbb{E}[\mathbf{w} \mathbf{w}^T] - 2\mathbf{m}_2 \mathbb{E}[\mathbf{w}^T] + \mathbf{m}_2 \mathbf{m}_2^T)) \\ &= -\frac{N}{2} \log 2\pi - \frac{1}{2} \log |\mathbf{P}_2| - \frac{1}{2} \text{Trace}(\mathbf{P}_2^{-1} (\mathbf{P}_1 + \mathbf{m}_1 \mathbf{m}_1^T - 2\mathbf{m}_2 \mathbf{m}_1^T + \mathbf{m}_2 \mathbf{m}_2^T)) \\ &= -\frac{N}{2} \log 2\pi - \frac{1}{2} \log |\mathbf{P}_2| - \frac{1}{2} (\mathbf{m}_1 - \mathbf{m}_2)^T \mathbf{P}_2^{-1} (\mathbf{m}_1 - \mathbf{m}_2) - \frac{1}{2} \text{Trace}(\mathbf{P}_2^{-1} \mathbf{P}_1) \\ &= \log \mathcal{N}(\mathbf{m}_1|\mathbf{m}_2, \mathbf{P}_2) - \frac{1}{2} \text{Trace}(\mathbf{P}_2^{-1} \mathbf{P}_1) \end{aligned} \tag{F.2}$$

Note that if $\mathbf{P}_1 = \mathbf{P}_2$ then $\text{Trace}(\mathbf{P}_2^{-1} \mathbf{P}_1) = N$.

F.3 Multiplication of two Gaussian pdfs

The product of two multivariate Gaussian pdfs $\mathcal{N}(\mathbf{w}|\mathbf{m}_1, \mathbf{P}_1)$ and $\mathcal{N}(\mathbf{w}|\mathbf{m}_2, \mathbf{P}_2)$ is obtained as (Eq. 371 in matrix cookbook [137])

$$\mathcal{N}(\mathbf{w}|\mathbf{m}_1, \mathbf{P}_1) \cdot \mathcal{N}(\mathbf{w}|\mathbf{m}_2, \mathbf{P}_2) = \mathcal{N}(\mathbf{m}_1|\mathbf{m}_2, \mathbf{P}_1 + \mathbf{P}_2) \mathcal{N}(\mathbf{w}|\bar{\mathbf{m}}, \bar{\mathbf{P}}) \quad (\text{F.3a})$$

$$\bar{\mathbf{m}} = (\mathbf{P}_1^{-1} + \mathbf{P}_2^{-1})^{-1} (\mathbf{P}_1^{-1} \mathbf{m}_1 + \mathbf{P}_2^{-1} \mathbf{m}_2) \quad (\text{F.3b})$$

$$\bar{\mathbf{P}} = (\mathbf{P}_1^{-1} + \mathbf{P}_2^{-1})^{-1} \quad (\text{F.3c})$$

Bibliography

- [1] R. Sandhu, D. Poirel, C. Pettit, M. Khalil, A. Sarkar, **Bayesian inference of nonlinear unsteady aerodynamics from aeroelastic limit cycle oscillations**, Journal of Computational Physics 316 (2016) 534 – 557.
- [2] R. Sandhu, C. Pettit, M. Khalil, D. Poirel, A. Sarkar, **Bayesian model selection using automatic relevance determination for nonlinear dynamical systems**, Computer Methods in Applied Mechanics and Engineering 320 (2017) 237 – 260.
- [3] M. Khalil, Bayesian Inference for Complex and Large-scale Engineering Systems, Ph.D. thesis, Carleton University (2013).
- [4] L. Ljung, **System Identification**, Birkhäuser Boston, Boston, MA, 1998, pp. 163–173.
- [5] J. Beck, **Bayesian system identification based on probability logic**, Structural Control and Health Monitoring 17 (7) (2010) 825–847.
- [6] S. Konishi, G. Kitagawa, Information Criteria and Statistical Modeling (Springer Series in Statistics), Springer, 2007.
- [7] K. P. Murphy, Machine learning : a probabilistic perspective, The MIT Press, 2012.
- [8] S. Tan, C. Fox, G. Nicholas, **Physics 707 (Inverse Problems Course) Lecture Notes** (2009).
URL <https://sites.google.com/site/szemengtan/>
- [9] M. Khalil, D. Poirel, A. Sarkar, **Probabilistic parameter estimation of a fluttering aeroelastic system in the transitional reynolds number regime**, Journal of Sound and Vibration 332 (15) (2013) 3670–3691.
- [10] E. M. Hernandez, **Optimal model-based state estimation in mechanical and structural systems**, Structural Control and Health Monitoring 20 (4) (2013) 532–543.
- [11] A. Desai, Scalable Domain Decomposition Algorithms for Uncertainty Quantification in High Performance Computing, Ph.D. thesis, Carleton University (2018).
- [12] R. E. Kalman, A New Approach to Linear Filtering and Prediction Problems 1, Journal Of Basic Engineering 82 (Series D) (1960) 35–45.
- [13] M. S. Grewal, **Kalman Filtering**, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 705–708.

