BAYESIAN UNCERTAINTY QUANTIFICATION OF COMPUTER MODELS WITH EFFICIENT CALIBRATION AND COMPUTATION

By

Vojtech Kejzlar

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Statistics – Doctor of Philosophy

2020

ABSTRACT

BAYESIAN UNCERTAINTY QUANTIFICATION OF COMPUTER MODELS WITH EFFICIENT CALIBRATION AND COMPUTATION

By

Vojtech Kejzlar

The use of mathematical models, typically implemented in the form of computer code, proliferates to solve complex problems in many scientific applications such as nuclear physics and climate research. The computational and statistical tools of Uncertainty Quantification (UQ) are instrumental in assessing how accurately a computer model describes a physical process. Bayesian framework for UQ has become the dominant approach, because it provides a principled way of quantifying uncertainty in the language of probabilities. The ever-growing access to high performance computing in scientific communities has meanwhile created the need to develop next-generation tools and theory for analysis of computer models. Motivated by practical research problems, this dissertations proposes novel computational tools and UQ methodology aimed to enhance the quality of computer models which leads to improved predictive capability and a more "honest" UQ.

First, we consider *model* uncertainty, which arises in situations when several competing models are available to describe the same or a similar physical phenomenon. One of the historically dominant methods to account for this source of uncertainty is Bayesian Model Averaging (BMA). We perform systematic analysis of prediction errors and show the use of BMA posterior mean predictor leads to mean squared error reduction. In a response to a recurrent research scenario in nuclear physics, BMA is extended to a situation where models are defined on non-identical study regions. We illustrate our methodology via pedagogical simulations and applications of forecasting nuclear observables, which exhibit improvements in both prediction error and empirical coverage probabilities.

In the second part of this dissertation, we concentrate on individual computer models with particular focus on those which are computationally too expensive to be used directly

for predictions. Furthermore, we consider computer models that need to be calibrated with experimental observations, because they depend on inputs whose values are generally unknown. We develop an efficient algorithm based on variational Bayes inference (VBI) for the calibration of computer models with Gaussian processes (GPs). To preserve the efficiency of VBI in the presence of dependent data, we adopt the pairwise decomposition of the data likelihood using vine copulas that separate the information on dependence structure in data from their marginal distribution. We provide both theoretical and empirical evidence for the computational scalability of our algorithm and demonstrate the opportunities given by our method on a real-data example through calibration of the Liquid Drop Model of nuclear binding energies.

As a fast and easy-to-implement alternative to the fully Bayesian treatment (such as the VBI approach), we propose an empirical Bayes approach to computer-enabled predictions of physical quantities. We offer a new perspective to the Bayesian calibration framework with GPs and provide its representation as a Bayesian hierarchical model. Consequently, a posterior consistency of the physical process is established, assuming certain smoothness properties of the GP priors and the existence of a strongly consistent estimator of a noise scale. A simulation study and a real-data example that support the consistency and efficiency of the empirical Bayes method are provided as well.

To Strýček for showing me it was possible.

ACKNOWLEDGEMENTS

I would like to take this opportunity to sincerely thank to my advisors, Prof. Tapabrata Maiti and Prof. Frederi Viens, for their support, guidance, and encouragement. Prof. Maiti's exemplary commitment to research will continue to serve as a guide in my own academic pursuits. Thank you, Prof. Viens, for consistently providing me with resources and opportunities over the past years.

My committee member Prof. Witold Nazarewicz who readily welcomed me to the world of physics and allowed me to participate in exciting interdisciplinary projects. I would also like to thank Prof. R.V. Ramamoorthi for being a part of my committee, for our discussions, and for his insight.

I would like to extend many thanks to my research partners, Dr. Léo Neufcourt, Dr. Shrijita Bhattacharya, Dr. Stefan Wild, Dr. P.-G. Reinhard, and Mookyong Son.

Lastly, I am grateful to my family. Thanks to my parents for their unequivocal support during all my studies and especially for their foresight in enrolling me in computer classes in elementary school despite my protests. Thanks to my brother, Jakub, for his willingness to meet up with me anywhere in the world, and thanks to my wife, May, for believing in me from the start and always cheering me on.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ALGORITHMS

KEY TO ABBREVIATIONS

BMA Bayesian model averaging

CM Computer model

CDF Cumulative distribution function

DFT Density functional theory

ECP Empirical coverage probability

EB Empirical Bayes

EDF Energy density functional

HPD Highest posterior density

GP Gaussian process

LDM Liquid Drop Model

MC Monte Carlo

MCMC Markov chain Monte Carlo

MH Metropolis-Hastings

MSE Mean square error

NUTS No-U-Turn sampler

PMSE Posterior mean square error

UQ Uncertainty quantification

RMSE Root mean square error

SGA Stochastic gradient ascend

SVI Stochastic variational inference

VI Variational inference

VBI Variational Bayes inference

CHAPTER 1

INTRODUCTION

With the advancements of computer architectures in the 21^{st} century, mathematical models implemented on a computer, which we shall refer to as *computer models*, have become the driving force behind the acceleration of the cycle of the scientific process. This is because computer models are typically much faster, safer, and economical to run than physical experiments. For example, experiments in high energy physics are conducted in particle colliders that cost billions of dollars and can take up to a decade to build. Moreover, some physical experiments associated with rare natural events such as volcanic eruptions or earthquakes are infeasible to carry out for all practical purposes.

Computer models, despite being an extremely useful tool (Box, 1976), are an imperfect representation of physical systems. The comprehensive study of the impact of all forms of modeling errors is called *uncertainty quantification* (UQ). Bayesian methodology of UQ, which is the main approach considered in this work, has been a heavily utilized statistical device due to its natural way to describe uncertainty in the language of probabilities; see Higdon et al. (2015); McDonnell et al. (2015), and King et al. (2019) for examples in nuclear physics, Sexton et al. (2012) and Pollard et al. (2016) for examples in climatology, and Williams et al. (2006); Lawrence et al. (2010), Plumlee et al. (2016) and Zhang et al. (2019) for applications in engineering, astrophysics, and medicine.

Meanwhile, the incoming era of exascale computing (systems capable of 10^{18} double precision floating point operations per second) has spawned the development of complex computer models that produce massive amounts of data. This consequently creates the need to bring the computational and statistical tools of UQ into the big-data age.

1.1 Computer models and sources of uncertainty

To illustrate various sources of uncertainties in computer models on a simple example, let us consider the 4-parameter Liquid Drop Model (LDM) (Weizsäcker, 1935; Bethe and Bacher, 1936; Myers and Swiatecki, 1966; Kirson, 2008; Benzaid et al., 2020) which is a global (across the whole nuclear chart) model of nuclear binding energies; the minimum energy needed to disassemble the nucleus of an atom into unbound protons and neutrons.

In principle, the LDM treats the nucleus like a drop of incompressible fluid of very high density. Despite this simplification, the LDM makes reasonable estimates of average properties of nuclei. The LDM is formulated through the semi-empirical mass formula as:

$$E_{\mathrm{B}}(N,Z) = a_{\mathrm{vol}}A - a_{surf}A^{2/3} - a_{sym}\frac{(N-Z)^2}{A} - a_{\mathrm{C}}\frac{Z(Z-1)}{A^{1/3}}, \tag{1.1}$$

where Z is the proton number, N is the neutron number, and $A = Z + N$ is the mass number of the nucleus. The model parameters are $(a_{\mathrm{vol}}, a_{surf}, a_{sym}, a_{\mathrm{C}})$ representing the volume, surface, symmetry and Coulomb energy, respectively. These parameters have specific physical meaning, where a_{vol} is for instance proportional to the volume of the nucleus. See Krane (1987) for more details. We can identify the following sources of uncertainty as proposed by Kennedy and O'Hagan (2001).

Parameter uncertainty: The model is a function of fixed but unknown parameters $(a_{\mathrm{vol}}, a_{surf}, a_{sym}, a_{\mathrm{C}})$. These parameters are context specific and need to be estimated with reported standard errors. The process of model fitting is also known as *calibration*.

Observation error: In estimating the unknown model parameters, we will be making use of experimental data from the actual physical process. These measurements typically contain some observation error that should be accounted for.

Model inadequacy: As we already mentioned at the beginning of this chapter, computer models are an imperfect representation of physical systems. Even if we know the true values

the model parameters, the LDM predictions will not equal the true values of the physical process. This uncertainty (error) is often interpreted as "missing physics" in the model and is differentiated from the observation error by its systematic nature.

Parametric uncertainty: Note that the LDM is a linear function of the parameter vector $(a_{\text{vol}}, a_{surf}, a_{sym}, a_{\text{C}})$. It is possible that one or more predictor variables are highly linearly correlated (multicollinearity) and the LDM can be reduced to a model with less parameters.

Model uncertainty: The LDM is not the only model of nuclear binding energies. In fact, there are many alternative and competing models. In order to conduct comprehensive UQ of modeling framework, we should allow for this possibility.

The subsequent two sections describe Bayesian formalisms that provide statistically principled ways to account for the various sources of uncertainty described above with exception of the parametric uncertainty. The parametric uncertainty is not the focus of this dissertation, and we refer the reader to Jaganathen et al. (2017) and Kejzlar et al. (2020) for some examples in nuclear physics.

1.2 Bayesian calibration of imperfect computer models

Let us consider observations $\boldsymbol{y} = (y_1, \ldots, y_n)$ of a physical process $\zeta(\boldsymbol{t}_i)$ depending on a known set of inputs $\boldsymbol{t}_i \in \boldsymbol{\Omega} \subset \mathbb{R}^p$ following the relationship

$$y_i = \zeta(\boldsymbol{t}_i) + \sigma\epsilon_i, \quad i = 1, \ldots, n, \tag{1.2}$$

where σ represent the scale of observation error (noise), typically $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$. Our aim is to establish statistically principled predictions of new values $\boldsymbol{y}^* = (y_1^*, \ldots, y_J^*)$ of the physical process ζ at, yet to be observed, inputs $(\boldsymbol{t}_1^*, \ldots, \boldsymbol{t}_J^*)$ using $\boldsymbol{y}$ and a computer model f_m defined as a mapping $(\boldsymbol{t}, \boldsymbol{\theta}) \mapsto f_m(\boldsymbol{t}, \boldsymbol{\theta})$. As we can see, the computer model depends on an additional set of inputs $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$ that we call calibration parameters. These are considered fixed but unknown quantities common to all the observations y_i and all the

instances of the physical process that we intend to predict using the calibrated computer model. The calibration parameters represent inherent properties of the physical process that cannot be directly measured or controlled in an experiment. In the most rudimentary form, one can think of the calibration parameters as parameters in standard regression problems. To this extent, we suppose the relationship between the observations y_i, the physical process ζ, and the computer model f_m as proposed by Kennedy and O'Hagan (2001)

$$y_i = f_m(\boldsymbol{t}_i, \boldsymbol{\theta}) + \delta(\boldsymbol{t}_i) + \sigma\epsilon_i, \tag{1.3}$$

where $\delta(\boldsymbol{t}_i)$ represents an unknown systematic error between the computer model and the physical process. While $\delta(\boldsymbol{t}_i)$ is intrinsically deterministic, a non-parametric approach using Gaussian process prior model is typically imposed for Bayesian inference.

Definition 1. *$\delta(\boldsymbol{t})$ has a Gaussian process distribution if for every $i = 1, 2, 3\ldots$ the joint distribution of $\delta(\boldsymbol{t}_1), \ldots \delta(\boldsymbol{t}_i)$ is multivariate normal. It is fully characterized by mean function $m(\boldsymbol{t}) = \mathbb{E}[\delta(\boldsymbol{t})]$ and covariance function $k(\boldsymbol{t}, \boldsymbol{t}') = \mathbb{C}ov[\delta(\boldsymbol{t}), \delta(\boldsymbol{t}')]$. We write*

$$\delta(\boldsymbol{t}) \sim \mathcal{GP}(m_\delta(\boldsymbol{t}), k_\delta(\boldsymbol{t}, \boldsymbol{t}')).$$

Gaussian processes are a convenient way of placing a distribution over a space of functions with the covariance function characterizing the relationship of the process at different inputs. Typically, the mean function is chosen to be zero or some dense family of basis functions (wavelets, Fourier, polynomials) across the input domain:

$$m(\cdot) = \boldsymbol{h}(\cdot)^T\boldsymbol{\beta},$$

where $\boldsymbol{h}(\cdot) = (h_1(\cdot), \ldots h_p(\cdot))$ are the basis functions and $\boldsymbol{\beta}$ is a *hyperparameter*. A typical choice for the covariance function is a stationary covariance function that depends on the inputs through $\boldsymbol{t} - \boldsymbol{t}'$. For example, a Gaussian kernel covariance function (also called squared exponential or radial basis function kernel) takes the form

$$k(\boldsymbol{t}, \boldsymbol{t}') = \eta^2 \exp\bigg(-\frac{1}{2}(\boldsymbol{t} - \boldsymbol{t}')^T M(\boldsymbol{t} - \boldsymbol{t}')\bigg),$$

where M corresponds to a positive definite diagonal matrix of hyperparameters and η is a scaling parameter. We refer to the case of $M = \frac{1}{\ell^2} I$, for some $\ell > 0$, as an *isotropic* version of the kernel, because it is invariant to the rotation. The case of M with different diagonal terms is called an *anisotropic* version of the kernel. Other popular choices for stationary covariance functions are Matérn kernels, polynomial kernels, or exponential kernels (Rasmussen and Williams, 2006). See Figure 1.1 to visualize realizations of a Gaussian process.

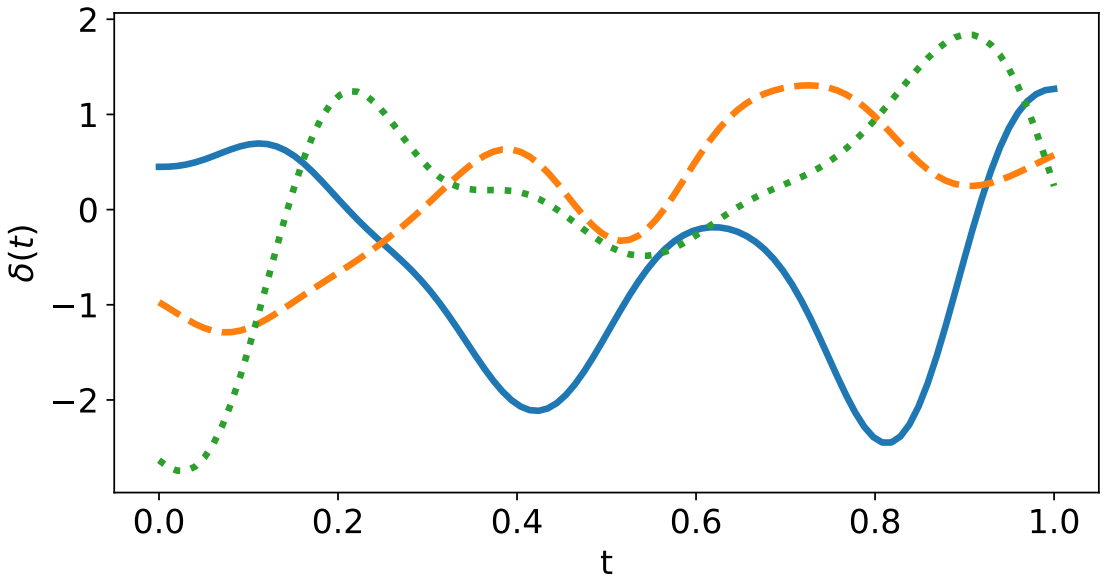

Figure 1.1: Realizations of a Gaussian process with zero mean and squared exponential covariance function with $\eta = 1$ and $\ell = 0.1$.

It is often the case the evaluation of the computer model f_m is too expensive in terms of both time and space (memory). It is common practice to reduce the number of necessary computer model evaluations by considering a Gaussian process prior model

$$f_m(\boldsymbol{t}, \boldsymbol{\theta}) \sim \mathcal{GP}(m_f(\boldsymbol{t}, \boldsymbol{\theta}), k_f((\boldsymbol{t}, \boldsymbol{\theta}), (\boldsymbol{t}', \boldsymbol{\theta}'))).$$

In this setup, the data also include a set of model evaluations $\boldsymbol{z} = (z_1, \ldots, z_s)$ over a grid $\{(\widetilde{\boldsymbol{t}}_1, \widetilde{\boldsymbol{\theta}}_1), \ldots, (\widetilde{\boldsymbol{t}}_s, \widetilde{\boldsymbol{\theta}}_s)\}$. These are usually selected using some space-filling design such as a uniform or Latin hypercube design (Morris and Mitchell, 1995), which is a design that has a good coverage of the space with evenly distributed points in each one-dimensional

projection. The complete data set $\boldsymbol{d}$ in the case of computationally expensive models consists of n observations y_i from the physical process ζ and s evaluations z_j of the computer model f_m, i.e. $\boldsymbol{d} = (d_1, \ldots, d_{n+s}) := (\boldsymbol{y}, \boldsymbol{z})$. We shall denote the set of unknown parameters as $\boldsymbol{\phi} = (\boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma)$ with $\boldsymbol{\gamma}$ denoting the set of hyperparameters of Gaussian processes' mean and covariance functions. Consequently, the complete dataset $\boldsymbol{d}$ conditioned on $(\boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma)$ follows the multivariate normal distribution

$$\boldsymbol{d}|\boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma \sim \mathcal{N}(M(\boldsymbol{\theta}, \boldsymbol{\gamma}), K(\boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma)), \tag{1.4}$$

where

$$M(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \begin{pmatrix} M_f(T_y(\boldsymbol{\theta})) + M_\delta(T_y) \\ M_f(T_z(\widetilde{\boldsymbol{\theta}})) \end{pmatrix} \tag{1.5}$$

and

$$K(\boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma) = \begin{pmatrix} K_f(T_y(\boldsymbol{\theta}), T_y(\boldsymbol{\theta})) + K_\delta(T_y, T_y) + \sigma^2 I_n & K_f(T_y(\boldsymbol{\theta}), T_z(\widetilde{\boldsymbol{\theta}})) \\ K_f(T_z(\widetilde{\boldsymbol{\theta}}), T_y(\boldsymbol{\theta})) & K_f(T_z(\widetilde{\boldsymbol{\theta}}), T_z(\widetilde{\boldsymbol{\theta}})) \end{pmatrix} \tag{1.6}$$

Here, $K_f(T_y(\boldsymbol{\theta}), T_y(\boldsymbol{\theta}))$ is the matrix with (i, j) element $k_f((\boldsymbol{t}_i, \boldsymbol{\theta}), (\boldsymbol{t}_j, \boldsymbol{\theta}))$, $K_\delta(T_y, T_y)$ is the matrix with (i, j) element $k_\delta(\boldsymbol{t}_i, \boldsymbol{t}_j)$, and $K_f(T_z(\widetilde{\boldsymbol{\theta}}), T_z(\widetilde{\boldsymbol{\theta}}))$ is the matrix with (i, j) element $k_f((\widetilde{\boldsymbol{t}}_i, \widetilde{\boldsymbol{\theta}}_i), (\widetilde{\boldsymbol{t}}_j, \widetilde{\boldsymbol{\theta}}_j))$. We can similarly define $K_f(T_y(\boldsymbol{\theta}), T_z(\widetilde{\boldsymbol{\theta}}))$ with the kernel k_f.

Under this framework, the Bayesian calibration consists of deriving the full posterior distribution $p(\boldsymbol{\phi}|\boldsymbol{d})$ given by the Bayes' theorem, namely

$$p(\boldsymbol{\phi}|\boldsymbol{d}) = \frac{p(\boldsymbol{d}|\boldsymbol{\phi})p(\boldsymbol{\phi})}{\int p(\boldsymbol{d}|\boldsymbol{\phi})p(\boldsymbol{\phi})\,\mathrm{d}\boldsymbol{\phi}} \propto p(\boldsymbol{d}|\boldsymbol{\phi})p(\boldsymbol{\phi}), \tag{1.7}$$

where $p(\boldsymbol{\phi})$ expresses our prior uncertainty about the unknown parameters. The Bayesian predictions of $\boldsymbol{y}^*$ are specified by the posterior predictive distribution $p(\boldsymbol{y}^*|\boldsymbol{d})$. This is given by integrating the conditional density of $\boldsymbol{y}^*$, given $\boldsymbol{\phi}$ and the data $\boldsymbol{d}$, against the posterior density $p(\boldsymbol{\phi}|\boldsymbol{d})$:

$$p(\boldsymbol{y}^*|\boldsymbol{d}) = \int p(\boldsymbol{y}^*|\boldsymbol{d}, \boldsymbol{\phi})p(\boldsymbol{\phi}|\boldsymbol{d})\,\mathrm{d}\boldsymbol{\phi}. \tag{1.8}$$

The conditional density $p(\boldsymbol{y}^*|\boldsymbol{d}, \boldsymbol{\phi})$ is a multivariate normal density given directly by the statistical model (1.3) and the specification of the Gaussian processes. We postpone the detailed description of this likelihood to Chapter 3.

Here we point out a few caveats of the framework described above. First, the calibration parameter $\boldsymbol{\theta}$ is in general non-identifiable. Indeed, $\delta(\boldsymbol{t}_i) = \zeta(\boldsymbol{t}_i) - f_m(\boldsymbol{t}_i, \boldsymbol{\theta})$ yields the same distribution for y_i for any choice of $\boldsymbol{\theta}$. Several authors have pointed this out and proposed various methods to mitigate the problem including Brynjarsdóttir and O'Hagan (2014); Plumlee (2017); Tuo and Wu (2015, 2016); Bayarri et al. (2007). Our main goal here, nonetheless, is not the correct identification of $\boldsymbol{\theta}$, but a prediction. Second, the posterior distribution $p(\boldsymbol{\phi}|\boldsymbol{d})$ does not have a closed form and needs to be approximated. The traditionally used Markov Chain Monte Carlo (MCMC) methods that approximate $p(\boldsymbol{\phi}|\boldsymbol{d})$—such as the Metropolis-Hastings (MH) algorithm (Chib and Greenberg, 1995) or more advanced ones including the Hamiltonian Monte Carlo or the No-U-Turn samplers (NUTS) (Homan and Gelman, 2014)—work only with a relatively small sample size because of the computational costs associated with the evaluation of $p(\boldsymbol{d}|\boldsymbol{\phi})$. This clearly calls for the development of computationally efficient alternatives to the traditional approaches.

1.3 Bayesian model averaging

Bayesian model averaging (BMA) is the natural Bayesian framework in scenarios with several competing models $\mathcal{M}_1, \ldots, \mathcal{M}_K$ when one is not comfortable selecting a single model at the desired level of certainty (Bernardo and Smith, 1994; Kass and Raftery, 1995; Hoeting et al., 1999; Wasserman, 2000). The seminal review work by Geweke (1999) introduced BMA in econometrics and later in other fields such as political and social sciences; BMA has also been applied to the medical sciences (Balasubramanian et al., 2014; Schorning et al., 2016), ecology and evolution (Silvestro et al., 2014; Hooten and Hobbs, 2015), genetics (Wei et al., 2011; Wen, 2015), astrophysics (Parkinson and Liddle, 2013), fluid dynamics (Radaideh et al., 2019), machine learning (Clyde et al., 2011; Hernández et al., 2018), and lately in

nuclear physics (Neufcourt et al., 2019, 2020a,b; Kejzlar et al., 2020).

For any quantity of interest $\mathcal{O}$, e.g., the value y^*, the BMA posterior density $p(\mathcal{O}|\boldsymbol{d})$ corresponds to the mixture of the posterior densities of the individual models:

$$p(\mathcal{O}|\boldsymbol{d}) = \sum_{k=1}^{K} p(\mathcal{O}|\boldsymbol{d}, \mathcal{M}_k) p(\mathcal{M}_k|\boldsymbol{d}), \tag{1.9}$$

where $\boldsymbol{d}$ are given datapoints. These datapoints are typically observations $\boldsymbol{y}$ unless we consider the specific scenario of computationally expensive computer models in Section 1.2, where we also include the set of model runs $\boldsymbol{z}$. For notation consistency, we shall denote the set of datapoints as $\boldsymbol{d}$ throughout this dissertation with the actual content of $\boldsymbol{d}$ clarified by the context in which it is considered. Formula (1.9) expresses the actual posterior probability of a quantity of interest $\mathcal{O}$ is the average of $\mathcal{O}$'s posterior distributions given each model, weighted by the model posterior probabilities. In other words, (1.9) is simply a mixture of K distributions, which makes sampling from the BMA posterior density immediate once we obtain the posterior samples under each model. The posterior model weights $p(\mathcal{M}_k|\boldsymbol{d})$ are the posterior probabilities that a given model is the hypothetical *true* model; it is given by a simple application of the Bayes' theorem:

$$p(\mathcal{M}_k|\boldsymbol{d}) = \frac{p(\boldsymbol{d}|\mathcal{M}_k)p(\mathcal{M}_k)}{\sum_{\ell=1}^{K} p(\boldsymbol{d}|\mathcal{M}_\ell)p(\mathcal{M}_\ell)}, \tag{1.10}$$

where $p(\mathcal{M}_k)$ represents the prior probability that $\mathcal{M}_k$ is the true model. The so called evidence (integrals) $p(\boldsymbol{d}|\mathcal{M}_k)$ are obtained by integrating the data likelihood against the prior density of the model parameters $\boldsymbol{\phi}_k$, namely

$$p(\boldsymbol{d}|\mathcal{M}_k) = \int p(\boldsymbol{d}|\boldsymbol{\phi}_k, \mathcal{M}_k) p(\boldsymbol{\phi}_k|\mathcal{M}_k) \, \mathrm{d}\boldsymbol{\phi}_k. \tag{1.11}$$

Additionally, the definition of expected value yields the posterior mean of $\mathcal{O}$ as

$$\mathbb{E}[\mathcal{O}|\boldsymbol{d}] = \sum_{k=1}^{K} \mathbb{E}[\mathcal{O}|\boldsymbol{d}, \mathcal{M}_k] p(\mathcal{M}_k|\boldsymbol{d}), \tag{1.12}$$

and the well-known conditional variance formula (Casella and Berger, 2002) yields the posterior variance of $\mathcal{O}$, given $\boldsymbol{d}$, as

$$\mathbb{V}ar[\mathcal{O}|\boldsymbol{d}] \;=\; \sum_{k=1}^{K} p(\mathcal{M}_k|\boldsymbol{d})\mathbb{V}ar[\Delta|\boldsymbol{d},\mathcal{M}_k] + \mathbb{V}ar[\mathbb{E}(\Delta|\boldsymbol{d},M)|\boldsymbol{d}]. \tag{1.13}$$

Note that the term $\mathbb{V}ar[\mathbb{E}(\mathcal{O}|\boldsymbol{d},M)|\boldsymbol{d}]$ is the variance of a function of the discrete random variable M (the set of all models being considered), which accounts for the model uncertainty. This model uncertainty is not accounted for by individual models. Its inclusion thus allows for a more honest UQ.

One of the challenges of BMA is that it becomes unclear how one should proceed in scenarios where alternative models are defined on different subsets of the same input space. This is, for example, a usual situation in nuclear physics, for instance for nuclear mass models; ab initio (also known as A-body) models range over lighter nuclei due to contemporary computational limitations, while Energy Density Functionals (EDF) can cover the whole nuclear chart (Klupfel et al., 2009; Kortelainen et al., 2010b).

1.4 Dissertation outline

The main content of this dissertation is organized as follows. Chapter 2 provides a survey of the remaining details for successful implementation of the BMA framework with particular focus on the calculation of the evidence integral (1.11). We perform a systematic analysis of the prediction errors, focusing on the fact that BMA is the optimal linear combination (projection) in the L^2 sense under the posterior probability distribution, among all the possible mixtures of models. Motivated by recurrent scenarios in nuclear physics, we subsequently extend BMA to the situations when the different models constrain different subsets of the data. Lastly, we present a set of pedagogical examples as well as real-data applications of the BMA methodology highlighting its benefits in terms of the improvement of the prediction accuracy and UQ. Some results from this chapter are also provided in Kejzlar et al. (2019).

Chapter 3 presents a novel and computationally efficient algorithm based on variational Bayes inference (VBI) for the calibration of computer models with Gaussian processes. We

provide both theoretical and empirical evidence for the computational scalability of our methodology and describe all the necessary details for an efficient implementation of the proposed algorithm. We demonstrate the opportunities given by our method for practitioners on a real data example through the calibration of the Liquid Drop Model of nuclear binding energies. The algorithmic development done in this chapter is also provided in Kejzlar and Maiti (2020).

Chapter 4 develops an empirical Bayes (EB) approach for the Bayesian calibration framework outlined in Section 1.2 that can be understood as an easy-to-implement and fast approximation of the fully Bayesian treatment. Firstly, we utilize the structural convenience of Gaussian processes and restate the calibration framework as a Bayesian hierarchical model. Secondly, we make use of this new representation and extend the results of Choi and Schervish (2007a) on non-parametric regression problems to theoretically investigate the proposed EB approach. A numerical simulation study and a real data example are also provided.

In Chapter 5, we discuss the likely future theoretical and computational extensions of the methodologies developed in Chapters 2 – 4.

For ease of readability, all proofs of lemmas, propositions, and theorems, altogether with technical details of numerical studies, are provided in the section titled "Technical details and supplementary results" at the end of respective chapter. We also provide fully documented Python code that reproduces all the results in this dissertation and can be easily modified and used by practitioners in a public repository at `https://github.com/kejzlarv`.

CHAPTER 2

SURVEY OF BAYESIAN MODEL AVERAGING WITH EXAMPLES AND EXTENSION TO DISCREPANT DOMAINS

Interest for model averaging arises, as discussed in Chapter 1, in situations when several competing models are available to solve the same or similar problem, and no single model can be selected at a desired level of certainty. For example, there is a multitude of competing computer models for numerical weather prediction including the American model (Global Forecast System) and the European model (European Centre for Medium-Range Weather Forecasts) (Lynch, 2008). In nuclear physics, alternative models arise through different theoretical strategies in modeling atomic nuclei such as the A-body modeling approach or the density functional theory (DFT) (Nazarewicz, 2016).

In this chapter, we consider a general situation where measurements $(\boldsymbol{t}_i, y_i)_{i=1}^n$ of a physical process $\boldsymbol{t} \mapsto \zeta(\boldsymbol{t})$ are used to predict new values $\boldsymbol{y}^*$ of the physical process ζ, where $\boldsymbol{t} \in \boldsymbol{\Omega} \subset \mathbb{R}^p$. Furthermore, we suppose there are K competing models $\mathcal{M}_1, \ldots, \mathcal{M}_K$ of observations y_i, where the k^{th} model is parametrized by a vector of unknown parameters $\boldsymbol{\phi}_k \in \mathbb{R}^{p_k}$ for $k = 1, \ldots, K$ and $p_k \geq 1$ (e.g., for the Bayesian calibration with Gaussian processes $\boldsymbol{\phi} = (\boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma)$). Given each model, we consider the data likelihood $p(\boldsymbol{d}|\boldsymbol{\phi}_k, \mathcal{M}_k)$ and the prior density $p(\boldsymbol{\phi}_k|\mathcal{M}_k)$; the dataset $\boldsymbol{d}$ typically consists of the experimental observations $\boldsymbol{y} = (y_1, \ldots, y_n)$ only, however, it can also include a set of computer model evaluations $\boldsymbol{z} = (z_1, \ldots, z_s)$ under the Bayesian calibration framework with computationally expensive models described in Section 1.2.

BMA provides a way of accounting for model uncertainty induced by the existence of alternative models. If $\mathcal{O}$ is the quantity of interest, e.g., the value y^*, the BMA posterior density $p(\mathcal{O}|\boldsymbol{d})$ corresponds to the mixture of the posterior densities of the individual models:

$$p(\mathcal{O}|\boldsymbol{d}) = \sum_{k=1}^{K} p(\mathcal{O}|\boldsymbol{d}, \mathcal{M}_k) p(\mathcal{M}_k|\boldsymbol{d}). \tag{2.1}$$

The posterior weights $p(\mathcal{M}_k|\boldsymbol{d})$ are given by a simple application of the Bayes' theorem:

$$p(\mathcal{M}_k|\boldsymbol{d}) = \frac{p(\boldsymbol{d}|\mathcal{M}_k)p(\mathcal{M}_k)}{\sum_{\ell=1}^{K} p(\boldsymbol{d}|\mathcal{M}_\ell)p(\mathcal{M}_\ell)}. \tag{2.2}$$

These weights are determined by two quantities that are the key to a successful implementation of the BMA framework. First, one needs to assign suitable prior probabilities $p(\mathcal{M}_k)$ that $\mathcal{M}_k$ is the true model. Hoeting et al. (1999) notes that,

> *When there is little prior information about the relative plausibility of the models considered, the assumption that all models are equally likely a priory is a reasonable "neutral" choice.*

One can, nevertheless, choose informative prior distributions when prior information about the plausibility of each model is available. Eliciting an informative prior is a non-trivial task, but Madigan et al. (1995) provide some guidance in the context of graphical models that can be applied in other settings as well.

The second key quantity is the evidence integral

$$p(\boldsymbol{d}|\mathcal{M}_k) = \int p(\boldsymbol{d}|\boldsymbol{\phi}_k, \mathcal{M}_k)p(\boldsymbol{\phi}_k|\mathcal{M}_k)\,\mathrm{d}\boldsymbol{\phi}_k. \tag{2.3}$$

The numerical evaluation of evidence integrals is challenging in practice, because a closed form solution is available only in elementary situations for the exponential family distributions with conjugate priors (see Section 2.4 for a simple example) and thus requires approximation. The simplest and most commonly used approximation in the literature, and which we have adopted in our applications, is to use the Monte Carlo (MC) integration estimate

$$\widehat{p}_{MC}(\boldsymbol{d}|\mathcal{M}_k) = \frac{1}{n_{MC}} \sum_i p(\boldsymbol{d}|\boldsymbol{\phi}_k^{(i)}, \mathcal{M}_k), \tag{2.4}$$

where $\boldsymbol{\phi}_k^{(i)}$ are i.i.d. samples from the prior $p(\boldsymbol{\phi}_k|\mathcal{M}_k)$ for $i = 1, \ldots, n_{MC}$. While this MC integration yields reasonable results, it requires separate evaluations of the likelihood at new samples from the prior $p(\boldsymbol{\phi}_k|\mathcal{M}_k)$, which can be very costly in computing time.

Another frequently used method is the Laplace approximation, which relies on the fact that the integration (2.3) has a closed form in the case of a linear regression with Gaussian noise. It corresponds to a second order Taylor expansion of the log-likelihood around its maximum, which makes the likelihood Gaussian. Namely the Laplace approximation is

$$\widehat{p}_L(\boldsymbol{d}|\mathcal{M}_k) = (2\pi)^{\frac{p_k}{2}} |\widetilde{\Sigma}_k|^{\frac{1}{2}} p(\boldsymbol{d}|\widetilde{\boldsymbol{\phi}}_k \mathcal{M}_k) p(\widetilde{\boldsymbol{\phi}}_k|\mathcal{M}_k), \tag{2.5}$$

where $\widetilde{\boldsymbol{\phi}}_k$ is the mode of $p(\boldsymbol{\phi}_k|\boldsymbol{d}, \mathcal{M}_k)$ and $\widetilde{\Sigma}_k = (-\boldsymbol{D}^2 l(\widetilde{\boldsymbol{\phi}}_k))^{-1}$ is the inverse of the Hessian matrix of second derivatives (evaluated at $\widetilde{\boldsymbol{\phi}}_k$) of $l(\boldsymbol{\phi}_k) = \log(p(\boldsymbol{d}|\boldsymbol{\phi}_k, \mathcal{M}_k) p(\boldsymbol{\phi}_k|\mathcal{M}_k))$. The Laplace method typically gives very good results for very peaked likelihoods. We refer the reader to Kass and Raftery (1995) for an exhaustive survey of classical methods used to compute the evidence integral. Also, more recently proposed Nested Sampling algorithm by Skilling (2006) and expanded by Feroz et al. (2009) provides another alternative to these classical approaches.

BMA, while a conceptually straightforward and natural approach to account for model uncertainty, becomes challenging in scenarios where alternative models are defined on different subsets of the same input space; this can typically arise with local models or with numerical models with different constraints. It is also a usual situation in nuclear physics, for instance for nuclear mass models; ab initio models range over lighter nuclei due to contemporary computational limitations, while EDFs can cover the whole nuclear chart (Klupfel et al., 2009; Kortelainen et al., 2010b). This also happens when one considers mixing models produced by the calibration of observables of different types – typically some nuclear models are fitted on nuclear binding energies, while others on binding energies and other observables such as rms charge radii (a measure of the size of an atomic nucleus). Surprisingly, we have not found in the literature a principled approach to adapt BMA to this situation, or how to compare models with similar, overlapping, but significantly non-identical domains. To address this "domain discrepancy", in Section 2.2 we present a method which relaxes the requirement that all models cover the same domain ($\boldsymbol{d}$ is common to all models considered). Other applications of our framework could include time series with missing data, or different

time scales, e.g. in a financial setting where additionally different classes of assets can be treated as observables.

The remaining sections of this chapter are organized as follows. Section 2.1 provides a systematic analysis of prediction errors under individual models as compared to the BMA framework. Section 2.2 develops the BMA methodology for models with discrepant domains. Section 2.3 contains an extensive collection of simulation studies, pedagogical examples as well as real-data applications highlighting the benefits of BMA in terms of the improvement of the prediction accuracy and UQ. All technical details and supplementary results are provided in section 2.4.

2.1 Optimality of BMA predictions

BMA is not the only way to deal with several alternative models and to account for model uncertainty, but it does have the property of reducing the Posterior Mean Square Error (PMSE) of prediction of a new observation y^*. In this section, we illustrate this property in a clear and concise way.

Let us, for simplicity of notation, consider two competing models $\mathcal{M}_1$ and $\mathcal{M}_2$ - the treatment of multiple models follows from a similar argument, and our verbal descriptions below in this section occasionally refer to the general case without further comment. Denote $\widehat{y_1^*} := \mathbb{E}[y^*|\boldsymbol{d}, \mathcal{M}_1]$ and $\widehat{y_2^*} := \mathbb{E}[y^*|\boldsymbol{d}, \mathcal{M}_2]$ as the posterior means of y^* under each model, and let $\widehat{y^*} := \mathbb{E}[y^*|\boldsymbol{d}]$. We also define $p_k := p(\mathcal{M}_k|\boldsymbol{d})$ for $k = 1, 2$ for the posterior probability of each model. Thus the BMA posterior mean estimator (1.12) can be written as $\widehat{y^*} = p_1\widehat{y_1^*} + p_2\widehat{y_2^*}$. The PMSE of y^* is then defined as $\mathbb{E}[(\widehat{y^*} - y^*)^2|\boldsymbol{d}]$ and has the following decomposition.

Lemma 1. *For every $\lambda_1, \lambda_2 \geq 0$ satisfying $\lambda_1 + \lambda_2 = 1$, we have*

$$\mathbb{E}[(y^* - \widehat{y^*})^2|\boldsymbol{d}] = \mathbb{E}[(y^* - \lambda_1\widehat{y_1^*} - \lambda_2\widehat{y_2^*})^2|\boldsymbol{d}] - [(\lambda_1 - p_1)\widehat{y_1^*} + (\lambda_2 - p_2)\widehat{y_2^*}]^2 \tag{2.6}$$

This Lemma shows explicitly that the PMSE of the BMA predictor is smaller than the PMSE associated with any convex combination $\lambda_1\widehat{y_1^*} + \lambda_2\widehat{y_2^*}$ of the each of the two

models' posterior means. It also measures how much smaller it is, and shows that equality holds as soon as the convex coefficients λ_k are equal to the posterior probabilities p_k of each model, $k = 1, 2$. Specifically, by applying Lemma 1 twice, with $(\lambda_1, \lambda_2) = (1, 0)$ and with $(\lambda_1, \lambda_2) = (0, 1)$, we obtain the following dual expression for the PMSE or the BMA predictor, involving each individual model's PMSE, showing how much smaller the former is compared to the two latter:

$$\mathbb{E}[(y^* - \widehat{y_1^*})^2|\boldsymbol{d}] - p_2^2(\widehat{y_1^*} - \widehat{y_2^*})^2 = \mathbb{E}[(y^* - \widehat{y^*})^2|\boldsymbol{d}] = \mathbb{E}[(y^* - \widehat{y_2^*})^2|\boldsymbol{d}] - p_1^2(\widehat{y_1^*} - \widehat{y_2^*})^2. \tag{2.7}$$

The relationship (2.7) directly implies

$$\mathbb{E}[(y^* - \widehat{y^*})^2|\boldsymbol{d}] \leq \mathbb{E}[(y^* - \widehat{y_k^*})^2|\boldsymbol{d}], \qquad k = 1, 2. \tag{2.8}$$

This inequality clearly states that the BMA estimator (1.12) gives prediction error at least as small as the best of the models considered, in the PMSE sense. We interpret this as a translation of the fact that each model that goes into creating the BMA estimator necessarily ignores model uncertainty. Note that this says nothing about how the BMA estimator would compare to a model not used in its definition.

Moreover, since Lemma 1 covers all convex combinations of the original models, it shows that BMA achieves the following minimum

$$\big(p(\mathcal{M}_k|\boldsymbol{d})\big)_{k=1,2} = \underset{\lambda \in [0,1]^2 : \lambda_1 + \lambda_2 = 1}{\arg\min} \mathbb{E}[\big(y^* - (\lambda_1 \widehat{y_1^*} + \lambda_2 \hat{y_2^*})\big)^2|\boldsymbol{d}]. \tag{2.9}$$

Hence, the BMA estimator is actually optimal over all convex combinations of the individual estimators $\widehat{y_1^*}$ and $\widehat{y_2^*}$. The optimality of BMA can be also established from a decision-theoretic perspective, see Chapter 6 in Bernardo and Smith (1994) for details.

We can also express the reduction of the PMSE for the BMA estimator, compared to the best (lowest) PMSE among all of the individual models', as

$$r^2_{BMA} := 1 - \frac{\mathbb{E}[(\widehat{y^*} - y^*)^2|\boldsymbol{d}]}{\min_k \mathbb{E}[(\widehat{y_k^*} - y^*)^2|\boldsymbol{d}]}, \qquad k = 1, \ldots, K \tag{2.10}$$

In the specific case of two competing models, if we assume for instance that the 'best' model is $\mathcal{M}_2$, we can obtain an even more explicit expression for r^2_{BMA} which provides the relative gain attained by BMA, namely

$$r^2_{BMA} = p(\mathcal{M}_1|\boldsymbol{d})^2 \frac{(\widehat{y_1^*} - \widehat{y_2^*})^2}{\mathbb{E}[(\widehat{y_2^*} - y^*)^2|\boldsymbol{d}]}. \tag{2.11}$$

Below in Section 2.3, we denote the sample version of the expression in (2.10) as $\hat{r}^2_{BMA}$, which we will use to evaluate the performance of BMA quantitatively.

To finish this section, we decompose the quantity $\mathbb{E}[(\widehat{y^*} - y^*)^2|\boldsymbol{d}]$ against the residuals $(\widehat{y_k^*} - y^*)$, $k = 1, 2$, from each individual model assuming $p_1, p_2 > 0$. This is easily done by symmetrizing formula (2.7) via reintroducing y^* to identify these residuals, and then taking another conditional expectation with respect to $\boldsymbol{d}$ to avoid an expression which depends on unobserved data. We obtain

$$\begin{aligned}\mathbb{E}[(y^* - \widehat{y^*})^2|\boldsymbol{d}] = (p_1 - p_1^2)\mathbb{E}[(y^* - \widehat{y_1^*})^2|\boldsymbol{d}] + (p_2 - p_2^2)\mathbb{E}[(y^* - \widehat{y_2^*})^2|\boldsymbol{d}] \\ - (p_1^2 + p_2^2)\mathbb{E}\big[(\widehat{y_1^*} - y^*)(y^* - \widehat{y_2^*})|\boldsymbol{d}\big].\end{aligned} \tag{2.12}$$

Formula (2.12) shows that the PMSE of the BMA estimator is an explicit linear combination of the prediction errors of estimators for each constituent model, but that one must subtract a coupling correction term on the right hand side of (2.12).

It is interesting to note that the weights in the aforementioned linear combination can be interpreted as the variances of Bernoulli random variables with the posterior model probabilities p_1 and p_2 as their success probabilities. Also note that, since these variances $p_k - p_k^2 < p_k$, the linear combination is not convex, but is smaller. The correction term is not necessarily a subtraction of a positive term, but it is likely to be when both individual models have significant biases in opposite directions for prediction of y^*. This is particularly interesting when the two models have similar posterior performances. Both values of p_k will be in this situation close to $1/2$, which minimizes the values of $p_k - p_k^2$ for both $k = 1, 2$. This is a scenario where using BMA will significantly improve prediction errors even when each model is competitive compared to the other, regardless of how large the individual models' biases

are, and without knowing in what direction they go, as long as the two models are assumed to have significant defects that work in opposite directions.

A sanity check reveals an interesting characteristic of BMA: suppose that $p_1 = 1$, so that the BMA estimate is given by $\widehat{y^*} = \widehat{y_1^*}$. According to (2.12) we must have $\mathbb{E}[(y^* - \widehat{y^*})^2 | \boldsymbol{d}] = 0$, and further $\mathbb{E}[(y^* - \widehat{y^*})^2] = 0$, i.e. $y^* = \widehat{y^*} = \widehat{y_1^*}$ a.s. given $\boldsymbol{d}$, in other words, model 1 must provide a perfect description of the reality.

2.2 BMA with discrepant domains

Let us continue with the discussion about BMA of models with similar, overlapping, but significantly non-identical domains from the beginning of this chapter in a formal setting. Let us consider two models $\mathcal{M}_A$ and $\mathcal{M}_B$, which we will also denote by (A) and (B) or merely A and B for simplicity, and assume that they are respectively defined only on different strict subsets $\boldsymbol{t}^{(A)}$ and $\boldsymbol{t}^{(B)}$ of the data. We denote $\boldsymbol{d}^{(A)}$ and $\boldsymbol{d}^{(B)}$ the corresponding $\boldsymbol{d}$ data as well as $\boldsymbol{d}^{(-A)}$ and $\boldsymbol{d}^{(-B)}$ their respective complements in $\boldsymbol{d}$. The actual *Bayesian evidence* for each of these models are the probabilities $p(\boldsymbol{d}|A)$ and $p(\boldsymbol{d}|B)$, but these quantities are not clearly defined. On the other hand $p(\boldsymbol{d}^{(A)}|A)$ and $p(\boldsymbol{d}^{(B)}|B)$, where each model refers only to its original range of validity, are the evidences of the models corresponding to the classical BMA theory described in Section 1.3 and also at the beginning of this chapter. Nevertheless, we have the following expansion:

$$p(\boldsymbol{d}|A) = p(\boldsymbol{d}^{(A)}, \boldsymbol{d}^{(-A)}|A) = p(\boldsymbol{d}^{(A)}|A) p(\boldsymbol{d}^{(-A)}|\boldsymbol{d}^{(A)}, A). \tag{2.13}$$

This expression means that to obtain model (A)'s actual Bayesian evidence, $p(\boldsymbol{d}^{(A)}|A)$ must be multiplied by a corrective factor $p(\boldsymbol{d}^{(-A)}|\boldsymbol{d}^{(A)}, A)$ which represents the information one has on $\boldsymbol{d}^{(-A)}$ assuming that model (A) holds and that it does not provide any prediction at the data points in $\boldsymbol{d}^{(-A)}$. Note that the distribution $p(\boldsymbol{d}|A)$ is meaningful only to the extent that $\boldsymbol{d}$ – and thus $\boldsymbol{d}^{(-A)}$ – is measurable in the underlying probability space, which implies the existence of underlying distributions $p(\boldsymbol{d}^{(-A)})$ and subsequently of $p(\boldsymbol{d}^{(-A)}|\boldsymbol{d}^{(A)})$ and $p(\boldsymbol{d}^{(-A)}|\boldsymbol{d}^{(A)}, A)$. To that extent, the problem of averaging models with different domains

can be ill posed, if these distributions cannot be defined convincingly.