- [14] C. Chui, G. Chen, Kalman Filtering with Real-Time Applications, Vol. 6 of Springer Ser. Info. Sci., Vol. 17, Springer, 1987.
- [15] M. S. Grewal, A. P. Andrews, Kalman filtering: theory and practice using MATLAB, Vol. 3, John Wiley & Sons, Inc., 2001.
- [16] G. Evensen, Data Assimilation: The Ensemble Kalman Filter, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.
- [17] M. Khalil, A. Sarkar, S. Adhikari, **Tracking noisy limit cycle oscillation with nonlinear filters**, Journal of Sound and Vibration 329 (2) (2010) 150–170.
- [18] N. J. Gordon, D. J. Salmond, A. F. M. Smith, Novel approach to nonlinear/non-gaussian bayesian state estimation, IEE Proceedings F - Radar and Signal Processing 140 (2) (1993) 107–113.
- [19] M. Khalil, A. Sarkar, S. Adhikari, D. Poirel, **The estimation of time-invariant parameters of noisy nonlinear oscillatory systems**, Journal of Sound and Vibration 344 (2015) 81 – 100.
- [20] E. M. Hernandez, D. Bernal, **State estimation in structural systems with model uncertainties**, Journal of Engineering Mechanics 134 (3) (2008) 252–257.
- [21] K. P. Burnham, D. R. Anderson, Model selection and multimodel inference: a practical information-theoretic approach, 2nd Edition, Springer, 2002.
- [22] T. Bayes, An essay towards solving a problem in the doctrine of chances, Phil. Trans. of the Royal Soc. of London 53 (1763) 370–418.
- [23] J. Bernardo, A. Smith, Bayesian Theory, Wiley series in probability and statistics, Wiley, 2000.
- [24] J. O. Berger, Statistical Decision Theory and Bayesian Analysis, Springer Series in Statistics, 1985.
- [25] T. Hastie, R. Tibshirani, J. H. Friedman, **The elements of statistical learning: data mining, inference, and prediction, 2nd Edition**, Springer series in statistics, Springer, 2009.
- [26] H. Akaike, A new look at the statistical model identification, IEEE Transactions on Automatic Control 19 (6) (1974) 716–723.
- [27] G. Schwarz, Estimating the dimension of a model, The Annals of Statistics 6 (1978) 461–464.
- [28] A. van der Linde, **DIC in variable selection**, Statistica Neerlandica 59 (1) (2005) 45–56.
- [29] J. Rissanen, **Estimation of structure by minimum description length**, Circuits, Systems and Signal Processing 1 (3) (1982) 395–406.

- [30] J. Ding, V. Tarokh, Y. Yang, Model selection techniques: An overview, *IEEE Signal Processing Magazine* 35 (6) (2018) 16–34.
- [31] P. Stoica, Y. Selen, Model-order selection: a review of information criterion rules, *IEEE Signal Processing Magazine* 21 (4) (2004) 36–47.
- [32] I. Guyon, A. Elisseeff, [An introduction to variable and feature selection](#), *J. Mach. Learn. Res.* 3 (2003) 1157–1182.
- [33] J. B. Johnson, K. S. Omland, [Model selection in ecology and evolution](#), *Trends in Ecology & Evolution* 19 (2) (2004) 101 – 108.
- [34] N. Mangan, J. Kutz, S. Brunton, J. Proctor, Model selection for dynamical systems via sparse regression and information criteria, *Proceedings of The Royal Society of London, Series A: Mathematical and Physical Sciences* 473 (2204) (8 2017).
- [35] S. Kullback, R. A. Leibler, On information and sufficiency, *The Annals of Mathematical Statistics* 22 (1) (1951) 79–86.
- [36] G. Box, G. Tiao, *Bayesian inference in statistical analysis*, Vol. 40, John Wiley & Sons, 2011.
- [37] M. Kennedy, A. O’Hagan, [Bayesian calibration of computer models](#), *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63 (3) (2001) 425–464.
- [38] G. E. P. Box, [Science and statistics](#), *Journal of the American Statistical Association* 71 (356) (1976) 791–799.
- [39] J. A. Hoeting, D. Madigan, A. E. Raftery, C. T. Volinsky, [Bayesian model averaging: A tutorial](#), *Statistical Science* 14 (4) (1999) 382–401.
- [40] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.
- [41] J. Beck, L. Katafygiotis, Updating Models and Their Uncertainties. I: Bayesian Statistical Framework, *Journal of Engineering Mechanics* 124 (4) (1998) 455–461.
- [42] L. S. Katafygiotis, J. L. Beck, Updating models and their uncertainties. II: Model identifiability, *Journal of Engineering Mechanics* 124 (4) (1998) 463–467.
- [43] K. Yuen, [Bayesian Methods for Structural Dynamics and Civil Engineering](#), Wiley, 2010.
- [44] J. Beck, K. Yuen, [Model selection using response measurements: Bayesian probabilistic approach](#), *Journal of Engineering Mechanics* 41 (February) (2004) 192–203.
- [45] S. H. Cheung, J. Beck, [Calculation of Posterior Probabilities for Bayesian Model Class Assessment and Averaging from Posterior Samples Based on Dynamic System Data](#), *Computer-Aided Civil and Infrastructure Engineering* 25 (5) (2010) 304–321.

- [46] W. Gilks (Ed.), S. Richardson (Ed.), D. Spiegelhalter (Ed.), P. Van der Heijden, B. Morgan, N. Keiding, *Markov Chain Monte Carlo in Practice*, Chapman and Hall/CRC, 1996.
- [47] N. Friel, J. Wyse, *Estimating the evidence - a review*, *Statistica Neerlandica* 66 (3) (2012) 288–308.
- [48] J. Ching, Y. Chen, *Transitional Markov Chain Monte Carlo Method for Bayesian Model Updating, Model Class Selection, and Model Averaging*, *Journal of Engineering Mechanics* 133 (7) (2007) 816–832.
- [49] H. Haario, M. Laine, A. Mira, E. Saksman, *DRAM: efficient adaptive MCMC*, *Statistics and Computing* 16 (4) (2006) 339–354.
- [50] P. Green, *Reversible jump Markov chain Monte Carlo computation and Bayesian model determination*, *Biometrika* 82 (4) (1995) 711–732.
- [51] H. Haario, E. Saksman, *Adaptive proposal distribution for random walk Metropolis algorithm*, *Computational Statistics* 4 (1999) 1–32.
- [52] H. Walach, *Ockham's razor*, *Wiley Interdisciplinary Reviews Computational Statistics*, vol. 2 (01 2007).
- [53] D. J. C. MacKay, *Bayesian Interpolation*, *Neural Computation* 4 (3) (1992) 415–447.
- [54] M. Muto, J. Beck, *Bayesian Updating and Model Class Selection for Hysteretic Structural Models Using Stochastic Simulation*, *Journal of Vibration and Control* 14 (1-2) (2008) 7–34.
- [55] P. Bisailon, R. Sandhu, M. Khalil, C. Pettit, D. Poirel, A. Sarkar, *Bayesian parameter estimation and model selection for strongly nonlinear dynamical systems*, *Nonlinear Dynamics* 82 (3) (2015) 1061–1080.
- [56] C. P. Robert, G. Casella, *Monte Carlo Statistical Methods* (Springer Texts in Statistics), Springer-Verlag, Berlin, Heidelberg, 2005.
- [57] L. Tierney, *Markov chains for exploring posterior distributions*, *Ann. Statist.* 22 (4) (1994) 1701–1728.
- [58] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller, *Equation of state calculations by fast computing machines*, *The Journal of Chemical Physics* 21 (6) (1953) 1087–1092.
- [59] W. Hastings, *Monte carlo sampling methods using markov chains and their applications*, *Biometrika* 57 (1) (1970) 97–109.
- [60] S. Chib, E. Greenberg, *Understanding the Metropolis-Hastings Algorithm*, *American Statistician* 49 (4) (1995) 327–335.
- [61] A. Gelman, D. B. Rubin, *Inference from iterative simulation using multiple sequences*, *Statistical Science* 7 (4) (1992) 457–472.