If the data $\boldsymbol{d}^{(A)}$ and $\boldsymbol{d}^{(-A)}$ are independent, conditionally to model (A), in other words if no information can be gleaned about $\boldsymbol{d}^{(-A)}$ from $\boldsymbol{d}^{(A)}$ or from (A), i.e. $\boldsymbol{d}^{(A)}$ is unconstrained by (A) and by $\boldsymbol{d}^{(-A)}$, then it is legitimate to ignore the aforementioned correction factor which should be $p(\boldsymbol{d}^{(-A)}|\boldsymbol{d}^{(A)}, A) = 1$. In particular, this is the case if, given model (A), $\boldsymbol{d}^{(-A)}$ is considered deterministically equal to its sample value. Conversely, setting the corrective factor to $p(\boldsymbol{d}^{(-A)}|\boldsymbol{d}^{(A)}, A) = 1$ outside of this scope is an approximation to such extent, and not in general a fair evaluation of the information contained in the "globality" of a model. We shall refer to this case as *BMA with independent model domains*. Although it has been adopted as a natural matter of convenience, it raises serious safeguards for which we cannot find better words than Trotta's ascertainment (Trotta, 2008):

> *On the other hand, it is important to notice that the Bayesian evidence does not penalize models with parameters that are unconstrained by the data. It is easy to see that unmeasured parameters (i.e. parameters whose posterior is equal to the prior) do not contribute to the evidence integral, and hence model comparison does not act against them, awaiting better data.*

Let us point out as an extreme situation that occurs when model (A) predicts the values $\boldsymbol{d}^{(-A)}$ that have no physical meaning, e.g. in the case of nuclear mass models, this can be the mass of a nucleus which a model predicts not to exist, and therefore the mass has no physical meaning. In this case, the model (A) is actually strongly constrained by $\boldsymbol{d}^{(-A)}$, to the point that $p(\boldsymbol{d}^{(-A)}|A) = 0$, yielding $p(\boldsymbol{d}^{(-A)}|\boldsymbol{d}^{(A)}, A) = 0$, which rules the model (A) impossible as long as $\boldsymbol{d}^{(-A)}$ is not empty.

Another tempting option is to restrict the domain of interest to the domain common to all models and simply consider $p(\boldsymbol{d}^{(A)\cap(B)}|A)$ and $p(\boldsymbol{d}^{(A)\cap(B)}|B)$, which can be obtained in a standard way according to (2.3). As we ignore even more data, this approach is arguably worse than setting $p(\boldsymbol{d}^{(-A)}|\boldsymbol{d}^{(A)}, A) = 1$.

Let us illustrate how the assumption of independent model domains, namely setting $p(\boldsymbol{d}^{(-A)}|\boldsymbol{d}^{(A)}, A) = 1$, can fail to provide a satisfactory ranking of models in two examples where a model takes a shortcut by 'refusing' to predict challenging points.

Scenario 1. Consider the situation where one model $\mathcal{M}_0$ is empty so that $p(\mathcal{M}_0|\boldsymbol{d}) \propto p(\mathcal{M}_0)$. On the other hand, any other model which constrains any part of the data will have an evidence most likely lower than 1 which implies that the model will end up with lower posterior weights when starting from equal prior weights. Thus any predictive model will be deemed inferior to a non-predictive one.

Scenario 2. Take two deterministic models A and B with input space (domain of $\boldsymbol{t}$) $\{a, b\}$; assume model A has deviation 0 at location a and 10^{99} at location b, and that model B has deviation 1.001 at location a, but does not predict anything at location b. One can easily adjust the numbers to reach an extreme situation (e.g. making A's prediction at location b to be extremely poor) where model B ends up with a much higher Bayes evidence than model A, while the common sense by which no prediction is a form of extremely poor prediction, would always imply that model A is better than model B.

These examples show how important it is to acknowledge that a model's inability to make predictions in some locations is not a neutral property. The classical BMA approach offers no trade-off: a model withholding its predictions at the most difficult points will always improve its weight. We now introduce our "domain-corrected BMA" where we amend the model weights to account more fairly for the (in-)ability of a model to provide predictions at locations of interest.

2.2.1 Two models

Starting from (2.13), instead of setting $p(\boldsymbol{d}^{(-A)}|\boldsymbol{d}^{(A)}, A) = 1$ which removes the effect of a model's domain in its posterior weights, we propose the weaker assumption that

$p(\boldsymbol{d}^{(-A)}|\boldsymbol{d}^{(A)}, A)$ is independent from the model, i.e. we assume

$$p(\boldsymbol{d}^{(-A)}|\boldsymbol{d}^{(A)}, A) = p(\boldsymbol{d}^{(-A)}|\boldsymbol{d}^{(A)}). \tag{2.14}$$

This is quite natural if we consider that model (A) implies a distribution $p(\boldsymbol{d}^{(A)}|A)$ but provides no information on $\boldsymbol{d}^{(-A)}$, leaving $\boldsymbol{d}^{(-A)}$ unconstrained by (A) (see the introduction of this section). The evidence $p(\boldsymbol{d}|A)$ is now given by

$$p(\boldsymbol{d}|A) \propto_A p(\boldsymbol{d}^{(A)}|A)p(\boldsymbol{d}^{(-A)}|\boldsymbol{d}^{(A)}). \tag{2.15}$$

Our assumption $\boldsymbol{d}^{(A)} \cup \boldsymbol{d}^{(B)} = \boldsymbol{d}$ implies that $\boldsymbol{d}^{(-A)}$ can only be informed by (B). Hence

$$p(\boldsymbol{d}^{(-A)}|\boldsymbol{d}^{(A)}) = p(\boldsymbol{d}^{(-A)}|\boldsymbol{d}^{(A)}, B) = p(\boldsymbol{d}^{(-A)}|\boldsymbol{d}^{(A)\cap(B)}, B), \tag{2.16}$$

which can be written as an explicit integral with respect to model (B)'s parameter $\boldsymbol{\phi}_B$,

$$\int p(\boldsymbol{d}^{(-A)}|\boldsymbol{d}^{(A)\cap(B)}, \boldsymbol{\phi}_B, B)p(\boldsymbol{\phi}_B|\boldsymbol{d}^{(A)\cap(B)}, B)\,\mathrm{d}\boldsymbol{\phi}_B. \tag{2.17}$$

To approximate (2.17), one can use the same approximation methods as in the case of classical evidence integral (see the beginning of this section for more details).

2.2.2 K models

In the general case, each model $\mathcal{M}_k$ constrains a subset $\boldsymbol{d}^{(k)}$ of the data $\boldsymbol{d}$ (for $k = 1, \ldots, K$); as in the case of two models, $\boldsymbol{d}^{(-k)}$ denotes the complement subset of $\boldsymbol{d}^{(k)}$ in $\boldsymbol{d}$. We also introduce $\boldsymbol{d}^{(\oslash)} := \bigcap_k \boldsymbol{d}^{(k)}$ as the set of data common to all individual models. Moreover we assume that $\boldsymbol{d} = \bigcup_k \boldsymbol{d}^{(k)}$, i.e. every datapoint is covered by at least one model. We also assume, up to taking equivalence classes on models (see Section 2.4.3 for details), that for each pair of models there exists a chain of models joining them where each model $\mathcal{M}_k$ shares a data point in its domain $\boldsymbol{d}^{(k)}$ with each of its neighbours. Relying on the same principles described in Section 2.2.1, we set

$$p(\boldsymbol{d}^{(-k)}|\boldsymbol{d}^{(k)}, \mathcal{M}_k) = p(\boldsymbol{d}^{(-k)}|\boldsymbol{d}^{(k)}), \tag{2.18}$$

which leads to the model posterior probabilities of the form

$$p(\mathcal{M}_k|\boldsymbol{d}) \propto_k p(\boldsymbol{d}^{(-k)}|\boldsymbol{d}^{(k)})p(\mathcal{M}_k|\boldsymbol{d}^{(k)}). \tag{2.19}$$

Compared to the two-model case, the computation of the corrective factors poses additional difficulty that, when there is more than one model constraining $\boldsymbol{d}^{(-k)}$, the factor $p(\boldsymbol{d}^{(-k)}|\boldsymbol{d}^{(k)})$ is no longer equal to a single $p(\boldsymbol{d}^{(-k)}|\boldsymbol{d}^{(k)}, \mathcal{M}_k)$, but rather to the an average of all models constraining $\boldsymbol{d}^{(-k)}$. Hence our domain-corrected BMA corresponds to the intermediate solution where one replaces the factors of the likelihood corresponding to the missing model predictions by a geometric average of the likelihoods over the models which do produce predictions, based on the predictive models' posterior weights. We have found that similar ideas have been developed in the broader framework of evidence theory (Park and Grandhi, 2012, Section 2.2).

The notation for a given corrective factor can become cumbersome when model domains have very general intersections, but these corrective factors can still be computed recursively rather than directly. We relegate the calculations of the general case to Section 2.4.3.

2.3 Examples and applications

To illustrate the methodology described in Chapter 2, we present several examples in which BMA leads to the reduction in prediction error and improved UQ. Our first example is a simple yet sensible scenario of averaging two different models of proton potentials. The second example is an application of BMA methodology to state-of-the-art nuclear mass models and nuclear mass data. The third example is a BMA study of the LDM (1.1) published by Kejzlar et al. (2020). Lastly, we provide a pedagogical application of model averaging to a synthetic dataset which highlights the interest of the domain-corrected BMA.

The predictive improvement is measured in the examples as a relative reduction in the mean square error (MSE), a sample version of (2.10), which we denote as $\widehat{r}^2_{BMA}$. As a measure of UQ fidelity, we consider what is know as the empirical coverage probability

(ECP) (Gneiting et al., 2007; Gneiting and Raftery, 2007). Formally, it can be written as

$$\eta(\alpha) := \frac{1}{J}\sum_{i=1}^{J} \mathbb{1}_{y_i^* \in I_\alpha(\boldsymbol{t}_i^*)}, \tag{2.20}$$

where $\mathbb{1}$ is the indicator function, $I_\alpha(\boldsymbol{t}_i^*)$ is the $\alpha-$credibility interval produced by a given model at a new input $\boldsymbol{t}_i^*$, and y_i^*'s are the (new) testing data. The ECP represents the proportion of a model's prediction of independent testing points falling into the respective credibility intervals. This quantity is typically plotted against the credibility level α to form a so called ECP line (e.g., Figure 2.2). This line should theoretically follow the diagonal so that the actual fidelity of the interval corresponds to the nominal value. If the respective ECP line falls above the reference, credible intervals produced by a given model are too wide (UQ is conservative). Naturally, a model with an ECP line below the reference underestimates the uncertainty of predictions (UQ is liberal). While values of empirical proportions close to the reference curve are desirable, it is preferable to be conservative rather than liberal. Overly narrow credible intervals declare a level of assurance higher than it should be.

Each of the examples in this section looks at a situation with several competing models without any prior knowledge of which is better; thus we set the prior model weights to be uniform over the model space. All the posterior samples were computed using the NUTS. The evidence integrals were approximated using the MC integration. All the credible intervals discussed are the *highest posterior density* (HPD) credible intervals. Given a credibility level α, the α-HPD of a scalar quantity consists of the minimum width interval containing an α proportion of its MCMC posterior samples. Lastly, some of the supplementary results and modeling details are delegated to Section 2.4.4.

2.3.1 Averaging of proton potentials

In this first example we demonstrate the potential of BMA to improve both prediction accuracy and honesty of UQ in a favorable situation where we average two models associated with different proton potentials.

We consider two single-proton potentials describing the average interaction acting on a proton within the spatial range of a nucleus; namely, the Woods-Saxon (WS) potential V_1 representing respectively the strong nuclear forces between nucleons (protons and neutrons), and the Coulomb potential V_2 representing the electromagnetic interactions between protons. For a given nucleus, which we will take with proton and neutron numbers $Z = 100$ and $N = 150$ and mass number $A = 250$, they can be expressed as

$$V_1(r) = -V_{WS}\frac{1}{1 + e^{\frac{r-R_A}{a}}}, \tag{2.21}$$

$$V_2(r) = -V_C\frac{Z}{r}. \tag{2.22}$$

Here, $V_{WS} = 50$ MeV, $V_C = 0.5$ MeV fm, and $a = 0.5$ are fixed parameters, and $R_A = A^{1/3} \times 1.25$ fm is the radius of the nucleus of interest. These two models for energy potentials have the interesting property that both are non-decreasing and vanishing at infinity, while with different speeds, and can correspond to two phenomenons with different length scales. As a matter of fact, the strong interactions described by the WS potential are confined to the volume of atomic nuclei (several fm $= 10^{-15}$ m), i.e. they are short-ranged; in contrast the electrostatic ones are long-ranged, i.e. they act on much larger length scales ($> 10^{-10}$ m) and compete with the strong interactions in superheavy elements, causing the so-called Coulomb frustration (see Nazarewicz (2018)). This fact is reproduced in our example where we also expect that V_1 should be well constrained by a dataset of stable nuclei, while V_2 should play an important role in the description of short-lived superheavy nuclei. More generally, we have in mind a scenario where two models have been developed for different subsets of an input domain and are in competition on some common intermediate domain. Both of these modeling approaches are equally confident that they prevail on the intermediate domain, while the truth is somewhere in between. This situation is quite realistic despite its simplicity, and we can reasonably expect model mixing to have positive outcomes.

We simulate the experimental data $\{(r_i, y_i)\}_{i=1}^n$ at different spatial locations r_i, relatively

far from the nucleus ($r > R_A$) following a mixture of the two models. Namely

$$y_i = (1-\omega)V_1(r_i) + \omega V_2(r_i) + \epsilon_i, \tag{2.23}$$

where ϵ_i are standard normal errors, and we take $\omega = \frac{1}{2}$. Note that in reality, the observations of the potentials are not available as such, but can be inferred indirectly from experimental nucleonic densities measured in nucleon scattering experiments (Anni et al., 1995). In particular, we drew a dataset of 210 observations generated according to the model (2.23) with the locations r_i sampled uniformly over $(R_A, 10)$.

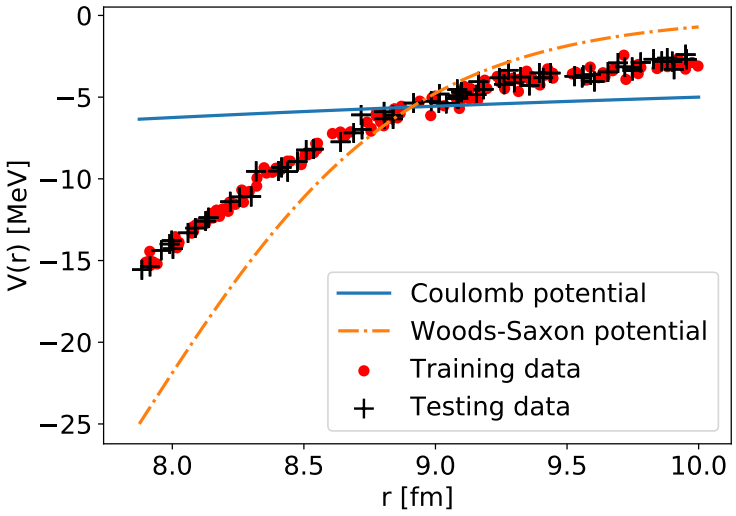

Figure 2.1: The Woods-Saxon potential and the Coulomb potential along with the training (140 observations) and the testing datasets (70 observations) generated from the mixture of the two potentials.

We further randomly divided the data into a training dataset of 140 observations and kept the remaining 70 observations for testing (see Figure 2.1). The two statistical models $\mathcal{M}_1$ and $\mathcal{M}_2$ considered here are given by the respective energy potentials (2.23) obtained with $\omega = 0$ and $\omega = 1$ and additive independent experimental errors distributed according to $\mathcal{N}(0, \sigma_j)$ for $j = 1, 2$. The prior distributions for standard deviations σ_j's were take to be the non-informative Inv-Gamma(1,30).

Table 2.1 shows the estimated root MSE (RMSE) for the testing dataset. We can see that this simple example gives significantly better predictions under the BMA posterior mean

predictor than each of the models individually. This is, of course, not a surprise and shows that BMA behaves as expected.

Model	RMSE	$P(\mathcal{M}_k\|y)$	$\widehat{r}^2_{BMA}$
$\mathcal{M}_1$	3.540	0.512	**0.930**
$\mathcal{M}_2$	3.607	0.488	**0.933**
$\mathcal{M}_{BMA}$	0.935	-	-

Table 2.1: RMSE (in MeV) and the improvement under the BMA posterior mean predictor calculated on the testing dataset ($n = 70$, $A = 250$).

More interesting results can be seen from the angle of the quality of the predictions' UQ in Figure 2.2. In contrast with the individual models, the ECP of the BMA posterior predictions matches closely the reference line and provides evidence that accounting for model uncertainty leads to the desired more honest UQ.

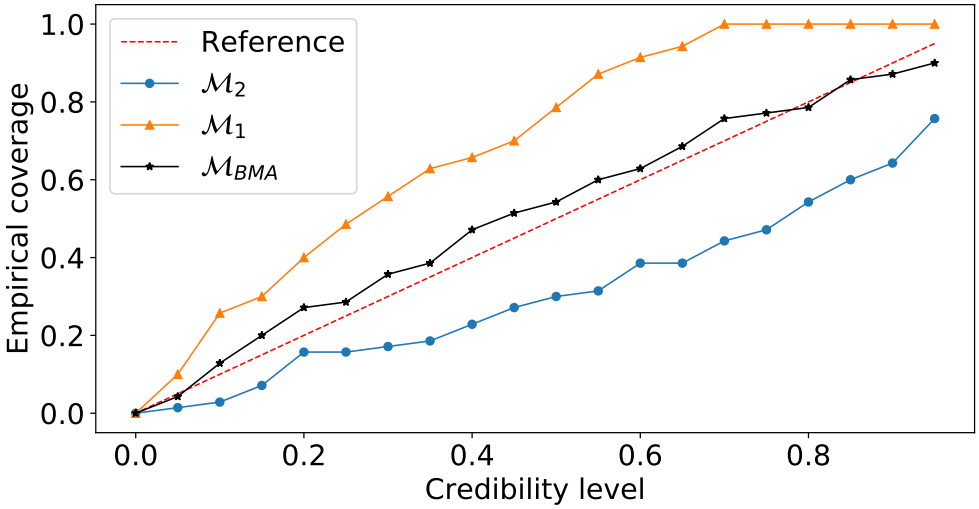

Figure 2.2: ECPs for the testing dataset ($m = 70$, $A = 250$).

2.3.2 Averaging of nuclear mass emulators in the Ca region

An important challenge in nuclear structure is to produce quantified predictions of nuclear observables, such as nuclear masses (McDonnell et al., 2015), for all possible pairs

(Z, N) of proton numbers Z and neutron numbers N which can be bound together in a nucleus. Such predictions are of direct interest to guide future nuclear experiments or to feed astrophysical calculations for the abundance of elements in the universe. The underlying astrophysical processes, such as the rapid neutron capture which produces heavy elements in stellar environments (Horowitz et al., 2019), take place far from the region of nuclear stability, where no experimental measurement are available, and these observables have to be extracted from extreme extrapolations of theoretical nuclear models.

In their recent work, Neufcourt et al. (2019) used GPs (see Section 1.2 for definition) to model the discrepancies between the experimental data and the theoretical calculations for several nuclear models based on the DFT, and obtained quantified extrapolations for nuclear masses in the Calcium region (at the frontier between experimental and theoretical limits). They computed a simplified BMA of 9 global mass models (Bartel et al., 1982; Dobaczewski et al., 1984; Chabanat et al., 1995; Klüpfel et al., 2009; Kortelainen et al., 2010a, 2012, 2014) listed in Table 2.2 defined across the full nuclear landscape from the light to the superheavy nuclei, thus suitable for extrapolations. Their weights, proportional to $p(y^* > 0|\boldsymbol{y}, \mathcal{M}_k)$, are based on each model's probability to assign a positive separation energy y^* to a testing set of nuclei which have been experimentally observed after 2003, thus independent from the training set of measured neutron separation energies $\boldsymbol{y}$ (separation energy is the energy needed to remove a neutron or proton form an atomic nucleus). Here, we compare the results of Neufcourt et al. (2019) to the full BMA analysis with model weights given by their posterior probabilities $p(\mathcal{M}_k|\boldsymbol{y})$. Note that all the physical models are taken here as calibrated and their parameter estimation is not part of our analysis.

We consider the same training dataset of one-neutron (S_{1n}) and two-neutron (S_{2n}) separation energies AME2003 (Audi et al., 2003) restricted to the calcium (Ca) region on the nuclear landscape with $Z \geq 14$ and $N \leq 22$ ($n = 139$). The predictive performances of each model augmented with the GP model for systematic discrepancies and the BMA posterior mean predictor are evaluated on both the training dataset and a testing dataset of

new measurements in AME2016 ($n = 14$) that we denote as AME2016 \ AME2003 (Wang et al., 2017). The predictive performances of each model augmented with a GP model for systematic discrepancies and the BMA posterior mean predictor are evaluated on both the training dataset and a testing dataset of new measurements in AME2016 ($n = 14$) (Wang et al., 2017). Similarly to Neufcourt et al. (2019), we calculate the model posterior probabilities independently over four non-overlapping nuclear domains according to the parity of numbers Z and N with uniform prior distribution over the model space. We assess the performance of BMA using the MSE improvement and the ECP. These were combined over odd and even parities of numbers Z and N in order to mitigate the relatively small size of each parity subset. The GP model specification and the sample sizes breakdown based on the parity of Z and N are given in Section 2.4.4.

	Model posterior weights				Errors			
	$\boldsymbol{S_{1n}}$ (odd N)		$\boldsymbol{S_{2n}}$ (even N)		Training		Testing	
Model	even Z	odd Z	even Z	odd Z	RMSE	$\widehat{r}^2_{BMA}$	RMSE	$\widehat{r}^2_{BMA}$
SLy4	0.000	0.000	0.000	0.008	0.076	-	0.713	0.313
SkP	0.000	0.000	0.000	0.000	0.127	0.308	0.989	0.642
SkM*	0.000	0.000	0.000	0.000	0.142	0.449	0.924	0.591
SV-min	0.000	0.000	0.000	0.001	0.107	0.023	0.840	0.505
UNEDF0	0.000	0.009	0.000	0.000	0.136	0.400	0.809	0.466
UNEDF1	0.845	0.669	0.000	0.089	0.110	0.077	0.550	-
UNEDF2	0.002	0.013	0.000	0.125	0.109	0.058	0.806	0.462
FRDM-2012	0.153	0.308	0.902	0.310	0.114	0.149	0.808	0.465
HFB-24	0.000	0.001	0.098	0.467	0.146	0.477	0.806	0.463
$\mathcal{M}_{BMA(prior)}$					0.110	0.045	0.641	0.078
$\mathcal{M}_{BMA(simple)}$					0.118	0.110	0.680	0.131
$\mathcal{M}_{BMA}$					0.105	-	0.591	-

Table 2.2: Model posterior weights for 9 nuclear mass models with the RMSE (in MeV) and the MSE improvement for the training and the testing datasets. The last three rows correspond to the averaging with the prior weights, the simplified BMA (Neufcourt et al., 2019), and the full BMA.

Table 2.2 presents the resulting posterior weights of the models, as well as the RMSE and the MSE improvement for both averaging procedures. The predictions based on the full BMA ($\mathcal{M}_{BMA}$) outperform the simplified method of Neufcourt et al. (2019) ($\mathcal{M}_{BMA(simple)}$) by

11% on the training dataset and 13% on the testing one, as measured by $\widehat{r}^2_{BMA}$. The lowest RMSE on the training dataset was attained by SLy4 and UNEDF1 respectively for AME2016 \ AME2003. This result should not discourage practitioner from using BMA posterior mean predictors, because the BMA methodology outlined in this paper allows for existence of a "best" model for a particular data domain. However, such a model does not account for modeling uncertainty whereas BMA does, and therefore the BMA posterior mean estimator performs consistently well irrespective of the dataset. In fact it attains the second lowest RMSE on both AME2003 and AME2016 \ AME2003.

Moreover, if we consider only a subset of the whole model space, the BMA attains the lowest RMSE. See Table 2.8 in Section 2.4.4 for the results with a restricted model space. Figure 2.3 shows the ECP of the averaged nuclear mass emulators. While it is not clear that the BMA has an improved ECP compared to each individual models, its ECP is certainly significantly better than the worst models and comparable to the models with highest fidelity.

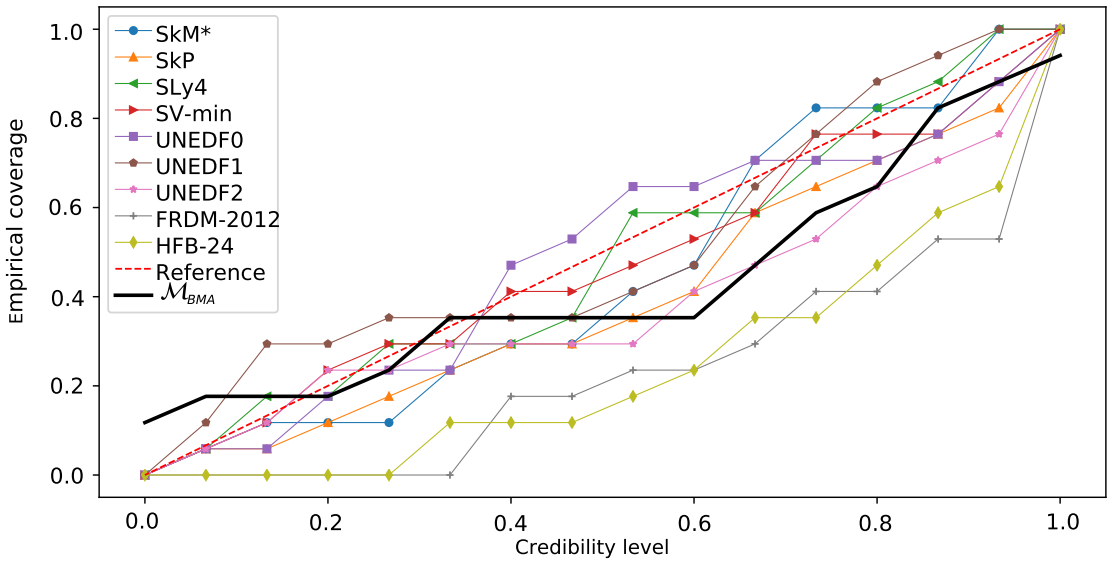

Figure 2.3: The ECPs calculated on the independent testing dataset (AME2016 \ AME2003).

2.3.3 Averaging of the Liquid Drop Model variants

In our second real-data example, we demonstrate the opportunities in nuclear theory offered by BMA through averaging of the LDM that has been optimized to various subsets of the nuclear domain. In the context of the following discussion, it is useful to clarify the notion of a "model". In this specific scenario, by model we understand the combination of the algebraic model formula, the dataset used for its parameter determination, and a statistical model that describes the error structure.

The parameters of the LDM are $(a_{\mathrm{vol}}, a_{surf}, a_{sym}, a_{\mathrm{C}})$ representing the volume, surface, symmetry and Coulomb energy, respectively. Because of its linearity and simplicity, the LDM has become a popular model for various statistical applications (Bertsch et al., 2005; Toivanen et al., 2008; Utama et al., 2016; Yuan, 2016; Bertsch and Bingham, 2017; Zhang et al., 2017; Cauchois et al., 2018; Shelley et al., 2014; Pastore, 2019).

To study the impact of the fitting domain on prediction accuracy, and UQ fidelity of nuclear mass models, we shall consider the experimental binding energies of 595 even-even nuclei of AME2003 (meaning both Z and N are even) divided into 3 domains according to Figure 2.4. Namely, we define the domain of *light* nuclei with $Z < 40$ and $N < 50$, *heavy* nuclei with $Z > 50$ and $N > 80$, and the *intermediate* domain $\mathcal{D}_{\mathcal{I}}$ consisting of the remaining even-even nuclei. To keep some of our results within computable ranges we will also consider 8 randomly selected nuclei in the *central* subset of the intermediate domain which we will denote $\mathcal{D}_{\mathcal{C}}$. By dividing nuclear domains according to A, we are trying to simulate the current theoretical strategy in modeling atomic nuclei: light nuclei are often described by different classes of models than heavy nuclei, with the intermediate domain being the testing ground for all approaches Nazarewicz (2016). Here we use, for testing, the same LDM expression in all domains. The models are distinguished merely by the fitting datasets.

In terms of these separated data domains, we consider four LDM variants fitted on specific regions of the nuclear landscape:

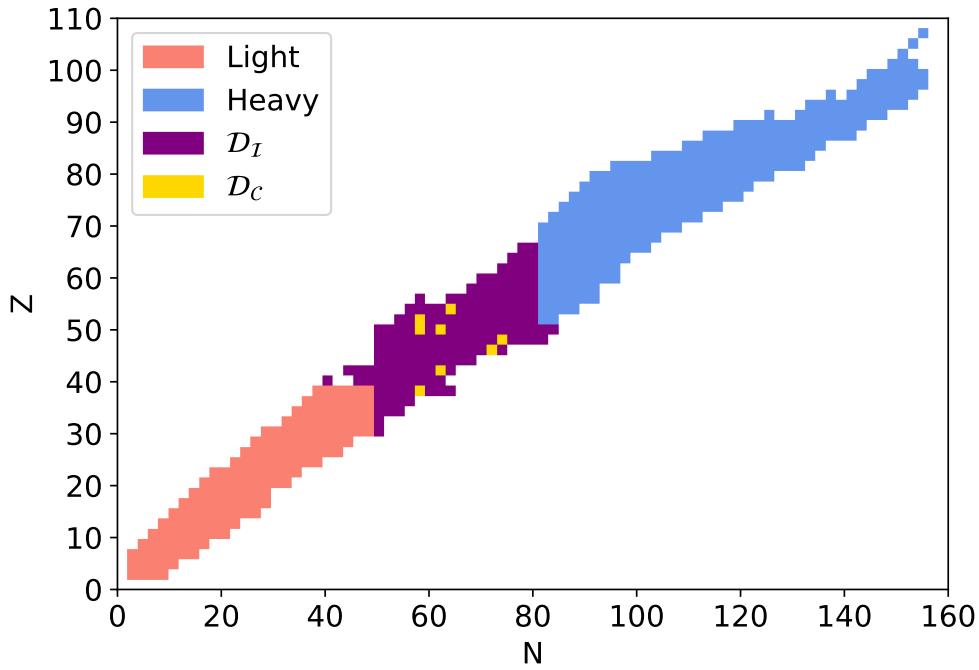

Figure 2.4: Even-even nuclei from AME2003 divided into the domains of light ($Z < 40$, $N < 50$), heavy ($Z > 50$, $N > 80$), and intermediate nuclei (remaining 155 nuclei). The subset of 8 randomly selected nuclei is also depicted (From Kejzlar et al. (2020)).

(i) LDM(A) – LDM fitted on all 595 even-even nuclei.

(ii) LDM(L) – LDM restricted to the light domain (153 nuclei).

(iii) LDM(H) – LDM restricted to the heavy domain (287 nuclei).

(iv) LDM(L + H) – LDM fitted on the both light and heavy domain (440 nuclei).

We emphasize that the intermediate domain $\mathcal{D}_\mathcal{I}$ (and $\mathcal{D}_\mathcal{C}$) is not used for training in variants (ii)-(iv), but kept aside as an independent testing domain where the different LDM variants compete. Thus we use the binding energies in the intermediate domain to evaluate the predictions and error bounds of these variants and their Bayesian averages. In short, this setup is designed to produce a scenario where two models, which have been optimized on their respective domains, compete to explain the data on a third disconnected domain.

Our statistical model for binding energies y_i is the standard

$$y_i = f_m(\boldsymbol{t}_i, \boldsymbol{\theta}) + \sigma\epsilon_i, \tag{2.24}$$

where the function $f_m(\boldsymbol{t}, \boldsymbol{\theta})$ represents the LDM prediction (1.1) with a given parameter vector $\boldsymbol{\theta} = (a_{\mathrm{vol}}, a_{surf}, a_{sym}, a_{\mathrm{C}})$ for a nucleus indexed by $\boldsymbol{t} = (Z, N)$. The errors are modeled as independent standard normal random variable ϵ_i with mean zero and unit variance, scaled by σ. For the LDM parameters $a_{\mathrm{vol}}, a_{surf}$ and a_{sym} we use independent normal prior distributions $\mathcal{N}(0, 100)$ with mean 0 and standard deviation 100, while for a_{C} we take $\mathcal{N}(0, 2)$. For σ we assume a gamma prior distribution Gamma(5,2) with shape parameter 5 and scale parameter 2. These are chosen to be *weakly informative*, i.e., distributions where hyperparameters are chosen to ensure that the prior distribution spans a much wider domain than the resulting posterior. Since the parameter estimation is not topic of this study, we refer the reader to Kejzlar et al. (2020) for more details about the posterior distributions of these parameters.

In this example, we wish to select a model's weight according to its predictive ability and also to avoid overfitting, in the same spirit as the approach implemented in Neufcourt et al. (2019, 2020a,b). To this end, we evaluate the evidence integrals over a set of binding energies $\boldsymbol{y}^*$ from the intermediate domain of Figure 2.4, which corresponds to integrating the posterior distribution of new predictions against the posterior distribution of the model parameters

$$p(\boldsymbol{y}^*|\boldsymbol{y}, \mathcal{M}_k) = \int p(\boldsymbol{y}^*|\boldsymbol{y}, \boldsymbol{\theta}_k, \sigma_k, \mathcal{M}_k) p(\boldsymbol{\theta}_k, \sigma_k|\boldsymbol{y}, \mathcal{M}_k)\, \mathrm{d}\boldsymbol{\theta}_k\, \mathrm{d}\sigma_k. \tag{2.25}$$

Given that posterior distribution of the parameters reflects the true distribution of the parameter more accurately than the prior, (2.25) more accurately represents the probability that $\mathcal{M}_k$ can explain data $\boldsymbol{y}$. To assess the impact of the number of evidence datapoints, we evaluate evidence integrals both on the full intermediate domain $\mathcal{D}_{\mathcal{I}}$ and a smaller central domain $\mathcal{D}_{\mathcal{C}}$.

The integral (2.25) can be estimated using the MC integration as

$$\widehat{p(\boldsymbol{y}^*|\boldsymbol{y},\mathcal{M}_k)} = \frac{1}{n_{MC}} \sum_{i=1}^{n_{MC}} p(\boldsymbol{y}^*|\boldsymbol{y},\boldsymbol{\theta}_k^{(i)},\sigma_k^{(i)},\mathcal{M}_k), \tag{2.26}$$

where $(\boldsymbol{\theta}_k^{(i)},\sigma_k^{(i)})$ are samples from the posterior distributions $p(\boldsymbol{\theta}_k,\sigma_k|\boldsymbol{y},\mathcal{M}_k)$.

Table 2.3 shows the posterior weights obtained under averaging scenarios with two (L and H) and three (L, H, and L+H) models. The corresponding RMSE values for individual models and the BMA posterior mean predictors are listed in Table 2.4.

		LDM(L)	LDM(H)	LDM(L+H)
$\mathcal{D}_\mathcal{I}$	BMA(L,H)	0.000	1.000	
	BMA(L,H,L+H)	0.000	0.000	1.000
$\mathcal{D}_\mathcal{C}$	BMA(L,H)	0.008	0.992	
	BMA(L,H,L+H)	0.002	0.255	0.743

Table 2.3: Posterior model weights under the averaging scenarios with two (L and H; left) and three (L, H, and L+H; right) models. The weights for the full intermediate domain of nuclei and the subset of 8 randomly selected nuclei are listed.

As expected, model (H) is selected in the two model variant, and the (L+H) variant dominates when it is included – this is true for both sets of evidence datasets $\mathcal{D}_\mathcal{C}$ and $\mathcal{D}_\mathcal{I}$. This is consistent with the RMSE of these models. It shall be emphasized that BMA performs a model selection in the two-model variant, where the RMSEs of the competing models are very different, and model averaging in the three-model variant, where the RMSE of (H) and (L+H) are close enough. Table 2.4 also shows how the RMSE of the BMA predictions compare with that of the individual models. In the two-model setup, BMA is very much like (H) and it has a similar RMSE. In the three-model setup, BMA performs much better than the worst model and very close to the best of the averaged models. When computed on the full test domain $\mathcal{D}_\mathcal{I}$, the RMSEs are systematically smaller for BMA than for all the individual models involved in the averaging (not considering LDM(A)). One may notice that the RMSE of BMA(L, H, L+H) is, perhaps unexpectedly, slightly worse than that of

LDM(L+H) on the small domain $\mathcal{D}_\mathcal{C}$. However, these values are based merely on 8 data points and should be viewed as a crude estimate of true predictive performance.

		LDM(A)	LDM(L)	LDM(H)	LDM(L+H)
$\mathcal{D}_\mathcal{I}$	$\mathcal{M}_k$	3.206	8.176	3.811	3.351
	BMA(L,H)			3.810	
	BMA(L,H,L+H)			3.223	
$\mathcal{D}_\mathcal{C}$	$\mathcal{M}_k$	1.930	6.825	3.292	1.881
	BMA(L,H)			3.300	
	BMA(L,H,L+H)			1.926	

Table 2.4: The RMSEs (in MeV) of the predictions from the 4 LDM variants as well as the values from BMA, calculated on the held-out data in the intermediate domain of even-even nuclei from AME2003.

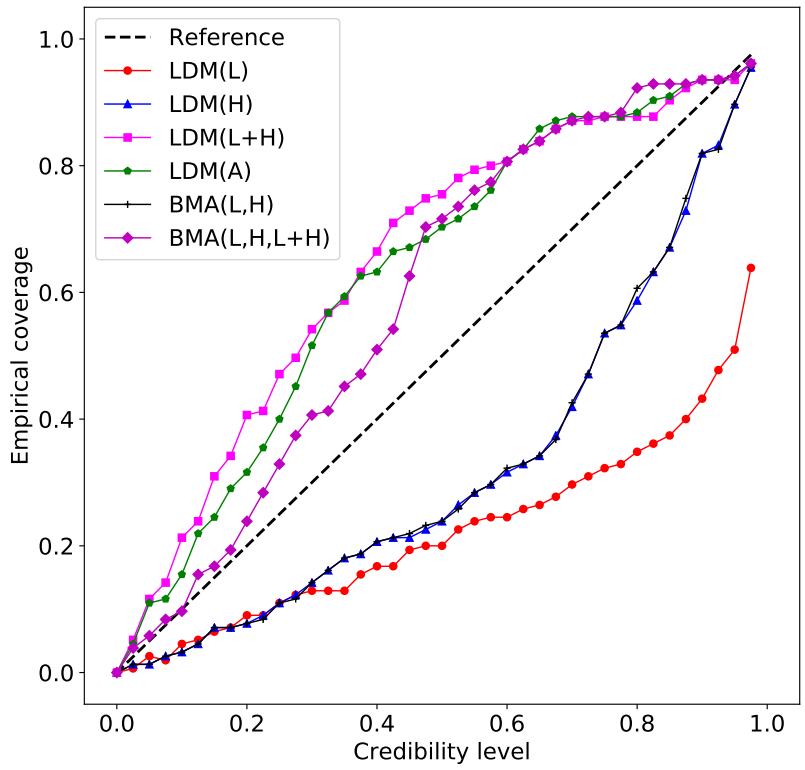

Figure 2.5: The ECPs for the four LDM variants used in our study and the averaging scenarios with two (L and H) and three models (L, H, and L+H) (From Kejzlar et al. (2020)).

Similarly to all the previous examples in Chapter 2, we also evaluate the models from UQ

quality perspective using the ECP curves. Figure 2.5 shows that the LDM variants fitted to the smaller domains (L or H) tend to underestimate the uncertainty of the predicted binding energies compared to the rather conservative UQ of the (L+H) variant and the LDM fitted to the entire AME2003 dataset. On the other hand, BMA(L,H,L+H) yields an ECP superior to all the LDM variants, including LDM(A), which aligns with our hypothesis that meaningful averaging can lead to an improved UQ.

2.3.4 Averaging of models with discrepant domains: a pedagogical example

In this example we study a simulated scenario where two models with $\boldsymbol{t}$-dynamics of the same order act in the opposite directions. We consider these models to be the realizations of GPs with means

$$m_i(\boldsymbol{t}) = \alpha_i \boldsymbol{t}^2 + \theta_i, \qquad i \in \{1, 2\}, \tag{2.27}$$

where $\alpha_1 = 0.5$ and $\alpha_2 = -0.5$, and θ_1 and θ_2 represent unknown parameters to be estimated. The covariance function used for the GPs is squared exponential kernel

$$k_i(\boldsymbol{t}, \boldsymbol{t}') = \eta_i^2 e^{-\frac{(\boldsymbol{t}-\boldsymbol{t}')^2}{2\ell_i^2}}, \qquad i \in \{1, 2\}. \tag{2.28}$$

The prior distributions for the unknown parameters $(\theta_i, \eta_i, \ell_i)$ are listed in Section 2.4.4. Overall, the two statistical models $\mathcal{M}_1$ and $\mathcal{M}_2$ considered here are given by the respective GPs and additive independent errors distributed according to $\mathcal{N}(0, \sigma_j)$ for $j = 1, 2$.

The two means (2.27) emulate a natural scenario of competition between models, similar to the proton potential example above, where we are uncertain about the nature of the physical law and resort to BMA in order to account for this uncertainty. To do so, we consider a synthetic dataset $\boldsymbol{y}$ of 18 observations drawn independently from $\mathcal{N}(0, 10^{-3})$ at input points $\boldsymbol{t} = \{\pm k, k = 1, 2, \dots 9\}$. Additionally, we study the impact of the domain correction by assigning a different training dataset $\boldsymbol{y}^{(k)}$ to the models $\mathcal{M}_1$ and $\mathcal{M}_2$, using seven different scenarios with proportions of shared observations (D_{shared}) ranging from 20%

to 80% according to the scheme in Table 2.5. Note the break of symmetry in the domain of $\boldsymbol{y}^{(k)}$ denoted by a circle, we shall refer to those as symmetric and asymmetric scenarios.

For each value of D_{shared}, we carried out the domain-corrected procedure detailed in Section 2.2 and computed the evidence integrals $p(\boldsymbol{y}^{(k)}|\mathcal{M}_k)$ as well as the corrective terms $p(\boldsymbol{y}^{(-k)}|\boldsymbol{y}^{(k)})$. Also note that the approximate computation of these terms (2.17) is more demanding than the computation of the evidence integrals, because it requires integration against the posterior distribution of parameters.

		Training dataset $\boldsymbol{y}^{(k)}$																
D_{shared}	Model	-9	-8	-7	-6	-5	-4	-3	-2	-1	1	2	3	4	5	6	7	8	9
0.2	$\mathcal{M}_1$	x	x	x	x	x	x	x	x	x	x								
	$\mathcal{M}_2$									x	x	x	x	x	x	x	x	x	x
0.3	$\mathcal{M}_1$	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗								
	$\mathcal{M}_2$								⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	
0.4	$\mathcal{M}_1$		x	x	x	x	x	x	x	x	x	x							
	$\mathcal{M}_2$								x	x	x	x	x	x	x	x	x	x	
0.5	$\mathcal{M}_1$		⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗							
	$\mathcal{M}_2$							⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗		
0.6	$\mathcal{M}_1$			x	x	x	x	x	x	x	x	x	x						
	$\mathcal{M}_2$							x	x	x	x	x	x	x	x	x	x		
0.7	$\mathcal{M}_1$			⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗						
	$\mathcal{M}_2$						⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗			
0.8	$\mathcal{M}_1$				x	x	x	x	x	x	x	x	x	x					
	$\mathcal{M}_2$						x	x	x	x	x	x	x	x	x	x			

Table 2.5: Scheme depicting the observations contained in the training dataset of the models according to the proportion of shared data. The crosses mark the values contained in the domain of each model.

Table 2.6 gives a quantitative summary of the simulation results in the asymmetric scenario, where the impact of the domain correction is stronger. See Table 2.9 in Section 2.4.4 for the symmetric case, where the impact of the domain correction is minor due to the symmetry of training data and the response functions. The RMSE was calculated based on the set of common observations ($\boldsymbol{t} \leq 5$). $BMA(Q)$ and $BMA(Q_0)$ represent respectively the domain corrected BMA and the BMA with independent model domains. Q denotes the posterior odds ratio $p(\boldsymbol{y}^{(-1)}|\boldsymbol{y}^{(1)})p(\mathcal{M}_1|\boldsymbol{y}^{(1)})/[p(\boldsymbol{y}^{(-2)}|\boldsymbol{y}^{(2)})p(\mathcal{M}_2|\boldsymbol{y}^{(2)})]$ used to draw samples

D_{shared}	Model	RMSE	$p(\boldsymbol{y}^{(k)}\|\mathcal{M}_k)$	$p(\boldsymbol{y}^{(-k)}\|\boldsymbol{y}^{(k)})$	Q_0	Q	$\widehat{r}^2_{BMA}$
0.3	$\mathcal{M}_1$	4.69	$2.69 \cdot 10^{-21}$	$2.13 \cdot 10^{-16}$	0.03	0.80	**0.495**
	$\mathcal{M}_2$	4.68	$7.78 \cdot 10^{-20}$	$9.25 \cdot 10^{-18}$			**0.494**
	$\mathcal{M}_{BMA(Q_0)}$	4.53	-	-			-
	$\mathcal{M}_{BMA(Q)}$	3.33	-	-			-
0.5	$\mathcal{M}_1$	4.63	$7.79 \cdot 10^{-20}$	$5.44 \cdot 10^{-13}$	0.02	0.61	**0.512**
	$\mathcal{M}_2$	4.38	$3.29 \cdot 10^{-18}$	$2.12 \cdot 10^{-14}$			**0.456**
	$\mathcal{M}_{BMA(Q_0)}$	4.29	-	-			-
	$\mathcal{M}_{BMA(Q)}$	3.23	-	-			-
0.7	$\mathcal{M}_1$	4.36	$3.23 \cdot 10^{-18}$	$1.13 \cdot 10^{-8}$	0.02	0.72	**0.593**
	$\mathcal{M}_2$	3.62	$1.45 \cdot 10^{-16}$	$3.49 \cdot 10^{-10}$			**0.410**
	$\mathcal{M}_{BMA(Q_0)}$	3.54	-	-			-
	$\mathcal{M}_{BMA(Q)}$	2.78	-	-			-

Table 2.6: Summary of the domain corrected BMA analysis in the asymmetric case of the pedagogical example.

from the mixture distribution (2.1) and Q_0 is the ratio $p(\mathcal{M}_1|\boldsymbol{y}^{(1)})/p(\mathcal{M}_2|\boldsymbol{y}^{(2)})$. The MSE improvement $\widehat{r}^2_{BMA}$ is with respect to the BMA with domain correction.