- [62] G. O. Roberts, A. Gelman, W. R. Gilks, [Weak convergence and optimal scaling of random walk Metropolis algorithms](#), The Annals of Applied Probability 7 (1) (1997) 110–120.
- [63] A. Mira, On Metropolis-Hastings algorithms with delayed rejection, Metron 59 (2001) 231–241.
- [64] H. Haario, E. Saksman, J. Tamminen, [An adaptive metropolis algorithm](#), Bernoulli 7 (2) (2001) 223–242.
- [65] A. Solonen, P. Ollinaho, M. Laine, H. Haario, J. Tamminen, H. Järvinen, Efficient MCMC for Climate Model Parameter Estimation: Parallel Adaptive Chains and Early Rejection, Analysis 7 (2) (2012) 1–22.
- [66] N. Friel, A. N. Pettitt, Marginal likelihood estimation via power posteriors, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70 (3) (2008) 589–607.
- [67] S. Chib, I. Jeliazkov, Marginal likelihood from the metropolis-hastings output, Journal of the American Statistical Association 96 (453) (2001) 270–281.
- [68] S. Chib, Marginal likelihood from the gibbs output, Journal of the American Statistical Association 90 (432) (1995) 1313–1321.
- [69] J. Skilling, Nested sampling for general Bayesian computation, Bayesian Analysis 1 (4) (2006) 833–860.
- [70] R. Neal, Annealed importance sampling, Statistics and Computing 11 (2) (2001) 125–139.
- [71] M. A. Newton, A. E. Raftery, Approximate Bayesian Inference with the Weighted Likelihood Bootstrap, Journal of the Royal Statistical Society. Series B (Methodological) 56 (1) (1994) 3–48.
- [72] D. A. P. Qing Liu, A note on gauss-hermite quadrature, Biometrika 81 (3) (1994) 624–629.
- [73] W. Betz, I. Papaioannou, D. Straub, Transitional markov chain monte carlo: Observations and improvements, Journal of Engineering Mechanics 142 (5) (2016) 04016016.
- [74] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society. Series B (Methodological) 58 (1) (1996) 267–288.
- [75] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression, Annals of Statistics 32 (2004) 407–499.
- [76] T. Park, G. Casella, The bayesian lasso, Journal of the American Statistical Association 103 (482) (2008) 681–686.
- [77] C. Bergmeir, J. M. Benitez, [On the use of cross-validation for time series predictor evaluation](#), Information Sciences 191 (2012) 192 – 213.

- [78] M. E. Tipping, [Sparse bayesian learning and the relevance vector machine](#), Journal of Machine Learning Research 1 (2001) 211–244.
- [79] S. D. Babacan, R. Molina, A. K. Katsaggelos, Bayesian compressive sensing using laplace priors, IEEE Transactions on Image Processing 19 (1) (2010) 53–63.
- [80] Y. Li, C. Campbell, M. Tipping, Bayesian automatic relevance determination algorithms for classifying gene expression data, Bioinformatics 18 (10) (2002) 1332–1339.
- [81] K. Sargsyan, C. Safta, H. N. Najm, B. J. Debusschere, D. Ricciuto, P. Thornton, Dimensionality reduction for complex models via Bayesian Compressive Sensing, International Journal for Uncertainty Quantification 4 (1) (2014) 63–93.
- [82] C. Oh, J. Beck, M. Yamada, Bayesian learning using automatic relevance determination prior with an application to earthquake early warning, Journal of Engineering Mechanics 134 (12) (2008) 1013–1020.
- [83] M. Raissi, P. Perdikaris, G. Karniadakis, [Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations](#), Journal of Computational Physics 378 (2019) 686 – 707.
- [84] T. W. Anderson, [An introduction to multivariate statistical analysis](#), 3rd Edition, Hoboken, NJ John Wiley, 2003.
- [85] D. Poirel, Y. Harris, A. Benaissa, [Self-sustained aeroelastic oscillations of a NACA0012 airfoil at low-to-moderate Reynolds numbers](#), Journal of Fluids and Structures 24 (5) (2008) 700 – 719.
- [86] R. Sandhu, M. Khalil, A. Sarkar, D. Poirel, [Bayesian model selection for nonlinear aeroelastic systems using wind-tunnel data](#), Computer Methods in Applied Mechanics and Engineering 282 (2014) 161 – 183.
- [87] M. E. Tipping, A. C. Faul, Fast marginal likelihood maximization for sparse bayesian models, in: AISTATS, 2003.
- [88] M. E. Tipping, [Bayesian Inference: An Introduction to Principles and Practice in Machine Learning](#), Springer Berlin Heidelberg, Berlin, Heidelberg, 2004, pp. 41–62.
- [89] D. W. Scott, Multivariate Density Estimation: Theory, Practice, and Visualization, John Wiley & Sons, New York, Chichester, 1992.
- [90] E. T. Jaynes, [Information Theory and Statistical Mechanics](#), Physical Review 106 (4) (1957) 620–630.
- [91] P. Bisailon, C. Manohar, R. Sandhu, M. Khalil, C. Pettit, D. Poirel, A. Sarkar, Combined selection of physics-based models and error models using Bayesian inference, Submitted to Journal of Computational Physics.
- [92] W. R. Gilks, D. Spiegelhalter, S. Richardson, Markov chain Monte Carlo in practice: Interdisciplinary Statistics, 1st Edition, Chapman & Hall, London, 1996.

- [93] R. Craiu, J. Rosenthal, C. Yang, Learn From Thy Neighbor: Parallel-Chain and Regional Adaptive MCMC, *Journal of the American Statistical Association* 104 (2009) 1454–1466.
- [94] L. Tierney, A. Mira, *Some adaptive monte carlo methods for Bayesian inference.*, *Statistics in medicine* 18 (1999) 2507–15.
- [95] R. E. Kass, A. E. Raftery, *Bayes Factors*, *Journal of the American Statistical Association* 90 (430) (1995) pp. 773–795.
- [96] L. Tierney, J. B. Kadane, Accurate approximations for posterior moments and marginal densities, *Journal of the American Statistical Association* 81 (393) (1986) pp. 82–86.
- [97] C. Pettit, Uncertainty quantification in aeroelasticity: Recent results and research challenges, *Journal of Aircraft* 41 (5) (2004) 1217 – 1229.
- [98] M. Gad-El-Hak, *Control of low-speed airfoil aerodynamics*, *AIAA Journal* 28 (9) (1990) 1537–1552.
- [99] R. F. Huang, C. L. Lin, *Vortex shedding and shear-layer instability of wing at low-Reynolds numbers*, *AIAA Journal* 33 (8) (1995) 1398–1403.
- [100] R.-F. Huang, W. Shy, S. Lin, F.-B. Hsiao, Influence of surface flow on aerodynamic loads of a cantilever wing, *AIAA Journal* 34 (1996) 527–532.
- [101] J. Hall, *Low Reynolds number aerodynamics for micro aerial vehicles*, Ph.D. thesis, University of Colorado at Boulder (2007).
- [102] D. Poirel, W. Yuan, *Aerodynamics of laminar separation flutter at a transitional Reynolds number*, *Journal of Fluids and Structures* 26 (7-8) (2010) 1174–1194.
- [103] P. Beran, C. Pettit, D. Millman, *Uncertainty quantification of limit-cycle oscillations*, *Journal of Computational Physics* 217 (1) (2006) 217 – 247.
- [104] B. Danowsky, J. Chrstos, D. Klyde, C. Farhat, M. Brenner, Evaluation of aeroelastic uncertainty analysis methods, *Journal of Aircraft* 47 (4) (2010) 1266 – 1273.
- [105] R. Sandhu, Bayesian model selection and parameter estimation for a nonlinear fluid-structure interaction problem, Master’s thesis, Carleton University (2012).
- [106] Y. Fung, *An Introduction to the Theory of Aeroelasticity*, Dover Books on Aeronautical Engineering, Dover Publications, 2008.
- [107] H. Wagner, *Über die Entstehung des dynamischen Auftriebes von Tragflügeln*, *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik* 5 (1) (1925) 17–35.
- [108] T. Theodorsen, *General Theory of Aerodynamic Instability and the Mechanism of Flutter*, National Advisory Committee for Aeronautics: Report, NACA, 1935.