As expected from our construction, BMA leads to a spectacular decrease of the MSE by about 50%. The BMA posterior mean predictor outperforms consistently the individual models, at all proportions of the shared training data. As the overlap between the two model domains increases, the RMSEs consistently decrease. The same observations hold in the symmetric case. The domain corrected BMA has consistently lower RMSE than the BMA with independent model domains across D_{shared}. We observe that the values of the corrective factors increase exponentially towards 1 as D_{shared} increases; indeed the extreme case $D_{shared} = 1$, where both models are defined on the same domain, corresponds to the classical BMA framework. The odds ratios stay expectedly close to 1, due to the fact that the deviations from out-of-domain data are comparable across the models; still the domain-corrected odds ratio Q has a consistently larger variability than Q_0, the difference vanishes as the proportion D_{shared} of data shared between the two models increases.

2.4 Technical details and supplementary results

2.4.1 A simple example of evidence integral with closed form solution

Let us suppose the following set of K models of experimental observations $(\boldsymbol{t}_i, y_i)_{i=1}^n$

$$y_i = f_k(\boldsymbol{t}_i) + \sigma_k \epsilon_i, \qquad k = 1, \ldots, K,$$

where $y_k(\boldsymbol{t})$ are known deterministic functions, ϵ_i are independent identically distributed standard normal random variables, and $\sigma_k^2 \sim$ Inv-Gamma(α_k, β_k). We can calculate the evidence integrals (2.3) explicitly as

$$\begin{aligned} p(\boldsymbol{y}|\mathcal{M}_k) &= \int_0^\infty \frac{1}{(2\pi\sigma_k^2)^{\frac{n}{2}}} e^{\left(-\frac{\sum_i (y(\boldsymbol{t}_i) - y_k(\boldsymbol{t}_i))^2}{2\sigma_k^2}\right)} \frac{\beta_k^{\alpha_k}}{\Gamma(\alpha_k)} \frac{1}{(\sigma_k^2)^{\alpha_k+1}} e^{\left(-\frac{\beta_k}{\sigma_k^2}\right)} \, \mathrm{d}\sigma_k^2 \\ &= \frac{\beta_k^{\alpha_k}}{(2\pi)^{\frac{n}{2}}\Gamma(\alpha_k)} \int_0^\infty \frac{1}{(\sigma_k^2)^{\frac{n}{2}+\alpha_k+1}} e^{\left(-\frac{\frac{1}{2}\sum_i (y(\boldsymbol{t}_i) - y_k(\boldsymbol{t}_i))^2 + \beta_k}{\sigma_k^2}\right)} \, \mathrm{d}\sigma_k^2 \\ &= \frac{\beta_k^{\alpha_k}\Gamma(\frac{n}{2} + \alpha_k)}{(2\pi)^{\frac{n}{2}}\Gamma(\alpha_k)(\frac{1}{2}\sum_i (y(\boldsymbol{t}_i) - y_k(\boldsymbol{t}_i))^2 + \beta_k)^{\frac{n}{2}+\alpha_k}}. \end{aligned}$$

2.4.2 Proofs

Proof of Lemma 1.

First, the standard factorization identities give the following expression:

$$\begin{aligned} &(y^* - \widehat{y^*})^2 - (y^* - \lambda_1 \widehat{y_1^*} - \lambda_2 \widehat{y_2^*})^2 \\ &= [2y^* - (\lambda_1 + p_1)\widehat{y_1^*} - (\lambda_2 + p_2)\widehat{y_2^*}][(\lambda_1 - p_1)\widehat{y_1^*} + (\lambda_2 - p_2)\widehat{y_2^*}]. \end{aligned}$$

To get the result of the Lemma, we now take the expectation of the expression above conditioned on $\boldsymbol{d}$ and notice that the right hand side is, with the exception of y^*, $\boldsymbol{d}$-measurable.

$$\begin{aligned} &E[(y^* - \widehat{y^*})^2|\boldsymbol{d}] - E[(y^* - \lambda_1 \widehat{y_1^*} - \lambda_2 \widehat{y_2^*})^2|\boldsymbol{d}] \\ &= [(p_1 - \lambda_1)\widehat{y_1^*} + (p_2 - \lambda_2)\widehat{y_2^*}][(\lambda_1 - p_1)\widehat{y_1^*} + (\lambda_2 - p_2)\widehat{y_2^*}]. \end{aligned}$$

□

2.4.3 Supplement for the general case of K models

Let us consider a dataset $\boldsymbol{d}$ and K models $\mathcal{M}_1, \ldots, \mathcal{M}_K$, and assume that each model $\mathcal{M}_k$ is defined on a subset $\boldsymbol{t}^{(k)}$ of the data inputs. Denote also $\boldsymbol{d}^{(k)}$ the subset of $\boldsymbol{d}$ corresponding to inputs $\boldsymbol{t}^{(k)}$ and $\boldsymbol{d}^{(-k)}$ the complementary subset as well as $\boldsymbol{d}^{(\oslash)} := \bigcap_k \boldsymbol{d}^{(k)}$. Suppose that all data locations are in the domain of at least one model so that $\boldsymbol{d} = \bigcup_k \boldsymbol{d}^{(k)}$.

Note that if the datasets are disjoint, there is simply no basis to compare the models. Given a set of models, one can define a unique *minimal* equivalence relationship $\star$ on the models (i.e. with a number of equivalence classes maximal) satisfying $\mathcal{M} \star \mathcal{M}'$ if $\mathcal{M}$ and $\mathcal{M}'$ share at least one data point, i.e. $\mathcal{M} \star \mathcal{M}'$ if and only if there exists $r \geq 0$ and a sequence of models $\mathcal{M} =: \mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_r := \mathcal{M}'$ such that $\mathcal{M}_i$ and $\mathcal{M}_{i+1}$ have a common data point for each $0 \leq i < r$. The computation of the posterior weights of the models can then be done within each class of equivalence, and we will therefore assume that there is only one such equivalence class.

In the standard BMA where all models share $\boldsymbol{d}$, one can express the posterior probabilities on the models $p(\mathcal{M}_k|\boldsymbol{d})$ using the Bayes formula

$$p(\mathcal{M}_k|\boldsymbol{d}) \propto_k p(\boldsymbol{d}|\mathcal{M}_k)p(\mathcal{M}_k) \tag{2.29}$$

and estimate the evidence integral $p(\boldsymbol{d}|\mathcal{M}_k)$ as at the beginning of this chapter. In our situation, however, the model $\mathcal{M}_k$ provides an expression $p(\boldsymbol{d}^{(k)}|\mathcal{M}_k)$ instead of $p(\boldsymbol{d}|\mathcal{M}_k)$, so that the standard procedure cannot be applied without a further argument.

Starting from (2.29), we expand $p(\boldsymbol{d}|\mathcal{M}_k)$ similarly to the two-model (2.13) case as

$$p(\boldsymbol{d}|\mathcal{M}_k) = p(\boldsymbol{d}^{(k)}, \boldsymbol{d}^{(-k)}|\mathcal{M}_k) = p(\boldsymbol{d}^{(-k)}|\boldsymbol{d}^{(k)}, \mathcal{M}_k)p(\boldsymbol{d}^{(k)}|\mathcal{M}_k). \tag{2.30}$$

Instead of setting $p(\boldsymbol{d}^{(-k)}|\boldsymbol{d}^{(k)}, \mathcal{M}_k) = 1$ which advantages models that withhold their predictions at difficult locations (see the example scenarios and discussion in Section 2.2), our domain-corrected BMA estimates

$$p(\boldsymbol{d}^{(-k)}|\boldsymbol{d}^{(k)}, \mathcal{M}_k) = p(\boldsymbol{d}^{(-k)}|\boldsymbol{d}^{(k)}). \tag{2.31}$$

This yields the evidence and the posterior weights given respectively by

$$p(\boldsymbol{d}|\mathcal{M}_k) = p(\boldsymbol{d}^{(-k)}|\boldsymbol{d}^{(k)})p(\boldsymbol{d}^{(k)}|\mathcal{M}_k) \tag{2.32}$$

$$p(\mathcal{M}_k|\boldsymbol{d}) \propto_k p(\boldsymbol{d}^{(-k)}|\boldsymbol{d}^{(k)})p(\boldsymbol{d}^{(k)}|\mathcal{M}_k)p(M_k), \tag{2.33}$$

similarly to the two-model case. All that is left now is to evaluate the $p(\boldsymbol{d}^{(-k)}|\boldsymbol{d}^{(k)})$.

Let $\mathcal{S}$ be the set of q indices of the models that constrain $\boldsymbol{d}^{(-k)}$, we can compute $p(\boldsymbol{d}^{(-k)}|\boldsymbol{d}^{(k)})$ by conditioning with respect to the models with indices in $\mathcal{S}$. Namely,

$$\begin{aligned} p(\boldsymbol{d}^{(-k)}|\boldsymbol{d}^{(k)}) &= p(\boldsymbol{d}^{(-k)}|\boldsymbol{d}^{(k)}, \cup_l^q[\mathcal{M} = \mathcal{M}_l; l \in \mathcal{S}]) \\ &= \frac{1}{\sum_{l\in\mathcal{S}} p(\mathcal{M}_l, \boldsymbol{d}^{(k)})} \sum_{l\in\mathcal{S}} p(\boldsymbol{d}^{(-k)}|\boldsymbol{d}^{(k)}, \mathcal{M}_l)p(\mathcal{M}_l, \boldsymbol{d}^{(k)}) \\ &= \frac{1}{\sum_{l\in\mathcal{S}} p(\mathcal{M}_l|\boldsymbol{d}^{(k)})} \sum_{l\in\mathcal{S}} p(\boldsymbol{d}^{(-k)}|\boldsymbol{d}^{(k)}, \mathcal{M}_l)p(\mathcal{M}_l|\boldsymbol{d}^{(k)}) \end{aligned}$$

The simplest case is when $\boldsymbol{d}^{(-k)}$ is *non-divisible*, in the sense that for every $l \in \mathcal{S}$, we have $\boldsymbol{d}^{(-k)} \subset \boldsymbol{d}^{(l)}$ or $\boldsymbol{d}^{(-k)} \cap \boldsymbol{d}^{(l)} = \emptyset$. Then $p(\boldsymbol{d}^{(-k)}|\boldsymbol{d}^{(k)}, \mathcal{M}_l)$ is given by (2.17) and the sum above have explicit expressions. In the general case, some models may be defined only on a strict subset of $y^{(-k)}$. In that case we have

$$\begin{aligned} p(\boldsymbol{d}^{(-k)}|\boldsymbol{d}^{(k)}, \mathcal{M}_l) &= p(\boldsymbol{d}^{(-k)\cap(l)}, \boldsymbol{d}^{(-k)\cap(-l)}|\boldsymbol{d}^{(k)}, \mathcal{M}_l) \\ &= p(\boldsymbol{d}^{(-k)\cap(l)}|\boldsymbol{d}^{(k)}, \mathcal{M}_l)p(\boldsymbol{d}^{(-k)\cap(-l)}|\boldsymbol{d}^{(k)}, \boldsymbol{d}^{(-k)\cap(l)}, \mathcal{M}_l) \\ &= p(\boldsymbol{d}^{(-k)\cap(l)}|\boldsymbol{d}^{(k)\cap(l)}, \mathcal{M}_l)p(\boldsymbol{d}^{(-k)\cap(-l)}|\boldsymbol{d}^{(l)}, \boldsymbol{d}^{(k)\cap(-l)}) \end{aligned}$$

The first term can be explicitly computed as (2.17) and $p(\boldsymbol{d}^{(-k)\cap(-l)}|\boldsymbol{d}^{(l)}, \boldsymbol{d}^{(k)\cap(-l)})$ can be computed recursively. For practical purposes, it is important to notice that the complexity of the underlying algorithm is at most exponential in the number of models, where each iteration requires the computation of a posterior predictive distributions of decreasing subsets of the data given decreasing subsets of the data, posterior model weights given decreasing subset of the data, and N computations of corrective likelihoods, where N is the number of non-divisible subsets.

2.4.4 Supplement for the examples and applications

Averaging of nuclear mass emulators in the Ca region. In this real data application, we follow the experimental framework of Neufcourt et al. (2019). Given a (known) theoretical nuclear model $f_m(\boldsymbol{t})$ for the one- and two-neutron separation energies, we consider the relationship between the experimental observations y_i and the nuclear model as

$$y_i = f_m(\boldsymbol{t}) + \delta(\boldsymbol{t}),$$

for $\boldsymbol{t} := (Z, N)$ ranging over the two-dimensional nuclear domain. We model the systematic discrepancy δ as the GP

$$\delta(Z, N) \sim \mathcal{GP}(0, k_{\eta,\ell}\{(Z, N), (Z', N')\}),$$

with the mean 0 and the quadratic exponential covariance kernel with three parameters

$$k_{\eta,\ell}\{(Z, N), (Z', N')\} = \eta^2 e^{-\frac{(Z-Z')^2}{2\ell_Z^2} - \frac{(N-N')^2}{2\ell_N^2}},$$

with independent gamma prior distributions with shape and scale parameters

$$\eta, \ell_Z, \ell_N \sim \text{Gamma}(a, b),$$

where $b = 1$ and a respectively set to 0.8, 0.5 and 1.8. Note that this corresponds to the framework of Bayesian calibration of imperfect computer models described in Section 1.2. The only difference is that here we consider models that were already calibrated, and we don't explicitly model the experimental error (the is common practice in the nuclear physics community since the experimental error is negligible compared with the systematic error for state-of-the-art models). See supplemental material to Neufcourt et al. (2019) for exhaustive description of the framework.

Averaging of models with discrepant domains: a pedagogical example. Table 2.9 gives a quantitative summary of the simulation results in the symmetric scenario. Figure 2.6

	Sample Size			
	$\boldsymbol{S_{1n}}$ (odd N)		$\boldsymbol{S_{2n}}$ (even N)	
Dataset	even Z	odd Z	even Z	odd Z
AME2003	41	31	39	28
AME2016 \ AME2003	3	3	3	5

Table 2.7: Sample size breakdown for the training (AME2003) and the testing (AME2016 \ AME2003) datasets of nuclear separation energies in the Ca region according to Z and N parities.

	Model posterior weights				Errors			
	$\boldsymbol{S_{1n}}$ (odd N)		$\boldsymbol{S_{2n}}$ (even N)		Training		Testing	
Model	even Z	odd Z	even Z	odd Z	RMSE	$\widehat{r}^2_{BMA}$	RMSE	$\widehat{r}^2_{BMA}$
SkM*	0.000	0.001	0.000	0.000	0.142	0.375	0.925	0.413
FRDM-2012	1.000	0.997	0.900	0.399	0.114	0.031	0.808	0.231
HFB-24	0.000	0.002	0.100	0.601	0.146	0.405	0.806	0.227
$\mathcal{M}_{BMA}$					0.112	-	0.709	-

Table 2.8: The model posterior weights, RMSE (in MeV) and MSE improvement calculated on both the training (AME2003) and the testing (AME2016 \ AME2003) datasets for 3 nuclear mass models.

(asymmetric design) and Figure 2.7 (symmetric design) show the posterior mean predictions for $\mathcal{M}_1$, $\mathcal{M}_2$, domain corrected BMA $\mathcal{M}_{BMA(Q)}$, and BMA with independent model domains $\mathcal{M}_{BMA(Q_0)}$. These were obtained for the pedagogical example 2.3.4 using the domain correction developed in Section 2.2. The RMSE for both $BMA(Q_0)$ and $BMA(Q)$ is almost identical here (up to a roundoff error) due to the symmetric nature of both the training dataset $\boldsymbol{y}^{(k)}$ and the response functions.

The prior distributions used in the example were

$$\theta_i \sim \mathcal{N}(0,1),$$

$$\sigma_i^2 \sim \text{Inv-Gamma}(10,1),$$

$$\ell_i \sim \text{Gamma}(1,10),$$

$$\eta_i \sim \text{Inv-Gamma}(10,1),$$

where both the gamma and the inverse gamma distributions are parametrized in terms of the shape and the rate parameters for $i \in \{1, 2\}$.

D_{shared}	Model	RMSE	$p(\boldsymbol{y}^{(k)}\|\mathcal{M}_k)$	$p(\boldsymbol{y}^{(-k)}\|\boldsymbol{y}^{(k)})$	Q_0	Q	$\widehat{r}^2_{BMA}$
0.2	$\mathcal{M}_1$	4.69	$2.78 \cdot 10^{-21}$	$1.98 \cdot 10^{-19}$	1.02	0.96	**0.512**
	$\mathcal{M}_2$	4.58	$2.73 \cdot 10^{-21}$	$2.11 \cdot 10^{-19}$			**0.488**
	$\mathcal{M}_{BMA(Q_0)}$	3.28	-	-			-
	$\mathcal{M}_{BMA(Q)}$	3.28	-	-			-
0.4	$\mathcal{M}_1$	4.64	$7.99 \cdot 10^{-20}$	$4.33 \cdot 10^{-16}$	1.01	1.10	**0.511**
	$\mathcal{M}_2$	4.53	$7.95 \cdot 10^{-20}$	$3.96 \cdot 10^{-16}$			**0.486**
	$\mathcal{M}_{BMA(Q_0)}$	3.24	-	-			-
	$\mathcal{M}_{BMA(Q)}$	3.25	-	-			-
0.6	$\mathcal{M}_1$	4.37	$3.32 \cdot 10^{-18}$	$8.59 \cdot 10^{-12}$	1.01	1.11	**0.504**
	$\mathcal{M}_2$	4.33	$3.29 \cdot 10^{-18}$	$7.84 \cdot 10^{-12}$			**0.495**
	$\mathcal{M}_{BMA(Q_0)}$	3.07	-	-			-
	$\mathcal{M}_{BMA(Q)}$	3.08	-	-			-
0.8	$\mathcal{M}_1$	3.61	$1.45 \cdot 10^{-16}$	$2.99 \cdot 10^{-6}$	1.02	1.03	**0.509**
	$\mathcal{M}_2$	3.56	$1.42 \cdot 10^{-16}$	$2.98 \cdot 10^{-6}$			**0.495**
	$\mathcal{M}_{BMA(Q_0)}$	2.53	-	-			-
	$\mathcal{M}_{BMA(Q)}$	2.53	-	-			-

Table 2.9: Summary of the domain corrected BMA analysis in the symmetric case of the pedagogical example.

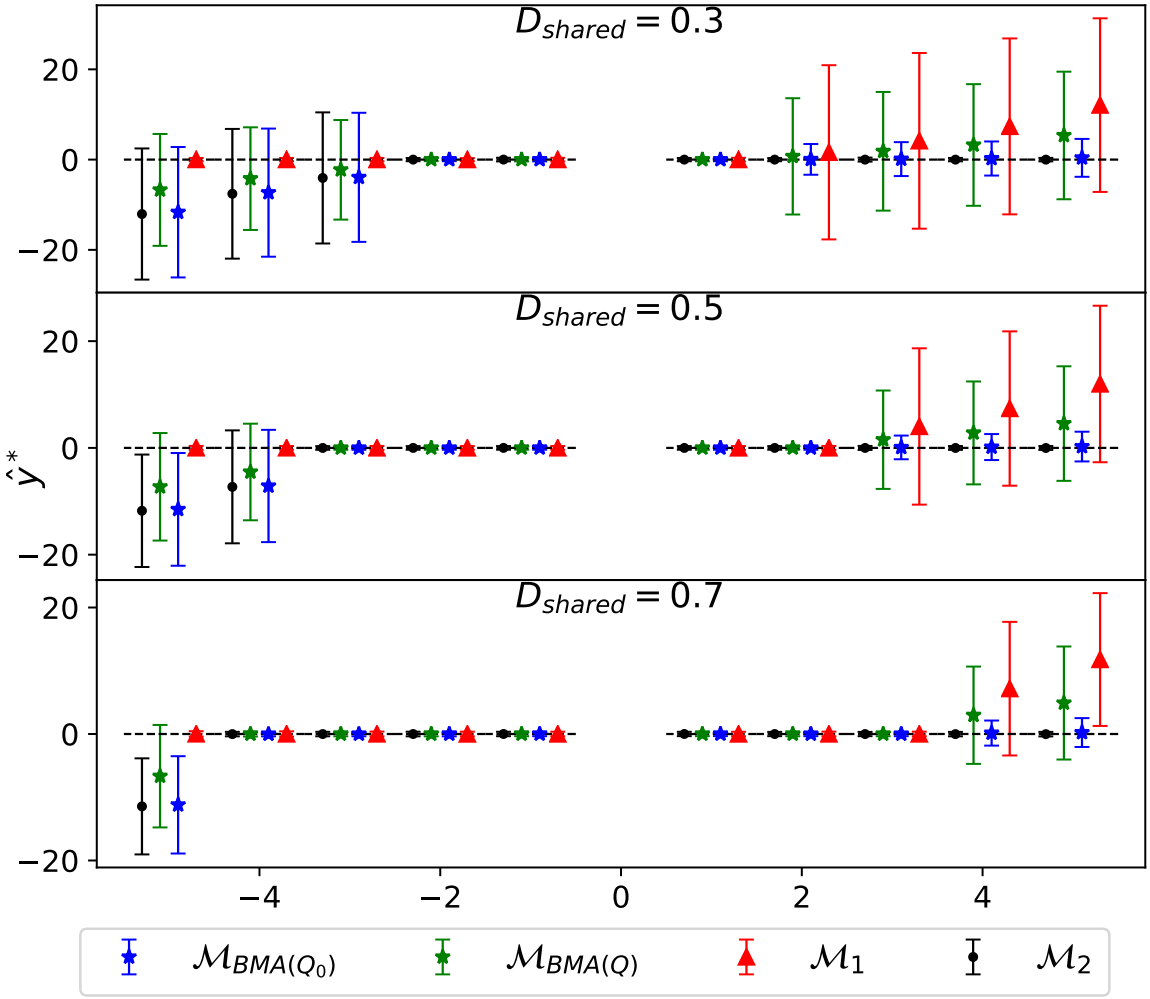

Figure 2.6: Posterior mean predictions (with 68% HPD credible intervals) for the 10 observations $\boldsymbol{y}$ for the two models in (2.27) as well as their BMA, with the domain correction and with the assumption of independent model domains. This is the asymmetric case. The dashed line segments represent the translated values of the original observations.

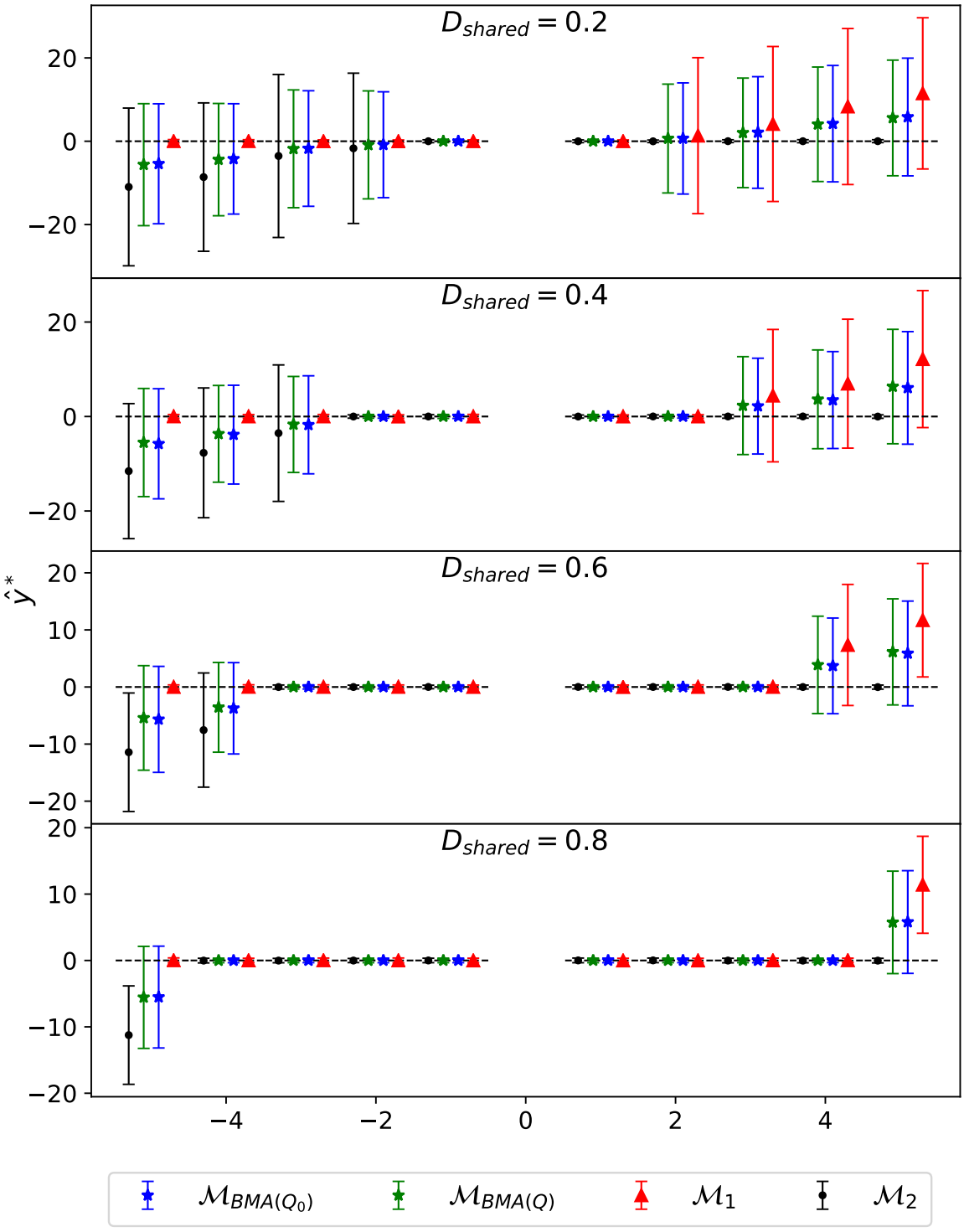

Figure 2.7: Posterior mean predictions (with 68% HPD credible intervals) for the 10 observations $\boldsymbol{y}$ for the two models in (2.27) as well as their BMA, with the domain correction and with the assumption of independent model domains. This is the symmetric case. The dashed line segments represent the translated values of the original observations.

CHAPTER 3

AN EFFICIENT ALGORITHM FOR BAYESIAN CALIBRATION OF COMPUTER MODELS VIA VARIATIONAL INFERENCE

With the ever-growing access to high performance computing in scientific communities, the use of computational models proliferates to solve complex problems in many scientific applications such as nuclear physics and climate research. An important class of such problems is making predictions, in order to aid the cycle of the scientific process. In particular, our task is to establish statistically principled predictions of new values $\boldsymbol{y}^*$ of a physical process ζ using a computer model f_m and a set of observations $\boldsymbol{y} = (y_1, \ldots, y_n)$ from this process. We would also like to account for various sources of uncertainty associated with individual models (see Section 1.1 for detailed discussion on UQ of computer models). The general framework that we shall follow and allows for predictions with UQ is called Bayesian calibration. It was originally developed by Kennedy and O'Hagan (2001) with extensions provided by Higdon et al. (2005, 2008); Bayarri et al. (2007); Plumlee (2017, 2019); Gu and Wang (2018) and Xie and Xu (2020), to name a few.

Formally, let $\boldsymbol{y} = (y_1, \ldots, y_n)$ be observations of a physical process $\zeta(\boldsymbol{t}_i)$ depending on a known set of inputs $\boldsymbol{t}_i \in \boldsymbol{\Omega} \subset \mathbb{R}^p$. Assume that y_i follows

$$y_i = \zeta(\boldsymbol{t}_i) + \sigma\epsilon_i, \tag{3.1}$$

where σ represent the scale of observation error $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$. As a mathematical description of ζ, we consider a computer model f_m defined as the mapping $(\boldsymbol{t}, \boldsymbol{\theta}) \mapsto f_m(\boldsymbol{t}, \boldsymbol{\theta})$ which depends on an additional set of inputs $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$ that we call calibration parameters. These are fixed but unknown quantities representing fundamental properties of the physical process that cannot be directly measured or controlled in an experiment. We assume a single value of calibration parameter $\boldsymbol{\theta}$ to be common among all the observations y_i and all the future instances of the physical process.

As we discussed in Chapter 1, a computer model is an imperfect description of the reality, and there often exists some systematic discrepancy (error) between the model and the physical process. To this extent, we assume that ζ satisfies $\zeta(\boldsymbol{t}) = f_m(\boldsymbol{t}, \boldsymbol{\theta}) + \delta(\boldsymbol{t})$, where $\delta(\boldsymbol{t})$ is the systematic discrepancy of the model whose form is generally unknown. The complete statistical model then reads as

$$y_i = f_m(\boldsymbol{t}_i, \boldsymbol{\theta}) + \delta(\boldsymbol{t}) + \sigma\epsilon_i. \tag{3.2}$$

The systematic discrepancy is modeled non-parametrically using a Gaussian process (GP) with the mean function $m_\delta(\boldsymbol{t})$ and the covariance function $k_\delta(\boldsymbol{t}, \boldsymbol{t}')$:

$$\delta(\boldsymbol{t}) \sim \mathcal{GP}(m_\delta(\boldsymbol{t}), k_\delta(\boldsymbol{t}, \boldsymbol{t}')). \tag{3.3}$$

The definition of a GP with examples is provided in Section 1.2.

In addition to the computer model being imperfect, it is often too expensive in terms of both computational time and memory to be used directly for inference. A common remedy is to consider the computer model as a realization of a GP with the mean function $m_f(\boldsymbol{t}, \boldsymbol{\theta})$ and the covariance function $k_f((\boldsymbol{t}, \boldsymbol{\theta}), (\boldsymbol{t}', \boldsymbol{\theta}'))$:

$$f_m(\boldsymbol{t}, \boldsymbol{\theta}) \sim \mathcal{GP}(m_f(\boldsymbol{t}, \boldsymbol{\theta}), k_f((\boldsymbol{t}, \boldsymbol{\theta}), (\boldsymbol{t}', \boldsymbol{\theta}'))). \tag{3.4}$$

In this situation, we generate an additional synthetic dataset of model runs $\boldsymbol{z} = (z_1, \ldots, z_s)$ over a fixed grid of inputs $\{(\widetilde{\boldsymbol{t}}_1, \widetilde{\boldsymbol{\theta}}_1), \ldots, (\widetilde{\boldsymbol{t}}_s, \widetilde{\boldsymbol{\theta}}_s)\}$ selected using a space-filling design such as a uniform or Latin hypercube design (Morris and Mitchell, 1995). The complete dataset $\boldsymbol{d}$ therefore consists of n observations y_i from the physical process ζ and s evaluations z_j of the computer model f_m, i.e. $\boldsymbol{d} = (d_1, \ldots, d_{n+s}) := (\boldsymbol{y}, \boldsymbol{z})$, and follows the multivariate normal distribution

$$\boldsymbol{d}|\boldsymbol{\phi} \sim \mathcal{N}(M(\boldsymbol{\phi}), K(\boldsymbol{\phi})), \tag{3.5}$$

where $\boldsymbol{\phi} = (\boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma)$ is the set of all unknown parameters with $\boldsymbol{\gamma}$ denoting the set of hyper-parameters of the GPs' mean and covariance functions. $M(\boldsymbol{\phi})$ (1.5) is the mean vector and $K(\boldsymbol{\phi})$ (1.6) is the covariance matrix given by the GPs' specifications.

Under this framework, the Bayesian predictions of $\boldsymbol{y}^*$ are given by the posterior predictive distribution $p(\boldsymbol{y}^*|\boldsymbol{d})$, namely

$$p(\boldsymbol{y}^*|\boldsymbol{d}) = \int p(\boldsymbol{y}^*|\boldsymbol{d}, \boldsymbol{\phi}) p(\boldsymbol{\phi}|\boldsymbol{d}) \, \mathrm{d}\boldsymbol{\phi}. \tag{3.6}$$

The conditional density $p(\boldsymbol{y}^*|\boldsymbol{d}, \boldsymbol{\phi})$ is a multivariate normal density given by the statistical model (1.3) and the specification of GPs (the explicit form is provided in Section 3.4). The posterior distribution of the unknown parameters $p(\boldsymbol{\phi}|\boldsymbol{d})$ is given by the Bayes' theorem

$$p(\boldsymbol{\phi}|\boldsymbol{d}) = \frac{p(\boldsymbol{d}|\boldsymbol{\phi}) p(\boldsymbol{\phi})}{\int p(\boldsymbol{d}|\boldsymbol{\phi}) p(\boldsymbol{\phi}) \, \mathrm{d}\boldsymbol{\phi}}. \tag{3.7}$$

The term "calibration" in the Bayesian paradigm includes both an estimation of $\boldsymbol{\phi}$ and a full evaluation of uncertainty for every parameter under a prior uncertainty expressed by $p(\boldsymbol{\phi})$. It is also worth noting that the posterior predictive density is rarely computed directly from (3.6). Instead, we first generate samples $\boldsymbol{\phi}^{(1)}, \ldots, \boldsymbol{\phi}^{(M)}$ from $p(\boldsymbol{\phi}|\boldsymbol{d})$ and then obtain samples $\boldsymbol{y}^{*(1)}, \ldots, \boldsymbol{y}^{*(M)}$ so that $\boldsymbol{y}^{*(i)} \sim p(\boldsymbol{y}^*|\boldsymbol{d}, \boldsymbol{\phi}^{(i)})$, $i = 1, \ldots, M$. The posterior predictive density is then approximated using the empirical density of samples $\boldsymbol{y}^{*(1)}, \ldots, \boldsymbol{y}^{*(M)}$.

As a consequence of this simple two-step algorithm, we are interested in effective sampling (approximation) from the posterior distribution $p(\boldsymbol{\phi}|\boldsymbol{d})$. *This becomes quickly infeasible with the increasing size of datasets, number of parameters, and model complexity.* Traditional MCMC methods that approximate $p(\boldsymbol{\phi}|\boldsymbol{d})$—such as the MH algorithm or more advanced ones including the Hamiltonian Monte Carlo or the NUTS—typically fail because of the computational costs associated with the evaluation of $p(\boldsymbol{d}|\boldsymbol{\phi})$. The conventional approaches to scalable Bayesian inference are in general not applicable here because of the highly correlated structure of $K(\boldsymbol{\phi})$ or the nature of calibration itself. Indeed, parallelization of MCMC (Neiswanger et al., 2014) works in the case of and independent $\boldsymbol{d}$, and GP approximation methods are developed in the context of regression problems (Quiñonero-Candela and Rasmussen, 2005; Titsias, 2009; Bauer et al., 2016).

This chapter presents a scalable and statistically principled approach to Bayesian calibration of computer models. We offer an alternative approximation to posterior densities using

variational Bayesian inference (VBI), which originated as a machine learning algorithm that approximates a target density through optimization. Statisticians and computer scientists (starting with Peterson and Anderson (1987); Jordan et al. (1999)) have been widely using variational techniques because they tend to be faster and easier to scale to massive datasets. Moreover, the recently published frequentist consistency of variational Bayes by Wang and Blei (2019) established VBI as a theoretically valid procedure. The scalability of VBI in modern applications hinges on the efficiency of stochastic optimization in scenarios with independent data points. This efficiency, however, diminishes in the case of Bayesian calibration of computer models due to the dependence structure in data (Robbins and Monro, 1951; Hoffman et al., 2013). To maintain the speed and scalability of VBI, we adopt a pairwise decomposition of data likelihood using vine copulas that separate the information on a dependence structure in data from their marginal distributions (Cooke and Kurowicka, 2006). Our specific contributions are as follows:

1. We propose a novel version of the black-box variational inference (Ranganath et al., 2014) for Bayesian calibration of computer models that preserves the efficiency of stochastic optimization in a scenario with dependent data.

2. We implement the Rao-Blackwellization, control variates, and importance sampling to reduce the variance of noisy gradient estimates involved in our algorithm.

3. We provide both theoretical and empirical evidence for scalability of our methodology and establish its superiority over the MH algorithm and the NUTS both in terms of time efficiency and memory requirements.

4. Finally, we demonstrate the opportunities in UQ given by the proposed algorithm on a real-word example in the field of nuclear physics.

The rest of this chapter is organized as follows. In Section 3.1, we give a general overview of VBI. In Section 3.2, we derive our proposed VBI approach to perform an inexpensive and

scalable calibration. We establish statistical validity of the method and provide theoretical justification for its scalability. Subsequently, in Section 3.3, we discuss the implementation details with focus on strategies to reduce the variance of the gradient estimators that are at the center of stochastic optimization for VBI. Section 3.4 presents a simulation study comparing our approach with state-of-the-art methods to approximate posterior distribution and illustrates our method on a real-data application. All technical details, proofs, and supplementary results are provided in section 3.5.

3.1 Variational Bayes inference

VBI is an optimization based method that approximates $p(\boldsymbol{\phi}|\boldsymbol{d})$ by a family of distributions $q(\boldsymbol{\phi}|\boldsymbol{\lambda})$ over latent variables with its own variational parameter $\boldsymbol{\lambda}$. Many commonly used families exist with the simplest mean-field family assuming independence of all the components in $\boldsymbol{\phi}$; see Wainwright and Jordan (2008); Hoffman and Blei (2015); Ranganath et al. (2016); Tran et al. (2015, 2017) for examples of more sophisticated families. The approximate distribution q^* is chosen to satisfy

$$q^* = \underset{q(\boldsymbol{\phi}|\boldsymbol{\lambda})}{\arg\min}\, KL(q(\boldsymbol{\phi}|\boldsymbol{\lambda})||p(\boldsymbol{\phi}|\boldsymbol{d})). \tag{3.8}$$

Here, KL denotes the Kullback-Leibler divergence of $q(\boldsymbol{\phi}|\boldsymbol{\lambda})$ from $p(\boldsymbol{\phi}|\boldsymbol{d})$. Finding q^* is done in practice by maximizing the *evidence lower bound (ELBO)*

$$\mathcal{L}(\boldsymbol{\lambda}) = \mathbb{E}_q\Big[\log p(\boldsymbol{d}|\boldsymbol{\phi})\Big] - KL(q(\boldsymbol{\phi}|\boldsymbol{\lambda})||p(\boldsymbol{\phi})), \tag{3.9}$$

which is a sum of the expected data log-likelihood $\log p(\boldsymbol{d}|\boldsymbol{\phi})$ and the KL divergence between the combined prior distribution $p(\boldsymbol{\phi})$ of calibration parameters, the error scale σ, and GP hyperparameters and the variational distribution $q(\boldsymbol{\phi}|\boldsymbol{\lambda})$. Note that we set $\mathcal{L}(\boldsymbol{\lambda}) := \mathcal{L}(q(\boldsymbol{\phi}|\boldsymbol{\lambda}))$ for the ease of notation. Minimizing the ELBO is equivalent to minimizing the original

objective function. Indeed,

$$\begin{aligned} KL(q(\boldsymbol{\phi}|\boldsymbol{\lambda})||p(\boldsymbol{\phi}|\boldsymbol{d})) &= \mathbb{E}_q\Big[\log q(\boldsymbol{\phi}|\boldsymbol{\lambda})\Big] - \mathbb{E}_q\Big[\log p(\boldsymbol{\phi}|\boldsymbol{d})\Big] \\ &= -\Big(\mathbb{E}_q\Big[\log p(\boldsymbol{d}|\boldsymbol{\phi})\Big] - KL(q(\boldsymbol{\phi}|\boldsymbol{\lambda})||p(\boldsymbol{\phi}))\Big) + \log p(\boldsymbol{d}). \end{aligned}$$

The ELBO can be optimized via the standard coordinate- or gradient-ascent methods. These techniques are inefficient for large datasets, because we must optimize the variational parameters globally for the whole dataset. Instead, it has become common practice to use a stochastic gradient ascent (SGA) algorithm, which Hoffman et al. (2013) named "stochastic variational inference" (SVI). Similarly to the traditional gradient ascent, SGA updates $\boldsymbol{\lambda}$ at the t^{th} iteration with

$$\boldsymbol{\lambda}_{t+1} \leftarrow \boldsymbol{\lambda}_t + \rho_t \tilde{l}(\boldsymbol{\lambda}_t). \tag{3.10}$$

Here, $\tilde{l}(\boldsymbol{\lambda})$ is a realization of the random variable $\tilde{\mathcal{L}}(\boldsymbol{\lambda})$, so that $\mathbb{E}(\tilde{\mathcal{L}}(\boldsymbol{\lambda})) = \nabla_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda})$, and Ranganath et al. (2014) showed that the gradient of ELBO with respect to the variational parameter $\boldsymbol{\lambda}$ can be written as

$$\nabla_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda}) = \mathbb{E}_q\Big[\nabla_{\boldsymbol{\lambda}} \log q(\boldsymbol{\phi}|\boldsymbol{\lambda})(\log p(\boldsymbol{d}|\boldsymbol{\phi}) - \log \frac{q(\boldsymbol{\phi}|\boldsymbol{\lambda})}{p(\boldsymbol{\phi})})\Big], \tag{3.11}$$

where $\nabla_{\boldsymbol{\lambda}} \log q(\boldsymbol{\phi}|\boldsymbol{\lambda})$ is the gradient of the variational log-likelihood with respect to $\boldsymbol{\lambda}$.

SGA converges to a local maximum of $\mathcal{L}(\boldsymbol{\lambda})$ (global for $\mathcal{L}(\boldsymbol{\lambda})$ concave (Bottou et al., 1997)) when the learning rate ρ_t follows the Robbins-Monro conditions (Robbins and Monro, 1951)

$$\sum_{t=1}^{\infty} \rho_t = \infty, \qquad \sum_{t=1}^{\infty} \rho_t^2 < \infty. \tag{3.12}$$

The bottleneck in the computation of the gradient $\nabla_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda})$ is the evaluation of the log-likelihood $\log p(\boldsymbol{d}|\boldsymbol{\phi})$, which makes the traditional gradient methods as hard to scale as MCMC methods. SGA algorithms address this challenge. If we consider N independent observations $d_i \sim p(d_i|\boldsymbol{\phi})$, then we can define a noisy estimate of the gradient $\nabla_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda})$ as

$$\tilde{\mathcal{L}}(\boldsymbol{\lambda}) := N\mathbb{E}_q\Big[\nabla_{\boldsymbol{\lambda}} \log q(\boldsymbol{\phi}|\boldsymbol{\lambda})(\log p(d_I|\boldsymbol{\phi}))\Big] - \mathbb{E}_q\Big[\nabla_{\boldsymbol{\lambda}} \log q(\boldsymbol{\phi}|\boldsymbol{\lambda}) \log \frac{q(\boldsymbol{\phi}|\boldsymbol{\lambda})}{p(\boldsymbol{\phi})}\Big], \tag{3.13}$$

where $I \sim U(1, \ldots, N)$ with $\mathbb{E}(\tilde{\mathcal{L}}(\boldsymbol{\lambda})) = \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda})$. Each update of $\boldsymbol{\lambda}$ computes the likelihood only for one observation d_i at a time and makes the SVI scalable for large datasets. One can easily see that, under the framework for Bayesian calibration, $\mathbb{E}(\tilde{\mathcal{L}}(\boldsymbol{\lambda})) \neq \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda})$ and that the corresponding the **SVI does not scale (the noisy estimates are biased)**.

3.2 Variational calibration of computer models

In this section, we derive the algorithm for scalable variational inference approach to Bayesian computer model calibration. The first step is finding a convenient decomposition of the likelihood $p(\boldsymbol{d}|\boldsymbol{\phi})$ that allows for an unbiased stochastic estimate of the gradient $\nabla_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda})$ that depends only on a small subset of data. Multivariate copulas, and specifically their pairwise construction which we shall introduce below, provide such a decomposition. We are not the first ones to use copulas in the context of VBI. For instance, Tran et al. (2015) and Smith et al. (2020) proposed a multivariate copula as a possible variational family. However, we are the first ones using copulas in the context of computer model calibration implementing via VBI.

3.2.1 Multivariate copulas and likelihood decomposition

Fundamentally, a copula separates the information on the dependence structure of $N > 1$ random variables $X_1, \ldots, X_N$ from their marginal distributions. Let us assume, for simplicity, that the marginal cumulative distribution functions (CDFs) $F_1, \ldots, F_N$ are continuous and possess the inverse functions $F_1^{-1}, \ldots, F_N^{-1}$. It follows from the probability integral transform that $U_i := F_i(X_i) \sim U(0, 1)$ and conversely that $X_i = F_i^{-1}(U_i)$. With this in mind, we have

$$P(X_1 \leq F_1^{-1}(x_1), \ldots, X_N \leq F_N^{-1}(x_N)) = P(U_1 \leq x_1, \ldots, U_N \leq x_N) := C(x_1, \ldots, x_N).$$

The function C is a distribution with support on $[0, 1]^N$, uniform marginals, and is called a copula. Under the above assumptions, a one-to-one correspondence exists between copula

C and the distribution of $\boldsymbol{X} = (X_1, \ldots, X_N)^T$, as stated in the following theorem due to Sklar (1959). To keep the notation consistency and readability, we re-state the theorem here.

Theorem 1 (Sklar (1959))**.** *Given the random v.* $X_1, \ldots, X_n$ *with continuous marginals* $F_1, \ldots, F_N$ *and the joint distribution function* F*, there exists a unique copula* C *such that for all* $\boldsymbol{x} = (x_1, \ldots, x_N)^T \in \mathbb{R}^n$*:* $F(x_1, \ldots, x_N) = C(F_1(x_1), \ldots, F_n(x_N))$*. Conversely, given the CDFs* $F_1, \ldots, F_N$ *and a copula* C*,* F *defined through* $C(F_1(x_1), \ldots, F_n(x_N))$ *is an* N*-variate distribution function with marginals* $F_1, \ldots, F_N$*.*

Consequently, one can write the joint probability density function (pdf) f of $\boldsymbol{X} = (X_1, \ldots, X_N)^T$ as

$$f(x_1, \ldots, x_N) = c(F_1(x_1), \ldots, F_n(x_N)) \prod_{i=1}^{N} f_i(x_i), \tag{3.14}$$

where c represents the copula density and f_i is the marginal pdf of X_i.