- [109] I. E. Garrick, **On Some Reciprocal Relations in the Theory of Nonstationary Flows**, National Advisory Committee for Aeronautics: Report, NACA, 1938.
- [110] E. H. Dowell, **A simple method for converting frequency-domain aerodynamics to the time domain**, NASA technical memorandum, National Aeronautics and Space Administration, Scientific and Technical Information Branch, 1980.
- [111] R. L. Bisplinghoff, H. Ashley, R. L. Halfman, *Aeroelasticity*, Courier Corporation, 2013.
- [112] R. T. Jones, **Operational Treatment of the Nonuniform-lift Theory in Airplane Dynamics**, Technical note, National Advisory Committee for Aeronautics, 1938.
- [113] W. N. P. Jones, **Aerodynamic Forces on Wings in Non-uniform Motion**, ARC technical report, H.M. Stationery Office, 1945.
- [114] C. Venkatesan, P. Friedmann, New approach to finite-state modeling of unsteady aerodynamics, *AIAA Journal* 24 (1986) 1889–1897.
- [115] R. Vepa, On the use of Pade approximants to represent unsteady aerodynamic loads for arbitrarily small motions of wings, in: *AIAA 14th Aerospace Sciences Meeting*, 1976.
- [116] E. H. Dowell, K. C. Hall, Modeling of fluid-structure interaction, *Annual Review of Fluid Mechanics* 33 (1) (2001) 445–490.
- [117] E. H. Dowell, M. Ilgamov, *Studies in Nonlinear Aeroelasticity*, Springer, 1988.
- [118] M. L. Facchinetti, E. De Langre, F. Biolley, Coupling of structure and wake oscillators in vortex-induced vibrations, *Journal of Fluids and Structures* 19 (2) (2004) 123–140.
- [119] R. D. Blevins, **Flow-Induced Vibration**, Krieger Pub Co, 2001.
- [120] P. Kloeden, E. Platen, **Numerical Solution of Stochastic Differential Equations**, Stochastic Modelling and Applied Probability, Springer Berlin Heidelberg, 2011.
- [121] M. E. Tipping, Sparse bayesian learning and the relevance vector machine, *J. Mach. Learn. Res.* 1 (2001) 211–244.
- [122] R. M. Neal, *Bayesian Learning for Neural Networks*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1996.
- [123] G. Casella, Illustrating empirical bayes methods, *Chemometrics and Intelligent Laboratory Systems* 16 (2) (1992) 107 – 125.
- [124] B. Carlin, T. Louis, Bayes and empirical bayes methods for data analysis, *Statistics and Computing* 7 (2) (1997) 153–154.
- [125] A. Wu, O. Koyejo, J. Pillow, **Dependent relevance determination for smooth and structured sparse regression**, *Journal of Machine Learning Research* 20 (89) (2019) 1–43.
- [126] R. E. Kass, L. Wasserman, **The Selection of Prior Distributions by Formal Rules**, *Journal of the American Statistical Association* 91 (435) (1996) pp. 1343–1370.

- [127] C. Sanderson, Armadillo: An open source C++ linear algebra library for fast prototyping and computationally intensive experiments, Technical Report NICTA (2010) 1–16.
- [128] M. Passing, I. Forum, MPI : A Message-Passing Interface Standard, Forum American Bar Association 8 (2009) 647.
- [129] C. Loken, D. Gruner, L. Groer, R. Peltier, N. Bunn, M. Craig, T. Henriques, J. Dempsey, C.-H. Yu, J. Chen, L. J. Dursi, J. Chong, S. Northrup, J. Pinto, N. Knecht, R. V. Zon, **Scinet: Lessons learned from building a power-efficient top-20 system and data centre**, Journal of Physics: Conference Series 256 (1) (2010) 012026.
- [130] Y. Bar-Shalom, X.-R. Li, Estimation with Applications to Tracking and Navigation, John Wiley & Sons, Inc., USA, 2001.
- [131] D. Reynolds, **Gaussian Mixture Models**, Springer US, Boston, MA, 2015, pp. 827–832.
- [132] Y. M. Marzouk, H. N. Najm, L. A. Rahn, **Stochastic spectral methods for efficient Bayesian solution of inverse problems**, Journal of Computational Physics 224 (2) (2007) 560 – 586.
- [133] Y. M. Marzouk, H. N. Najm, **Dimensionality reduction and polynomial chaos acceleration of Bayesian inference in inverse problems**, Journal of Computational Physics 228 (6) (2009) 1862 – 1902.
- [134] T. Cui, Y. M. Marzouk, K. E. Willcox, **Data-driven model reduction for the Bayesian solution of inverse problems**, International Journal for Numerical Methods in Engineering 102 (5) (2015) 966–990.
- [135] D. Galbally, K. Fidkowski, K. Willcox, O. Ghattas, **Non-linear model reduction for uncertainty quantification in large-scale inverse problems**, International Journal for Numerical Methods in Engineering 81 (12) (2010) 1581–1608.
- [136] A. C. Faul, M. E. Tipping, Analysis of sparse bayesian learning, in: Advances in Neural Information Processing Systems 14, MIT Press, 2001, pp. 383–389.
- [137] K. B. Petersen, M. S. Pedersen, **The matrix cookbook**, version 20121115 (nov 2012). URL <http://www2.imm.dtu.dk/pubdb/p.php?3274>
- [138] B. W. Silverman, Density Estimation for Statistics and Data Analysis, Chapman & Hall, London, 1986.
- [139] C. L. Pettit, D. K. Wilson, **Variational inference of cluster-weighted models for local and global sensitivity analysis**, International Journal of Reliability and Safety 8 (2-4) (2014) 196–227.
- [140] J. Nocedal, S. J. Wright, Numerical Optimization, 2nd Edition, Springer, New York, NY, USA, 2006.
- [141] G. Strang, **Linear algebra and its applications**, Thomson, Brooks/Cole, Belmont, CA, 2006.