The key reason for considering copulas is that one can decompose the N-dimensional copula density c into a product of bivariate copulas. The starting point for this construction is a recursive decomposition of the density f into a product of conditional densities

$$f(x_1, \ldots, x_N) = \prod_{i=2}^{N} f(x_i | x_1, \ldots, x_{i-1}) f(x_1). \tag{3.15}$$

For $N = 2$, the Sklar's theorem implies that

$$f(x_1, x_2) = c_{12}(F_1(x_1), F_2(x_2)) f_1(x_1) f_2(x_2), \tag{3.16}$$

and

$$f(x_1 | x_2) = c_{12}(F_1(x_1), F_2(x_2)) f_1(x_1), \tag{3.17}$$

where

$$c_{12} := c_{12}(F_1(x_1), F_2(x_2)) \tag{3.18}$$

is a density of $C(F_1(x_1), F_2(x_2)) = F(x_1, x_2)$. Using (3.17) for the decomposition of (X_1, X_t) given $X_2, \ldots, X_{t-1}$, we obtain

$$f(x_t | x_1, \ldots, x_{t-1}) = (\prod_{s=1}^{t-2} c_{s,t;s+1,\ldots,t-1}) c_{(t-1),t} \cdot f_t(x_t), \tag{3.19}$$

where

$$c_{i,j;i_1,\ldots,i_k} := c_{i,j;i_1,\ldots,i_k}(F(x_i|x_{i_1},\ldots,x_{i_k}), F(x_j|x_{i_1},\ldots,x_{i_k})) \tag{3.20}$$

and

$$F(x_i, x_j|x_{i_1},\ldots,x_{i_k}) := C_{i,j;i_1,\ldots,i_k}(F(x_i|x_{i_1},\ldots,x_{i_k}), F(x_j|x_{i_1},\ldots,x_{i_k})). \tag{3.21}$$

Using (3.15) and (3.19) with the specific index choices $s = i, t = i + j$, we have that

$$f(x_1,\ldots,x_N) = \Big[\prod_{j=1}^{N-1}\prod_{i=1}^{N-j} c_{i,(i+j);(i+1),\ldots,(i+j-1)}\Big]\prod_{k=1}^{N} f_k(x_k). \tag{3.22}$$

Note that $c_{i,j;i_1,\ldots,i_k}$ are two-dimensional copulas evaluated at the CDFs $F(x_i|x_{i_1},\ldots,x_{i_k})$ and $F(x_j|x_{i_1},\ldots,x_{i_k})$. The decomposition above is called a *D-vine distribution*. A similar class of decompositions is possible when one applies (3.17) on (X_{t-1}, X_t) given $X_1,\ldots,X_{t-2}$ and sets $j = t-k, j+i = t$ to get a *canonical vine (C-vine)* (Cooke and Kurowicka, 2006):

$$\begin{aligned} f(x_1,\ldots,x_N) &= f_1(x_1)\Big[\prod_{t=2}^{N}\prod_{k=1}^{t-1} c_{t-k,t;1,\ldots,(t-k-1)}\cdot f_t(x_t)\Big] \\ &= \Big[\prod_{j=1}^{N-1}\prod_{i=1}^{N-j} c_{j,(j+i);1,\ldots,(j-1)}\Big]\prod_{k=1}^{N} f_k(x_k). \end{aligned}$$

One can easily imagine that many such pair-copula decompositions exist. Bedford and Cooke (2002) observed that these can be represented graphically as a sequence of nested trees with undirected edges, which are referred to as *vine trees*. In order for a pair-copula decomposition to be feasible, Bedford and Cooke (2002) defined a *regular vine tree (R-vine)* on N variables consisting of connected trees $\mathcal{T}_1,\ldots,\mathcal{T}_{N-1}$ with nodes N_i and edges E_i satisfying the following conditions:

1. $\mathcal{T}_1$ has nodes $N_1 = \{1,\ldots,N\}$ and edges E_1.

2. For $i = 2,\ldots,N-1$ the tree T_i has nodes $N_i = E_{i-1}$ (i.e., edges in a tree become nodes in the next tree).

3. Two edges in $\mathcal{T}_i$ are joined in $\mathcal{T}_{i+1}$ if they share a common node in $\mathcal{T}_i$.

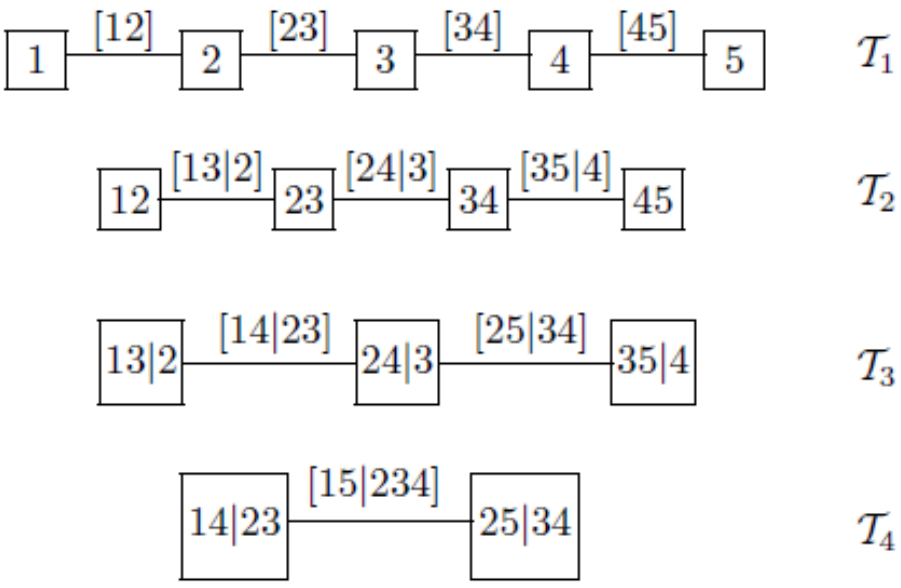

Figure 3.1: A D-vine tree representation of a copula with 5 variables.

Here, we focus exclusively on the D-vine and C-vine decompositions because they represent the most-studied instances of regular vines and provide an especially efficient notation. We note, however, that the following results can be extended to any regular vines.

Properties of vine copulas (Cooke and Kurowicka, 2006). The vine copula construction is particularly attractive for two reasons. First, each pair of variables occurs only once as a conditioning set. Second, the bivariate copulas involved in the decompositions have convenient form in the case of Gaussian likelihood f. In particular, let $\boldsymbol{X} = (X_1, \ldots, X_N)^T$ follows a multivariate normal distribution with $F_j = \Phi, j = 1, \ldots, N$, where Φ is the standard normal CDF. The bivariate copula density is

$$c_{i,j;i_1,\ldots,i_k}(u_i, u_j) = \frac{1}{\sqrt{1-\kappa^2}} \exp\{-\frac{\kappa^2(w_i^2 + w_j^2) - 2\kappa w_i w_j}{2(1-\kappa^2)}\}. \tag{3.23}$$

Here, $u_i = F(x_i|x_{i_1}, \ldots, x_{i_k})$, $u_j = F(x_j|x_{i_1}, \ldots, x_{i_k})$, $w_i = \Phi^{-1}(u_i)$, $w_j = \Phi^{-1}(u_j)$, and $\kappa = \rho_{i,j \cdot i_1,\ldots,i_k}$ is the partial correlation of variables i, j given $i_1, \ldots, i_k$. The D-vine and C-vine decompositions also involve conditional CDFs, for which we need further expressions. Let $v \in D$ and $D_{-v} := D \setminus v$ so that D contains more than one element, $F(x_j|\boldsymbol{x}_D)$ is

typically computed recursively as

$$F(x_j|\boldsymbol{x}_D) = h(F(x_j|\boldsymbol{x}_{D_{-v}}), F(x_v|\boldsymbol{x}_{D_{-v}})|\rho_{jv|D_{-v}}) \tag{3.24}$$

and the function h is for the Gaussian case given by

$$h(u_i, u_j|\rho_{i,j\cdot i_1,\ldots,i_k}) = \Phi\left(\frac{\Phi^{-1}(u_i) - \rho_{i,j\cdot i_1,\ldots,i_k}\Phi^{-1}(u_j)}{\sqrt{1-\rho^2_{i,j\cdot i_1,\ldots,i_k}}}\right). \tag{3.25}$$

Lastly, the partial correlation can be also computed recursively as

$$\rho_{i,j\cdot D} = \frac{\rho_{i,j\cdot D_{-v}} - \rho_{i,v\cdot D_{-v}}\rho_{v,j\cdot D_{-v}}}{\sqrt{1-\rho^2_{i,v\cdot D_{-v}}}\sqrt{1-\rho^2_{v,j\cdot D_{-v}}}}. \tag{3.26}$$

3.2.2 Scalable algorithm with truncated vine copulas

We now consider the data likelihood $p(\boldsymbol{d}|\boldsymbol{\phi})$ according to (3.5) and make use of vines to construct a noisy estimate of the gradient $\nabla_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda})$. We additionally assume that $N = n+s$, where n is the number of observations y_i from the physical process, and s is the number of computer model runs z_j. The log-likelihood $\log p(\boldsymbol{d}|\boldsymbol{\phi})$ can be rewritten according to the D-vine decomposition as

$$\log p(\boldsymbol{d}|\boldsymbol{\phi}) = \sum_{j=1}^{N-1}\sum_{i=1}^{N-j} p^D_{i,i+j}(\boldsymbol{\phi}), \tag{3.27}$$

where

$$p^D_{i,i+j}(\boldsymbol{\phi}) = \log c_{i,(i+j);(i+1),\ldots,(i+j-1)} + \frac{1}{n-1}\big(\log p_i(d_i|\boldsymbol{\phi}) + \log p_{i+j}(d_{i+j}|\boldsymbol{\phi})\big). \tag{3.28}$$

This can be conveniently used in the expression of the ELBO gradient. For a D-vine, we have that

$$\nabla_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda}) = \sum_{j=1}^{N-1}\sum_{i=1}^{N-j}\mathbb{E}_q\Big[\nabla_{\boldsymbol{\lambda}}\log q(\boldsymbol{\phi}|\boldsymbol{\lambda})(p^D_{i,i+j}(\boldsymbol{\phi}))\Big] - \mathbb{E}_q\left[\nabla_{\boldsymbol{\lambda}}\log q(\boldsymbol{\phi}|\boldsymbol{\lambda})\log\frac{q(\boldsymbol{\phi}|\boldsymbol{\lambda})}{p(\boldsymbol{\phi})}\right]. \tag{3.29}$$

The following proposition gives a noisy unbiased estimate $\tilde{\mathcal{L}}_D(\boldsymbol{\lambda})$ of the gradient (3.29). Similarly, we can derive a noisy estimate $\tilde{\mathcal{L}}_C(\boldsymbol{\lambda})$ of the gradient using a C-vine. We leave the details to Section 3.5.1.

Proposition 1. *Let $\tilde{\mathcal{L}}_D(\boldsymbol{\lambda})$ be an estimate of the ELBO gradient $\nabla_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda})$ defined as*

$$\tilde{\mathcal{L}}_D(\boldsymbol{\lambda}) = \frac{N(N-1)}{2}\mathbb{E}_q\Big[\nabla_{\boldsymbol{\lambda}} \log q(\boldsymbol{\phi}|\boldsymbol{\lambda})(p^D_{I_D(K)}(\boldsymbol{\phi}))\Big] - \mathbb{E}_q\Big[\nabla_{\boldsymbol{\lambda}} \log q(\boldsymbol{\phi}|\boldsymbol{\lambda}) \log \frac{q(\boldsymbol{\phi}|\boldsymbol{\lambda})}{p(\boldsymbol{\phi})}\Big],$$

where $K \sim U(1, \ldots, \frac{N(N-1)}{2})$, and I_D is the bijection

$$I_D : \{1, \ldots, \frac{N(N-1)}{2}\} \to \{(i, i+j) : i \in \{1, \ldots, N-j\} \text{ for } j \in \{1, \ldots N-1\}\},$$

then $\tilde{\mathcal{L}}_D(\boldsymbol{\lambda})$ is unbiased i.e., $\mathbb{E}(\tilde{\mathcal{L}}_D(\boldsymbol{\lambda})) = \nabla_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda})$.

As in the case of SVI for independent data, these noisy estimates allow to update the variational parameter $\boldsymbol{\lambda}$ without the need to evaluate the whole likelihood $p(\boldsymbol{d}|\boldsymbol{\phi})$. We need to consider only the data consisting of a copula's conditioning and conditioned sets. Unfortunately, both $\tilde{\mathcal{L}}_D(\boldsymbol{\lambda})$ and $\tilde{\mathcal{L}}_C(\boldsymbol{\lambda})$ can be relatively costly to compute for large datasets because of the recursive nature of calculations involved in the copula densities' evaluation. According to Brechmann et al. (2012); Dissmann et al. (2013), and Brechmann and Joe (2015), the most important and strongest dependencies among variables can be typically captured best by the pair copulas of the first trees. This notion motivates the use of *truncated vine copulas*, where the copulas associated with the higher-order trees are set to the independence copulas. From the definition of a regular vine, one can show that the joint density f can be decomposed as

$$f(d_1, \ldots, d_N) = \Big[\prod_{j=1}^{N-1}\prod_{e \in E_i} c_{j(e),k(e);D(e)}\Big]\prod_{k=1}^{N} f_k(d_k),$$

where $e = j(e), k(e); D(e) \in E_i$ is an edge in the i^{th} tree of the vine specification. We define the truncated regular vine copula as follows.

Definition 2 (Brechmann et al. (2012))**.** *Let $\boldsymbol{U} = \{U_1, \ldots, U_N\}$ be a random vector with uniform marginals, and let $l \in \{1, \ldots, N-1\}$ be the truncation level. Let Π denote the bivariate independence copula. Then, $\boldsymbol{U}$ is said to be distributed according to an N-dimensional l-truncated R-vine copula if C is an N-dimensional R-vine copula with*

$$C_{j(e),k(e);D(e)} = \Pi \quad \forall e \in E_i \quad i = l+1, \ldots, N-1.$$

For the case of an **l-truncated D-vine**, we have

$$f(d_1,\dots,d_N)=\Big[\prod_{j=1}^{l}\prod_{i=1}^{N-j}c_{i,(i+j);(i+1),\dots,(i+j-1)}\Big]\prod_{k=1}^{N}f_k(d_k), \tag{3.30}$$

and analogically to the case of D-vine with no truncation, the log-likelihood $p(\boldsymbol{d}|\boldsymbol{\phi})$ can be written as a sum of unique elements given in Proposition 2.

Proposition 2. *If the copula of $p(\boldsymbol{d}|\boldsymbol{\phi})$ is distributed according to an l-truncated D-vine, we can rewrite*

$$\log p(\boldsymbol{d}|\boldsymbol{\phi})=\sum_{j=1}^{l}\sum_{i=1}^{N-j}p^{D_l}_{i,i+j}(\boldsymbol{\phi}), \tag{3.31}$$

where

$$p^{D_l}_{i,i+j}(\boldsymbol{\phi})=\log c_{i,(i+j);(i+1),\dots,(i+j-1)}+\frac{1}{a_i}\log p_i(d_i|\boldsymbol{\phi})+\frac{1}{b_{i+j}}\log p_{i+j}(d_{i+j}|\boldsymbol{\phi}), \tag{3.32}$$

and

$$a_i=2l-\Big[(l+1-i)\mathbb{1}_{i\le l}+(l-N+i)\mathbb{1}_{i>N-l}\Big],$$
$$b_{i+j}=2l-\Big[(l+1-j-i)\mathbb{1}_{i+j\le l}+(l-N+j+i)\mathbb{1}_{i+j>N-l}\Big].$$

The main idea for the scalable variational calibration (VC) of computer models is replacing the full log-likelihood $\log(\boldsymbol{d}|\boldsymbol{\phi})$ in the definition of ELBO with the likelihood based on a truncated vine copula. This yields the *l-truncated ELBO* for the l-truncated D-vine

$$\mathcal{L}_{D_l}(\boldsymbol{\lambda})=\mathbb{E}_q\Big[\sum_{j=1}^{l}\sum_{i=1}^{N-j}p^{D_l}_{i,i+j}(\boldsymbol{\phi})\Big]-KL(q(\boldsymbol{\phi}|\boldsymbol{\lambda})||p(\boldsymbol{\phi})) \tag{3.33}$$

with its gradient

$$\nabla_{\boldsymbol{\lambda}}\mathcal{L}_{D_l}(\boldsymbol{\lambda})=\sum_{j=1}^{l}\sum_{i=1}^{N-j}\mathbb{E}_q\Big[\nabla_{\boldsymbol{\lambda}}\log q(\boldsymbol{\phi}|\boldsymbol{\lambda})(p^{D_l}_{i,i+j}(\boldsymbol{\phi}))\Big]-\mathbb{E}_q\Big[\nabla_{\boldsymbol{\lambda}}\log q(\boldsymbol{\phi}|\boldsymbol{\lambda})\log\frac{q(\boldsymbol{\phi}|\boldsymbol{\lambda})}{p(\boldsymbol{\phi})}\Big].$$

The following proposition gives a noisy unbiased estimate $\tilde{\mathcal{L}}_{D_l}(\boldsymbol{\lambda})$ of the gradient $\nabla_{\boldsymbol{\lambda}}\mathcal{L}_{D_l}(\boldsymbol{\lambda})$. We can analogously derive an unbiased estimate $\tilde{\mathcal{L}}_{C_l}(\boldsymbol{\lambda})$ of the gradient using C-vine (see Section 3.5.1).

Proposition 3. *Let* $\tilde{\mathcal{L}}_{D_l}(\boldsymbol{\lambda})$ *be an estimate of the ELBO gradient* $\nabla_{\boldsymbol{\lambda}}\mathcal{L}_{D_l}(\boldsymbol{\lambda})$ *defined as*

$$\tilde{\mathcal{L}}_{D_l}(\boldsymbol{\lambda}) = \frac{l(2N-(l+1))}{2}\mathbb{E}_q\left[\nabla_{\boldsymbol{\lambda}}\log q(\boldsymbol{\phi}|\boldsymbol{\lambda})(p^{D_l}_{I_{D_l}(K)}(\boldsymbol{\phi}))\right] - \mathbb{E}_q\left[\nabla_{\boldsymbol{\lambda}}\log q(\phi|\boldsymbol{\lambda})\log\frac{q(\phi|\boldsymbol{\lambda})}{p(\phi)}\right],$$

where $K \sim U(1,\ldots,\frac{l(2N-(l+1))}{2})$*, and* I_{D_l} *is the bijection*

$$I_{D_l} : \{1,\ldots,\frac{l(2N-(l+1))}{2}\} \to \{(i,i+j) : i \in \{1,\ldots,N-j\} \text{ for } j \in \{1,\ldots l\}\},$$

then $\tilde{\mathcal{L}}_{D_l}(\boldsymbol{\lambda})$ *is unbiased i.e.,* $\mathbb{E}(\tilde{\mathcal{L}}_{D_l}(\boldsymbol{\lambda})) = \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{D_l}(\boldsymbol{\lambda})$.

Considering the l-truncated ELBO defined above, our proposed algorithm for variational calibration of computer models with truncated vine copulas is stated in Algorithm 3.1. Note that $\tilde{\mathcal{L}}_{D_l}(\boldsymbol{\lambda})$ does not have closed form expression in general due to expectations involved in the computation. Therefore, we resort to a MC approximation of the gradient estimate $\tilde{\mathcal{L}}_{D_l}(\boldsymbol{\lambda})$ using samples from the variational distribution.

Algorithm 3.1: Variational calibration with truncated D-vine copulas.

Input: Data $\boldsymbol{d}$, mean and covariance functions for GPs in Kennedy-O'Hagan framework, variational family $q(\boldsymbol{\phi}|\boldsymbol{\lambda})$, **truncation level l**

$\lambda \leftarrow$ random initial value
$t \leftarrow 1$
repeat
  for $s = 1$ *to* S **do**
    $\boldsymbol{\phi}[s] \sim q(\boldsymbol{\phi}|\boldsymbol{\lambda})$ // Random sample from q
  $K \leftarrow U(1,\ldots,\frac{l(2N-(l+1))}{2})$
  $\rho \leftarrow t^{\text{th}}$ value of a Robbins-Monro sequence
  $\boldsymbol{\lambda} \leftarrow \boldsymbol{\lambda} + \rho\frac{1}{S}\sum_{s=1}^{S}\left[\frac{l(2N-(l+1))}{2}\nabla_{\boldsymbol{\lambda}}\log q(\boldsymbol{\phi}[s]|\boldsymbol{\lambda})\big(p^{D_l}_{I_{D_l}(K)}(\boldsymbol{\phi}[s]) - \frac{2}{l(2N-(l+1))}\log\frac{q(\boldsymbol{\phi}[s]|\boldsymbol{\lambda})}{p(\boldsymbol{\phi}[s])}\big)\right]$
  $t \leftarrow t+1$
until *change of* $\boldsymbol{\lambda}$ *is less than* ϵ

Scalability Discussion. The complexity of a bivariate copula evaluation depends on the size of the conditioning dataset due to the recursive nature of the calculations involved (Cooke and Kurowicka, 2006). From the vine tree construction, the cardinality of the conditioning

set for D-vine and C-vine is in the worst case $N-2$. Nevertheless, on average, we can do better.

Lemma 2. *Let X be the cardinality of the conditioning set in $p^D_{I_D(K)}(\boldsymbol{\phi})$ or $p^C_{I_C(K)}(\boldsymbol{\phi})$, then*

$$P(X=i) = \frac{N-(i+1)}{\binom{N}{2}} \qquad \text{for } i \in \{0,\dots,N-2\} \tag{3.34}$$

and $\mathbb{E}(X) = \frac{N-2}{3}$.

The cardinality of conditioning set in Lemma 2 is on average roughly $N/3$. On the other hand, the cardinality of conditioning set for the case of Algorithm 3.1 is at most $l-1$ with the average given by the following lemma.

Lemma 3. *Let X be the cardinality of the conditioning set in $p^{D_l}_{I_{D_l}(K)}(\boldsymbol{\phi})$ or $p^{C_l}_{I_{C_l}(K)}(\boldsymbol{\phi})$, then*

$$P(X=i) = \frac{N-(i+1)}{\frac{l(2N-(l+1))}{2}} \qquad \text{for} \qquad i \in \{0,\dots,l-1\}, \tag{3.35}$$

and

$$\mathbb{E}(X) = \frac{(l-1)(3N-2l-2)}{3(2N-l-1)}.$$

As a consequence of Lemma 3, $\mathbb{E}(X) \approx 2$ for $N = 10^5$ and truncation level $l = 5$, which is a significant improvement to the average case $p^D_{I_D(K)}(\boldsymbol{\phi})$ and $p^C_{I_C(K)}(\boldsymbol{\phi})$ (≈ 33333 for $N = 10^5$). This provides a heuristic yet convincing argument for the scalability.

3.3 Implementation details

3.3.1 Selection of truncation level

Selection of the truncation level l is an important element in effective approximation of the posterior distribution $p(\boldsymbol{\phi}|\boldsymbol{d})$ under Algorithm 3.1. Dissmann et al. (2013) propose a sequential approach for selection of l in the case of vine estimation. One sequentially fits models with an increasing truncation level until the quality of fit stays stable or computational resources are depleted. We adopt similar idea for the case of VC of computer

models with vine copulas. Let $\lambda(l)$ represents the value of variational parameter estimated with Algorithm 3.1 for a fixed truncation level l. One can then sequentially increase l until $\Delta(\lambda(l+1), \lambda(l)) < \epsilon$ for some distance metric Δ and a desired tolerance ϵ.

3.3.2 Variance reduction of Monte Carlo approximations

The computational convenience of simple MC approximations of the gradient estimators based on the l-truncated D-vine and C-vine copulas $\tilde{\mathcal{L}}_{D_l}(\boldsymbol{\lambda})$ and $\tilde{\mathcal{L}}_{C_l}(\boldsymbol{\lambda})$ (see Section 3.2.2) is typically accompanied by their large variance. The consequence in practice is the need for small step size ρ_t in the SGA portion of Algorithm 3.1 which results in a slower convergence. In order to reduce the variance of MC approximations, we adopt the same approach as Ruiz et al. (2016) and use the Rao-Blackwellization (Casella and Robert, 1996) in combination with the control variates (Ross, 2006) and importance sampling. The reminder of this section focuses on the case of D-vine decomposition, see Section 3.5.1 for the derivations for C-vines.

Rao-Blackwellization. The idea here is to replace the noisy estimate of gradient with its conditional expectation with respect to a subset of $\boldsymbol{\phi}$. For simplicity, let us consider a situation with $\boldsymbol{\phi} = (\phi_1, \phi_2) \in \mathbb{R}^2$ and variational family $q(\boldsymbol{\phi}|\boldsymbol{\lambda})$ that factorizes into $q(\phi_1|\boldsymbol{\lambda}_1)q(\phi_2|\boldsymbol{\lambda}_2)$. Additionally, let $\hat{\mathcal{L}}_{\boldsymbol{\lambda}}(\phi_1, \phi_2)$ be the MC approximation of the gradient $\nabla_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda})$. Now, the conditional expectation $\mathbb{E}[\hat{\mathcal{L}}_{\boldsymbol{\lambda}}(\phi_1, \phi_2)|\phi_1]$ is also an unbiased estimate of $\nabla_{\boldsymbol{\lambda}}\mathcal{L}(\lambda)$ since $\mathbb{E}_q(\mathbb{E}[\hat{\mathcal{L}}_{\boldsymbol{\lambda}}(\phi_1, \phi_2)|\phi_1]) = \mathbb{E}_q(\hat{\mathcal{L}}_{\boldsymbol{\lambda}}(\phi_1, \phi_2))$ and

$$\mathbb{V}ar_q(\mathbb{E}[\hat{\mathcal{L}}_{\boldsymbol{\lambda}}(\phi_1, \phi_2)|\phi_1]) = \mathbb{V}ar_q(\hat{\mathcal{L}}_{\boldsymbol{\lambda}}(\phi_1, \phi_2)) - \mathbb{E}[(\hat{\mathcal{L}}_{\boldsymbol{\lambda}}(\phi_1, \phi_2) - \mathbb{E}[\hat{\mathcal{L}}_{\boldsymbol{\lambda}}(\phi_1, \phi_2)|\phi_1])^2]$$

shows that $\mathbb{V}ar_q(\mathbb{E}[\hat{\mathcal{L}}_{\boldsymbol{\lambda}}(\phi_1, \phi_2)|\phi_1]) \leq \mathbb{V}ar_q(\hat{\mathcal{L}}_{\boldsymbol{\lambda}}(\phi_1, \phi_2))$. The factorization of the variational family also makes the conditional expectation straightforward to compute as

$$\mathbb{E}[\hat{\mathcal{L}}_{\boldsymbol{\lambda}}(\phi_1, \phi_2)|\phi_1] = \int_{\phi_2} \mathbb{E}[\hat{\mathcal{L}}_{\boldsymbol{\lambda}}(\phi_1, \phi_2)] \frac{q(\phi_1|\boldsymbol{\lambda}_1)q(\phi_2|\boldsymbol{\lambda}_2)}{q(\phi_1|\boldsymbol{\lambda}_1)} \, \mathrm{d}\phi_2 = \mathbb{E}_{q(\phi_2|\boldsymbol{\lambda}_2)}(\hat{\mathcal{L}}_{\boldsymbol{\lambda}}(\phi_1, \phi_2)),$$

i.e., we just need to integrate out some variables. Let us consider the MC approximation of the gradient estimator $\tilde{\mathcal{L}}_{D_l}(\boldsymbol{\lambda})$. The j^{th} entry of the Rao-Blackwellized estimator is

$$\frac{1}{S}\sum_{s=1}^{S}\Big[\frac{l(2N-(l+1))}{2}\nabla_{\boldsymbol{\lambda}_j}\log q(\phi_j[s]|\boldsymbol{\lambda}_j)\big(\tilde{p}_{(j)}(\boldsymbol{\phi}[s])-\frac{2}{l(2N-(l+1))}\log\frac{q(\phi_j[s]|\boldsymbol{\lambda}_j)}{p(\phi_j[s])}\big)\Big],$$

where $\tilde{p}_{(j)}(\boldsymbol{\phi})$ are the components of $p^{D_l}_{I_{D_l}(K)}(\boldsymbol{\phi})$ that include ϕ_j.

Control Variates. To further reduce the variance of the MC approximations we will replace the Rao-Blackwellized estimate above with a function that has the same expectation but again smaller variance. For illustration, let us first consider a target function $\xi(\boldsymbol{\phi})$ whose variance we want to reduce, and a function $\psi(\boldsymbol{\phi})$ with finite expectation. Define

$$\hat{\xi}(\boldsymbol{\phi})=\xi(\boldsymbol{\phi})-a(\psi(\boldsymbol{\phi})-\mathbb{E}_q[\psi(\boldsymbol{\phi})]), \tag{3.36}$$

where a is a scalar and $\mathbb{E}_q(\hat{\xi}(\boldsymbol{\phi}))=\mathbb{E}_g[\xi(\boldsymbol{\phi})]$. The variance of $\hat{\xi}(\boldsymbol{\phi})$ is

$$\mathbb{V}ar_q(\hat{\xi}(\boldsymbol{\phi}))=\mathbb{V}ar_q(\xi(\boldsymbol{\phi}))+a^2\mathbb{V}ar_q(\psi(\boldsymbol{\phi}))-2a\mathbb{C}ov_q(\xi(\boldsymbol{\phi}),\psi(\boldsymbol{\phi})). \tag{3.37}$$

This shows that a good choice for function $\psi(\boldsymbol{\phi})$ is one that has high covariance with $\xi(\boldsymbol{\phi})$. Moreover, the value of a that minimizes (3.37) is

$$a^*=\frac{\mathbb{C}ov_q(\xi(\boldsymbol{\phi}),\psi(\boldsymbol{\phi}))}{\mathbb{V}ar_q(\psi(\boldsymbol{\phi}))}. \tag{3.38}$$

Let us place the CV back into the context of calibration. Meeting the above described criteria, Ranganath et al. (2014) propose $\psi(\boldsymbol{\phi})$ to be $\nabla_{\boldsymbol{\lambda}}\log q(\boldsymbol{\phi}|\boldsymbol{\lambda})$, because it depends only on the variational distribution and has expectation zero. We can now set the target function $\xi(\boldsymbol{\phi})$ to be

$$\frac{l(2N-(l+1))}{2}\nabla_{\boldsymbol{\lambda}_j}\log q(\phi_j|\boldsymbol{\lambda}_j)\big(\tilde{p}_{(j)}(\boldsymbol{\phi})-\frac{2}{l(2N-(l+1))}\log\frac{q(\phi_j|\boldsymbol{\lambda}_j)}{p(\phi_j)}\big),$$

which gives the following j^{th} entry of the MC approximation of the gradient estimator $\tilde{\mathcal{L}}_{D_l}(\boldsymbol{\lambda})$ with CV

$$\tilde{\mathcal{L}}_{D_l}^{CV(j)}(\boldsymbol{\lambda})$$
$$= \frac{1}{S}\sum_{s=1}^{S}\Bigg[\frac{l(2N-(l+1))}{2}\nabla_{\boldsymbol{\lambda}_j}\log q(\phi_j[s]|\boldsymbol{\lambda}_j)\big(\tilde{p}_{(j)}(\boldsymbol{\phi}[s]) - \frac{2(\log\frac{q(\phi_j[s]|\boldsymbol{\lambda}_j)}{p(\phi_j[s])} + \hat{a}_j^D)}{l(2N-(l+1))}\big)\Bigg],$$

where $\hat{a}_j^D$ is the estimate of a^* based on additional independent draws from the variational approximation (otherwise the estimator would be biased).

Importance sampling. Here, we outline the last variance reduction technique that makes use of importance sampling. We refer to Ruiz et al. (2016) for full description of the method and illustration of its efficiency in the VBI framework. Fundamentally, instead of taking samples from the variational family $q(\boldsymbol{\phi}|\boldsymbol{\lambda})$ to carry out the MC approximation of the ELBO gradient estimate, we will take samples from an overdispersed distribution $r(\boldsymbol{\phi}|\boldsymbol{\lambda},\tau)$ in the same family that depends on an additional dispersion parameter $\tau > 1$. Namely, we can write the estimate $\tilde{\mathcal{L}}_{D_l}(\lambda)$ as

$$E_{r(\boldsymbol{\phi}|\boldsymbol{\lambda},\tau)}\Bigg[\frac{l(2N-(l+1))}{2}\nabla_{\boldsymbol{\lambda}}\log q(\boldsymbol{\phi}|\boldsymbol{\lambda})(p^{D_l}_{I_{D_l}(K)}(\boldsymbol{\phi}) - \frac{2}{l(2N-(l+1))}\log\frac{q(\boldsymbol{\phi}|\boldsymbol{\lambda})}{p(\boldsymbol{\phi})})w(\boldsymbol{\phi})\Bigg],$$

where $w(\boldsymbol{\phi}) = q(\boldsymbol{\phi}|\boldsymbol{\lambda})/r(\boldsymbol{\phi}|\boldsymbol{\lambda},\tau)$ is the importance weight which guarantees the estimator to be unbiased. The reason to formulate the $\tilde{\mathcal{L}}_{D_l}(\lambda)$ this way comes from the fact the optimal proposal (Robert and Casella, 2005) distribution to form the MC estimate is not $q(\boldsymbol{\phi}|\boldsymbol{\lambda})$, but rather

$$r^*(\boldsymbol{\phi}) \propto q(\boldsymbol{\phi}|\boldsymbol{\lambda})|\xi(\boldsymbol{\phi})|, \tag{3.39}$$

where

$$\xi(\boldsymbol{\phi}) = \frac{l(2N-(l+1))}{2}\nabla_{\boldsymbol{\lambda}}\log q(\boldsymbol{\phi}|\boldsymbol{\lambda})\big(p^{D_l}_{I_{D_l}(K)}(\boldsymbol{\phi}) - \frac{2}{l(2N-(l+1))}\log\frac{q(\boldsymbol{\phi}|\boldsymbol{\lambda})}{p(\boldsymbol{\phi})}\big). \tag{3.40}$$

However, the normalizing constant for the optimal $r^*(\boldsymbol{\phi})$ is intractable, and so Ruiz et al. (2016) propose that an overdispersed version of the variational family that assigns higher

probability to the tails of $q(\boldsymbol{\phi}|\boldsymbol{\lambda})$ is closer to the optimum than $q(\boldsymbol{\phi}|\boldsymbol{\lambda})$ itself. For example, if the value of $\boldsymbol{\lambda}$ makes the variational family a poor fit, then the samples $\boldsymbol{\phi}[s] \sim q(\boldsymbol{\phi}|\boldsymbol{\lambda})$ have a high value for the variational distribution but low for the true posterior. On the other hand, $r^*(\boldsymbol{\phi})$ proposes values of $\boldsymbol{\phi}[s]$ for which $\xi(\boldsymbol{\phi})$ is large that are in the tails of $p(\boldsymbol{\phi}|\boldsymbol{\lambda})$.

To see how the importance sampling leads to the reduction of variance of the MC estimates, let us consider the following estimator

$$\widehat{\mathcal{L}}_{MC} = \frac{1}{S}\sum_{s=1}^{S} \xi(\boldsymbol{\phi}[s]), \qquad \boldsymbol{\phi}[s] \sim p(\boldsymbol{\phi}|\boldsymbol{\lambda}), \tag{3.41}$$

then

$$\mathbb{V}ar\big[\widehat{\mathcal{L}}_{MC}\big] = \frac{1}{S}\mathbb{E}_q\big[\xi^2(\boldsymbol{\phi})\big] - \frac{1}{S}\big[\tilde{\mathcal{L}}_{D_l}(\lambda)\big]^2. \tag{3.42}$$

Similarly, we can derived the variance of the MC estimator with the importance weights

$$\widehat{\mathcal{L}}^{O}_{MC} = \frac{1}{S}\sum_{s=1}^{S} \xi(\boldsymbol{\phi}[s])\frac{q(\boldsymbol{\phi}[s]|\boldsymbol{\lambda})}{r(\boldsymbol{\phi}[s]|\boldsymbol{\lambda},\tau)}, \qquad \boldsymbol{\phi}[s] \sim r(\boldsymbol{\phi}|\boldsymbol{\lambda},\tau), \tag{3.43}$$

as

$$\mathbb{V}ar\big[\widehat{\mathcal{L}}^{O}_{MC}\big] = \frac{1}{S}\mathbb{E}_q\big[\xi^2(\boldsymbol{\phi})\frac{q(\boldsymbol{\phi}|\boldsymbol{\lambda})}{r(\boldsymbol{\phi}|\boldsymbol{\lambda},\tau)}\big] - \frac{1}{S}\big[\tilde{\mathcal{L}}_{D_l}(\lambda)\big]^2. \tag{3.44}$$

Now, if we choose the distribution $r(\boldsymbol{\phi}|\boldsymbol{\lambda},\tau)$ such that

$$\mathbb{E}_q\big[\xi^2(\boldsymbol{\phi})\frac{q(\boldsymbol{\phi}|\boldsymbol{\lambda})}{r(\boldsymbol{\phi}|\boldsymbol{\lambda},\tau)}\big] \leq \mathbb{E}_q\big[\xi^2(\boldsymbol{\phi})\big], \tag{3.45}$$

the variance reduction will be achieved. The optimal r^* obviously satisfies the condition (3.45). Ruiz et al. (2016) show that the choice of overdispersed version of the variational family $q(\boldsymbol{\phi}|\boldsymbol{\lambda})$ has similar effect on the variance reduction as the optimal r^*. The details on the form of overdispersed families for specific variational families are discussed later in Section 3.3.4.

Combining the ideas of the Rao-Blackwellization, CV, and importance sampling, we have the following j^{th} entry of the MC approximation of the gradient estimator $\tilde{\mathcal{L}}_{D_l}(\boldsymbol{\lambda})$

$$\begin{aligned}
&\tilde{\mathcal{L}}^{OCV(j)}_{D_l}(\boldsymbol{\lambda}) \\
&= \sum_{s=1}^{S}\Bigg[\frac{l(2N-(l+1))}{2S}\nabla_{\boldsymbol{\lambda}_j}\log q(\phi_j[s]|\boldsymbol{\lambda}_j)(\tilde{p}_{(j)}(\phi[s]) - \frac{2(\log\frac{q(\phi_j[s]|\boldsymbol{\lambda}_j)}{p(\phi_j[s])} + \tilde{a}^D_j)}{l(2N-(l+1))})w(\phi_j[s])\Bigg],
\end{aligned}$$

where $\boldsymbol{\phi}[s] \sim r(\boldsymbol{\phi}|\boldsymbol{\lambda}, \tau)$ and

$$\tilde{a}_j^D = \frac{\widehat{\mathbb{C}ov}_r\left(\frac{l(2N-(l+1))w(\phi_j)}{2}\nabla_{\boldsymbol{\lambda}_j}\log q(\phi_j|\boldsymbol{\lambda}_j)(\tilde{p}_{(j)}(\boldsymbol{\phi}) - \frac{2\log\frac{q(\phi_j|\boldsymbol{\lambda}_j)}{p(\phi_j)}}{l(2N-(l+1))}), \nabla_{\boldsymbol{\lambda}_j}\log q(\phi_j|\boldsymbol{\lambda}_j)w(\phi_j)\right)}{\widehat{\mathbb{V}ar}_r(\nabla_{\boldsymbol{\lambda}_j}\log q(\phi_j|\boldsymbol{\lambda}_j)w(\phi_j))}.$$

The extension of Algorithm 3.1 with the variance reductions of the MC approximations due to the Rao-Blackwellization, control variates, and importance sampling is in Algorithm 3.2.

Algorithm 3.2: Variational calibration with truncated D-vine copulas II.

Input: Data $\boldsymbol{d}$, mean and covariance functions for GPs in Kennedy-O'Hagan framework, variational family $q(\boldsymbol{\phi}|\boldsymbol{\lambda})$, dispersion parameter τ **truncation level l**

$\lambda \leftarrow$ random initial value
$t \leftarrow 1$
repeat
  for $s = 1$ *to* S **do**
    $\boldsymbol{\phi}[s] \sim r(\boldsymbol{\phi}|\boldsymbol{\lambda}, \tau)$ // Random sample from r
  $K \leftarrow U(1, \ldots, \frac{l(2N-(l+1))}{2})$
  $\rho \leftarrow t^{\text{th}}$ value of a Robbins-Monro sequence
  $\boldsymbol{\lambda} \leftarrow \boldsymbol{\lambda} + \boldsymbol{\rho}\sum_{s=1}^{S}\left[\frac{l(2N-(l+1))}{2S}\nabla_{\boldsymbol{\lambda}_j}\log q(\phi_j[s]|\boldsymbol{\lambda}_j)\big(\tilde{p}_{(j)}(\boldsymbol{\phi}[s]) - \frac{2(\log\frac{q(\phi_j[s]|\boldsymbol{\lambda}_j)}{p(\phi_j[s])} + \tilde{a}_j^D)}{l(2N-(l+1))}\big)w(\phi_j[s])\right]$
  $t \leftarrow t + 1$
until *change of* $\boldsymbol{\lambda}$ *is less than* ϵ

3.3.3 Choice of the learning rate

Even though the SGA is straightforward in its general definition, the choice of learning rate ρ_t can be challenging in practice. Ideally, one would want the rate to be small in the situations where the noisy estimates of the gradient have large variance and vice-versa. The elements of variational parameter $\boldsymbol{\lambda}$ can also differ in scale, and one needs to set the learning

rate so that the SGA can accommodate even the smallest scales. The rapidly increasing usage of machine learning techniques in recent years produced various algorithms for element-wise adaptive-scale learning rates. We use the adaptive gradient (AdaGrad) algorithm (Duchi et al., 2011) which has been considered in similar problems before, e.g., Ranganath et al. (2014), however, there are other popular algorithms such as the ADADELTA (Zeiler, 2012) or the RMSProp (Tieleman and Hinton, 2012). Let $\boldsymbol{g}_T$ be the gradient used in the T^{th} step of the SGA algorithm, and $\boldsymbol{G}_t$ be the matrix consisting of the sum of the outer products of these gradients across the first t iterations, namely

$$\boldsymbol{G}_t = \sum_{T=1}^{t} \boldsymbol{g}_T \boldsymbol{g}_T^T. \tag{3.46}$$

The AdaGrad defines the element-wise adaptive scale learning rate as

$$\boldsymbol{\rho}_t = \eta \cdot \text{diag}(\boldsymbol{G}_t)^{-1/2}, \tag{3.47}$$

where η is the initial learning rate. It is a common practice, however, to add a small constant value to $\text{diag}(\boldsymbol{G}_t)$ (typically of order 10^{-6}) to avoid division by zero.

3.3.4 Parametrizations

Variational families. We use a Gaussian distribution for real valued components of $\boldsymbol{\phi}$ and a gamma distribution for positive variables. Both of these families are parametrized in terms of their mean and standard deviation. Moreover, in order to avoid constrained optimization, we transform all the positive variational parameters $\boldsymbol{\lambda}$ to $\tilde{\boldsymbol{\lambda}} = \log(e^{\boldsymbol{\lambda}} - 1)$ and optimize with respect to $\tilde{\boldsymbol{\lambda}}$.

Overdispersed families. Given a fixed dispersion coefficient τ, the overdispersed Gaussian distribution with mean μ and standard deviation σ is a Gaussian distribution with mean μ and standard deviation $\sigma\sqrt{\tau}$. The overdispersed gamma distribution with mean μ and standard deviation σ is a gamma distribution with mean $\mu + (\tau - 1)\frac{\sigma^2}{\mu}$ and standard deviation $\sigma \times \frac{\sqrt{\tau\mu^2 + \tau\sigma^2(\tau-1)}}{\mu}$ (Ruiz et al., 2016).

3.4 Applications

This section empirically establishes the efficiency of our methodology for the VBI based calibration of computer models. First, we conduct an extensive simulation study, where we focus both on the fidelity of variational approximation and prediction accuracy. Second, we demonstrate the opportunities in UQ given by the proposed methodology on calibration of the Liquid Drop Model.

The Bayesian predictions of new observations from the physical process ζ at input locations $(\boldsymbol{t}_1^*, \ldots, \boldsymbol{t}_J^*)$ are obtained according to (3.6). The conditional distribution $p(\boldsymbol{y}^*|\boldsymbol{d}, \boldsymbol{\phi})$ is a multivariate normal distribution with the mean vector

$$M_{y^*}(\boldsymbol{\phi}) = M_f(T_y^*(\boldsymbol{\theta})) + M_\delta(T_y^*) + C_* K(\boldsymbol{\phi})^{-1}(\boldsymbol{d} - M(\boldsymbol{\phi})), \tag{3.48}$$

and the covariance matrix

$$K_{y^*}(\boldsymbol{\phi}) = K_f(T_y^*(\boldsymbol{\theta}), T_y(\boldsymbol{\theta})) + K_\delta(T_y^*, T_y) + \sigma^2 I_m - C_* K(\boldsymbol{\phi})^{-1} C_*^T, \tag{3.49}$$

where

$$C_* = \begin{pmatrix} K_f(T_y^*(\boldsymbol{\theta}), T_y(\boldsymbol{\theta})) + K_\delta(T_y^*, T_y) & K_f(T_y^*(\boldsymbol{\theta}), T_z(\widetilde{\boldsymbol{\theta}})) \end{pmatrix}. \tag{3.50}$$

Here, $M(\boldsymbol{\phi})$ and $K(\boldsymbol{\phi})$ is the mean vector and the covariance matrix of the data likelihood $p(\boldsymbol{d}|\boldsymbol{\phi})$, $K_f(T_y^*(\boldsymbol{\theta}), T_y(\boldsymbol{\theta}))$ is the matrix with (i,j) element $k_f((\boldsymbol{t}_i^*, \boldsymbol{\theta}), (\boldsymbol{t}_j, \boldsymbol{\theta}))$ and $K_\delta(T_y^*, T_y)$ is the matrix with (i,j) element $k_\delta(\boldsymbol{t}_i^*, \boldsymbol{t}_j)$. We can similarly define $K_f(T_y^*(\boldsymbol{\theta}), T_z(\widetilde{\boldsymbol{\theta}}))$ with the kernel k_f.

3.4.1 Simulation study

In this section, we study Algorithm 3.2 in a simulated scenario, where we first demonstrate the method's fidelity in approximating the posterior distribution of calibration parameters $p(\boldsymbol{\theta}|\boldsymbol{d})$ and substantiate the indispensability of the variance reduction techniques described in Section 3.3.2 in order to achieve convergence. Second, we show the scalability of our method in comparison to the popular MH algorithm and the NUTS.

Let us consider a simple scenario following the model (3.4) with a two-dimensional calibration parameter $\boldsymbol{\theta} = (0.39, 0.60)$ that was obtained as a sample from its prior distribution $p(\boldsymbol{\theta})$ and a two-dimensional input variable $\boldsymbol{t} = (t_1, t_2)$. We model $f_m(\boldsymbol{t}, \boldsymbol{\theta})$ and $\delta(\boldsymbol{t})$ with GPs according to the specifications in Table 3.1 with the particular choices of $\eta_f = \frac{1}{30}$, $l_t = 1$, $l_\theta = 1$, $\eta_\delta = \frac{1}{30}$, $l_\delta = \frac{1}{2}$, and $\beta_\delta = 0.15$.

	GP mean	GP covariance function
f_m	$\theta_1 cos(t_1) + \theta_2 sin(t_2)$	$\eta_f \cdot \exp(-\frac{\|\|\boldsymbol{t}-\boldsymbol{t'}\|\|^2}{2l_t^2} - \frac{\|\|\boldsymbol{\theta}-\boldsymbol{\theta'}\|\|^2}{2l_\theta^2})$
δ	β_δ	$\eta_\delta \cdot \exp(-\frac{\|\|\boldsymbol{t}-\boldsymbol{t'}\|\|^2}{2l_\delta^2})$

Table 3.1: The specification of GPs for the simulation study.

We choose the variational family to be the mean-field family with Gaussian distributions for real valued parameters and gamma distributions for positive variables following the parametrization discussed in Section 3.3.4. The variational parameters are initialized to match the prior distributions, and we use the AdaGrad for the learning rate updates.

Calibration. For the purpose of model calibration, we sampled the data $\boldsymbol{d}$ jointly from the prior with the experimental noise following $\mathcal{N}(0, \frac{1}{100})$. The calibration parameter values for the model runs $\boldsymbol{z}$ were selected on a uniform grid over $[0, 1]^2$ and the inputs $\boldsymbol{t}$ over $[0, 3]^2$. For the first set of experiments, the size of the dataset was $N = 225$ with $n = 144$ and $s = 81$. We used 50 samples from the variational family to approximate the expectations in Algorithm 3.2 and 10 samples to implement the control variates.

Figure 3.2 demonstrates the quality of the variational approximation (Algorithm 3.2) in comparison to the MH algorithm and the NUTS. We can see that our method was able to accurately match both MCMC-based approximations with a minor deviation in θ_1. It is important to note, however, that the variance reduction through the combination of the Rao-Blackwellization, control variates, and importance sampling was necessary to achieve meaningful convergence.

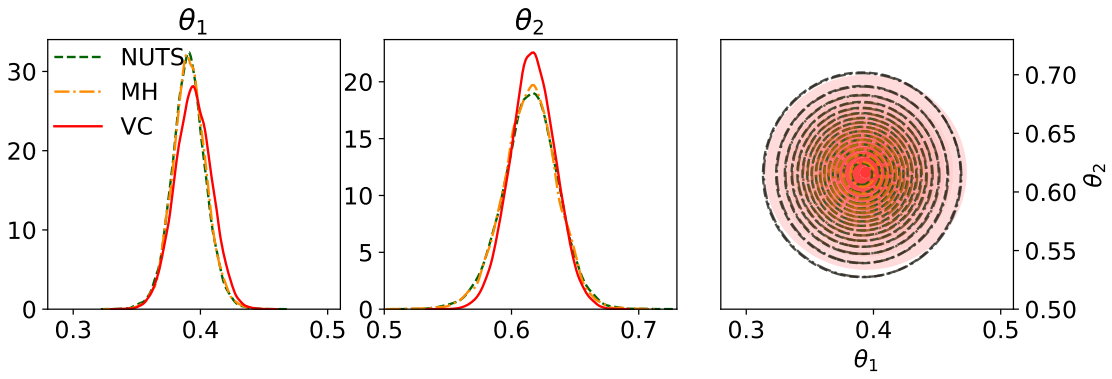

Figure 3.2: The approximate posterior distributions for the target calibration parameters. The VC (Algorithm 3.2) was carried out using $l = 3$ truncated D-vine and compared with the results from the NUTS and the MH algorithm.

In particular, Figure 3.3 shows the MSE of the posterior predictive means, evaluated on an independently generated set of 50 data points, based on the VC with cumulatively implemented variance reduction techniques. Algorithm 3.2 which employs the importance sampling clearly outperforms the calibration with only the Rao-Blackwellization and the calibration with control variates. In fact, each additional attempt to reduce the variance tends to decrease the MSE by one order of magnitude. There is naturally a time and space (memory) cost associated with each variance reduction technique. Figure 3.3 shows that the control variates and the importance sampling practically double the time per iteration of the algorithm. This additional complexity is, however, outweighed by the gain in the MSE reduction. The increase in memory consumption is less significant and is due to the storage of dispersion coefficients used for importance sampling and samples needed to compute control variates. Note that the memory consumed by the algorithms rises over time, because we chose to store the values of variational parameters during each step; the memory demands can be dramatically reduced if we drop these intermediate results.

For completeness, in Table 3.2, we also compare the MSE of the MCMC approximations and the VC at the point of convergence of the algorithms. The resulting errors in the predictions were, for all practical purposes, equivalent.

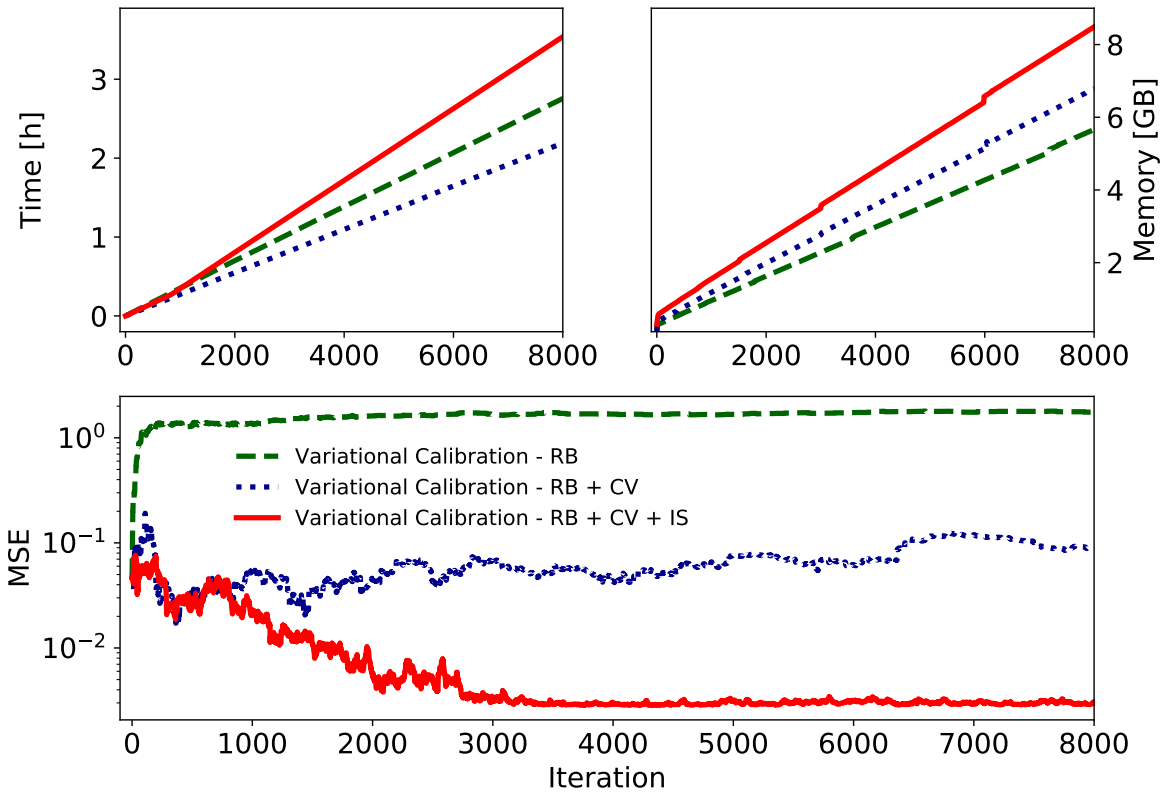

Figure 3.3: The evolution of MSE of the posterior predictive means based on the VC with cumulatively implemented variance reduction techniques described in Section 3.3.2. The figure is based on an independently generated set of 50 testing points. Time and memory demands for each of the implementations are also plotted the VC (Algorithm 3.2) was carried out using $l = 3$ truncated D-vine.

Algorithm	MSE
Variational Calibration - RB + CV + IS	2.9×10^{-3}
Metropolis-Hastings	3.0×10^{-3}
No-U-Turn	3.0×10^{-3}

Table 3.2: Comparison of the MSE for the simple scenario using the MH, the NUTS, and the VC algorithms.

Scalability. We now significantly increase the size of the dataset from $N = 225$ to 0.5×10^4 and eventually to 2×10^4 with the simulated experimental measurements and the model runs split equally ($n = s$). For better numerical stability, we expand the space of the input variables to $\boldsymbol{t} \in [0, 10]^2$ and select those using the Latin hypercube design. We also enlarge the testing dataset to 200 points. All the remaining simulation parameters are

unchanged. The conventional MCMC methods are already impractical for the purpose of Bayesian calibration with these moderately large amounts of data. We were able to obtain only around 600 posterior samples in the case of $N = 1 \times 10^4$ and about 120 for $N = 2 \times 10^4$ in 25 hours of sampling using the MH algorithm (significantly less with the NUTS).

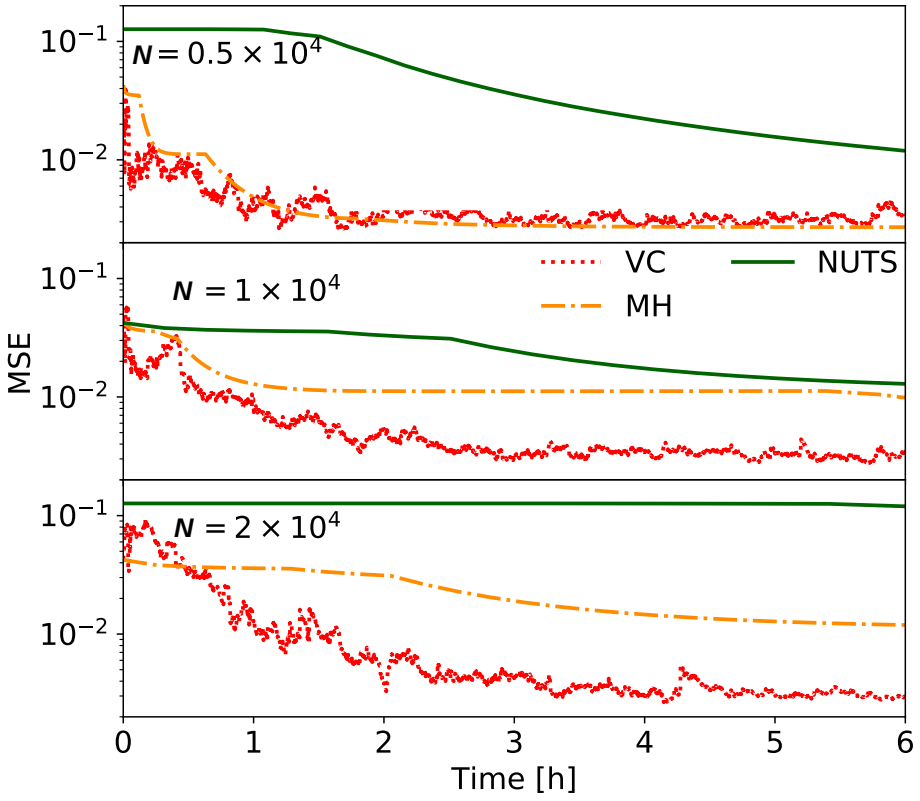

Figure 3.4: The evolution of the MSE of the posterior predictive means based on the VC (Algorithm 3.2), the MH algorithm, and the NUTS. The figure is based on an independently generated set of 200 testing points. The VC (Algorithm 3.2) was carried out using $l = 5$ truncated D-vine.

Figure 3.4 demonstrates that Algorithm 3.2 (D-vine with truncation $l = 5$) converges to the predictive MSE of about 0.003 under 4 hours for $N = 2 \times 10^4$ and 2 hours for $N = 0.5 \times 10^4$. It took similar time for the MH to achieve this MSE value for $N = 0.5 \times 10^4$ but almost 25 hours for the NUTS. Once we increased the data size to 2×10^4, neither the NUTS nor the MH were able to achieve a similar predictive MSE as the VC within the 25 hour window allotted for sampling. In fact, they were by an order of magnitude larger. It

is important to mention that both MCMC-based algorithms have also substantially larger memory demands than the VC as depicted in Figure 3.5. These memory profiles were recorded during a one hour period of running the algorithms. The MH algorithm and the NUTS were implemented in Python 3.0 using the PyMC3 module version 3.5. The memory profiles were measured using the memory-profiler module version 0.55.0 in Python 3.0. The VC was also implemented in Python 3.0.

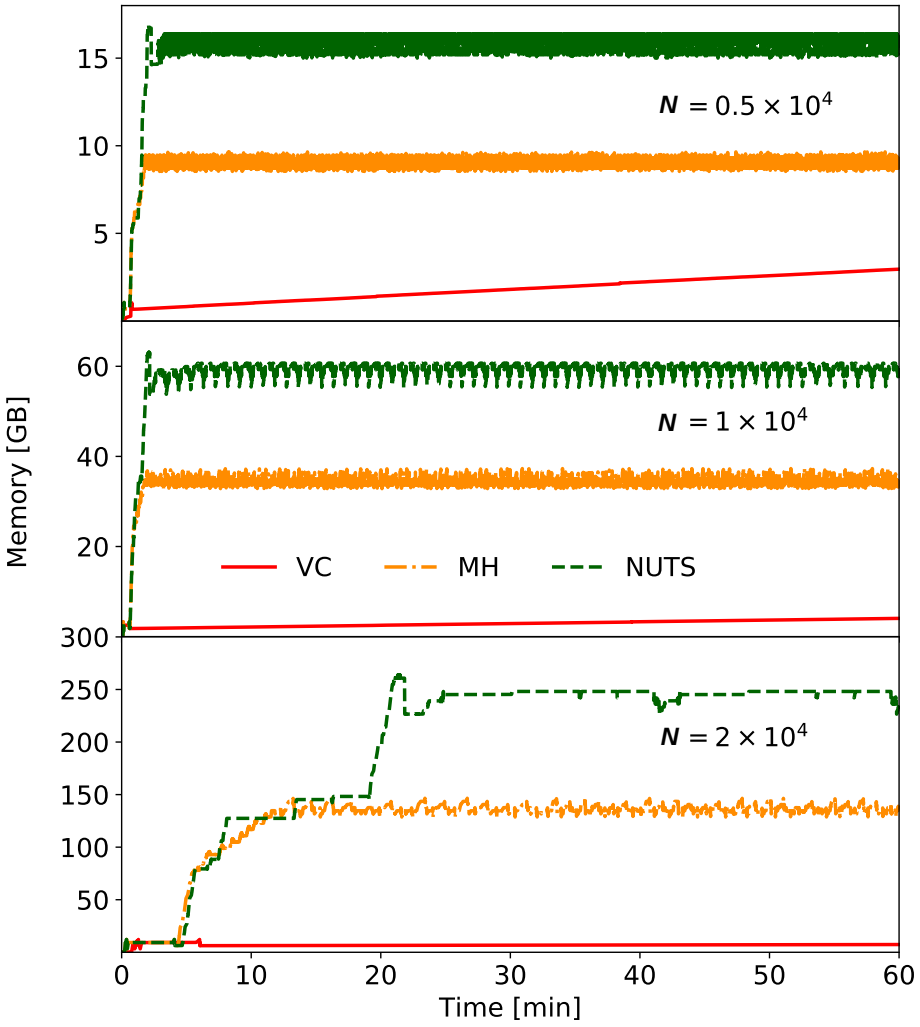

Figure 3.5: Recorded memory profiles of Algorithm 3.2, the MH algorithm, and the NUTS for the duration of 1 hour under the simulation scenario.

3.4.2 Calibration of the Liquid Drop Model

Over the past decade or so, the statistical tools of UQ have experienced a robust ramp-up in use in the field of nuclear physics (Ireland and Nazarewicz, 2015). Bayesian calibration has been especially popular because it enhances the understanding of a nuclear model's structure through parameter estimation and potentially advances the quality of nuclear modeling by accounting for systematic errors. In this context, we use our variational Algorithm 3.2 to calibrate the 4-parameter LDM. Since we discussed the LDM several times in this dissertation, we refer reader to Section 1.1 for a detailed description of the model.

Here we also note that this is by no means the first case when Bayesian calibration methodology is applied to study the LDM. In fact, the LDM is a popular model for statistical applications (Bertsch et al., 2005; Yuan, 2016; Bertsch and Bingham, 2017) which is why we choose the model to illustrate our methodology as well. The LDM also generally performs better on heavy nuclei as compared to the light nuclei which alludes to the existence of a significant systematic discrepancy between the model and the experimental binding energies (Reinhard et al., 2006; Kejzlar et al., 2020). Namely, we consider the following statistical model

$$y = E_{\mathrm{B}}(N, Z) + \delta(N, Z) + \sigma\epsilon, \tag{3.51}$$

where $\delta(N, Z)$ represents the unknown systematic discrepancy between the semi-empirical mass formula $E_{\mathrm{B}}(N, Z)$ and the experimental binding energies y. The parameter σ is as usual the scale of observation error $\epsilon \sim \mathcal{N}(0, 1)$. The nuclear physics community often (Dobaczewski et al., 2014) considers the least squares (LS) estimator of $\boldsymbol{\theta}$ defined as

$$\hat{\boldsymbol{\theta}}_{L_2} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(y_i - E_{\mathrm{B}}(N_i, Z_i)\right)^2, \tag{3.52}$$

which is also the maximum likelihood estimate of $\boldsymbol{\theta}$ in the case of $\delta = 0$. The benefit of this estimator is that it is fast, easy-to-compute, and allows for analysis under the standard linear regression theory. It, however, neglects some sources of uncertainty that are accounted for in the Bayesian calibration framework.

To this end, we shall consider a GP prior with the mean zero and the squared exponential covariance function for the systematic discrepancy $\delta(Z, N)$. Since the main purpose of the example is to provide a canonical illustration of the methodology in a real data scenario, we also set a GP prior for the LDM and treat $E_B(Z, N)$ as an unknown function. We use 2000 experimental binding energies randomly selected from the AME2003 dataset (Audi et al., 2003) (publicly available at `http://amdc.impcas.ac.cn/web/masstab.html`) for calibration, see Figure 3.6, and an additional set of 10^4 model evaluations. The calibration inputs were generated with the Latin hypercube design so that all the reasonable values of $(a_{\text{vol}}, a_{\text{surf}}, a_{\text{sym}}, a_{\text{C}})$ given by the literature are covered (Weizsäcker, 1935; Bethe and Bacher, 1936; Myers and Swiatecki, 1966; Kirson, 2008; Benzaid et al., 2020). The model inputs (Z, N) were selected from the set of 2000 experimental binding energies, duplicated five-fold, and randomly permutated among the generated calibration inputs to span only the set of relevant nuclei. This relatively large number of model runs was chosen so that the

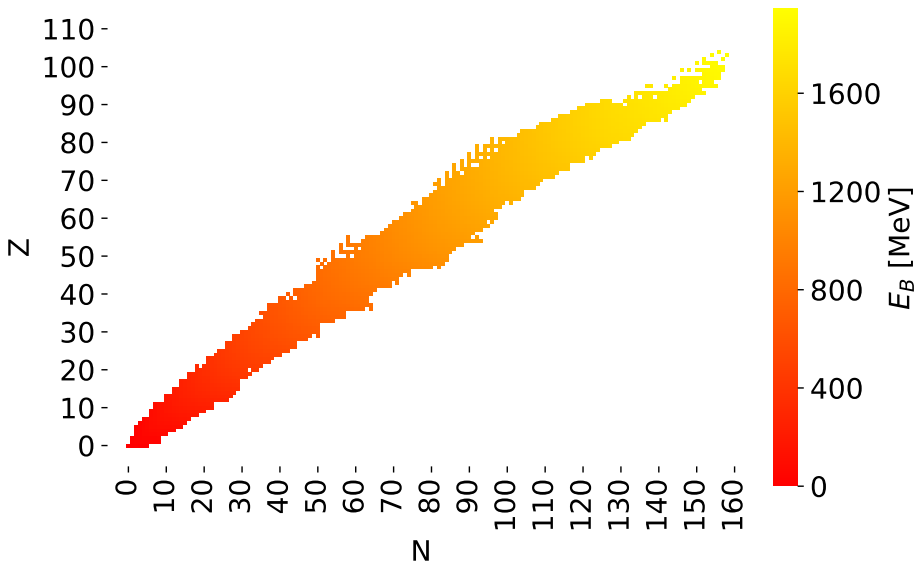

Figure 3.6: Experimental binding energies of nuclei in AME2003 dataset (2225 observations).

combined 6 dimensional space of calibration parameters and model inputs is sufficiently covered considering the existence of a non-trivial systematic discrepancy. In fact, the uniform

experimental design would amount only to 4-5 points per dimension.

Independent Gaussian distributions centered at the LS estimates $\hat{\boldsymbol{\theta}}_{L_2}$ (in Table 3.3) with standard deviations large enough to cover the space of inputs used for generating the model runs were selected to represent the prior knowledge about the calibration parameters. Independent gamma distributions were used as the prior models for the hyperparameters of the GP's covariance functions. We choose the variational family to be fully-factorized with the Gaussian distributions for real valued parameters and the gamma distributions for positive variables. The means of variational families were initialized as random samples from their respective prior distributions and the variances were set to match those of the prior distributions. We used the AdaGrad for stochastic optimization. See Section 3.5.3 for further discussion on the prior distributions and the experimental design.

Results. Including the generated model runs, the overall size of the training dataset is 1.2×10^4 which already makes the MCMC-based calibration impractical, as illustrated by the simulation study in Section 3.4.1. We therefore asses the quality of variational approximation only against the standard LS estimation and do not consider the MCMC methods. In particular, we consider the testing dataset of the remaining 225 experimental binding energies in AME2003 that were excluded from the training data. The predictions $\hat{\boldsymbol{y}}^*$ of these testing binding energies $\boldsymbol{y}^*$ were calculated, under the variational approximation, as the posterior means of $\boldsymbol{y}^*$ conditioned on the 1.2×10^4 binding energies from the training data set, i.e., the posterior means of the predictive distribution $p(\boldsymbol{y}^*|\boldsymbol{d})$. The predictions under the LS estimates $\hat{\boldsymbol{\theta}}_{L_2}$ were given by the semi-empirical mass formula (1.1).

Table 3.3 gives the RMSEs for both methods under consideration. The VC (Algorithm 3.2) results are based on a 24 hour window dedicated to running the algorithm with 50 samples used to approximate the expectations, 10 samples used to implement the control variates, and the truncation level selected to be $l = 3$. By using GPs to account for the systematic discrepancies of the semi-empirical mass formula and the uncertainty of the LDM

itself, we were able to significantly reduce the RMSE approx. 57% compared to the LS benchmark. Table 3.3 additionally shows the calibration parameter estimates and their standard errors. The estimates under the VC are given by the means of their variational families. Both the methods calibrate the LDM around the same values with notably low standard errors of the LS estimates. This is, however, expected since $\hat{\boldsymbol{\theta}}_{L_2}$ are ordinary LS estimates that in the presence of heteroscedasticity (see Figure 3.7) become inefficient and tend to significantly underestimate the true variance (Goldberger, 1966; Johnston, 1976).

Method	Parameter estimate and standard errors				Testing error
	a_{vol}	a_{surf}	a_{sym}	a_{C}	RMSE (MeV)
LS	15.42 (0.027)	16.91 (0.086)	22.47 (0.070)	0.69 (0.002)	3.54
VC	15.78 (0.198)	15.99 (0.681)	21.94 (0.510)	0.68 (0.018)	1.52

Table 3.3: The RMSE of the VC (Algorithm 3.2) after 24 hours dedicated to running the algorithm compared with the RMSE based on the LS estimates. The parameter estimates (and their standard errors) are also displayed.

The residual plot in Figure 3.7, showing the difference between $\boldsymbol{y}^*$ and $\hat{\boldsymbol{y}}^*$ as a function of the nuclear mass number A, clearly demonstrates a better fit of the testing data with our methodology than is achieved by the simple LS fit. The majority of the residuals appear to be randomly spread around 0 which strongly supports the efficiency of GPs in accounting for the systematic discrepancy δ.

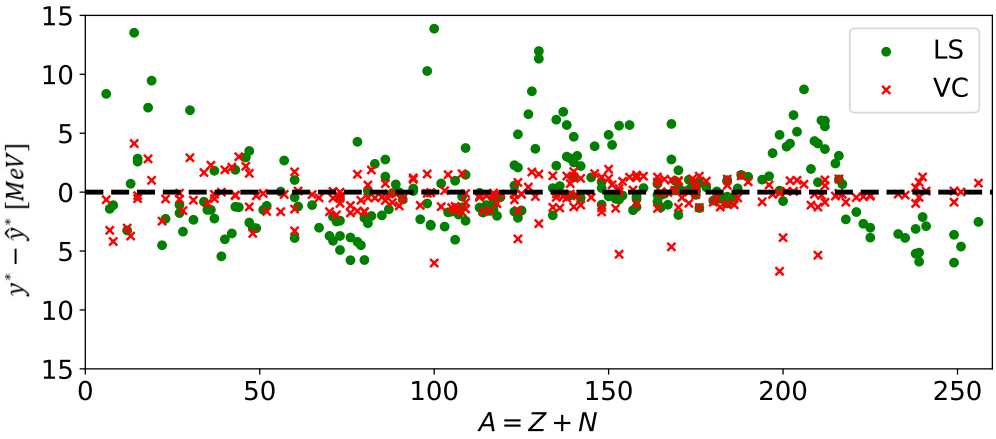

Figure 3.7: The residual plot for 225 experimental binding energies in the testing dataset.

3.5 Technical details and supplementary results

3.5.1 Scalable algorithm with truncated vine copulas: C-vine

Here we present the details of the C-vine based versions of Algorithm 3.1 and Algorithm 3.2. First, we can decompose the log-likelihood $\log p(\boldsymbol{d}|\boldsymbol{\phi})$ using a C-vine as

$$\log p(\boldsymbol{d}|\boldsymbol{\phi}) = \sum_{j=1}^{N-1} \sum_{i=1}^{N-j} p^C_{j,j+i}(\boldsymbol{\phi}), \tag{3.53}$$

where

$$p^C_{j,j+i}(\boldsymbol{\phi}) = \log c_{j,(j+i);1,\ldots,(j-1)} + \frac{1}{N-1}\big(\log p_j(d_j|\boldsymbol{\phi}) + \log p_{j+i}(d_{j+i}|\boldsymbol{\phi})\big). \tag{3.54}$$

This now yields the following expression for the ELBO gradient:

$$\nabla_\lambda \mathcal{L}(\lambda) = \sum_{j=1}^{N-1} \sum_{i=1}^{N-j} E_q\Big[\nabla_{\boldsymbol{\lambda}} \log q(\boldsymbol{\phi}|\boldsymbol{\lambda})(p^C_{j,j+i}(\boldsymbol{\phi}))\Big] - E_q\Big[\nabla_{\boldsymbol{\lambda}} \log q(\boldsymbol{\phi}|\boldsymbol{\lambda}) \log \frac{q(\boldsymbol{\phi}|\boldsymbol{\lambda})}{p(\boldsymbol{\phi})}\Big]. \tag{3.55}$$

Equivalently to Proposition 1, we have the following proposition that establishes the noisy unbiased estimate of the gradient (3.55) using the C-vine copula decomposition.

Proposition 4. *Let $\tilde{\mathcal{L}}_C(\boldsymbol{\lambda})$ be an estimate of the ELBO gradient $\nabla_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda})$ defined as*

$$\tilde{\mathcal{L}}_C(\boldsymbol{\lambda}) = \frac{N(N-1)}{2}\mathbb{E}_q\Big[\nabla_{\boldsymbol{\lambda}} \log q(\boldsymbol{\phi}|\boldsymbol{\lambda})(p^C_{I_C(K)}(\boldsymbol{\phi}))\Big] - \mathbb{E}_q\Big[\nabla_{\boldsymbol{\lambda}} \log q(\boldsymbol{\phi}|\boldsymbol{\lambda}) \log \frac{q(\boldsymbol{\phi}|\boldsymbol{\lambda})}{p(\boldsymbol{\phi})}\Big],$$

where $K \sim U(1,\ldots,\frac{N(N-1)}{2})$, and I_C is the bijection

$$I_C : \{1,\ldots,\frac{N(N-1)}{2}\} \to \{(j, j+i) : i \in \{1,\ldots,N-j\} \text{ for } j \in \{1,\ldots N-1\}\},$$

then $\tilde{\mathcal{L}}_C(\boldsymbol{\lambda})$ is unbiased i.e., $\mathbb{E}(\tilde{\mathcal{L}}_C(\boldsymbol{\lambda})) = \nabla_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda})$.

Again, $\tilde{\mathcal{L}}_C(\boldsymbol{\lambda})$ can be relatively costly to compute for large datasets due to the recursive nature of the copula density computations. We now carry out exactly the same development an using l-truncated C-vine as in the case of Proposition 2 and Proposition 3.

Proposition 5. *If the copula of $p(\boldsymbol{d}|\boldsymbol{\phi})$ is distributed according to an l-truncated C-vine, we can rewrite*

$$\log p(\boldsymbol{d}|\boldsymbol{\phi}) = \sum_{j=1}^{l}\sum_{i=1}^{N-j} p_{i,i+j}^{C_l}(\boldsymbol{\phi}), \tag{3.56}$$

where

$$p_{i,i+j}^{c_l}(\boldsymbol{\phi}) = \log c_{j,(j+i);1,\ldots,(j-1)} + \frac{1}{a_j}\log p_j(d_j|\boldsymbol{\phi}) + \frac{1}{b_{j+i}}\log p_{j+i}(d_{j+i}|\boldsymbol{\phi}), \tag{3.57}$$

and

$$a_j = N-1,$$

$$b_{j+i} = (N-1-l)\mathbb{1}_{j+i\leq l} + l.$$

Let us now replace the full log-likelihood $\log(\boldsymbol{d}|\boldsymbol{\phi})$ in the definition of ELBO with the likelihood based on a truncated vine copula. This yields the l-truncated ELBO for the l-truncated C-vine

$$\mathcal{L}_{C_l}(\boldsymbol{\lambda}) = \mathbb{E}_q\Bigg[\sum_{j=1}^{l}\sum_{i=1}^{N-j} p_{j,j+i}^{C_l}(\boldsymbol{\phi})\Bigg] - KL(q(\boldsymbol{\phi}|\boldsymbol{\lambda})||p(\boldsymbol{\phi})) \tag{3.58}$$

with its gradient

$$\nabla_{\boldsymbol{\lambda}}\mathcal{L}_{C_l}(\boldsymbol{\lambda}) = \sum_{j=1}^{l}\sum_{i=1}^{N-j}\mathbb{E}_q\Bigg[\nabla_{\boldsymbol{\lambda}}\log q(\boldsymbol{\phi}|\boldsymbol{\lambda})(p_{j,j+i}^{C_l}(\boldsymbol{\phi}))\Bigg] - \mathbb{E}_q\Bigg[\nabla_{\boldsymbol{\lambda}}\log q(\boldsymbol{\phi}|\boldsymbol{\lambda})\log\frac{q(\boldsymbol{\phi}|\boldsymbol{\lambda})}{p(\boldsymbol{\phi})}\Bigg].$$

Consequently, we get the following proposition that establishes the noisy unbiased estimate of $\nabla_{\boldsymbol{\lambda}}\mathcal{L}_{C_l}(\boldsymbol{\lambda})$.

Proposition 6. *Let $\tilde{\mathcal{L}}_{C_l}(\boldsymbol{\lambda})$ be an estimate of the ELBO gradient $\nabla_{\boldsymbol{\lambda}}\mathcal{L}_{C_l}(\boldsymbol{\lambda})$ defined as*

$$\tilde{\mathcal{L}}_{C_l}(\boldsymbol{\lambda}) = \frac{l(2N-(l+1))}{2}\mathbb{E}_q\Bigg[\nabla_{\boldsymbol{\lambda}}\log q(\boldsymbol{\phi}|\boldsymbol{\lambda})(p_{I_{C_l}(K)}^{C_l}(\boldsymbol{\phi}))\Bigg] - \mathbb{E}_q\Bigg[\nabla_{\boldsymbol{\lambda}}\log q(\phi|\boldsymbol{\lambda})\log\frac{q(\phi|\boldsymbol{\lambda})}{p(\phi)}\Bigg],$$

where $K \sim U(1,\ldots,\frac{l(2N-(l+1))}{2})$, and I_{C_l} is the bijection

$$I_{C_l} : \{1,\ldots,\frac{l(2N-(l+1))}{2}\} \to \{(j,j+i) : i \in \{1,\ldots,N-j\} \text{ for } j \in \{1,\ldots l\}\},$$

then $\tilde{\mathcal{L}}_{C_l}(\boldsymbol{\lambda})$ is unbiased i.e., $\mathbb{E}(\tilde{\mathcal{L}}_{C_l}(\boldsymbol{\lambda})) = \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{C_l}(\boldsymbol{\lambda})$.

Algorithm 3.3 postulates the version of Algorithm 3.1 based on the truncated C-vine decomposition.

Algorithm 3.3: Variational calibration with truncated C-vine copulas.

Input: Data $\boldsymbol{d}$, mean and covariance functions for GPs in Kennedy-O'Hagan framework, variational family $q(\boldsymbol{\phi}|\boldsymbol{\lambda})$, **truncation level l**

$\lambda \leftarrow$ random initial value
$t \leftarrow 1$
repeat
  for $s = 1$ *to* S **do**
    $\boldsymbol{\phi}[s] \sim q(\boldsymbol{\phi}|\boldsymbol{\lambda})$ // Random sample from q
  $K \leftarrow U(1, \ldots, \frac{l(2N-(l+1))}{2})$
  $\rho \leftarrow t^{\text{th}}$ value of a Robbins-Monro sequence
  $\boldsymbol{\lambda} \leftarrow \boldsymbol{\lambda} + \rho \frac{1}{S} \sum_{s=1}^{S} \Big[\frac{l(2N-(l+1))}{2} \nabla_{\boldsymbol{\lambda}} \log q(\boldsymbol{\phi}[s]|\boldsymbol{\lambda}) \big(p^{C_l}_{I_{C_l}(K)}(\boldsymbol{\phi}[s]) - \frac{2}{l(2N-(l+1))} \log \frac{q(\boldsymbol{\phi}[s]|\boldsymbol{\lambda})}{p(\boldsymbol{\phi}[s])} \big) \Big]$
  $t \leftarrow t + 1$
until *change of* $\boldsymbol{\lambda}$ *is less than* ϵ

Variance reduction of MC estimates. Let us now consider the MC approximation of the gradient estimator $\tilde{\mathcal{L}}_{C_l}(\boldsymbol{\lambda})$, the j^{th} entry of the Rao-Blackwellized estimator is

$$\frac{1}{S} \sum_{s=1}^{S} \Big[\frac{l(2N-(l+1))}{2} \nabla_{\boldsymbol{\lambda}_j} \log q(\phi_j[s]|\boldsymbol{\lambda}_j) \big(\tilde{p}_{(j)}(\phi[s]) - \frac{2}{l(2N-(l+1))} \log \frac{q(\phi_j[s]|\boldsymbol{\lambda}_j)}{p(\phi_j[s])} \big) \Big],$$

where $\tilde{p}_{(j)}(\boldsymbol{\phi})$ are here the components of $p^{C_l}_{I_{C_l}(K)}(\boldsymbol{\phi})$ that include ϕ_j.

We can again use the control variates to reduce the variance of MC approximation of the gradient estimator $\tilde{\mathcal{L}}_{C_l}(\boldsymbol{\lambda})$. In particular, we consider the following j^{th} entry of the Rao-Blackwellized MC approximation of the gradient estimator $\tilde{\mathcal{L}}_{C_l}(\boldsymbol{\lambda})$ with control variates

$$\tilde{\mathcal{L}}^{CV(j)}_{C_l}(\boldsymbol{\lambda}) = \sum_{s=1}^{S} \Big[\frac{l(2N-(l+1))}{2S} \nabla_{\boldsymbol{\lambda}_j} \log q(\phi_j[s]|\boldsymbol{\lambda}_j) (\tilde{p}_{(j)}(\phi[s]) - \frac{2(\log \frac{q(\phi_j[s]|\boldsymbol{\lambda}_j)}{p(\phi_j[s])} + \hat{a}^C_j)}{l(2N-(l+1))}) \Big],$$

where $\hat{a}^C_j$ is the estimate of the optimal control variate scalar a^* based on S (or fever) independent draws from the variational distribution. Namely,

$$\hat{a}^C_j = \frac{\widehat{\mathbb{C}ov}_q(\frac{l(2N-(l+1))}{2} \nabla_{\boldsymbol{\lambda}_j} \log q(\phi_j|\boldsymbol{\lambda}_j) (\tilde{p}_{(j)}(\boldsymbol{\phi}) - \frac{2 \log q(\phi_j|\boldsymbol{\lambda}_j)}{l(2N-(l+1)) \log p(\phi_j)}), \nabla_{\boldsymbol{\lambda}_j} \log q(\phi_j|\boldsymbol{\lambda}_j))}{\widehat{\mathbb{V}ar}_q(\nabla_{\boldsymbol{\lambda}_j} \log q(\phi_j|\boldsymbol{\lambda}_j))}.$$

As in the case of the D-vine, we now derive the ultimate version of Algorithm 3.3. Again, instead of taking the samples from $q(\boldsymbol{\phi}|\boldsymbol{\lambda})$ to approximate the gradient estimates, we will take samples from an overdispersed distribution $r(\boldsymbol{\phi}|\boldsymbol{\lambda},\tau)$. Combining the Rao-Blackwellization, control variates, and importance sampling, we have the following j^{th} entry of the MC approximation of the gradient estimator $\tilde{\mathcal{L}}_{C_l}(\lambda)$

$$\tilde{\mathcal{L}}_{C_l}^{OCV(j)}(\boldsymbol{\lambda})$$
$$= \sum_{s=1}^{S}\left[\frac{l(2N-(l+1))}{2S}\nabla_{\boldsymbol{\lambda}_j}\log q(\phi_j[s]|\boldsymbol{\lambda}_j)(\tilde{p}_{(j)}(\phi[s]) - \frac{2(\log\frac{q(\phi_j[s]|\boldsymbol{\lambda}_j)}{p(\phi_j[s])} + \tilde{a}_j^C)}{l(2N-(l+1))})w(\phi_j[s])\right],$$

where $\boldsymbol{\phi}[s] \sim r(\boldsymbol{\phi}|\boldsymbol{\lambda},\tau)$ and $w(\boldsymbol{\phi}[s]) = q(\boldsymbol{\phi}[s]|\boldsymbol{\lambda})/r(\boldsymbol{\phi}[s]|\boldsymbol{\lambda},\tau)$ with

$$\tilde{a}_j^C =$$
$$\frac{\widehat{\mathbb{C}ov}_r(\frac{l(2N-(l+1))w(\phi_j)}{2}\nabla_{\boldsymbol{\lambda}_j}\log q(\phi_j|\boldsymbol{\lambda}_j)(\tilde{p}_{(j)}(\boldsymbol{\phi}) - \frac{2\log\frac{q(\phi_j|\boldsymbol{\lambda}_j)}{p(\phi_j)}}{l(2N-(l+1))}), \nabla_{\boldsymbol{\lambda}_j}\log q(\phi_j|\boldsymbol{\lambda}_j)w(\phi_j))}{\widehat{\mathbb{V}ar}_r(\nabla_{\boldsymbol{\lambda}_j}\log q(\phi_j|\boldsymbol{\lambda}_j)w(\phi_j))}.$$

Algorithm 3.4: Variational calibration with truncated C-vine copulas II.

Input: Data $\boldsymbol{d}$, mean and covariance functions for GPs, variational family $q(\boldsymbol{\phi}|\boldsymbol{\lambda})$, dispersion parameter τ **truncation level l**

$\lambda \leftarrow$ random initial value
$t \leftarrow 1$
repeat
  for $s = 1$ *to* S **do**
    $\boldsymbol{\phi}[s] \sim r(\boldsymbol{\phi}|\boldsymbol{\lambda},\tau)$ // Random sample from r
  $K \leftarrow U(1,\ldots,\frac{l(2N-(l+1))}{2})$
  $\rho \leftarrow t^{\text{th}}$ value of a Robbins-Monro sequence
  $\boldsymbol{\lambda} \leftarrow \boldsymbol{\lambda} + \boldsymbol{\rho}\sum_{s=1}^{S}\left[\frac{l(2N-(l+1))}{2S}\nabla_{\boldsymbol{\lambda}_j}\log q(\phi_j[s]|\boldsymbol{\lambda}_j)\big(\tilde{p}_{(j)}(\boldsymbol{\phi}[s]) - \frac{2(\log\frac{q(\phi_j[s]|\boldsymbol{\lambda}_j)}{p(\phi_j[s])}+\tilde{a}_j^C)}{l(2N-(l+1))}\big)w(\phi_j[s])\right]$
  $t \leftarrow t+1$
until *change of* $\boldsymbol{\lambda}$ *is less than* ϵ

3.5.2 Proofs

Proof of Proposition 1.

Since $P(K=k)=\frac{2}{N(N-1)}$, we have directly from the definition of expectation

$$\mathbb{E}(\tilde{\mathcal{L}}_D(\boldsymbol{\lambda})) = \frac{N(N-1)}{2} \sum_{k=1}^{\frac{N(N-1)}{2}} \frac{2}{N(N-1)} \mathbb{E}_q\Big[\nabla_{\boldsymbol{\lambda}} \log q(\boldsymbol{\phi}|\boldsymbol{\lambda})(p^D_{I_D(k)}(\boldsymbol{\phi}))\Big] - \mathbb{E}_q\Big[\nabla_{\boldsymbol{\lambda}} \log q(\boldsymbol{\phi}|\boldsymbol{\lambda}) \log \frac{q(\boldsymbol{\phi}|\boldsymbol{\lambda})}{p(\boldsymbol{\phi})}\Big] = \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda}).$$

The final equality is the consequence of the uniqueness of the pairs of variables in the conditioned sets of the copula density $c_{i,(i+j);(i+1),\ldots,(i+j-1)}$, and that $\frac{N(N-1)}{2}$ is the number of unordered pairs of N variables. □

Proof of Proposition 2.

It is sufficient to show that for $l \in \{1,\ldots,N-1\}$ the following equality holds:

$$\sum_{j=1}^{l} \sum_{i=1}^{N-j} \left[\frac{1}{a_i} \log p_i(d_i|\boldsymbol{\phi}) + \frac{1}{b_{i+j}} \log p_{i+j}(d_{i+j}|\boldsymbol{\phi})\right] = \sum_{k=1}^{N} \log p(\boldsymbol{d}_k|\boldsymbol{\phi}), \tag{3.59}$$

where

$$a_i = 2l - \Big[(l+1-i)\mathbb{1}_{i\leq l} + (l-N+i)\mathbb{1}_{i>N-l}\Big],$$
$$b_{i+j} = 2l - \Big[(l+1-j-i)\mathbb{1}_{i+j\leq l} + (l-N+j+i)\mathbb{1}_{i+j>N-l}\Big].$$

To show this, let us consider the summation

$$\sum_{j=1}^{l} \sum_{i=1}^{N-j} \Big[\log p_i(d_i|\boldsymbol{\phi}) + \log p_{i+j}(d_{i+j}|\boldsymbol{\phi})\Big]$$
$$= \sum_{j=1}^{l} \Big[(\log p_1(d_1|\boldsymbol{\phi}) + \log p_{1+j}(d_{1+j}|\boldsymbol{\phi})) + \cdots + (\log p_{N-j}(d_{N-j}|\boldsymbol{\phi}) + \log p_N(d_N|\boldsymbol{\phi}))\Big].$$

For $l = 1$, we get

$$\sum_{j=1}^{l}\sum_{i=1}^{N-j}\Big[\log p_i(d_i|\boldsymbol{\phi}) + \log p_{i+j}(d_{i+j}|\boldsymbol{\phi})\Big]$$
$$= (\log p_1(d_1|\boldsymbol{\phi}) + \log p_2(d_2|\boldsymbol{\phi})) + \cdots + (\log p_{N-1}(d_{N-1}|\boldsymbol{\phi}) + \log p_N(d_N|\boldsymbol{\phi})),$$

and for $l \geq 2$

$$\sum_{j=1}^{l}\sum_{i=1}^{N-j}\Big[\log p_i(d_i|\boldsymbol{\phi}) + \log p_{i+j}(d_{i+j}|\boldsymbol{\phi})\Big]$$
$$= \Big[(\log p_1(d_1|\boldsymbol{\phi}) + \log p_2(d_2|\boldsymbol{\phi})) + \cdots + (\log p_{N-1}(d_{N-1}|\boldsymbol{\phi}) + \log p_N(d_N|\boldsymbol{\phi}))\Big]$$
$$+ \cdots + \Big[(\log p_1(d_1|\boldsymbol{\phi}) + \log p_{1+l}(d_{1+l}|\boldsymbol{\phi})) + \cdots + (\log p_{N-l}(d_{N-l}|\boldsymbol{\phi}) + \log p_N(d_N|\boldsymbol{\phi}))\Big].$$

Note that in the case of $l = N - 1$, the last summation consists of only one element $\log p_1(d_1|\boldsymbol{\phi}) + \log p_{1+l}(d_{1+l}|\boldsymbol{\phi})$. By careful examination of the two cases above, we get the following results. For $2l \leq N$:

$$\sum_{j=1}^{l}\sum_{i=1}^{N-j}\Big[\log p_i(d_i|\boldsymbol{\phi}) + \log p_{i+j}(d_{i+j}|\boldsymbol{\phi})\Big]$$
$$= \sum_{k=1}^{l}(l+k-1)\log p_k(d_k|\boldsymbol{\phi}) + \sum_{k=l+1}^{N-l} 2l \log p_k(d_k|\boldsymbol{\phi}) + \sum_{k=N-l+1}^{N}(N-i+l)\log p_k(d_k|\boldsymbol{\phi}),$$

where the middle term disappears in the case $2l = N$, and for $2l > N$:

$$\sum_{j=1}^{l}\sum_{i=1}^{N-j}\Big[\log p_i(d_i|\boldsymbol{\phi}) + \log p_{i+j}(d_{i+j}|\boldsymbol{\phi})\Big]$$
$$= \sum_{k=1}^{N-l}(l+k-1)\log p_k(d_k|\boldsymbol{\phi}) + \sum_{k=N-l+1}^{l}(N-1)\log p_k(d_k|\boldsymbol{\phi})$$
$$+ \sum_{k=l+1}^{N}(N-i+l)\log p_k(d_k|\boldsymbol{\phi}).$$

If we now check that a_i equals to the factors in front of the log-likelihoods in the two cases above, the proof of Proposition 2 is complete. Note that once we check the equality for a_i,

the same directly translates to b_{i+j} since b_{i+j} is a_i with indices set to $i+j$ instead of i. Indeed, for $2l \leq N$

$$a_i = \begin{cases} l+i-1 & i \leq l \\ 2l & l < i \leq N-l \\ N-i+l & N-l < i \end{cases},$$

and for $2l > N$

$$a_i = \begin{cases} l+i-1 & i \leq N-l \\ N-1 & N-l < i \leq l \\ N-i+l & l < i \end{cases}.$$

□

Proof of Proposition 3.