- [142] B. N. Datta, Numerical Linear Algebra and Applications, Second Edition, 2nd Edition, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2010.
- [143] A. R. Conn, N. I. M. Gould, Ph. L. Toint, Trust-Region Methods, SIAM, Philadelphia, PA, USA, 2000.
- [144] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. Jarrod Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. Carey, Í. Polat, Y. Feng, E. W. Moore, J. Vand erPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, S. . . Contributors, SciPy 1.0—Fundamental Algorithms for Scientific Computing in Python, arXiv e-prints (2019) arXiv:1907.10121.
- [145] N. Wiener, **The Homogeneous Chaos**, American Journal of Mathematics 60 (4) (1938) 897–936.
- [146] R. G. Ghanem, P. D. Spanos, Stochastic Finite Elements: A Spectral Approach, Springer-Verlag, Berlin, Heidelberg, 1991.
- [147] B. Sudret, Global sensitivity analysis using polynomial chaos expansions, Rel. Eng. & Sys. Safety 93 (7) (2008) 964–979.
- [148] D. Xiu, G. E. Karniadakis, **The wiener–askey polynomial chaos for stochastic differential equations**, SIAM J. Sci. Comput. 24 (2) (2002) 619–644.
- [149] G. Blatman, B. Sudret, Adaptive sparse polynomial chaos expansion based on least angle regression, J. Comput. Physics 230 (6) (2011) 2345–2367.
- [150] M. Raisee, D. Kumar, C. Lacor, **A non-intrusive model reduction approach for polynomial chaos expansion using proper orthogonal decomposition**, International Journal for Numerical Methods in Engineering 103 (4) (2015) 293–312.
- [151] K. Cheng, Z. Lu, **Sparse polynomial chaos expansion based on d-morph regression**, Applied Mathematics and Computation 323 (2018) 17 – 30.
- [152] L. Le Gratiet, S. Marelli, B. Sudret, Metamodel-Based Sensitivity Analysis: Polynomial Chaos Expansions and Gaussian Processes, Springer International Publishing, Cham, 2017, pp. 1289–1325.
- [153] Q. Shao, A. Younes, M. Fahs, T. A. Mara, **Bayesian sparse polynomial chaos expansion for global sensitivity analysis**, Computer Methods in Applied Mechanics and Engineering 318 (2017) 474 – 496.
- [154] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, S. Tarantola, **Global Sensitivity Analysis: The Primer**, Wiley, 2008.
- [155] A. Matachniouk, Parametric uncertainty quantification in coalescence flutter, Master’s thesis, Carleton University (2013).

- [156] L. J. Rocha Da Costa, Nonlinear aeroelastic modeling of a flexible wing and comparison with experiments, Master's thesis, Carleton University (2017).
- [157] B. M. G. Robinson, Aeroelastic Oscillations of a Pitching Cantilever Wing with Structural Geometric Nonlinearities: Theory, Numerical Simulation and Global Sensitivity Analysis, Master's thesis, Carleton University (2018).
- [158] E. T. Jaynes, **Information theory and statistical mechanics**, Phys. Rev. 106 (1957) 620–630.
- [159] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, Journal of the Royal Statistical Society, Series B 67 (2005) 301–320.
- [160] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Rubin, Bayesian Data Analysis, 2nd Edition, Chapman and Hall/CRC, 2004.
- [161] D. Fink, A compendium of conjugate priors, Technical Report. (01 1997).
- [162] R. Vrabel, A note on the matrix determinant lemma, International Journal of Pure and Applied Mathematics 111 (2016) 643–646.
- [163] S. Ji, Y. Xue, L. Carin, Bayesian compressive sensing, IEEE Transactions on Signal Processing 56 (6) (2008) 2346–2356.
- [164] C. C. Liu, M. Aitkin, **Bayes factors: Prior sensitivity and model generalizability**, Journal of Mathematical Psychology 52 (6) (2008) 362 – 375.
- [165] G. O. Roberts, J. S. Rosenthal, **Optimal Scaling for Various Metropolis-Hastings Algorithms**, Statistical Science 16 (4) (2001) pp. 351–367.
- [166] A. Gelman, G. Roberts, W. Gilks, Efficient Metropolis jumping rules, in: Bayesian statistics, 5 (Alicante, 1994), Oxford Sci. Publ., Oxford Univ. Press, New York, 1996, pp. 599–607.
- [167] T. Chan, G. Golub, R. LeVeque, S. U. C. D. O. C. SCIENCE., Updating formulae and a pairwise algorithm for computing sample variances, Annals of Physics 54 (1979) 258.
- [168] S. P. Brooks, A. Gelman, General methods for monitoring convergence of iterative simulations, Journal of computational and graphical statistics 7 (1998) 434–455.
- [169] D. C. M. Poirel, Random dynamics of a structurally nonlinear airfoil in turbulent flow, Ph.D. thesis, McGill University (2001).