By the construction of R-vine (see Section 3.2.1), each tree $\mathcal{T}_i$, for $i = 1, \ldots, N-1$ has exactly $N-i$ edges (these are the unique conditioned variable pairs). For any R-vine truncated at level $l \in \{1, \ldots, N-1\}$, we get the number of edges to be

$$\sum_{i=1}^{l}(N-i) = lN - \frac{l(l+1)}{2} = \frac{l(2N-(l+1))}{2}$$

The rest of the proof is identical with that of Proposition 1 due to the uniqueness of the conditioned variable pairs in the copula density $c_{i,(i+j);(i+1),\ldots,(i+j-1)}$, but in this case $P(K=k) = \frac{2}{l(2N-(l+1))}$. □

Proof of Proposition 4.

The proof is identical with that of Proposition 1 since each conditioned pair in the copula density $c_{j,(j+i);1,\ldots,(j-1)}$ is unique as well. □

Proof of Proposition 5.

It is sufficient to show that for $l \in \{1, \ldots, N-1\}$ the following equality holds:

$$\sum_{j=1}^{l} \sum_{i=1}^{N-j} \left[\frac{1}{a_j} \log p_j(d_j|\boldsymbol{\phi}) + \frac{1}{b_{j+i}} \log p_{j+i}(d_{j+i}|\boldsymbol{\phi}) \right] = \sum_{k=1}^{N} \log p(\boldsymbol{d}_k|\boldsymbol{\phi}), \tag{3.60}$$

where

$$a_j = N - 1,$$

$$b_{j+i} = (N - 1 - l)\mathbb{1}_{j+i \leq l} + l.$$

To show this, let us consider the following summation

$$\begin{aligned}
&\sum_{j=1}^{l} \sum_{i=1}^{N-j} \left[\log p_j(d_j|\boldsymbol{\phi}) + \log p_{j+i}(d_{j+i}|\boldsymbol{\phi}) \right] \\
&= \sum_{j=1}^{l} \left[(N-j) \log p_j(d_j|\boldsymbol{\phi}) + \sum_{i=1}^{N-j} \log p_{j+i}(d_{j+i}|\boldsymbol{\phi}) \right] \\
&= \sum_{j=1}^{l} (N-j) \log p_j(d_j|\boldsymbol{\phi}) + \sum_{j=1}^{l} \left[\log p_{j+1}(d_{j+i}|\boldsymbol{\phi})) + \cdots + \log p_N(d_N|\boldsymbol{\phi}) \right].
\end{aligned}$$

Now, for $l = 1$, we have

$$\sum_{j=1}^{l} \left[\log p_{j+1}(d_{j+1}|\boldsymbol{\phi})) + \cdots + \log p_N(d_N|\boldsymbol{\phi}) \right] = \log p_2(d_2|\boldsymbol{\phi}) + \ldots \log p_N(d_N|\boldsymbol{\phi}).$$

For $l \geq 2$, we have

$$\begin{aligned}
&\sum_{j=1}^{l} \left[\log p_{j+1}(d_{j+1}|\boldsymbol{\phi})) + \cdots + \log p_N(d_N|\boldsymbol{\phi}) \right] \\
&= \left[\log p_2(d_2|\boldsymbol{\phi}) + \ldots \log p_N(d_N|\boldsymbol{\phi}) \right] + \cdots + \left[\log p_{l+1}(d_{l+1}|\boldsymbol{\phi}) + \ldots \log p_N(d_N|\boldsymbol{\phi}) \right].
\end{aligned}$$

Therefore we can rewrite

$$\begin{aligned}
&\sum_{j=1}^{l} \left[\log p_{j+1}(d_{j+1}|\boldsymbol{\phi})) + \cdots + \log p_N(d_N|\boldsymbol{\phi}) \right] \\
&= \sum_{j=1}^{l} (j-1) \log p_j(d_j|\boldsymbol{\phi}) + \sum_{j=l+1}^{N} l \log p_j(d_j|\boldsymbol{\phi})
\end{aligned}$$

Overall,

$$\begin{aligned}
&\sum_{j=1}^{l}\sum_{i=1}^{N-j}\Big[\log p_j(d_j|\boldsymbol{\phi})+\log p_{j+i}(d_{j+i}|\boldsymbol{\phi})\Big]\\
&=\sum_{j=1}^{l}(N-j)\log p_j(d_j|\boldsymbol{\phi})+\sum_{j=1}^{l}(j-1)\log p_j(d_j|\boldsymbol{\phi})+\sum_{j=l+1}^{N} l\log p_j(d_j|\boldsymbol{\phi})\\
&=\sum_{k=1}^{l}(N-1)\log p_k(d_k|\boldsymbol{\phi})+\sum_{k=l+1}^{N} l\log p_k(d_k|\boldsymbol{\phi}).
\end{aligned}$$

Since $j\in\{1,\dots,l\}$ and

$$b_{j+i}=\begin{cases}N-1 & j+i\le l\\ l & j+i>l\end{cases},$$

the equality 3.60 holds. □

Proof of Proposition 6.

The proof is identical with that of Proposition 3 since each conditioned pair in the copula density $c_{j,(j+i);1,\dots,(j-1)}$ is unique, and a C-vine is a special case of R-vine. □

Proof of Lemma 2.

As we discussed in the proof of Proposition 3, the construction of R-vine implies that each tree $\mathcal{T}_i$, for $i=1,\dots,N-1$ has exactly $N-i$ edges (pairs of conditioned variables). Moreover, each tree $\mathcal{T}_i$ corresponds to copulas with the conditioning set of size $i-1$. Therefore, for X being the cardinality of the conditioning set, we get

$$P(X=i)=\frac{N-(i+1)}{\binom{N}{2}}\qquad \text{for } i\in\{0,\dots,N-2\}.$$

Now

$$\begin{aligned}
\mathbb{E}(X)&=\sum_{i=0}^{N-2} i\frac{N-(i+1)}{\binom{N}{2}}=\frac{2}{N(N-1)}\sum_{i=0}^{N-2}[i(N-1)-i^2]\\
&=\frac{2}{N(N-1)}\left[(N-1)[\frac{(N-2)(N-1)}{2}]-\frac{(N-2)(N-1)(2N-3)}{6}\right]=\frac{N-2}{3}.
\end{aligned}$$

Where the equality on the second line is due to the standard algebraic results on the sum of powers of the first first N integers. □

Proof of Lemma 3.

Analogically to the proof of Lemma 2, while recalling the number of edges for any l-truncated R-vine provided in the proof of Proposition 3, we have for the cardinality of the conditioning set X:

$$P(X = i) = \frac{N - (i+1)}{\frac{l(2N-(l+1))}{2}} \qquad \text{for } i \in \{0, \dots, l-1\}.$$

Now

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{i=0}^{l-1} i \frac{N-(i+1)}{\frac{l(2N-(l+1))}{2}} = \frac{2}{l(2N-(l+1))} \sum_{i=0}^{l-1} [i(N-1) - i^2] \\
&= \frac{2}{l(2N-(l+1))} \left[(N-1)[\frac{(l-1)l}{2}] - \frac{(l-1)l(2l-1)}{6} \right] \\
&= \frac{(l-1)(3N-2l-2)}{3(2N-l-1)}.
\end{aligned}$$

□

3.5.3 Supplement for the calibration of the Liquid Drop Model

GP specifications. In the case of the LDM $E_{\mathrm{B}}(Z, N)$, we consider the GP prior with the mean zero and the covariance function

$$\begin{aligned}
\eta_E \cdot \exp\Bigg(&- \frac{(Z-Z')^2}{2\nu_Z^2} - \frac{(N-N')^2}{2\nu_N^2} - \frac{(a_{\mathrm{vol}} - a'_{\mathrm{vol}})^2}{2\nu_1^2} \\
&- \frac{(a_{\mathrm{surf}} - a'_{\mathrm{surf}})^2}{2\nu_2^2} - \frac{(a_{\mathrm{sym}} - a'_{\mathrm{sym}})^2}{2\nu_3^2} - \frac{(a_{\mathrm{C}} - a'_{\mathrm{C}})^2}{2\nu_4^2} \Bigg).
\end{aligned}$$

Similarly, we consider the GP prior for the systematic discrepancy $\delta(Z, N)$ with mean zero and covariance function

$$\eta_\delta \cdot \exp\left(-\frac{(Z-Z')^2}{2l_Z^2} - \frac{(N-N')^2}{2l_N^2} \right).$$

Experimental design. Kennedy and O'Hagan (2001) recommend to select the calibration inputs for the model runs so that any plausible value $\boldsymbol{\theta}$ of the true calibration parameter is covered. In this context, we consider the space of calibration parameters to be centered at the values of least squares estimates $\hat{\boldsymbol{\theta}}_{L_2}$ and broad enough to contain the majority of values provided by the nuclear physics literature (Weizsäcker, 1935; Bethe and Bacher, 1936; Myers and Swiatecki, 1966; Kirson, 2008; Benzaid et al., 2020). Table 3.4 gives the lower and upper bounds for the parameter space so that Lower bound $= \hat{\theta}_{L_2} - 15 \times SE(\hat{\theta}_{L_2})$ and Upper bound $= \hat{\theta}_{L_2} + 15 \times SE(\hat{\theta}_{L_2})$. Here $SE(\hat{\theta}_{L_2})$ is given by the standard linear regression theory.

Parameter	Lower bound	Upper bound
a_{vol}	15.008	15.829
a_{surf}	15.628	18.193
a_{sym}	21.435	23.505
a_{C}	0.665	0.72

Table 3.4: The space of calibration parameters used for generating the outputs of the semi-empirical mass formula (1.1).

Prior distributions. First, we consider the independent Gaussian distributions centered at the LS estimates $\hat{\boldsymbol{\theta}}_{L_2}$ (in Table 3.3) with standard deviations $7.5 \times SE(\hat{\boldsymbol{\theta}}_{L_2})$ so that the calibration parameters used for generating the model runs are covered roughly within two standard deviations of the priors. Namely,

$$a_{\text{vol}} \sim \mathcal{N}(15.42, 0.203),$$

$$a_{\text{surf}} \sim \mathcal{N}(16.91, 0.645),$$

$$a_{\text{sym}} \sim \mathcal{N}(22.47, 0.525),$$

$$a_{\text{C}} \sim \mathcal{N}(0.69, 0.015).$$

The prior distributions for hyperparameters of the GPs were selected as Gamma(α, β) with the shape parameter α and scale parameter β, so that they represent a vague knowledge about the scale of these parameters given by the literature on nuclear mass models (Weizsäcker,

1935; Bethe and Bacher, 1936; Myers and Swiatecki, 1966; Fayans, 1998; Kirson, 2008; McDonnell et al., 2015; Kortelainen et al., 2010a, 2012, 2014; Benzaid et al., 2020; Kejzlar et al., 2020). In particular, the error scale σ is in the majority of nuclear applications within units of MeV, therefore we set

$$\sigma \sim \text{Gamma}(2, 1),$$

with the scale of the systematic error being

$$\eta_\delta \sim \text{Gamma}(10, 1),$$

to allow for this quantity to range between the units and tens of MeV. It is also reasonable to assume that the mass of a given nucleus is correlated mostly with its neighbours on the nuclear chart. We express this notion through these reasonably wide prior distributions

$$\begin{aligned} l_Z &\sim \text{Gamma}(10, 1), \\ l_N &\sim \text{Gamma}(10, 1), \\ \nu_Z &\sim \text{Gamma}(10, 1), \\ \nu_N &\sim \text{Gamma}(10, 1), \\ \nu_i &\sim \text{Gamma}(10, 1), \qquad i = 1, 2, 3, 4. \end{aligned}$$

Finally, the majority of the masses in the training dataset of 2000 experimental binding energies fall into the range of $[1000, 2000]$ MeV (1165 of masses precisely). We consider the following prior distribution for the parameter η_f to reflect on the scale of the experimental binding energies:

$$\eta_f \sim \text{Gamma}(110, 10).$$

CHAPTER 4

EMPIRICAL BAYES CALIBRATION OF COMPUTER MODELS WITH CONSISTENT PREDICTIONS

Up to this point, we have seen that the Bayesian framework for computer-model-aided inference, described in detail in Section 1.2 and at the beginning of Chapter 3, provides a statistically principled way to account for various sources of uncertainty and leads to better predictions. It can be especially powerful in scenarios where computer models under consideration are complex and computationally too expensive to be used directly for predictions with quantified uncertainties, because each evaluation of such models often takes several days. Despite these advantages, we have also identified many challenges that make the implementation of the Kennedy and O'Hagan (2001) framework challenging in practice.

Let us recall that under a fully Bayesian treatment, the predictions of new values $\boldsymbol{y}^*$ of a physical ζ using a computer model f_m are specified by the posterior predictive distribution $p(\boldsymbol{y}^*|\boldsymbol{d})$. The dataset $\boldsymbol{d}$ here and for the rest of this chapter consists of n observations y_i from the physical process ζ and s evaluations z_j of the computer model f_m, i.e. $\boldsymbol{d} = (d_1, \ldots, d_{n+s}) := (\boldsymbol{y}, \boldsymbol{z})$, and follows the multivariate normal distribution (1.4). The predictive distribution $p(\boldsymbol{y}^*|\boldsymbol{d})$ is obtained by integrating the conditional density $p(\boldsymbol{y}^*|\boldsymbol{d}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma)$, which is a multivariate normal density given by the statistical model (1.3) and the specification of GPs, against the posterior density $p(\boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma|\boldsymbol{d})$, namely

$$p(\boldsymbol{y}^*|\boldsymbol{d}) = \int_{\boldsymbol{\phi}} p(\boldsymbol{y}^*|\boldsymbol{d}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma) p(\boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma|\boldsymbol{d}) \, \mathrm{d}\boldsymbol{\theta} \, \mathrm{d}\boldsymbol{\gamma} \, \mathrm{d}\sigma. \tag{4.1}$$

An analogical relationship also holds for the predictions of new realizations of the physical process $\boldsymbol{\zeta}^*$. The posterior density $p(\boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma|\boldsymbol{d})$, however, does not have a closed form in general and one typically resorts to MCMC methods for approximation. Additionally, the nature of the likelihood $p(\boldsymbol{d}|\boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma)$ makes the problem hard to scale due to the complex structure of the covariance matrix $K(\boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma)$ (see (1.6)). In Chapter 3, we developed a novel VBI algorithm that provides an efficient and scalable alternative to the traditional

MCMC methods. Nevertheless, the practical implementation of either the MCMC or our VBI approach can be a non-trivial task and requires some practical experience.

As an easy-to-implement alternative that avoids the difficulties described above, we propose an empirical Bayes approach for fast and statistically principled predictions of physical quantities using imperfect computer models which instead of placing a (prior) distribution on $(\boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma)$ estimates these parameters directly form the data. One can therefore utilize the convenience of GPs to obtain closed form, simple, and fast predictions given by the conditional distribution $p(\boldsymbol{y}^*|\boldsymbol{d}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma)$ (or $p(\boldsymbol{\zeta}^*|\boldsymbol{d}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma)$). The proposed approach can be viewed as an approximation of the fully Bayesian treatment that neglects some of the uncertainty associated with the unknown parameters.

Our contributions are the following. First, we present a fast and easy to implement framework for computer-model-enabled predictions and provide two alternative plug-in estimators for all the unknown quantities involved. Second, we offer a new perspective on the Kennedy and O'Hagan (2001) framework and provide its representation as a Bayesian hierarchical model. This alternative representation allows us to discuss the framework in the context of non-parametric regression problems with GP priors and establish our methods' theoretical validity through a posterior consistency result. Lastly, we validate the empirical Bayes approach empirically through a simulation study, and illustrate our methodology on a real data application in nuclear physics.

The rest of this chapter is organized as follows. In Section 4.1, we show the equivalence of the general framework for Bayesian calibration of computer models with a Bayesian hierarchical model. Then, in Section 4.2, we discuss the theoretical properties of our approach and establish its posterior consistency. Section 4.3 defines two plug-in estimators for GP model parameters and a consistent estimator of a noise scale component. Section 4.4 contains a simulation study that empirically validates the methodology in this chapter. A real-data application is also included in Section 4.4.

4.1 Hierarchical model for Bayesian calibration of computer models

Here we show that we can represent the model of Kennedy and O'Hagan (2001) described in Section 1.2, hierarchically, as the following hypotheses about the observations y_i, the computer model evaluations z_j, and a set of prior distributions.

Model for data:

$$y_i = \zeta(\boldsymbol{t}_i) + \sigma\epsilon_i \qquad i = 1, \ldots, n, \tag{4.2}$$

$$z_j = f_m(\widetilde{\boldsymbol{t}}_j, \widetilde{\boldsymbol{\theta}}_j), \qquad j = 1, \ldots, s, \tag{4.3}$$

$$\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma). \tag{4.4}$$

Priors:

$$\delta(\boldsymbol{t}) \sim \mathcal{GP}_\delta(m_\delta(\boldsymbol{t}), k_\delta(\boldsymbol{t}, \boldsymbol{t}')), \qquad \text{given } \boldsymbol{\gamma} \text{ and independent of } \epsilon \text{ and } \sigma,$$

$$f_m(\boldsymbol{t}, \boldsymbol{\theta}) \sim \mathcal{GP}_f(m_f(\boldsymbol{t}, \boldsymbol{\theta}), k_f((\boldsymbol{t}, \boldsymbol{\theta}), (\boldsymbol{t}', \boldsymbol{\theta}'))), \qquad \text{given } \boldsymbol{\gamma} \text{ and independent of } \epsilon_i,\ \sigma, \text{ and } \delta,$$

$$\zeta(\boldsymbol{t})|\boldsymbol{\theta}, \boldsymbol{\gamma} \sim \mathcal{GP}_f + \mathcal{GP}_\delta.$$

Under this model, the conditional likelihoods for y_i and z_j are

$$p(y_i|\zeta(\boldsymbol{t}_i), \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(- \frac{(y_i - \zeta(\boldsymbol{t}_i))^2}{2\sigma^2} \right), \tag{4.5}$$

$$p(z_j|f_m(\widetilde{\boldsymbol{t}}_j, \widetilde{\boldsymbol{\theta}}_j)) = 1_{z_j = f_m(\widetilde{\boldsymbol{t}}_j, \widetilde{\boldsymbol{\theta}}_j)}(z_j), \tag{4.6}$$

where $p(z_j|f_m(\widetilde{\boldsymbol{t}}_j, \widetilde{\boldsymbol{\theta}}_j))$ is a likelihood with the point mass at $z_j = f_m(\widetilde{\boldsymbol{t}}_j, \widetilde{\boldsymbol{\theta}}_j)$. Consequently, the equivalence of the two formulations is given through the equality between the likelihood (1.4) and the integral

$$\begin{aligned}
\int_{\boldsymbol{\zeta}} \int_{\tilde{f}_m} p(\boldsymbol{\zeta}, \tilde{f}_m, \boldsymbol{d}|\boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma)\, \mathrm{d}\tilde{f}_m\, \mathrm{d}\boldsymbol{\zeta} &= \int_{\boldsymbol{\zeta}} \int_{\tilde{f}_m} p(\boldsymbol{d}|\boldsymbol{\zeta}, \tilde{f}_m, \boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma) p(\boldsymbol{\zeta}, \tilde{f}_m|\boldsymbol{\theta}, \boldsymbol{\gamma})\, \mathrm{d}\tilde{f}_m\, \mathrm{d}\boldsymbol{\zeta} \\
&= \int_{\boldsymbol{\zeta}} \int_{\tilde{f}_m} \prod_i^n p(y_i|\zeta_i, \sigma) \prod_j^s p(z_j|\tilde{f}_{m,j}) p(\boldsymbol{\zeta}, \tilde{f}_m|\boldsymbol{\theta}, \boldsymbol{\gamma})\, \mathrm{d}\tilde{f}_m\, \mathrm{d}\boldsymbol{\zeta} \\
&= \int_{\boldsymbol{\zeta}} \prod_i^n p(y_i|\zeta_i, \sigma) p(\boldsymbol{\zeta}, \boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{\gamma})\, \mathrm{d}\boldsymbol{\zeta},
\end{aligned}$$

where $\boldsymbol{\zeta} = (\zeta(\boldsymbol{t}_1), \dots, \zeta(\boldsymbol{t}_n)) = (\zeta_1, \dots, \zeta_n)$ and $\tilde{f}_m = (f_m(\widetilde{\boldsymbol{t}}_1, \widetilde{\boldsymbol{\theta}}_1), \dots, f_m(\widetilde{\boldsymbol{t}}_s, \widetilde{\boldsymbol{\theta}}_s))$. The likelihood $p(\boldsymbol{\zeta}, \boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{\gamma})$ is the multivariate normal distribution with the mean $M(\boldsymbol{\theta}, \boldsymbol{\gamma})$ (see (1.5)) and the covariance

$$K_p(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \begin{pmatrix} K_f(T_y(\boldsymbol{\theta}), T_y(\boldsymbol{\theta})) + K_\delta(T_y, T_y) & K_f(T_y(\boldsymbol{\theta}), T_z(\widetilde{\boldsymbol{\theta}})) \\ K_f(T_z(\widetilde{\boldsymbol{\theta}}), T_y(\boldsymbol{\theta})) & K_f(T_z(\widetilde{\boldsymbol{\theta}}), T_z(\widetilde{\boldsymbol{\theta}})) \end{pmatrix}.$$

Again, $K_f(T_y(\boldsymbol{\theta}), T_y(\boldsymbol{\theta}))$ is the matrix with (i, j) element $k_f((\boldsymbol{t}_i, \boldsymbol{\theta}), (\boldsymbol{t}_j, \boldsymbol{\theta}))$, $K_\delta(T_y, T_y)$ is the matrix with (i, j) element $k_\delta(\boldsymbol{t}_i, \boldsymbol{t}_j)$, and $K_f(T_z(\widetilde{\boldsymbol{\theta}}), T_z(\widetilde{\boldsymbol{\theta}}))$ is the matrix with (i, j) element $k_f((\widetilde{\boldsymbol{t}}_i, \widetilde{\boldsymbol{\theta}}_i), (\widetilde{\boldsymbol{t}}_j, \widetilde{\boldsymbol{\theta}}_j))$. $K_f(T_y(\boldsymbol{\theta}), T_z(\widetilde{\boldsymbol{\theta}}))$ is defined analogically with the kernel k_f.

We leave the details of the integral computation for Section 4.5.1. This representations of the model is crucial for the theoretical results obtained in the subsequent section. It reframes the Bayesian model as a version of a non-parametric regression problem with GP prior for $\zeta(\boldsymbol{t})$ and an additive noise. Additionally, we can gain a further insight into the role of the set of model runs $\boldsymbol{z}$. Let us consider a function space $\mathcal{F}$ and a subset $\widetilde{\mathcal{F}} \subset \mathcal{F}$, then

$$p(\zeta \in \widetilde{\mathcal{F}}|\boldsymbol{d}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma) \propto \int_{\widetilde{\mathcal{F}}} \prod_i^n p(y_i|\zeta_i, \sigma) p(\boldsymbol{\zeta}|\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\gamma}) \, \mathrm{d}\boldsymbol{\zeta}. \tag{4.7}$$

One can therefore interpret the model runs $\boldsymbol{z}$ as an additional information provided by the computer model f_m that enhances the GP prior $p(\boldsymbol{\zeta}|\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\gamma})$ for the physical process ζ, having the mean function

$$m_\zeta(\boldsymbol{t}) = m_f(\boldsymbol{t}, \theta) + m_\delta(\boldsymbol{t}) + \sum_{i,j=1}^m \kappa_{j,i} \Big[k_f((\boldsymbol{t}, \boldsymbol{\theta}), (\widetilde{\boldsymbol{t}}_j, \widetilde{\boldsymbol{\theta}}_j)) \Big] \Big[z_i - m_f(\widetilde{\boldsymbol{t}}_i, \widetilde{\boldsymbol{\theta}}_i) \Big], \tag{4.8}$$

and the covariance function

$$\begin{aligned} k_\zeta(\boldsymbol{t}, \boldsymbol{t}') = {} & k_f((\boldsymbol{t}, \boldsymbol{\theta}), (\boldsymbol{t}', \boldsymbol{\theta})) + k_\delta(\boldsymbol{t}, \boldsymbol{t}') \\ & - \sum_{i,j=1}^m \kappa_{j,i} \Big[k_f((\boldsymbol{t}, \boldsymbol{\theta}), (\widetilde{\boldsymbol{t}}_j, \widetilde{\boldsymbol{\theta}}_j)) \Big] \Big[k_f((\widetilde{\boldsymbol{t}}_i, \widetilde{\boldsymbol{\theta}}_i), (\boldsymbol{t}', \boldsymbol{\theta})) \Big], \end{aligned} \tag{4.9}$$

where $\kappa_{j,i}$ is the (j, i) element of the matrix $K_f(T_z(\widetilde{\boldsymbol{\theta}}), T_z(\widetilde{\boldsymbol{\theta}}))^{-1}$.

4.2 Posterior consistency, a theoretical validation

The revealing consequence of the previous section is that the Kennedy and O'Hagan (2001) framework is equivalent to the non-parametric regression model of an unknown function $\zeta(\boldsymbol{t})$ with the prior distribution $p(\boldsymbol{\zeta}|\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\gamma})$. This is not only a new perspective on the popular framework, but also happens to be the key step that allows us to validate our empirical Bayes approach theoretically and establish the posterior consistency of the physical process when the prior $p(\boldsymbol{\zeta}|\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\gamma})$ satisfies certain properties.

In the reminder of this section, we assume that the true physical process ζ_0 is a continuously differentiable function on the compact and convex set $\boldsymbol{\Omega} \subset \mathbb{R}^p$. Without loss of generality, we take $\boldsymbol{\Omega} = [0,1]^p$. Additionally, we shall assume the hyperparameters $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ take values in a set Υ. For any $\nu > 0$, we aim to establish, under suitable conditions, the following:

$$\sup_{(\boldsymbol{\theta},\boldsymbol{\gamma})\in\Upsilon} p(\zeta \in W_{\nu,n}^{C} | y_1, \ldots, y_n, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \hat{\sigma}_n) \xrightarrow[\text{n}]{} 0 \quad \text{a.s. } P_0, \tag{4.10}$$

where P_0 denotes the joint conditional distribution of $\{y_i\}_{i=1}^{\infty}$ given true ζ_0 and σ_0, $\hat{\sigma}_n$ is a strongly consistent estimator of σ_0, and

$$W_{\nu,n} = \left\{ \zeta : \int |\zeta(\boldsymbol{t}) - \zeta_0(\boldsymbol{t})| \, \mathrm{d}Q_n(\boldsymbol{t}) \leq \nu \right\}, \tag{4.11}$$

with Q_n being the empirical measure on the design points given as $Q_n(\boldsymbol{t}) = n^{-1} \sum_{i=1}^{n} \mathbb{1}_{\boldsymbol{t}_i}(\boldsymbol{t})$.

In Theorem 2, we first present a general result on the consistency of non-parametric regression problems and subsequently discuss the theorem's conditions in the context of the model described in Section 4.1. This is based on the work of Choi and Schervish (2007a) and Choi (2007), where the authors assume σ is included in $W_{\nu,n}$, and the posterior consistency is derived jointly for ζ and σ. On the other hand, the consistency of ζ conditioned on $\hat{\sigma}_n$ requires a non-trivial modification of their original results. The proof of Theorem 2 is given in Section 4.5.2.

Theorem 2. *Let $\{y_i\}_{i=1}^{\infty}$ be independently and normally distributed with the mean $\zeta(\boldsymbol{t}_i)$ and the standard deviation σ with respect to a common σ-finite measure, where ζ belongs to a*

space of continuously differentiable functions on $[0,1]^p$ *denoted as* $\mathcal{F}$*, and* $\sigma > 0$*. Let* $\zeta_0 \in \mathcal{F}$ *and let* P_0 *denotes the joint conditional distribution of* $\{y_i\}_{i=1}^{\infty}$ *given true* ζ_0 *and* σ_0*. Let* $\{U_n\}_{n=1}^{\infty}$ *be a sequence of subsets of* $\mathcal{F}$*. Let* ζ *have a prior* $\Pi(\cdot|\boldsymbol{\theta},\boldsymbol{\gamma})$ *where* $(\boldsymbol{\theta},\boldsymbol{\gamma})$ *take values in a set* Υ*. For any* $0 < \epsilon < 1$ *and* $\zeta_0(\boldsymbol{t}_i) = \zeta_{0,i}$ *define:*

$$\Lambda_i(\zeta_0,\zeta) = \log \frac{p(y_i|\zeta_{0,i},\sigma_0)}{p(y_i|\zeta_i,\sigma_0(1-\epsilon))},$$
$$K_i(\zeta_0,\zeta) = \mathbb{E}_{\zeta_0,\sigma_0}(\Lambda_i(\zeta_0,\zeta)),$$
$$V_i(\zeta_0,\zeta) = \mathbb{V}ar_{\zeta_0,\sigma_0}(\Lambda_i(\zeta_0,\zeta)).$$

If the following assumptions are satisfied:

(A1) Suppose there exists a set B *with* $\Pi(B|\boldsymbol{\theta},\boldsymbol{\gamma}) > 0$ *and for any* $\Delta > 0$ *a constant* $0 < \tilde{\epsilon}_1 < 1$*, so that for any* $\epsilon < \tilde{\epsilon}_1$*:*

(i) $\sum_{i=1}^{\infty} \frac{V_i(\zeta_0,\zeta)}{i^2} < \infty$, $\forall \zeta \in B$,

(ii) $\Pi(B \cap \{\zeta : K_i(\zeta_0,\zeta) < \Delta \text{ for all } i\}|\boldsymbol{\theta},\boldsymbol{\gamma}) > 0$.

(A2) Suppose there exist tests $\{\Phi_n\}_{n=1}^{\infty}$*, sets* $\{\mathcal{F}_n\}_{n=1}^{\infty}$*, and constants* $C_2, C_1, c_1 > 0$ *and* $0 < \tilde{\epsilon}_2 < 1$ *so that:*

(i) $\sum_{n=1}^{\infty} \mathbb{E}_{\zeta_0,\sigma_0}\Phi_n < \infty$

(ii) $\sup_{(\boldsymbol{\theta},\boldsymbol{\gamma})\in\Upsilon} \Pi(\mathcal{F}_n^C|\boldsymbol{\theta},\boldsymbol{\gamma}) < C_1 e^{-c_1 n}$

(iii) There exists a constant $c_\epsilon > 0$ such that for any $0 < \epsilon < \tilde{\epsilon}_2$ the inequality $c_\epsilon + \log(1-\epsilon) - \log(1+\epsilon) > 0$ holds and

$$\sup_{\zeta \in U_n^C \cap \mathcal{F}_n} \mathbb{E}_{\zeta,\sigma_0(1+\epsilon)}(1-\Phi_n) \leq C_2 e^{-c_\epsilon n}.$$

(A3) $\hat{\sigma}_n$ *is strongly consistent, i.e* $\hat{\sigma}_n \xrightarrow[n]{} \sigma_0$ *a.s.* P_0*.*

Then

$$\sup_{(\boldsymbol{\theta},\boldsymbol{\gamma})\in\Upsilon} p(\zeta \in U_n^C|y_1,\ldots,y_n,\boldsymbol{\theta},\boldsymbol{\gamma},\hat{\sigma}_n) \xrightarrow[\text{n}]{} 0 \quad \text{a.s. } P_0.$$

For the purpose of generality of Theorem 2, we do not explicitly condition on the set of model runs $\boldsymbol{z}$. It is clear from our previous discussions (see (4.7) in particular) that the model runs play the role of fixed constants in the prior distribution over ζ. The dependence on $\boldsymbol{z}$ in (4.10) arises by setting $\Pi(\boldsymbol{\zeta}|\boldsymbol{\theta}, \boldsymbol{\gamma}) := p(\boldsymbol{\zeta}|\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\gamma})$.

We now consider the conditions of Theorem 2 in the context of the model in Section 4.1. These conditions fall into two general categories; one group of conditions is related to the existence of the test functions Φ_n, and the second group revolves around the conditions for the prior distributions.

Our approach to establish the existence of test functions $\{\Phi_n\}_{n=1}^{\infty}$ that satisfy the conditions (i) and (iii) in Theorem 2 is similar to that of Theorem 2 in Choi and Schervish (2007a). We consider a sieve $\mathcal{F}_n$ which grows to the space of continuously differentiable functions on $[0,1]^p$. Namely, let

$$\mathcal{F}_n = \left\{ \zeta : \| \zeta \|_\infty < M_n, \;\; \| \frac{\partial}{\partial t_i} \zeta \|_\infty < M_n, \;\; i = 1, \cdots, p \right\} \tag{4.12}$$

where $M_n = \mathcal{O}(n^\alpha)$ for some $\alpha \in (\frac{1}{2}, 1)$. Also, $\| \cdot \|_\infty$ denotes the supremum norm. Each test is defined as a combination of tests over finitely many elements in the covering of $\mathcal{F}_n$. The existence of tests in the specific case of $W_{n,\nu}$ is given in Theorem 3 with its prove provided in Section 4.5.2.

Theorem 3. *Let $\mathcal{F}_n$ be the sieves defined in* (4.12)*. For any $\nu > 0$ there exist tests $\{\Phi_n\}_{n=1}^{\infty}$ and constants C and $0 < \tilde{\epsilon} < 1$ so that:*

(i) $\sum_{n=1}^{\infty} \mathbb{E}_{\zeta_0, \sigma_0} \Phi_n < \infty$

(ii) There exists a constant $c_\epsilon > 0$ such that for any $0 < \epsilon < \tilde{\epsilon}$ the inequality $c_\epsilon + \log(1 - \epsilon) - \log(1 + \epsilon) > 0$ holds and

$$\sup_{\zeta \in W_{n,\nu}^C \cap \mathcal{F}_n} \mathbb{E}_{\zeta, \sigma_0(1+\epsilon)} (1 - \Phi_n) \leq C e^{-c_\epsilon n}.$$

To verify conditions (A1) of Theorem 2, it is sufficient to show that the GP prior for ζ assigns positive probability to the following set for any $\delta > 0$:

$$B_\delta = \{\zeta : \| \zeta - \zeta_0 \|_\infty < \delta\}. \tag{4.13}$$

For any $0 < \epsilon < 1$, a short calculation leads to

$$\begin{aligned} K_i(\zeta_0, \zeta) &= \log(1-\epsilon) - \frac{1}{2}\left(1 - \frac{1}{(1-\epsilon)^2}\right) + \frac{[\zeta_0(\boldsymbol{t}_i) - \zeta(\boldsymbol{t})]^2}{2\sigma_0^2(1-\epsilon)^2} \\ &\leq \log(1-\epsilon) - \frac{1}{2}\left(1 - \frac{1}{(1-\epsilon)^2}\right) + \frac{\| \zeta_0(\boldsymbol{t}_i) - \zeta(\boldsymbol{t}) \|_\infty^2}{2\sigma_0^2(1-\epsilon)^2}. \end{aligned}$$

Let $a(\epsilon) = \log(1-\epsilon) - 1/2 + 1/[2(1-\epsilon)^2]$, it is easy to see that $a(\epsilon)$ is positive and continuous at $\epsilon = 0$. Therefore, for every $\Delta > 0$, there exist $\delta > 0$ and $0 < \tilde{\epsilon} < 1$ so that $K_i(\zeta_0, \zeta) < \Delta$ for all i and any $\epsilon < \tilde{\epsilon}$.

Additionally, for any $\epsilon < \tilde{\epsilon}$ and any $\delta > 0$

$$\begin{aligned} V_i(\zeta_0, \zeta) &= \frac{1}{2}\left[\frac{1}{(1-\epsilon)^2} - 1\right]^2 + \left[\frac{[\zeta_0(\boldsymbol{t}_i) - \zeta(\boldsymbol{t})]}{(1-\epsilon)^2}\right]^2 \\ &< \infty \quad \text{uniformly in } i, \end{aligned}$$

and as a result, for all $\zeta \in B_\delta$, $\sum_{i=1}^{\infty} \frac{V_i(\zeta_0,\zeta)}{i^2} < \infty$. The prior condition (ii) of (A2) for the sieve $\mathcal{F}_n$ (4.12) is addressed in Lemma 4 (for proof see Section 4.5.2).

Lemma 4. *Let the mean function $m_\zeta(\cdot)$ of the GP prior for ζ defined on $[0,1]^p$ be continuously differentiable, and the covariance function $k_\zeta(\cdot,\cdot)$ has mixed partial derivatives up to order 4 that are continuous. Define,*

$$\begin{aligned} \rho_0^2(\boldsymbol{\theta}, \boldsymbol{\gamma}) &= \sup_{\boldsymbol{t} \in [0,1]^p} \mathbb{V}ar\left(\zeta(\boldsymbol{t}) | \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\gamma}\right), \\ \rho_i^2(\boldsymbol{\theta}, \boldsymbol{\gamma}) &= \sup_{\boldsymbol{t} \in [0,1]^p} \mathbb{V}ar\left(\frac{\partial}{\partial t_i}\zeta(\boldsymbol{t}) \middle| \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\gamma}\right), \qquad i = 1, \ldots, p. \end{aligned}$$

Suppose that Υ is a compact set, and ρ_i^2 are continuous functions of $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ for all $(\boldsymbol{\theta}, \boldsymbol{\gamma}) \in \Upsilon$, $i = 0, \ldots, p$. Then there exist constants $C, c > 0$ so that

$$\sup_{(\boldsymbol{\theta}, \boldsymbol{\gamma}) \in \Upsilon} p(\mathcal{F}_n^C | \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\gamma}) < Ce^{-cn},$$

where $\mathcal{F}_n$ are the sieves defined in (4.12).

Below we present the almost sure consistency result 4.10 as a corollary of Theorem 2, Theorem 3, and Lemma 4.

Corollary 1. *Let P_0 denotes the joint conditional distribution of $\{y_i\}_{i=1}^{\infty}$ given true ζ_0 and σ_0. Let $m_\zeta(\cdot)$ and $k_\zeta(\cdot,\cdot)$ be the mean and covariance functions of the GP prior for ζ satisfying the conditions of Lemma 4. Assume Υ is a compact set, and for any $\delta > 0$, $p(B_\delta|\boldsymbol{z},\theta,\boldsymbol{\gamma}) > 0$. If $\hat{\sigma}_n$ is a strongly consistent estimator of σ_0, then for any $\nu > 0$*

$$\sup_{(\boldsymbol{\theta},\boldsymbol{\gamma})\in\Upsilon} p(\zeta \in W^C_{\nu,n}|y_1,\ldots,y_n,\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\gamma},\hat{\sigma}_n) \xrightarrow[n]{} 0 \quad a.s.\ P_0. \tag{4.14}$$

Prior conditions. The prior positivity condition requiring $p(B_\delta|\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\gamma}) > 0$ for any δ was extensively studied by Ghosal and Roy (2006) and Tokdar and Ghosh (2007). Theorem 4 of Ghosal and Roy (2006) implies that this condition is satisfied for a GP with continuous sample paths and continuous mean and covariance functions, as long as ζ_0 and the m_ζ belong to the reproducing kernel Hilbert space (RKHS) of k_ζ. The continuity of GP's sample paths is given by the application of Theorem 5 in Ghosal and Roy (2006) which requires the same continuity conditions as Lemma 4 in this section (excluding those on ρ_i^2). It should be clear from (4.8) and (4.9) that m_ζ is continuously differentiable on $[0,1]^p$, and k_ζ has continuous mixed partial derivatives up to 4^{th} order on $[0,1]^p$, as long as the same holds about m_f and m_δ and respectively k_f and k_δ. Tokdar and Ghosh (2007) show that the RKHS of k_ζ spans the space of continuously differentiable functions on $[0,1]^p$, if k_ζ is a product of p isotropic and integrable univariate covariance functions with continuous mixed partial derivatives up to order 4. For example, the squared exponential covariance function satisfies these requirements including the continuity of ρ_i^2 for $i = 0,\ldots,p$.

This, of course, does not directly imply that such choices for m_f and m_δ, and k_f and k_δ respectively, result in the conditional mean m_ζ and covariance k_ζ functions satisfying these sufficient conditions. For larger applicability of our results, we note that further investigation of specific choices for mean and covariance functions that satisfy the desired conditions is needed. We intend to address this in our future work. Nevertheless, the simulation study

conducted in Section 4.4.1 strongly suggests that choosing the squared exponential kernel leads to consistent predictions.

4.3 Parameter estimation and prediction

Thus far, we established that the empirical Bayesian framework provides a principled approach for inference and enjoys good theoretical properties, all this assuming a (strongly) consistent estimator of σ_0, smoothness of the prior mean and covariance function, and the GP hyperparameters $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ taking values in some compact set.

In this section, we first propose a strongly consistent estimator of the true noise scale σ_0 and two different plug-in estimators of $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ as minimizers of two alternative loss functions. In particular, we consider negative data log-likelihood and negative predictive log-likelihood combined with K-fold cross-validation. Second, we provide the complete empirical Bayes algorithm for simple and fast predictions of physical quantities using (imperfect) computer models.

Let us consider n observations y_i from the physical process under the model (4.2), we propose the following estimator of the noise variance σ_0^2:

$$\hat{\sigma}_n^2 = \frac{\sum_{i=1}^{n-1}(y_{i+1} - y_i)^2}{2(n-1)} \tag{4.15}$$

Theorem 4. *Suppose $\zeta_0(\boldsymbol{t})$ represents the true physical process and σ_0^2 be the true value of the experimental error variance, where $\boldsymbol{t} \in \boldsymbol{\Omega}$ is a compact and convex subset of $\mathbb{R}^p$, and ζ_0 is continuously differentiable on $\boldsymbol{\Omega}$. Let P_0 denotes the joint conditional distribution of $\{y_i\}_{i=1}^{\infty}$ given true ζ_0 and σ_0^2. Also assume the following holds about the design points $\boldsymbol{t}_i$:*

$$\sup_{i \in \{1,\ldots,n\}, j \in \{1,\ldots,p\}} |t_{i+1,j} - t_{i,j}| \xrightarrow[n]{} 0, \tag{AD}$$

then

$$\hat{\sigma}_n^2 \xrightarrow[n]{} \sigma_0^2 \quad a.s.\ P_0. \tag{4.16}$$

The proof of Theorem 4 is given in Section 4.5.2. The continuous mapping theorem directly implies the following.

Corollary 2. *Under the assumptions of Theorem 4,*

$$\hat{\sigma}_n = \sqrt{\hat{\sigma}_n^2} \xrightarrow[\text{n}]{} \sigma_0 \quad \text{a.s. } P_0. \tag{4.17}$$

Remark 1. *The assumption* (AD) *is satisfied by a design that contains at least one point in each hypercube H in $\boldsymbol{\Omega}$ with its Lebesgue measure $\lambda(H) \geq \frac{1}{Kn}$, for some constant $0 < K \leq 1$. This is, for example, the case of equally spaced design.*

4.3.1 Estimation of hyperparameters

4.3.1.1 Marginal data likelihood

We first consider estimates of $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ as minimizers of a loss function that is reminiscent of the standard maximum likelihood approach, namely

$$L_{MLE}(\boldsymbol{\theta}, \boldsymbol{\gamma}) = -\log p(\boldsymbol{d}|\boldsymbol{\theta}, \boldsymbol{\gamma}, \hat{\sigma}_n), \tag{4.18}$$

with the negative log-likelihood being

$$\begin{aligned} -\log p(\boldsymbol{d}|\boldsymbol{\theta}, \boldsymbol{\gamma}, \hat{\sigma}_n) = \frac{1}{2}(\boldsymbol{d} - M(\boldsymbol{\theta}, \boldsymbol{\gamma}))^T K(\boldsymbol{\theta}, \boldsymbol{\gamma}, \hat{\sigma}_n)(\boldsymbol{d} - M(\boldsymbol{\theta}, \boldsymbol{\gamma})) \\ + \frac{1}{2}\log|K(\boldsymbol{\theta}, \boldsymbol{\gamma}, \hat{\sigma}_n)| + \frac{n+s}{2}\log 2\pi. \end{aligned}$$

We can readily interpret the minimizer of L_{MLE} as a trade-off between the data-fit $\frac{1}{2}(\boldsymbol{d} - M(\boldsymbol{\theta}, \boldsymbol{\gamma}))^T K(\boldsymbol{\theta}, \boldsymbol{\gamma}, \hat{\sigma}_n)(\boldsymbol{d} - M(\boldsymbol{\theta}, \boldsymbol{\gamma}))$ and the model complexity penalty $\frac{1}{2}\log|K(\boldsymbol{\theta}, \boldsymbol{\gamma}, \hat{\sigma}_n)|$ that depends only on the model parameters and the variable inputs.

4.3.1.2 Predictive likelihood with K-fold cross-validation

Another viable approach to estimating the parameters $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ is to base these on a model's predictive performance on unseen data. Cross-validation is a popular and robust approach to estimate this predictive performance that has been utilized across many statistical applications. See Sundararajan and Keerthi (2001); Rasmussen and Williams (2006); Martino et al.

(2017) for applications with GPs. Here, we consider a K-fold cross-validation where the basic idea is to randomly partition the training detest into K subsets of equal size. We then select $K-1$ subsets for training and the hold-out data as a proxy for estimating the predictive performance. This is then repeated until we exhaust all the K subsets for the purpose of validation with typical choices for K being 3, 5, 10, or n (leave-one-out cross-validation).

Formally, let $\boldsymbol{y}_i$ represent the i^{th} subset of the observations $\boldsymbol{y}$ and $\boldsymbol{y}_{-i} = \boldsymbol{y} \setminus \boldsymbol{y}_i$. The negative predictive log-likelihood under the K-fold cross-validation is

$$L_{CV(K)}(\boldsymbol{\theta}, \boldsymbol{\gamma}) = -\sum_{i}^{K} \log p(\boldsymbol{y}_i | \boldsymbol{y}_{-i}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \hat{\sigma}_n), \tag{4.19}$$

The cross-validation should be more robust against model miss-specification and overfitting (Wahba, 1990).

4.3.2 Algorithm for predictions

One of the main benefits of the empirical Bayes approach is that once we estimate the unknown parameters $(\boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma)$, we can obtain a closed form predictive distribution given these estimates. Formally, let us consider a set of new inputs $(\boldsymbol{t}_1^*, \ldots, \boldsymbol{t}_J^*)$ at which we want to obtain the predictions according to the model (1.3). As discussed in Section 3.4, the joint normality between $\boldsymbol{d}$ and $\boldsymbol{y}^*$ implies that the conditional distribution $p(\boldsymbol{y}^* | \boldsymbol{d}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma)$ is a multivariate normal distribution with the mean vector

$$M_{y^*}(\boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma) = M_f(T_y^*(\boldsymbol{\theta})) + M_\delta(T_y^*) + C_* K(\boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma)^{-1}(\boldsymbol{d} - M(\boldsymbol{\theta}, \boldsymbol{\gamma})), \tag{4.20}$$

and the covariance matrix

$$K_{y^*}(\boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma) = K_f(T_y^*(\boldsymbol{\theta}), T_y(\boldsymbol{\theta})) + K_\delta(T_y^*, T_y) + \sigma^2 I_m - C_* K(\boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma)^{-1} C_*^T, \tag{4.21}$$

where

$$C_* = \left(K_f(T_y^*(\boldsymbol{\theta}), T_y(\boldsymbol{\theta})) + K_\delta(T_y^*, T_y) \quad K_f(T_y^*(\boldsymbol{\theta}), T_z(\widetilde{\boldsymbol{\theta}})) \right). \tag{4.22}$$

Similarly to the conditional covariance matrices discussed previously, $K_f(T_y^*(\boldsymbol{\theta}), T_y(\boldsymbol{\theta}))$ is the matrix with (i,j) element $k_f((\boldsymbol{t}_i^*, \boldsymbol{\theta}), (\boldsymbol{t}_j, \boldsymbol{\theta}))$ and $K_\delta(T_y^*, T_y)$ is the matrix with (i,j)

element $k_\delta(\boldsymbol{t}_i^*, \boldsymbol{t}_j)$. The matrix $K_f(T_y^*(\boldsymbol{\theta}), T_z(\widetilde{\boldsymbol{\theta}}))$ is defined accordingly with the kernel k_f. Analogical relationship holds for the conditional distribution of the new realizations from the physical process $p(\boldsymbol{\zeta}^*|\boldsymbol{d}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma)$, where the mean vector is identical with (4.20) and the covariance matrix is

$$K_{\zeta^*}(\boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma) = K_f(T_y^*(\boldsymbol{\theta}), T_y(\boldsymbol{\theta})) + K_\delta(T_y^*, T_y) - C_* K(\boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma)^{-1} C_*^T, \tag{4.23}$$

Algorithm 4.1 summarizes the procedure for predictions of physical quantities using imperfect and computationally expensive computer models.

Algorithm 4.1: Empirical Bayes algorithm for predictions of physical quantities.

Input: Data $\boldsymbol{d}$, mean and covariance functions for GPs, and new inputs $(\boldsymbol{t}_1^*, \ldots, \boldsymbol{t}_J^*)$

1 Use the experimental observations $y_1, \ldots, y_n$ to compute $\hat{\sigma}_n = \sqrt{\hat{\sigma}_n^2}$

2 Minimize either $L_{MLE}(\boldsymbol{\theta}, \boldsymbol{\gamma})$ or $L_{CV(K)}(\boldsymbol{\theta}, \boldsymbol{\gamma})$ to obtain the estimates $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}})$

3 Compute $M_{y^*}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}, \hat{\sigma}_n)$ and $K_{y^*}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}, \hat{\sigma}_n)$ or $M_{\zeta^*}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}, \hat{\sigma}_n)$ and $K_{\zeta^*}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}, \hat{\sigma}_n)$ respectively to get the posterior predictive distribution

4.4 Applications

The main objective of this section is to empirically establish the efficiency of the empirical Bayes method in Algorithm 4.1 and support the consistency result presented in section 4.2. All this while sacrificing minimally in terms of the fidelity of UQ as compared to the fully Bayesian treatment. To this extent, we consider a simulation study where we compare our method (under both L_{MLE} and $L_{CV(K)}$) to a fully Bayesian treatment with posterior samples obtained using the standard MH algorithm. Finally, we revisit the LDM and illustrate our methodology in a real data scenario.

4.4.1 Transverse harmonic wave

Let us consider a simple computer model representing a periodic wave disturbance that moves through a medium and causes displacement of individual atoms or molecules in the medium. This is called a transverse harmonic wave, where the displacement $f_m((t, x), \boldsymbol{\theta})$ of

a particle at location x over time t is given by

$$f_m((t,x),\boldsymbol{\theta}) = \theta_1 \sin\big(kx - \theta_2 t + \psi\big), \tag{4.24}$$

where θ_1 represents the amplitude of the wave, and θ_2 is the frequency of the wave. The model also depends on the wave number k, which is reciprocal to the wave length, and the phase constant ψ. For the purpose of this example, we shall consider these to be known values with $k = 5$ and $\psi = 1$, and define the model inputs (t, x) over the space $[0, 1]^2$ (we assume that the length and time units are all equal to one). The true physical process is modeled according to

$$\zeta_0(t,x) = f_m((t,x),\boldsymbol{\theta}) + \delta(t,x) = \theta_1 \sin\big(5x - \theta_2 t + 1\big) + \beta, \tag{4.25}$$

where $\beta = 1$ is a constant systematic error of the model, and $\boldsymbol{\theta} = (\theta_1, \theta_2)$ are arbitrarily set to be (1.2, 1.8). We generate the experimental observation according to the model (1.3) with the true value of the observation error scale $\sigma_0 = 0.2$, where the model inputs (t, x) are chosen using the Latin hypercube design over the full space $[0, 1]^2$. The space filling properties of the design guarantee decreasing bias of the estimator $\hat{\sigma}_n$ with an increasing sample size. Additionally, we assume that the computer model for the periodic wave disturbance is computationally expensive and generate the set of model runs $\boldsymbol{z}$ using again the Latin hypercube design, now over $[0, 1]^2 \times [0, 2]^2$. We define the GP priors for f_m and δ to have zero means and the covariance functions

$$k_f(\{t,x,\boldsymbol{\theta}\},\{t',x',\boldsymbol{\theta}'\}) = \eta_f \cdot \exp(-\frac{||t-t'||^2}{2\ell_t^2} - \frac{||x-x'||^2}{2\ell_x^2} - \frac{||\theta_1-\theta_1'||^2}{2\ell_{\theta_1}^2} - \frac{||\theta_2-\theta_2'||^2}{2\ell_{\theta_2}^2})$$

$$k_\delta(\{t,x\},\{t',x'\}) = \eta_\delta \cdot \exp(-\frac{||t-t'||^2}{2\nu_t^2} - \frac{||x-x'||^2}{2\nu_x^2}).$$

The hyperparameters in this scenario are therefore $\boldsymbol{\gamma} = (\eta_f, \ell_t, \ell_x, \ell_{\theta_1}, \ell_{\theta_2}, \eta_\delta, \nu_t, \nu_x)$. For the case of the fully Bayesian treatment, we choose inverse gamma priors with mean 1/2 and variance 1/4 for $(\sigma, \eta_f, \eta_\delta)$, gamma priors with mean 1/3 and variance 1/9 for the length scales, and independent Gaussian distributions with mean 0 and variance 4 for the

calibration parameters (θ_1, θ_2). These are non-informative priors given the spans of both the input space $[0,1]^2$ and the parameter space $[0,2]^2$. Table 4.1 shows the RMSEs of predictions of new realizations from the true physical process (4.25) evaluated on a testing dataset of 225 realizations over a uniform grid on $[0,1]^2$. The predictions are taken to be the posterior predictive means under each method. We consider the estimates of hyperparameters using the L_{MLE} loss and the 10-fold cross-validation predictive loss function. The noise scale parameter was estimated using the consistent estimator $\hat{\sigma}_n$ defined in Section 4.3.

	RMSE values on the testing dataset		
	L_{MLE}	$L_{CV(10)}$	Metropolis-Hastings
$n = 125$ $s = 125$	**0.048**	0.071	0.049
$n = 250$ $s = 250$	**0.019**	0.030	0.037
$n = 500$ $s = 500$	**0.010**	0.019	0.021

Table 4.1: The RMSE comparison of the empirical Bayes approach and the fully Bayesian treatment. The GP hyperparameters were estimated using Algorithm 4.1.

The proposed empirical Bayes approach closely matches the fully Bayesian treatment. In fact, the RMSE under the L_{MLE} loss is consistently the lowest and monotonously decreases with the increasing size of the dataset. This is a desirable outcome since the empirical Bayes fit can be readily obtained in several minutes using standard numerical solvers while sampling from posterior distributions can take hours. It took approximately 2 hours to obtain 10^4 samples in the scenario with the largest sample size on a standard PC with 4 cores.

Parameter	$n = 125$, $s = 125$			$n = 250$, $s = 250$			$n = 500$, $s = 500$		
	L_{MLE}	$L_{CV(10)}$	MH	L_{MLE}	$L_{CV(10)}$	MH	L_{MLE}	$L_{CV(10)}$	MH
θ_1	1.197	1.217	1.251	1.160	1.251	1.166	1.207	1.206	1.208
θ_2	1.781	1.787	1.771	1.805	1.799	1.765	1.792	1.818	1.765
σ	0.328		0.208	0.259		0.206	0.228		0.198

Table 4.2: The estimates of calibration parameters and the noise scale under each method.

For completeness, we also show the estimates of calibration parameters and the noise scale under each method in Table 4.2. Posterior means were taken as the estimates of the fully Bayesian solution. We can see again a close match between the approximate empirical Bayes method and the MH algorithm. The only notable difference is in terms of the noise scale estimate $\hat{\sigma}_n$. This is expected since the estimate is asymptotically unbiased.

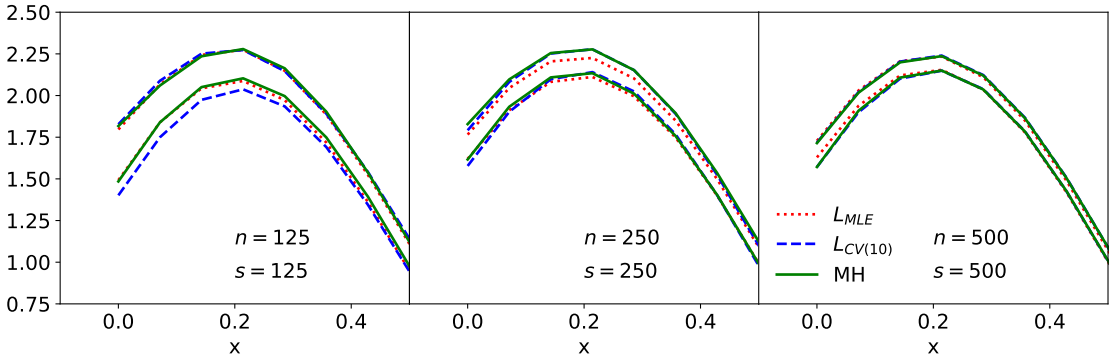

Figure 4.1: Detail of 95% credible bands plotted at $t = 0.21$.

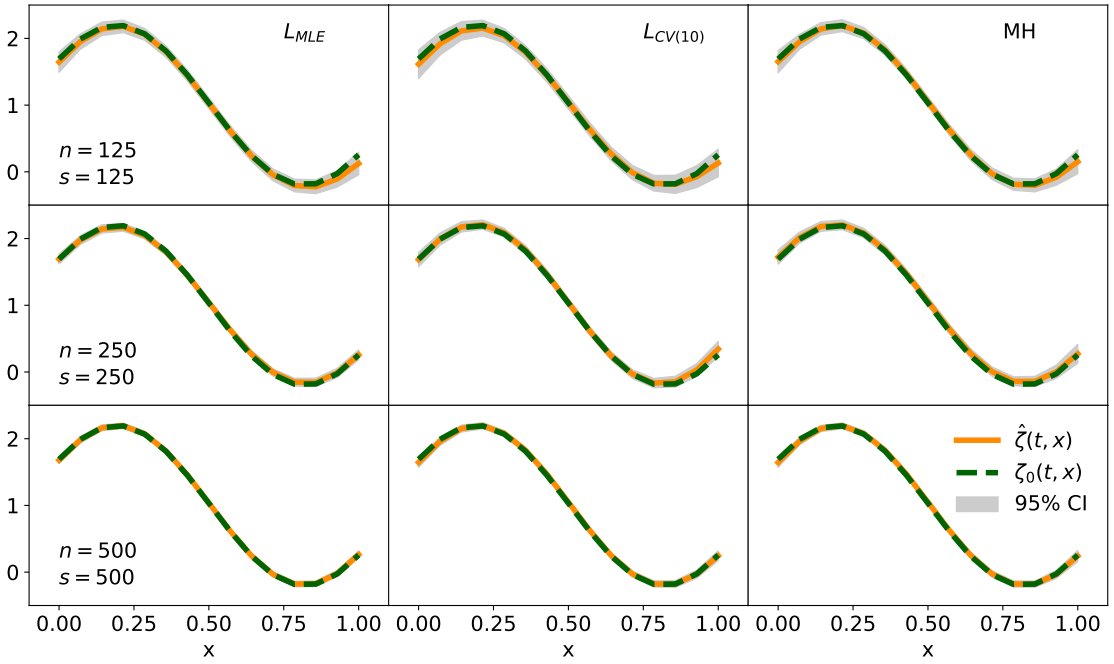

Figure 4.2: Comparison of the convergence to the true physical process. The curves with 95% credible intervals are plotted at $t = 0.21$.

Figure 4.1 and Figure 4.2 show the loss in terms of UQ is negligible for all practical pur-

poses. We can see that the empirical Bayes approach slightly overestimates the uncertainty for smaller sample size, but this quickly diminishes as the sample size increases. This is likely the consequence of the inflation of the noise scale given by the bias of $\hat{\sigma}_n$ which diminishes with the increasing sample size as expected. See Section 4.5.3 for additional figures of the empirical Bayes fit at the time locations $t = 0$, $t = 0.43$, $t = 0.71$, and $t = 1$.

4.4.2 The Liquid Drop Model revisited

To illustrate our empirical Bayes framework for computer-enabled predictions on a real data example, we yet again consider the 4-parameter LDM of nuclear binding energies (see Section 1.1 for details).

We now present an analysis of 595 experimental binding energies of even-even nuclei from the AME2003 dataset (Audi et al., 2003) (publicly available at `http://amdc.impcas.ac.cn/web/masstab.html`) randomly divided into a training set of 450 nuclei and a testing set of the remaining 145 nuclei, see Figure 4.3. We consider the statistical model (1.3) and model

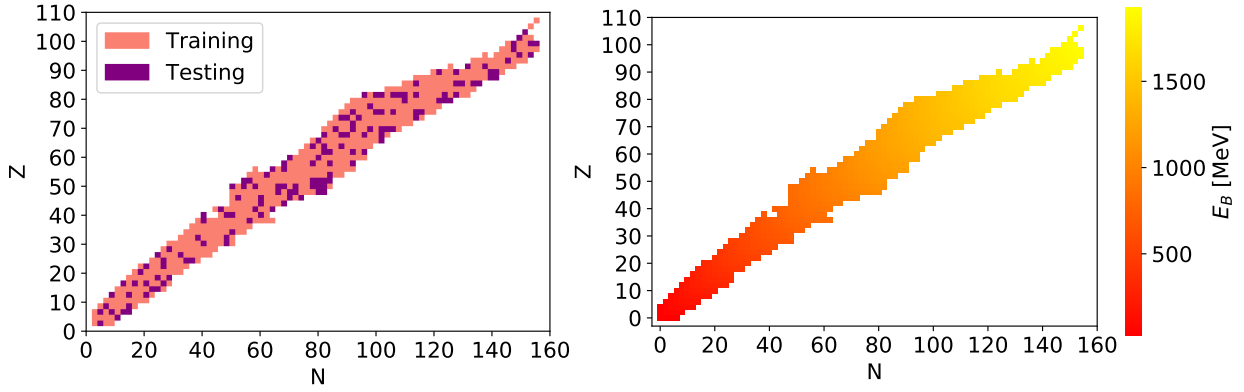

Figure 4.3: Binding energies of even-even nuclei in AME2003 dataset divided into a testing and a training dataset.

the systematic discrepancy δ with zero mean GP and the isotropic squared exponential covariance function. For the purpose of this example, we also assume that the LDM is computationally expensive (or not directly accessible) and regard it is an unknown function of (Z, N) and $\boldsymbol{\theta}$. Similarly to the discrepancy δ, we assign a GP prior to $E_{\mathrm{B}}(N, Z)$ with

zero mean and the isotropic squared exponential covariance function. To this extent, we additionally generated a set of 900 model evaluations using the Latin hypercube design over the space spanning all reasonable values of the parameters $\boldsymbol{\theta}$ as given by the nuclear physics literature similarly to our previous analysis in Section 3.5.3. Corresponding nuclear configurations, the inputs (Z, N), were randomly assigned to the generated values of $\boldsymbol{\theta}$ from a set of two times duplicated training nuclei.

Results. The predictions of nuclear binding energies were computed as the means of the posterior predictive distribution (4.20) conditioned on the estimates of the calibration parameters $\boldsymbol{\theta}$, GP's hyperparameters $\boldsymbol{\gamma}$, and the noise scale $\hat{\sigma}_n$. The estimates for $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ were obtained numerically as the minimizers of L_{MLE} and $L_{CV(10)}$. The priors for the GP hyperparameters were chosen according to Section 3.5.3 in the case of the fully Bayesian treatment.

	Parameter estimates				Testing error
	a_{vol}	a_{surf}	a_{sym}	a_{C}	RMSE (MeV)
L_{MLE}	15.07	15.58	22.00	0.68	1.16
$L_{CV(10)}$	15.08	16.08	21.19	0.67	1.26
MH	15.32	16.09	22.09	0.70	1.16

Table 4.3: The RMSEs of the predictions evaluated on 145 even-even nuclei from the AME2003 dataset. The parameter estimates are also listed. The posterior means are shown in the case of the MH algorithm.

Table 4.3 gives the RMSE values calculated on the testing set of 145 even-even nuclei for the empirical Bayes approach and also the MH algorithm. The calibration parameter estimates are also provided with values that do not significantly differ between the methods considered. The resulting RMSEs are $1.1 - 1.3$ MeV which is a consistent result with our previous study in Section 3.4.2 that was conducted on the whole AME2003 dataset using the VBI approach. We also carried out a simple least squares fit of the LDM with the resulting RMSE of 4.10 MeV evaluated on the same testing set of even-even nuclei. This is

an improvement that is consistent with our previous study on the full dataset using the VBI algorithm. Overall, this is quite a remarkable result given the considerable effort that needs to be put forth to implement the fully Bayesian solution and to obtain sufficient amount of posterior samples.

4.5 Technical details and supplementary results

4.5.1 Equivalency of hierarchical model

To establish the equivalency between the Bayesian model given by the data likelihood $p(\boldsymbol{d}|\boldsymbol{\theta},\boldsymbol{\gamma},\sigma)$ and the hierarchical model (see Section 4.1), we need to show that the following equality holds

$$p(\boldsymbol{d}|\boldsymbol{\theta},\boldsymbol{\gamma},\sigma) = \int_{\boldsymbol{\zeta}} \prod_{i}^{n} p(y_i|\zeta_i,\sigma) p(\boldsymbol{\zeta},\boldsymbol{z}|\boldsymbol{\theta},\boldsymbol{\gamma})\,\mathrm{d}\boldsymbol{\zeta}, \tag{4.26}$$

where $\boldsymbol{\zeta} = (\zeta(\boldsymbol{t}_1),\ldots,\zeta(\boldsymbol{t}_n)) = (\zeta_1,\ldots,\zeta_n)$ and $p(\boldsymbol{\zeta},\boldsymbol{z}|\boldsymbol{\theta},\boldsymbol{\gamma})$ is the multivariate normal distribution with the mean $M(\boldsymbol{\theta},\boldsymbol{\gamma})$ (see (1.5)) and the covariance

$$K_p(\boldsymbol{\theta},\boldsymbol{\gamma}) = \begin{pmatrix} K_f(T_y(\boldsymbol{\theta}),T_y(\boldsymbol{\theta})) + K_\delta(T_y,T_y) & K_f(T_y(\boldsymbol{\theta}),T_z(\widetilde{\boldsymbol{\theta}})) \\ K_f(T_z(\widetilde{\boldsymbol{\theta}}),T_y(\boldsymbol{\theta})) & K_f(T_z(\widetilde{\boldsymbol{\theta}}),T_z(\widetilde{\boldsymbol{\theta}})) \end{pmatrix} = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}.$$

For the ease of notation, let us now assume $M(\boldsymbol{\theta},\boldsymbol{\gamma}) = (M_y^T, M_z^T)^T$. Then

$$\begin{aligned}
&\int_{\boldsymbol{\zeta}} \prod_{i}^{n} p(y_i|\zeta_i,\sigma) p(\boldsymbol{\zeta},\boldsymbol{z}|\boldsymbol{\theta},\boldsymbol{\gamma})\,\mathrm{d}\boldsymbol{\zeta} = \int_{\boldsymbol{\zeta}} \frac{1}{(2\pi)^{n/2}|\sigma^2 I_n|^{1/2}} \exp\Big(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\zeta})^T(\sigma^2 I_n)^{-1}(\boldsymbol{y}-\boldsymbol{\zeta})\Big) \\
&\quad \times \frac{1}{(2\pi)^{(n+m)/2}|K_p|^{1/2}} \exp\Big(-\frac{1}{2}\begin{pmatrix}\boldsymbol{\zeta}-M_y \\ \boldsymbol{z}-M_z\end{pmatrix}^T K_p^{-1} \begin{pmatrix}\boldsymbol{\zeta}-M_y \\ \boldsymbol{z}-M_z\end{pmatrix}\Big)\,\mathrm{d}\boldsymbol{\zeta} \\
&= \frac{1}{(2\pi)^{(n+m)/2}|K|^{1/2}} \exp\Big(-\frac{1}{2}\begin{pmatrix}\boldsymbol{y}-M_y \\ \boldsymbol{z}-M_z\end{pmatrix}^T K^{-1} \begin{pmatrix}\boldsymbol{y}-M_y \\ \boldsymbol{z}-M_z\end{pmatrix}\Big) \\
&\quad \times \int_{\boldsymbol{\zeta}} \frac{|K|^{1/2}}{(2\pi)^{n/2}|\sigma^2 I_n|^{1/2}|K_p|^{1/2}} \exp\Big(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\zeta})^T(\sigma^2 I_n)^{-1}(\boldsymbol{y}-\boldsymbol{\zeta})\Big) \\
&\quad \times \exp\Big(-\frac{1}{2}\begin{pmatrix}\boldsymbol{\zeta}-M_y \\ \boldsymbol{z}-M_z\end{pmatrix}^T K_p^{-1} \begin{pmatrix}\boldsymbol{\zeta}-M_y \\ \boldsymbol{z}-M_z\end{pmatrix}\Big) + \frac{1}{2}\begin{pmatrix}\boldsymbol{y}-M_y \\ \boldsymbol{z}-M_z\end{pmatrix}^T K^{-1} \begin{pmatrix}\boldsymbol{y}-M_y \\ \boldsymbol{z}-M_z\end{pmatrix}\Big)\,\mathrm{d}\boldsymbol{\zeta} \\
&= \frac{1}{(2\pi)^{(n+m)/2}|K|^{1/2}} \exp\Big(-\frac{1}{2}\begin{pmatrix}\boldsymbol{y}-M_y \\ \boldsymbol{z}-M_z\end{pmatrix}^T K^{-1} \begin{pmatrix}\boldsymbol{y}-M_y \\ \boldsymbol{z}-M_z\end{pmatrix}\Big) \times 1.
\end{aligned}$$

The integral is equal to 1 since it is an integration of multivariate normal probability density function over $\boldsymbol{\zeta}$ with the covariance function $((\sigma^2 I_n)^{-1} + (C_{11} - C_{12}C_{22}^{-1}C_{21})^{-1})^{-1}$. Namely,

$$\begin{aligned}
\frac{|K|^{1/2}}{|\sigma^2 I_n|^{1/2}|K_p|^{1/2}} &= \frac{|C_{22}|^{1/2}|C_{11} + \sigma^2 I_n - C_{12}C_{22}^{-1}C_{21}|^{1/2}}{|\sigma^2 I_n|^{1/2}|C_{22}|^{1/2}|C_{11} - C_{12}C_{22}^{-1}C_{21}|^{1/2}} \\
&= \frac{|C_{11} + \sigma^2 I_n - C_{12}C_{22}^{-1}C_{21}|^{1/2}}{|\sigma^2 I_n|^{1/2}|C_{11} - C_{12}C_{22}^{-1}C_{21}|^{1/2}} \\
&= \frac{|A+B|^{1/2}}{|A|^{1/2}|B|^{1/2}} = \frac{1}{|A|^{1/2}|B|^{1/2}|A+B|^{-1/2}} = \frac{1}{(|A^{-1}||B^{-1}||A+B|)^{-1/2}} \\
&= \frac{1}{|A^{-1}B^{-1}A + A^{-1}B^{-1}B|^{-1/2}} = \frac{1}{|A^{-1}B^{-1}A + A^{-1}|^{-1/2}} \\
&= \frac{1}{|A^{-1}(B^{-1} + A^{-1})A|^{-1/2}} = \frac{1}{(|A^{-1}||(B^{-1} + A^{-1})||A|)^{-1/2}} \\
&= \frac{1}{|(B^{-1} + A^{-1})^{-1}|^{1/2}}
\end{aligned}$$

where we used the Schur complement identity for determinants in the first equality and

$$\begin{aligned}
A &= C_{11} - C_{12}C_{22}^{-1}C_{21}, \\
B &= \sigma^2 I_n.
\end{aligned}$$

Lastly, considering the notation

$$K_p^{-1} = \begin{pmatrix} C_{11}^- & C_{12}^- \\ C_{21}^- & C_{22}^- \end{pmatrix}$$

we have

$$\begin{aligned}
&\exp\bigg(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\zeta})^T(\sigma^2 I_n)^{-1}(\boldsymbol{y} - \boldsymbol{\zeta}) - \frac{1}{2}\begin{pmatrix} \boldsymbol{\zeta} - M_y \\ \boldsymbol{z} - M_z \end{pmatrix}^T K_p^{-1}\begin{pmatrix} \boldsymbol{\zeta} - M_y \\ \boldsymbol{z} - M_z \end{pmatrix}\bigg) \\
&\times \exp\bigg(\frac{1}{2}\begin{pmatrix} \boldsymbol{y} - M_y \\ \boldsymbol{z} - M_z \end{pmatrix}^T K^{-1}\begin{pmatrix} \boldsymbol{y} - M_y \\ \boldsymbol{z} - M_z \end{pmatrix}\bigg) \\
&\propto \exp\bigg(-\frac{1}{2}\boldsymbol{\zeta}^T(\sigma^2 I_n)^{-1}\boldsymbol{\zeta} + \boldsymbol{\zeta}^T(\sigma^2 I_n)^{-1}\boldsymbol{y} - \frac{1}{2}\boldsymbol{y}^T(\sigma^2 I_n)^{-1}\boldsymbol{y}\bigg) \\
&\times \exp\bigg(-\frac{1}{2}[(\boldsymbol{\zeta} - M_y)^T C_{11}^- + (\boldsymbol{z} - M_z)^T C_{21}^-, (\boldsymbol{\zeta} - M_y)^T C_{12}^- + (\boldsymbol{z} - M_z)^T C_{22}^-]\begin{pmatrix} \boldsymbol{\zeta} - M_y \\ \boldsymbol{z} - M_z \end{pmatrix}\bigg) \\
&\propto \exp\bigg(-\frac{1}{2}\boldsymbol{\zeta}^T((\sigma^2 I_n)^{-1} + C_{11}^-)\boldsymbol{\zeta} + \boldsymbol{\zeta}^T\boldsymbol{b}\bigg)
\end{aligned}$$

where $C_{11}^{-} = C_{11} - C_{12}C_{22}^{-1}C_{21}$ due to the Schur complement identity for matrix inverse, and $\boldsymbol{b}$ is a constant column vector. This shows that integral is indeed equal to 1 as stated, and the equality (4.26) holds.

4.5.2 Proofs

Proof of Theorem 2

Note that for any $\epsilon > 0$, the posterior probability of interest $p(\zeta \in U_n^C | y_1, \ldots, y_n, \boldsymbol{\theta}, \boldsymbol{\gamma}, \hat{\sigma}_n)$ can be bound from the above as

$$p(\zeta \in U_n^C | y_1, \ldots, y_n, \boldsymbol{\theta}, \boldsymbol{\gamma}, \hat{\sigma}_n) \leq p(\zeta \in U_n^C | y_1, \ldots, y_n, \boldsymbol{\theta}, \boldsymbol{\gamma}, \hat{\sigma}_n) 1_{\{\left|\frac{\hat{\sigma}_n}{\sigma_0} - 1\right| \leq \epsilon\}} + 1_{\{\left|\frac{\hat{\sigma}_n}{\sigma_0} - 1\right| > \epsilon\}},$$

where

$$\begin{aligned}
&p(\zeta \in U_n^C | y_1, \ldots, y_n, \boldsymbol{\theta}, \boldsymbol{\gamma}, \hat{\sigma}_n) 1_{\{\left|\frac{\hat{\sigma}_n}{\sigma_0} - 1\right| \leq \epsilon\}} \\
&\leq \Phi_n + \frac{(1 - \Phi_n) \int_{U_n^c \cap \mathcal{F}_n} \prod_{i=1}^n \frac{p(y_i|\zeta_i, \hat{\sigma}_n)}{p(y_i|\zeta_{0,i}, \sigma_0)} 1_{\{\left|\frac{\hat{\sigma}_n}{\sigma_0} - 1\right| \leq \epsilon\}} \, \mathrm{d}\Pi(\boldsymbol{\zeta}|\boldsymbol{\theta}, \boldsymbol{\gamma})}{\int_{\mathcal{F}} \prod_{i=1}^n \frac{p(y_i|\zeta_i, \hat{\sigma}_n)}{p(y_i|\zeta_{0,i}, \sigma_0)} \, \mathrm{d}\Pi(\boldsymbol{\zeta}|\boldsymbol{\theta}, \boldsymbol{\gamma})} \\
&\quad + \frac{\int_{U_n^c \cap \mathcal{F}_n^C} \prod_{i=1}^n \frac{p(y_i|\zeta_i, \hat{\sigma}_n)}{p(y_i|\zeta_{0,i}, \sigma_0)} 1_{\{\left|\frac{\hat{\sigma}_n}{\sigma_0} - 1\right| \leq \epsilon\}} \, \mathrm{d}\Pi(\boldsymbol{\zeta}|\boldsymbol{\theta}, \boldsymbol{\gamma})}{\int_{\mathcal{F}} \prod_{i=1}^n \frac{p(y_i|\zeta_i, \hat{\sigma}_n)}{p(y_i|\zeta_{0,i}, \sigma_0)} \, \mathrm{d}\Pi(\boldsymbol{\zeta}|\boldsymbol{\theta}, \boldsymbol{\gamma})} \\
&= \Phi_n + \frac{\mathbf{I}_{1n}(y_1, \ldots, y_n, \boldsymbol{\theta}, \boldsymbol{\gamma}, \hat{\sigma}_n, \epsilon)}{\mathbf{I}_{3n}(y_1, \ldots, y_n, \boldsymbol{\theta}, \boldsymbol{\gamma}, \hat{\sigma}_n)} + \frac{\mathbf{I}_{2n}(y_1, \ldots, y_n, \boldsymbol{\theta}, \boldsymbol{\gamma}, \hat{\sigma}_n, \epsilon)}{\mathbf{I}_{3n}(y_1, \ldots, y_n, \boldsymbol{\theta}, \boldsymbol{\gamma}, \hat{\sigma}_n)}.
\end{aligned}$$

Since the assumption (A3) implies that $1_{\{\left|\frac{\hat{\sigma}_n}{\sigma_0} - 1\right| > \epsilon\}} \underset{\text{n}}{\longrightarrow} 0$ a.s. P_0, it is enough to show that there exists $\epsilon > 0$ so that

$$\sup_{(\boldsymbol{\theta}, \boldsymbol{\gamma}) \in \Upsilon} \Phi_n \underset{\text{n}}{\longrightarrow} 0 \text{ a.s. } P_0, \tag{4.27}$$

$$\sup_{(\boldsymbol{\theta}, \boldsymbol{\gamma}) \in \Upsilon} e^{\beta_1 n} \mathbf{I}_{1n}(y_1, \ldots, y_n, \boldsymbol{\theta}, \boldsymbol{\gamma}, \hat{\sigma}_n, \epsilon) \underset{\text{n}}{\longrightarrow} 0 \text{ a.s. } P_0 \quad \text{for some } \beta_1 > 0, \tag{4.28}$$

$$\sup_{(\boldsymbol{\theta}, \boldsymbol{\gamma}) \in \Upsilon} e^{\beta_2 n} \mathbf{I}_{2n}(y_1, \ldots, y_n, \boldsymbol{\theta}, \boldsymbol{\gamma}, \hat{\sigma}_n, \epsilon) \underset{\text{n}}{\longrightarrow} 0 \text{ a.s. } P_0 \quad \text{for some } \beta_2 > 0, \tag{4.29}$$

$$\sup_{(\boldsymbol{\theta}, \boldsymbol{\gamma}) \in \Upsilon} e^{\beta_3 n} \mathbf{I}_{3n}(y_1, \ldots, y_n, \boldsymbol{\theta}, \boldsymbol{\gamma}, \hat{\sigma}_n) \underset{\text{n}}{\longrightarrow} \infty \text{ a.s. } P_0 \quad \text{for some } \beta_3 > 0, \tag{4.30}$$

where $\beta_3 \leq \min\{\beta_1, \beta_2\}$.

The rest of the proof follows the general steps of the proof of Theorem 1 in Choi and Schervish (2007a) and Theorem 9 in Choi (2007) with some non-trivial treatment of the constant ϵ. We shall provide step by step details below.

Step 1). By Markov inequality, for any $\delta > 0$

$$\sum_{n=1}^{\infty} P_0(\Phi_n > \delta) \leq \frac{1}{\delta} \sum_{n=1}^{\infty} \mathbb{E}_{\zeta_0,\sigma_0} \Phi_n,$$

which due to the condition (i) of (A2) and the first Borel-Cantelli Lemma yields

$$\Phi_n \underset{\text{n}}{\longrightarrow} 0 \text{ a.s. } P_0.$$

Since this does not depend on $(\boldsymbol{\theta}, \boldsymbol{\gamma})$, it implies (4.27).

Step 2). By Fubini's theorem and for any $0 < \epsilon < \tilde{\epsilon}_2$

$$\begin{aligned}
&\mathbb{E}_{\zeta_0,\sigma_0}(\mathbf{I}_{1n}(y_1, \ldots, y_n, \boldsymbol{\theta}, \boldsymbol{\gamma}, \hat{\sigma}_n, \epsilon)) \\
&= \mathbb{E}_{\zeta_0,\sigma_0}\left[(1 - \Phi_n) \int_{U_n^c \cap \mathcal{F}_n} \prod_{i=1}^{n} \frac{p(y_i|\zeta_i, \hat{\sigma}_n)}{p(y_i|\zeta_{0,i}, \sigma_0)} 1_{\{\left|\frac{\hat{\sigma}_n}{\sigma_0} - 1\right| \leq \epsilon\}} \, \mathrm{d}\Pi(\boldsymbol{\zeta}|\boldsymbol{\theta}, \boldsymbol{\gamma})\right] \\
&= \int_{U_n^c \cap \mathcal{F}_n} \int (1 - \Phi_n) \prod_{i=1}^{n} \frac{p(y_i|\zeta_i, \hat{\sigma}_n)}{p(y_i|\zeta_{0,i}, \sigma_0)} 1_{\{\left|\frac{\hat{\sigma}_n}{\sigma_0} - 1\right| \leq \epsilon\}} \, \mathrm{d}P_0 \, \mathrm{d}\Pi(\boldsymbol{\zeta}|\boldsymbol{\theta}, \boldsymbol{\gamma}) \\
&\leq \left(\frac{\sigma_0(1-\epsilon)}{\sigma_0(1+\epsilon)}\right)^{-n} \int_{U_n C \cap \mathcal{F}_n} \mathbb{E}_{\boldsymbol{\zeta},\sigma_0(1+\epsilon)}[(1 - \Phi_n)] \, \mathrm{d}\Pi(\boldsymbol{\zeta}|\boldsymbol{\theta}, \boldsymbol{\gamma}) \\
&\leq \left(\frac{1-\epsilon}{1+\epsilon}\right)^{-n} \sup_{\zeta \in U_n^C \cap \mathcal{F}_n} \mathbb{E}_{\zeta,\sigma_0(1+\epsilon)}[(1 - \Phi_n)] \\
&\leq \left(\frac{1-\epsilon}{1+\epsilon}\right)^{-n} C_2 e^{-c_\epsilon n} = C_2 e^{-\tilde{c}_\epsilon n},
\end{aligned}$$

where $\tilde{c}_\epsilon = c_\epsilon + \log(1-\epsilon) - \log(1+\epsilon)$ together with condition (iii) of (A2) implies $\tilde{c}_\epsilon > 0$. Thus

$$P_0\left\{\mathbf{I}_{1n}(y_1, \ldots, y_n, \boldsymbol{\theta}, \boldsymbol{\gamma}, \hat{\sigma}_n, \epsilon) \geq e^{-\tilde{c}_\epsilon \frac{n}{2}}\right\} \leq C_1 e^{\tilde{c}_\epsilon \frac{n}{2}} e^{-\tilde{c}_\epsilon n} = C_1 e^{-\tilde{c}_\epsilon \frac{n}{2}}.$$

Therefore, for any $\epsilon > 0$ so that $\epsilon < \tilde{\epsilon}_2$ there exists a constant $\tilde{c}_\epsilon$ for which the first Borel-Cantelli Lemma implies

$$e^{\tilde{c}_\epsilon \frac{n}{4}} \mathbf{I}_{1n}(y_1, \ldots, y_n, \boldsymbol{\theta}, \boldsymbol{\gamma}, \hat{\sigma}_n, \epsilon) \underset{\text{n}}{\longrightarrow} 0 \text{ a.s. } P_0.$$

Since this does not depend on $(\boldsymbol{\theta}, \boldsymbol{\gamma})$, it implies (4.28).

Step 3). If we proceed as in the step 2), the Fubini's theorem implies

$$\begin{aligned}
&\mathbb{E}_{\zeta_0,\sigma_0}(\mathbf{I}_{2n}(y_1, \ldots, y_n, \boldsymbol{\theta}, \boldsymbol{\gamma}, \hat{\sigma}_n, \epsilon)) \\
&= \mathbb{E}_{\zeta_0,\sigma_0}\left[\int_{U_n^c \cap \mathcal{F}_n} \prod_{i=1}^n \frac{p(y_i|\zeta_i, \hat{\sigma}_n)}{p(y_i|\zeta_{0,i}, \sigma_0)} 1_{\{\left|\frac{\hat{\sigma}_n}{\sigma_0} - 1\right| \leq \epsilon\}} \, \mathrm{d}\Pi(\boldsymbol{\zeta}|\boldsymbol{\theta}, \boldsymbol{\gamma}) \right] \\
&\leq \left(\frac{\sigma_0(1-\epsilon)}{\sigma_0(1+\epsilon)} \right)^{-n} \int_{U_n C \cap \mathcal{F}_n^C} \mathbb{E}_{\zeta, \sigma_0(1+\epsilon)}[1] \, \mathrm{d}\Pi(\boldsymbol{\zeta}|\boldsymbol{\theta}, \boldsymbol{\gamma}) \\
&\leq \left(\frac{1-\epsilon}{1+\epsilon} \right)^{-n} \Pi(\mathcal{F}_n^C|\boldsymbol{\theta}, \boldsymbol{\gamma}).
\end{aligned}$$

The condition (ii) of (A2) and the first Borel-Cantelli Lemma implies that for any $\epsilon < \frac{1-e^{-c_1}}{1+e^{-c_1}}$:

$$\sup_{(\boldsymbol{\theta},\boldsymbol{\gamma}) \in \Upsilon} e^{\tilde{k}_\epsilon \frac{n}{4}} \mathbf{I}_{2n}(y_1, \ldots, y_n, \boldsymbol{\theta}, \boldsymbol{\gamma}, \hat{\sigma}_n, \epsilon) \underset{\text{n}}{\longrightarrow} 0 \text{ a.s. } P_0,$$

where $\tilde{k}_\epsilon = c_1 + \log(1-\epsilon) - \log(1+\epsilon)$.

Step 4). To prove (4.30), given any $0 < \rho < 1$, we first observe the following:

$$\begin{aligned}
\mathbf{I}_{3n}(y_1, \ldots, y_n, \boldsymbol{\theta}, \boldsymbol{\gamma}, \hat{\sigma}_n) &\geq \mathbf{I}_{3n}(y_1, \ldots, y_n, \boldsymbol{\theta}, \boldsymbol{\gamma}, \hat{\sigma}_n) 1_{\{\left|\frac{\hat{\sigma}_n}{\sigma_0} - 1\right| \leq \rho\}} \\
&\geq \left(\frac{1-\rho}{1+\rho} \right)^n \int_{\mathcal{F}} \prod_{i=1}^n \frac{p(y_i|\zeta_i, \sigma_0(1-\rho))}{p(y_i|\zeta_{0,i}, \sigma_0)} \, \mathrm{d}\Pi(\boldsymbol{\zeta}|\boldsymbol{\theta}, \boldsymbol{\gamma}).
\end{aligned}$$

Let us now define $\log_+(x) = \max\{0, log(x)\}$ and $\log_-(x) = -\min\{0, log(x)\}$ as well as

$$\begin{aligned}
W_i &= \log_+ \frac{p(y_i|\zeta_{0,i}, \sigma_0)}{p(y_i|\zeta_i, \sigma_0(1-\rho))}, \\
K_i^+(\zeta_0, \zeta) &= \int p(y_i|\zeta_{0,i}, \sigma_0) \log_+ \frac{p(y_i|\zeta_{0,i}, \sigma_0)}{p(y_i|\zeta_i, \sigma_0(1-\rho))} \, \mathrm{d}y_i, \\
K_i^-(\zeta_0, \zeta) &= \int p(y_i|\zeta_{0,i}, \sigma_0) \log_- \frac{p(y_i|\zeta_{0,i}, \sigma_0)}{p(y_i|\zeta_i, \sigma_0(1-\rho))} \, \mathrm{d}y_i.
\end{aligned}$$

Then we get

$$\begin{aligned}
\mathbb{V}ar_{\zeta_0,\sigma_0}(W_i) &= \mathbb{E}_{\zeta_0,\sigma_0}(W_i^2) - \{K_i^+(\zeta_0,\zeta)\}^2 \\
&\leq \mathbb{E}_{\zeta_0,\sigma_0}(W_i^2) - \{K_i(\zeta_0,\zeta)\}^2 \\
&\leq \mathbb{E}_{\zeta_0,\sigma_0}(W_i^2) + \int p(y_i|\zeta_{0,i},\sigma_0)\left(\log_- \frac{p(y_i|\zeta_{0,i},\sigma_0)}{p(y_i|\zeta_i,\sigma_0(1-\rho))}\right)^2 \mathrm{d}y_i - \{K_i(\zeta_0,\zeta)\}^2 \\
&= \int p(y_i|\zeta_{0,i},\sigma_0)\left(\log \frac{p(y_i|\zeta_{0,i},\sigma_0)}{p(y_i|\zeta_i,\sigma_0(1-\rho))}\right)^2 \mathrm{d}y_i - \{K_i(\zeta_0,\zeta)\}^2 \\
&= V_i(\zeta_0,\zeta).
\end{aligned}$$

Hence, by condition (i) of (A1) for any $\rho < \tilde{\epsilon}_1$ and $\zeta \in B$

$$\sum_{i=1}^{n=\infty} \frac{\mathbb{V}ar_{\zeta_0,\sigma_0}(W_i)}{i^2} \leq \sum_{i=1}^{n=\infty} \frac{V_i(\zeta_0,\zeta)}{i^2} < \infty,$$

and by the Kolmogorov's strong law of large numbers for independent non-identically distributed random variables (e.g. Shiryaev (1996), Chapter 3),

$$\frac{1}{n}\sum_{i=1}^{n}(W_i - K_i^+(\zeta_0,\zeta)) \underset{\mathrm{n}}{\longrightarrow} 0 \text{ a.s. } P_0.$$

As a result, for every $\zeta \in B$, with P_0 probability 1

$$\begin{aligned}
\liminf_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n}\log\frac{p(y_i|\zeta_i,\sigma_0(1-\rho))}{p(y_i|\zeta_{0,i},\sigma_0)}\right) &= -\liminf_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n} -\log\frac{p(y_i|\zeta_i,\sigma_0(1-\rho))}{p(y_i|\zeta_{0,i},\sigma_0)}\right) \\
&= -\liminf_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n}\log\frac{p(y_i|\zeta_{0,i},\sigma_0)}{p(y_i|\zeta_i,\sigma_0(1-\rho))}\right) \\
&\geq -\limsup_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n}\log_+\frac{p(y_i|\zeta_{0,i},\sigma_0)}{p(y_i|\zeta_i,\sigma_0(1-\rho))}\right) \\
&= -\limsup_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n}K_i^+(\zeta_0,\zeta)\right) \\
&\geq -\limsup_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n}K_i(S_0,S) + \frac{1}{n}\sum_{i=1}^{n}\sqrt{\frac{K_i(\zeta_0,\zeta)}{2}}\right) \\
&\geq -\limsup_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n}K_i(\zeta_0,\zeta) + \sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{K_i(\zeta_0,\zeta)}{2}}\right).
\end{aligned}$$

The fourth line follows from the almost sure convergence proved in the previous paragraph, the second to last line follows from Amewou-Atisso et al. (2003). We now make use of the condition (ii) of (A1). Let us consider $\beta > 0$ and select Δ so that $\Delta + \sqrt{\frac{\Delta}{2}} \leq \frac{\beta}{8}$ and also $C = B \cap \{\zeta : K_i(\zeta_0, \zeta) < \Delta \text{ for all } i\}$. By (A1) there exists $\tilde{\epsilon}_1$ so that for all $0 < \rho < \tilde{\epsilon}_1$ implies $\Pi(C|\boldsymbol{\theta}, \boldsymbol{\gamma}) > 0$. Therefore, for each $\zeta \in C$

$$\begin{aligned}\liminf_{n\to\infty} \left(\frac{1}{n} \sum_{i=1}^{n} log \frac{p(y_i|\zeta_i, \sigma_0(1-\rho))}{p(y_i|\zeta_{0,i}, \sigma_0)} \right) &\geq - \limsup_{n\to\infty} \left(\frac{1}{n} \sum_{i=1}^{n} K_i(\zeta_0, \zeta) + \sqrt{\frac{1}{n} \sum_{i=1}^{n} \frac{K_i(\zeta_0, \zeta)}{2}} \right) \\ &\geq -(\Delta + \sqrt{\frac{\Delta}{2}}),\end{aligned}$$

since $\frac{1}{n} \sum_{i=1}^{n} K_i(\zeta_0, \zeta) < \Delta$ for all $\zeta \in C$. Finally, for any $\rho < \min\{\tilde{\epsilon}_1, \frac{1-e^{\frac{-\beta}{8}}}{1+e^{\frac{-\beta}{8}}}\}$

$$\begin{aligned}&\liminf_{n\to\infty} e^{\frac{2n\beta}{8}} \mathbf{I}_{3n}(y_1, \ldots, y_n, \boldsymbol{\theta}, \boldsymbol{\gamma}, \hat{\sigma}_n) \\ &\qquad \geq \liminf_{n\to\infty} e^{\frac{2n\beta}{8}} \left(\frac{1-\rho}{1+\rho} \right)^n \int_{\mathcal{F}} \prod_{i=1}^{n} \frac{p(y_i|\zeta_i, \sigma_0(1-\rho))}{p(y_i|\zeta_{0,i}, \sigma_0)} \, \mathrm{d}\Pi(\boldsymbol{\zeta}|\boldsymbol{\theta}, \boldsymbol{\gamma}) \\ &\qquad \geq \liminf_{n\to\infty} e^{\frac{2n\beta}{8}} \left(\frac{1-\rho}{1+\rho} \right)^n \int_{C} \prod_{i=1}^{n} \frac{p(y_i|\zeta_i, \sigma_0(1-\rho))}{p(y_i|\zeta_{0,i}, \sigma_0)} \, \mathrm{d}\Pi(\boldsymbol{\zeta}|\boldsymbol{\theta}, \boldsymbol{\gamma}) \\ &\qquad \geq \int_{C} \liminf_{n\to\infty} e^{\frac{2n\beta}{8}} \left(\frac{1-\rho}{1+\rho} \right)^n \prod_{i=1}^{n} \frac{p(y_i|\zeta_i, \sigma_0(1-\rho))}{p(y_i|\zeta_{0,i}, \sigma_0)} \, \mathrm{d}\Pi(\boldsymbol{\zeta}|\boldsymbol{\theta}, \boldsymbol{\gamma}) \\ &\qquad = \infty.\end{aligned}$$

Note that the actual bound on $\mathbf{I}_{3n}$ does not depend on $(\boldsymbol{\theta}, \boldsymbol{\gamma})$. Taking $\epsilon < \min\{\tilde{\epsilon}_2, \frac{1-e^{-c_1}}{1+e^{-c_1}}\}$ concludes the proof. □

Proof of Theorem 3

We shall first define some notation. Let $0 < r < \frac{\nu}{2}$ and $t = \frac{r}{4}$. Let $N_t = N(t, \mathcal{F}_n, \|\cdot\|_\infty)$ be the covering number of $\mathcal{F}_n$. In Theorem 2.7.1, van der Vaart and Wellner (1996) show that there exist a constant K so that $\log N_t \leq \frac{KM_n}{t^p}$ and therefore $N_t = \mathcal{O}(M_n)$, where $M_n = \mathcal{O}(n^\alpha)$ for $\alpha \in (\frac{1}{2}, 1)$ according to the definition of the sieves. Let us consider $\tau \in (\frac{\alpha}{2}, \frac{1}{2})$ and define $c_n = n^\tau$ so that $\log(N_t) = o(c_n^2)$. Moreover, let $\zeta^1, \ldots, \zeta^{N_t} \in \mathcal{F}_n$ be

finitely many elements of the sieve so that for every $\zeta \in \mathcal{F}_n$ there is $i \in \{1, \ldots, N_t\}$ satisfying $\| \zeta - \zeta^i \|_\infty < t$. This implies that if $\zeta \in \mathcal{F}_n$ such that $\int |\zeta(\boldsymbol{t}) - \zeta_0(\boldsymbol{t})| \, \mathrm{d}Q_n(\boldsymbol{t}) > \nu$, then $\int |\zeta^i(\boldsymbol{t}) - \zeta_0(\boldsymbol{t})| \, \mathrm{d}Q_n(\boldsymbol{t}) > \frac{\nu}{2}$.

The next step in the proof is to construct a test for each ζ^i with the resulting functions Φ_n defined as a combination of the individual tests and showing that the probabilities of type I and type II errors satisfies the properties of the theorem. Let us recall that $\zeta_j = \zeta(\boldsymbol{t}_j)$ and $\zeta_{0,j} = \zeta_0(\boldsymbol{t}_j)$. For an arbitrary $\zeta \in \mathcal{F}_n$ such that $\| \zeta - \zeta^i \|_\infty < t$, let us define $\zeta_{1,j} = \zeta^i(\boldsymbol{t}_j)$ and $b_j = 1$ if $\zeta_{1,j} > \zeta_{0,j}$ and -1 otherwise. For any $\nu > 0$, let $\Psi_n[\zeta, \nu]$ be the indicator of set A defined as follows

$$A = \left\{ \sum_{j=1}^{n} b_j \left(\frac{y_j - \zeta_{0,j}}{\sigma_0} \right) > 2c_n \sqrt{n} \right\}.$$

The test functions Φ_n are then

$$\Phi_n = \max_{1 \le j \le N_t} \Psi_n[\zeta^j, \frac{\nu}{2}].$$

Type I error. The Mill's ratio implies

$$\begin{aligned}
\mathbb{E}_{\zeta_0, \sigma_0}(\Psi_n) &= P_0 \left[\sum_{j=1}^{n} b_j \left(\frac{y_j - \zeta_{0,j}}{\sigma_0} \right) > 2c_n \sqrt{n} \right] \\
&= 1 - \Phi(2c_n) \\
&\le \frac{1}{2c_n \sqrt{2\pi}} e^{-2c_n^2} \\
&\le e^{-2c_n^2}.
\end{aligned}$$

The function $\Phi(\cdot)$ is the CDF of the standard normal distribution. Consequently, we have

$$\begin{aligned}
\mathbb{E}_{\zeta_0, \sigma_0}(\Phi_n) &\le \sum_{j=1}^{N_t} \mathbb{E}_{\zeta_0, \sigma_0}(\Psi_n[\zeta^j, \frac{\nu}{2}]) \\
&\le N_t e^{-2c_n^2} = e^{\log(N_t) - 2c_n^2} \\
&\le e^{-c_n^2},
\end{aligned}$$

and

$$\sum_{n=1}^{\infty} \mathbb{E}_{\zeta_0,\sigma_0} \Phi_n < \infty.$$

Type II error. It is sufficient to find i for which the probability of type II error of $\Psi_n[\zeta^i, \frac{\nu}{2}]$, given an arbitrary ζ in $W^C_{\nu,n} \cap \mathcal{F}_n$, is sufficiently small. This is because the probability of type II error for the composite test Φ_n is no larger than the smallest of $\Psi_n[\zeta^i, \frac{\nu}{2}]$. Note that here we assume $\int |\zeta(\boldsymbol{t}) - \zeta_0(\boldsymbol{t})| \, \mathrm{d}Q_n(\boldsymbol{t}) > \nu$, and then $\int |\zeta^i(\boldsymbol{t}) - \zeta_0(\boldsymbol{t})| \, \mathrm{d}Q_n(\boldsymbol{t}) > \frac{\nu}{2}$. For every $r < \frac{\nu}{2}$, Choi and Schervish (2007b) show that

$$\sum_{j=1}^{n} |\zeta_{1,j} - \zeta_{0,j}| > rn.$$

Let n be large enough so that $4\sigma_0 c_n < r\sqrt{n}$, then for any $0 < \epsilon < 1$

$$\begin{aligned}
\mathbb{E}_{\zeta,\sigma_0(1+\epsilon)}(1 - \Psi_n[\zeta^i, \frac{\nu}{2}]) &= P_{\zeta,\sigma_0(1+\epsilon)}\left[\sum_{j=1}^{n} b_j \left(\frac{y_j - \zeta_{0,j}}{\sigma_0}\right) \le 2c_n\sqrt{n}\right] \\
&= P_{\zeta,\sigma_0(1+\epsilon)}\left[\sum_{j=1}^{n} b_j \left(\frac{y_j - \zeta_j + \zeta_j - \zeta_{1,j} + \zeta_{1,j} + \zeta_{0,j}}{\sigma_0}\right) \le 2c_n\sqrt{n}\right] \\
&= P_{\zeta,\sigma_0(1+\epsilon)}\left[\frac{1}{\sqrt{n}}\sum_{j=1}^{n} b_j \left(\frac{y_j - \zeta_j}{\sigma_0}\right) + \frac{1}{\sqrt{n}}\sum_{j=1}^{n} b_j \left(\frac{\zeta_j - \zeta_{1,j}}{\sigma_0}\right)\right. \\
&\quad \left. + \frac{1}{\sqrt{n}}\sum_{j=1}^{n} \left|\frac{\zeta_{1,j} - \zeta_{0,j}}{\sigma_0}\right| \le 2c_n\right] \\
&\le P_{\zeta,\sigma_0(1+\epsilon)}\left[\frac{1}{\sqrt{n}}\sum_{j=1}^{n} b_j \left(\frac{y_j - \zeta_j}{\sigma_0}\right) \le \frac{r\sqrt{n}}{4\sigma_0} - \frac{r\sqrt{n}}{\sigma_0} + 2c_n\right] \\
&\le P_{\zeta,\sigma_0(1+\epsilon)}\left[\frac{1}{\sqrt{n}}\sum_{j=1}^{n} b_j \left(\frac{y_j - \zeta_j}{\sigma_0(1+\epsilon)}\right) \le -\frac{r\sqrt{n}}{4\sigma_0(1+\epsilon)}\right] \\
&= \Phi\left(-\frac{r\sqrt{n}}{4\sigma_0(1+\epsilon)}\right) \\
&\le \frac{4\sigma_0(1+\epsilon)}{r\sqrt{2\pi n}} e^{-\frac{nr^2}{32\sigma_0^2(1+\epsilon)^2}}.
\end{aligned}$$

To establish the part (ii) of the theorem, we need to show that there exists a constant

$0 < \tilde{\epsilon} < 1$ so that for any $\epsilon < \tilde{\epsilon}$

$$\frac{r^2}{32\sigma_0^2(1+\epsilon)^2} + \log\left(\frac{1-\epsilon}{1+\epsilon}\right) > 0. \tag{4.31}$$

Take $\kappa = \frac{r^2}{32\sigma_0^2}$ and define $b(\epsilon)$ to be the left hand side of (4.31),

$$b(\epsilon) = \kappa\left(\frac{1}{(1+\epsilon)^2} + \frac{1}{\kappa}\log\left(\frac{1-\epsilon}{1+\epsilon}\right)\right).$$

The function $b(\epsilon)$ is clearly continuous at $\epsilon = 0$. Hence, for each $\kappa > 0$, there exists $\tilde{\epsilon}$ such that for all $0 < \epsilon < \tilde{\epsilon}$, $b(\epsilon) > 0$. □

Proof of Lemma 4

Theorem 5 of Ghosal and Roy (2006) implies that there exist positive constants $C, d_1, \ldots, d_p$ so that for $i = 1, \ldots, p$

$$P\left(\sup_{\boldsymbol{t}\in[0,1]^p} |\zeta(\boldsymbol{t})| > M_n \middle| \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \right) \leq Ce^{-d_0\frac{M_n^2}{\rho_0^2(\boldsymbol{\theta},\boldsymbol{\gamma})}},$$

$$P\left(\sup_{\boldsymbol{t}\in[0,1]^p} \left|\frac{\partial}{\partial t_i}\zeta(\boldsymbol{t})\right| > M_n | \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \right) \leq Ce^{-d_i\frac{M_n^2}{\rho_i^2(\boldsymbol{\theta},\boldsymbol{\gamma})}}.$$

The continuity of $\rho_i^2(\boldsymbol{\theta}, \boldsymbol{\gamma})$, for $i = 0, \cdots, p$, on a compact set Υ implies that they are uniformly bounded. Therefore, there exist universal constants $(\Xi_{0,1}, \Xi_{0,2}), \cdots, (\Xi_{p,1}, \Xi_{p,2})$ such that for $i = 0, \cdots, p$,

$$0 < \Xi_{i,1} \leq \sup_{(\boldsymbol{\theta},\boldsymbol{\gamma})\in\Upsilon} |\rho_i^2(\boldsymbol{\theta}, \boldsymbol{\gamma})| \leq \Xi_{i,2}.$$

Hence, for $i = 0, \cdots, p$,

$$\sup_{(\boldsymbol{\theta},\boldsymbol{\gamma})\in\Upsilon} P\left(\sup_{\boldsymbol{t}\in[0,1]^p} |\zeta(\boldsymbol{t})| > M_n \middle| \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \right) \leq Ce^{-d_0\frac{M_n^2}{\Xi_{0,1}}},$$

$$\sup_{(\boldsymbol{\theta},\boldsymbol{\gamma})\in\Upsilon} P\left(\sup_{\boldsymbol{t}\in[0,1]^p} \left|\frac{\partial}{\partial t_i}\zeta(\boldsymbol{t})\right| > M_n | \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \right) \leq Ce^{-d_i\frac{M_n^2}{\Xi_{i,1}}}.$$

□

Proof of Theorem 4

First, we show that $\hat{\sigma}_n^2$ is asymptotically unbiased. Note that

$$\begin{aligned}\mathbb{E}[(y_{i+1}-y_i)^2] &= [\zeta_0(\boldsymbol{t}_{i+1})-\zeta_0(\boldsymbol{t}_i)]^2 + \sigma_0^2\mathbb{E}[(\epsilon_{i+1}-\epsilon_i)^2] \\ &= [\zeta_0(\boldsymbol{t}_{i+1})-\zeta_0(\boldsymbol{t}_i)]^2 + 2\sigma_0^2,\end{aligned}$$

because $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. Consequently

$$\mathbb{E}(\hat{\sigma}_n^2) = \frac{\sum_{i=1}^{n-1}[\zeta_0(\boldsymbol{t}_{i+1})-\zeta_0(\boldsymbol{t}_i)]^2}{2(n-1)} + \sigma_0^2. \tag{4.32}$$

Since ζ_0 is continuously differentiable on the compact and convex set $\boldsymbol{\Omega}$, it is also (globally) Lipschitz on $\boldsymbol{\Omega}$ (e.g. Schaeffer and Cain (2016), Corollary 3.2.4), and there exists a real constant K so that

$$|\zeta_0(\boldsymbol{t}_{i+1})-\zeta_0(\boldsymbol{t}_i)| \leq K\sum_{j=1}^{p}|t_{i+1,j}-t_{i,j}|.$$

Therefore, due to the design assumption (AD)

$$0 \leq \frac{\sum_{i=1}^{n-1}[\zeta_0(\boldsymbol{t}_{i+1})-\zeta_0(\boldsymbol{t}_i)]^2}{2(n-1)} \leq \frac{K^2p^2}{2}\left[\sup_{i\in\{1,\ldots,n\},j\in\{1,\ldots,p\}}|t_{i+1,j}-t_{i,j}|\right]^2 \underset{\text{n}}{\longrightarrow} 0, \tag{4.33}$$

and the combination of (4.32) with (4.33) implies

$$\mathbb{E}(\hat{\sigma}_n^2) \underset{\text{n}}{\longrightarrow} \sigma_0^2. \tag{4.34}$$

To show the almost sure convergence of $\hat{\sigma}_n^2$, let us now denote $x_i = (y_{i+1}-y_i)^2$ and rewrite the estimator $\hat{\sigma}_n^2$ as a sum of two estimators, each consisting of a sum of independent variables:

$$\hat{\sigma}_n^2 = \frac{\frac{1}{2}\sum_{i=1}^{\frac{n-1}{2}}x_{2i}}{2\left(\frac{n-1}{2}\right)} + \frac{\frac{1}{2}\sum_{j=1}^{\frac{n-1}{2}}x_{2j-1}}{2\left(\frac{n-1}{2}\right)} = \hat{\sigma}_{n,e}^2 + \hat{\sigma}_{n,o}^2.$$

Without loss of generality, we assumed that n is an odd integer. Lastly note that $\mathbb{V}ar(x_i) \leq C < \infty$ uniformly in i. This is because the differences $\zeta_0(\boldsymbol{t}_{i+1})-\zeta_0(\boldsymbol{t}_i)$ are uniformly bounded on the compact set $\boldsymbol{\Omega}$ due to the continuity of ζ_0. Additionally, $y_{i+1}-y_i$ are Gaussian and have bounded moments. We can now apply the Kolmogorov's strong law of large numbers

for independent non-identically distributed random variables (e.g. Shiryaev (1996), Chapter 3),

$$\hat{\sigma}^2_{n,e} \xrightarrow[\text{n}]{} \frac{1}{2}\sigma_0^2 \quad \text{a.s. } P_0,$$
$$\hat{\sigma}^2_{n,0} \xrightarrow[\text{n}]{} \frac{1}{2}\sigma_0^2 \quad \text{a.s. } P_0,$$

and as a result

$$\hat{\sigma}^2_n = \hat{\sigma}^2_{n,e} + \hat{\sigma}^2_{n,o} \xrightarrow[\text{n}]{} \sigma_0^2 \quad \text{a.s. } P_0.$$

□

4.5.3 Supplement for the transverse harmonic wave simulation

This section contains some additional figures comparing the empirical Bayes fit with the fully Bayesian approach under the posterior samples obtained via MH algorithm.

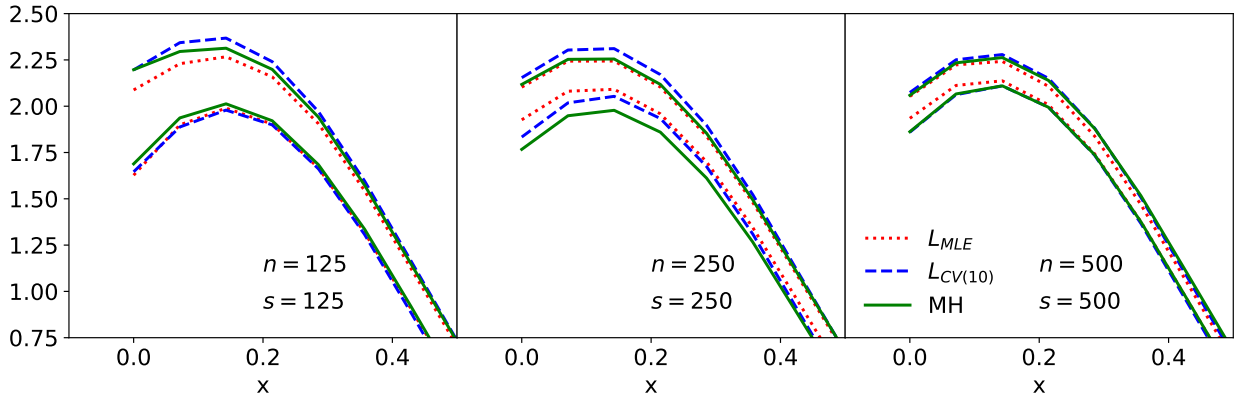

Figure 4.4: Detail of 95% credible bands plotted at $t = 0.00$.

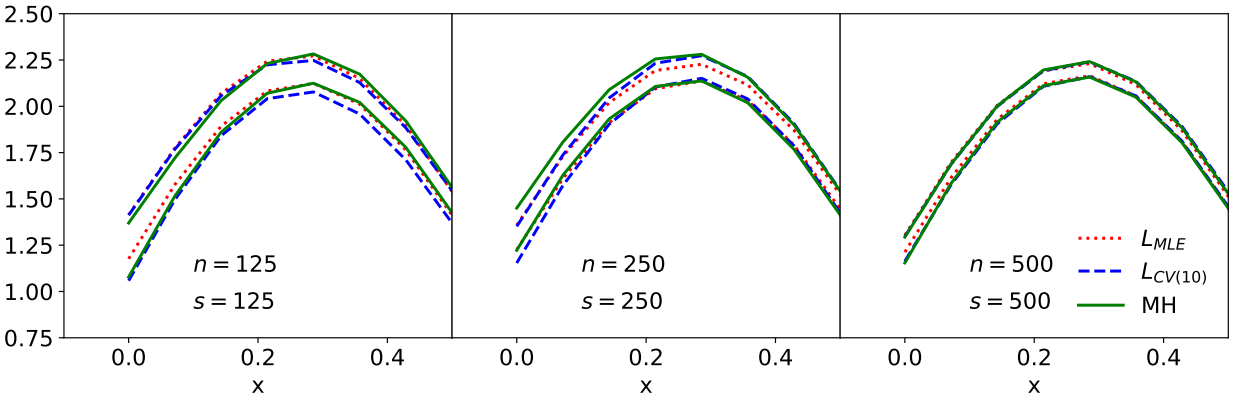

Figure 4.5: Detail of 95% credible bands plotted at $t = 0.43$.

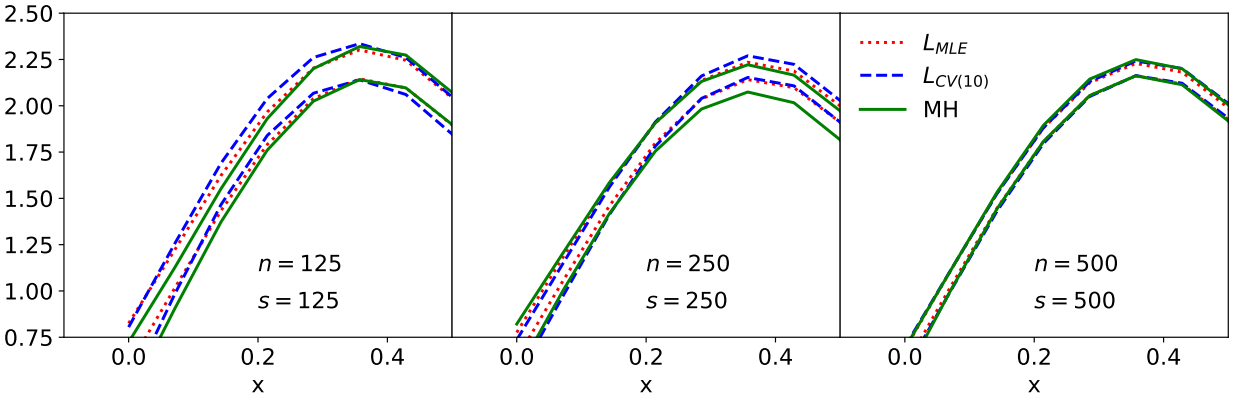

Figure 4.6: Detail of 95% credible bands plotted at $t = 0.71$.

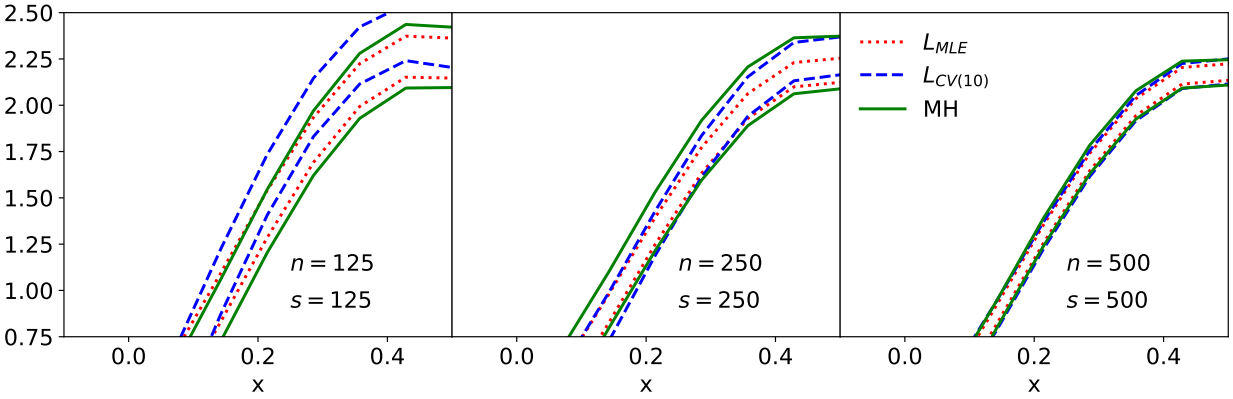

Figure 4.7: Detail of 95% credible bands plotted at $t = 1.00$.

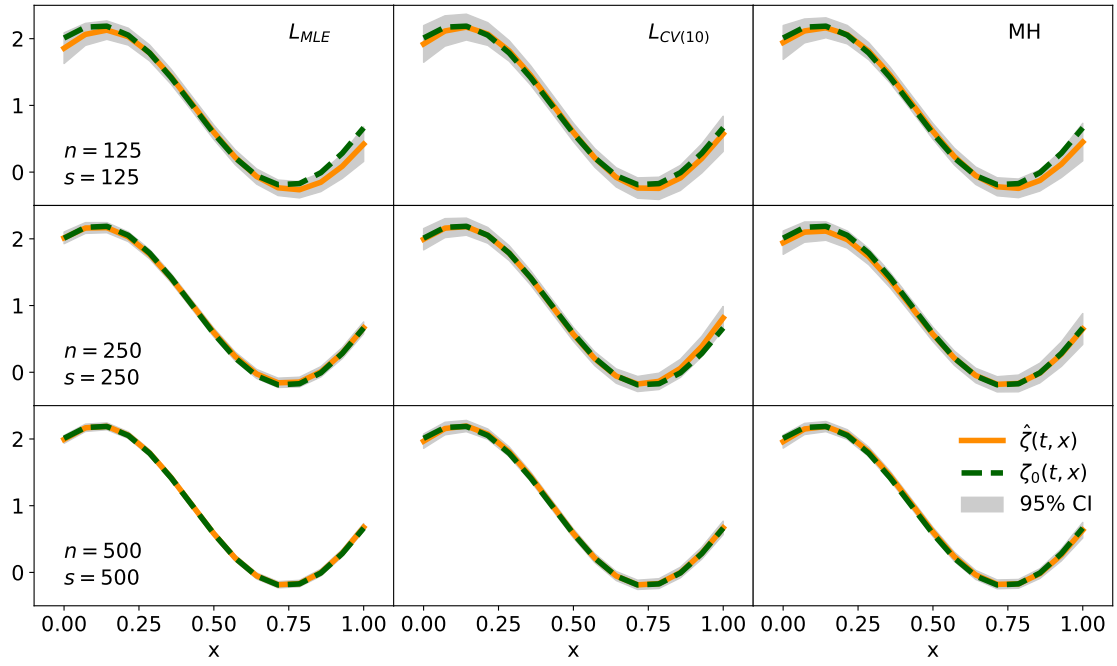

Figure 4.8: Comparison of the convergence to the true physical process $\zeta_0(t, x)$. The curves with 95% credible intervals are plotted at $t = 0.00$.

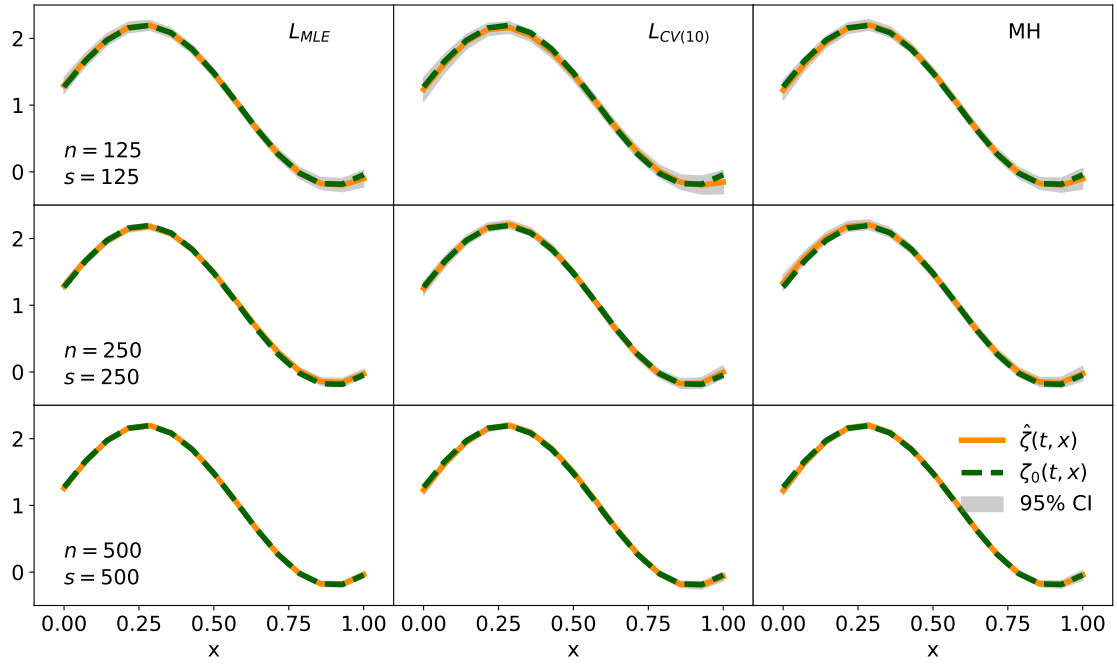

Figure 4.9: Comparison of the convergence to the true physical process. The curves with 95% credible intervals are plotted at $t = 0.43$.

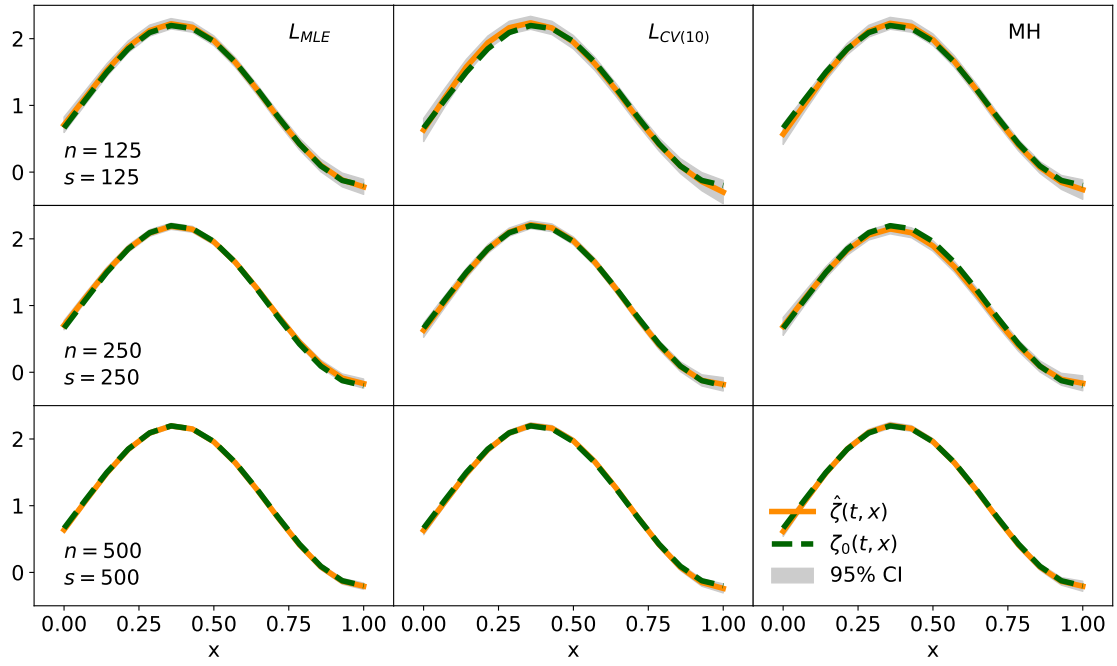

Figure 4.10: Comparison of the convergence to the true physical process. The curves with 95% credible intervals are plotted at $t = 0.71$.

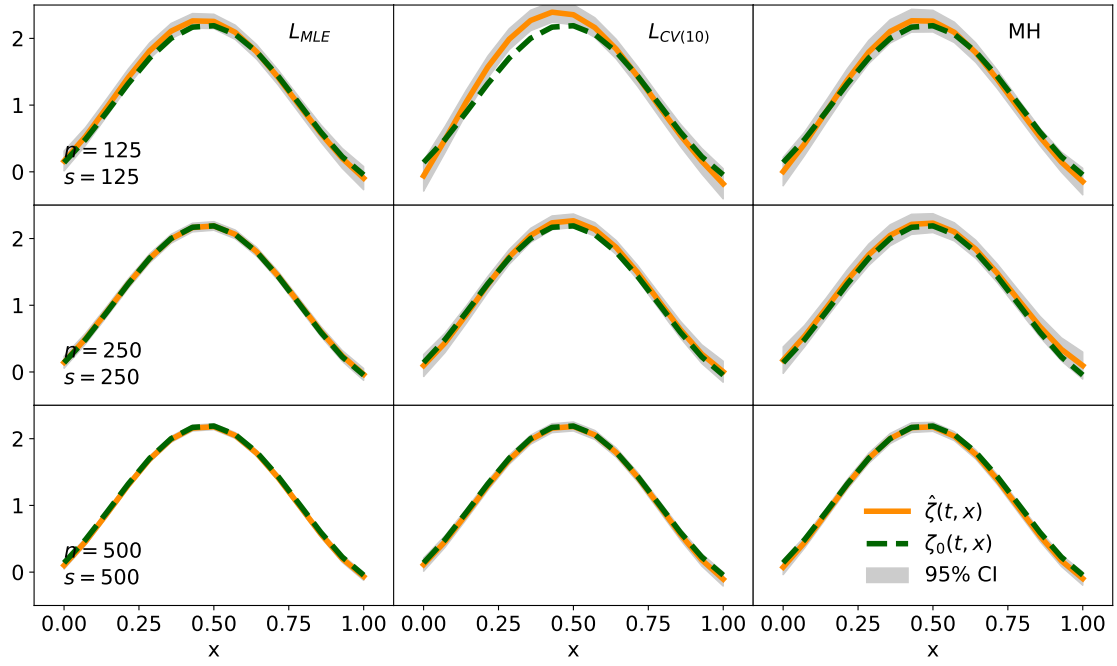

Figure 4.11: Comparison of the convergence to the true physical process. The curves with 95% credible intervals are plotted at $t = 1.00$.

CHAPTER 5

CONCLUSION

We devote the final chapter of this dissertation to the summary of the advances in computational statistics and the developments of new statistical tools of UQ that were made in Chapters 2, 3, and 4. We also provide an overview of the new and exciting avenues this work opens for future research.

In Chapter 2, we studied BMA, the natural Bayesian framework to account for the model uncertainty that arises in situations when multiple competing models are available to describe the same or similar physical process. Motivated by a recurrent scenario in the field of nuclear physics, we extended BMA to the scenario where competing models are defined on non-identical study regions. We gave a theoretical justification for the use of BMA posterior mean predictor in terms of PMSE reduction. While this predictor does not guarantee a universal improvement in predictive ability, on average, it performs at least as well as the best model under consideration. Finally, we applied the methodology outlined in Chapter 2 under several scenarios that lead to better predictions and improved UQ; one simple and transparent exercise of averaging of proton potentials, and a pedagogical example of domain-corrected averaging with a synthetic dataset. We also provided a full-scale BMA analysis of 9 state-of-the-art nuclear mass models and a study of the LDM of nuclear binding energies trained on discrepant domains of the nuclear chart.

In Chapter 3, we developed a novel VBI approach to Bayesian calibration of computationally complex and many-parameter computer models. We exploited the probabilistic theory of approximation coupled with pairwise construction of multivariate copulas to create a computationally efficient and scalable algorithm for calibration. In addition, we proposed the Rao-Blackwellization, control variates, and importance sampling to reduce the variance of noisy gradient estimates involved in the stochastic approximation. The theoretical justification for scalability was also established. In our examples, we first carried out an extensive

simulation study that provided empirical evidence for the accuracy and scalability of our method in scenarios where the traditional MCMC-based approaches become impractical. We established the superiority of variational calibration over the MH algorithm and NUTS in terms of time efficiency and memory requirements. We also demonstrated the opportunities given by our method for practitioners on a real data example through the calibration of the LDM.

In Chapter 4, we proposed an empirical Bayes approach to model-enabled predictions of physical quantities as a fast and easy-to-implement alternative to the fully Bayesian treatment (also discussed in Chapter 3). A new hierarchical model representation of the Bayesian model for calibration of computer models was presented. Theoretical study of the proposed methodology was provided under this new representation. In particular, we established the posterior consistency of the physical process, assuming smoothness of the mean and covariance function of GP priors and existence of a strongly consistent estimator of the noise scale. Consequently, we proposed two plug-in estimators for GP model hyperparameters and a strongly consistent estimator of the noise scale parameter. A simulation study that established the efficiency of the method and empirically verified the consistency was provided. Lastly, we revisited the LDM of binding energies and showed that our method yields comparable results to the fully Bayesian treatment.

5.1 Future research

The extension of BMA to the situation with models defined over non-overlapping input domains addresses only one of many practical challenges in Bayesian model mixing. From methodology perspective, developing a principled approach to average models locally, with model wights depending on input values, would mitigate the tendency of BMA to perform global model selection when one of the models significantly dominates on some (small) part of the input space. Computationally, BMA is a two step procedure, when one needs to first obtain samples from posterior distributions under individual models and consequently sample

from the BMA posterior density. A direct approximation of the BMA posterior, potentially using variational methods, would considerably improve the ease of implementation.

A natural next step to enhance the impact of the VBI approach for calibration of computer models that we proposed in Chapter 3, would be to examine its theoretical properties. For example, one could pursue similar frequentist consistency result as Wang and Blei (2019). If we establish the conditions under which the ELBO $\mathcal{L}(\boldsymbol{\lambda})$ and the l-truncated ELBO $\mathcal{L}_{D_l}(\boldsymbol{\lambda})$ (respectively $\mathcal{L}_{C_l}(\boldsymbol{\lambda})$) are equivalent in limit, namely $\mathcal{L}_{D_l}(\boldsymbol{\lambda}) = \mathcal{L}(\boldsymbol{\lambda}) + o_p(1)$, the asymptotic properties of Wang and Blei (2019) will directly extend to our methodology. Besides theoretical investigations, a procedure that avoids the current sequential approach to select the truncation level would be beneficial. For instance, using fit indices for finding sufficient truncation appears to be a promising approach as discussed by Brechmann and Joe (2015).

When it comes to the empirical Bayes approach to model-enabled prediction, we have already noted the need for further investigation of specific mean and covariance functions of GP priors that satisfy the smoothness conditions for posterior consistency. Most importantly, the hierarchical model representation of Kennedy and O'Hagan (2001) framework together with the theoretical developments in Section 4.2 constitute a solid foundation to establish the posterior consistency of the physical process ζ in the fully Bayesian regime; that is, in a scenario with suitable prior distributions over the hyperparameters of Gaussian process priors and the calibration parameters.

BIBLIOGRAPHY

BIBLIOGRAPHY

Amewou-Atisso, M., Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. (2003). Posterior consistency for semi-parametric regression problems. *Bernoulli*, 9(2):291–312.

Anni, R., Co', G., and Pellegrino, P. (1995). Nuclear charge density distributions from elastic electron scattering data. *Nuclear Physics A*, 584:35–59.

Audi, G., Wapstra, A., and Thibault, C. (2003). The AME2003 atomic mass evaluation: (ii). tables, graphs and references. *Nuclear Physics A*, 729:337–676.

Balasubramanian, J. B., Visweswaran, S., Cooper, G. F., and Gopalakrishnan, V. (2014). Selective model averaging with Bayesian rule learning for predictive biomedicine. *AMIA Joint Summits on Translational Science proceedings AMIA Summit on Translational Science*, 2014:17–22.

Bartel, J., Quentin, P., Brack, M., Guet, C., and Håkansson, H.-B. (1982). Towards a better parametrisation of Skyrme-like effective forces: a critical study of the SkM force. *Nuclear Physics A*, 386(1):79–100.

Bauer, M., van der Wilk, M., and Rasmussen, C. E. (2016). Understanding probabilistic sparse gaussian process approximations. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NeurIPS'16, pages 1533–1541.

Bayarri, M. J., Berger, J. O., Paulo, R., Sacks, J., Cafeo, J. A., Cavendish, J., Lin, C.-H., and Tu, J. (2007). A framework for validation of computer models. *Technometrics*, 49:138–154.

Bedford, T. and Cooke, R. M. (2002). Vines–a new graphical model for dependent random variables. *Annals of Statistics*, 30(4):1031–1068.

Benzaid, D., Bentridi, S., Kerraci, A., and Amrani, N. (2020). Bethe–Weizsäcker semiempirical mass formula coefficients 2019 update based on AME2016. *Nuclear Science and Techniques*, 31:9.

Bernardo, J. M. and Smith, A. F. M. (1994). *Reference analysis*, chapter Inference. Wiley.

Bertsch, G. F. and Bingham, D. (2017). Estimating parameter uncertainty in binding-energy models by the frequency-domain bootstrap. *Physical Review Letters*, 119:252501.

Bertsch, G. F., Sabbey, B., and Uusnäkki, M. (2005). Fitting theories of nuclear binding energies. *Physical Review C*, 71:054311.

Bethe, H. A. and Bacher, R. F. (1936). Nuclear physics a. stationary states of nuclei. *Reviews of Modern Physics*, 8:82–229.

Bottou, L., Le Cun, Y., and Bengio, Y. (1997). Global training of document processing systems using graph transformer networks. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, pages 489–493. IEEE.

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799.

Brechmann, E. C., Czado, C., and Aas, K. (2012). Truncated regular vines in high dimensions with application to financial data. *The Canadian Journal of Statistics*, 40(1):68–85.

Brechmann, E. C. and Joe, H. (2015). Truncation of vine copulas using fit indices. *Journal of Multivariate Analysis*, 138:19–33.

Brynjarsdóttir, J. and O'Hagan, A. (2014). Learning about physical parameters: the importance of model discrepancy. *Inverse Problems*, 30:114007.

Casella, G. and Berger, R. (2002). *Statistical Inference*. Duxbury advanced series in statistics and decision sciences. Thomson Learning, second edition.

Casella, G. and Robert, C. P. (1996). Rao-blackwellisation of sampling schemes. *Biometrika*, 83(1):81–94.

Cauchois, B., Lü, H., Boilley, D., and Royer, G. (2018). Uncertainty analysis of the nuclear liquid drop model. *Physical Review C*, 98:024305.

Chabanat, E., Bonche, P., Haensel, P., Meyer, J., and Schaeffer, R. (1995). New Skyrme effective forces for supernovae and neutron rich nuclei. *Physica Scripta*, 1995(T56):231.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49:327–335.

Choi, T. (2007). Alternative posterior consistency results in nonparametric binary regression using gaussian process priors. *Journal of Statistical Planning and Inference*, 137(9):2975 – 2983.

Choi, T. and Schervish, M. J. (2007a). On posterior consistency in nonparametric regression problems. *Journal of Multivariate Analysis*, 98(10):1969 – 1987.

Choi, T. and Schervish, M. J. (2007b). Posterior Consistency in Nonparametric Regression Problems under Gaussian Process Priors.

Clyde, M. A., Ghosh, J., and Littman, M. L. (2011). Bayesian adaptive sampling for variable selection and model averaging. *Journal of Computational and Graphical Statistics*, 20:80–101.

Cooke, R. and Kurowicka, D. (2006). *Uncertainty Analysis With High Dimensional Dependence Modelling*. Wiley.

Dissmann, J., Brechmann, E., Czado, C., and Kurowicka, D. (2013). Selecting and estimating regular vine copulae and application to financial returns. *Computational Statistics & Data Analysis*, 59:52–69.

Dobaczewski, J., Flocard, H., and Treiner, J. (1984). Hartree-Fock-Bogolyubov description of nuclei near the neutron-drip line. *Nuclear Physics A*, 422(1):103–139.

Dobaczewski, J., Nazarewicz, W., and Reinhard, P.-G. (2014). Error estimates of theoretical models: a guide. *Journal of Physics G: Nuclear and Particle Physics*, 41(7):074001.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159.

Fayans, S. A. (1998). Towards a universal nuclear density functional. *Journal of Experimental and Theoretical Physics Letters*, 68(3):169–174.

Feroz, F., Hobson, M. P., and Bridges, M. (2009). Multinest: an efficient and robust bayesian inference tool for cosmology and particle physics. *Monthly Notices of the Royal Astronomical Society*, 398:1601–1614.

Geweke, J. (1999). Using simulation methods for Bayesian econometric models: inference, development, and communication. *Econometric Reviews*, 18:1–73.

Ghosal, S. and Roy, A. (2006). Posterior consistency of gaussian process prior for nonparametric binary regression. *Annals of Statistics*, 34(5):2413–2429.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *The Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 69:243–268.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378.

Goldberger, A. (1966). *Econometric theory*. Wiley publications in statistics. J. Wiley.

Gu, M. and Wang, L. (2018). Scaled Gaussian stochastic process for computer model calibration and prediction. *SIAM/ASA Journal on Uncertainty Quantification*, 6(4):1555–1583.

Hernández, B., Raftery, A. E., Pennington, S. R., and Parnell, A. C. (2018). Bayesian additive regression trees using Bayesian model averaging. *Statistics and Computing*, 28:869–890.

Higdon, D., Gattiker, J., Williams, B., and Rightley, M. (2008). Computer model calibration using high-dimensional output. *Journal of the American Statistical Association*, 103:570–583.

Higdon, D., Kennedy, M., Cavendish, J. C., Cafeo, J. A., and Ryne, R. D. (2005). Combining field data and computer simulations for calibration and prediction. *SIAM Journal on Scientific Computing*, 26:448–466.

Higdon, D., McDonnell, J. D., Schunck, N., Sarich, J., and Wild, S. M. (2015). A Bayesian approach for parameter estimation and prediction using a computationally intensive model. *Journal of Physics G: Nuclear and Particle Physics*, 42(3):034009.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, 14:382–401.

Hoffman, M. and Blei, D. (2015). Stochastic structured variational inference. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38, pages 361–369, San Diego, CA. PMLR.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347.

Homan, M. D. and Gelman, A. (2014). The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15:1351–1381.

Hooten, M. B. and Hobbs, N. T. (2015). A guide to Bayesian model selection for ecologists. *Ecological Monographs*, 85:3–28.

Horowitz, C. J., Arcones, A., Côté, B., Dillmann, I., Nazarewicz, W., et al. (2019). r-process nucleosynthesis: connecting rare-isotope beam facilities with the cosmos. *Journal of Physics G: Nuclear and Particle Physics*, 46(8):083001.

Ireland, D. G. and Nazarewicz, W. (2015). Enhancing the interaction between nuclear experiment and theory through information and statistics. *Journal of Physics G: Nuclear and Particle Physics*, 42(3):030301.

Jaganathen, Y., Betan, R. M. I., Michel, N., Nazarewicz, W., and Płoszajczak, M. (2017). Quantified gamow shell model interaction for *psd*-shell nuclei. *Physical Review C*, 96:054316.

Johnston, J. (1976). *Econometric Methods*. McGraw-Hill.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90:773–795.

Kejzlar, V. and Maiti, T. (2020). Variational inference with vine copulas: An efficient approach for bayesian computer model calibration. *arxiv.org/2003.12890*.

Kejzlar, V., Neufcourt, L., Maiti, T., and Viens, F. (2019). Bayesian averaging of computer models with domain discrepancies: a nuclear physics perspective. *arxiv.org/1904.04793*.

Kejzlar, V., Neufcourt, L., Nazarewicz, W., and Reinhard, P.-G. (2020). Statistical aspects of nuclear mass models. *Journal of Physics G: Nuclear and Particle Physics. URL* `https://doi.org/10.1088/1361-6471/ab907c`.

Kennedy, M. C. and O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63:425–464.

King, G. B., Lovell, A. E., Neufcourt, L., and Nunes, F. M. (2019). Direct comparison between Bayesian and frequentist uncertainty quantification for nuclear reactions. *Physical Review Letters*, 122:232502.

Kirson, M. W. (2008). Mutual influence of terms in a semi-empirical mass formula. *Nuclear Physics A*, 798(1):29 – 60.

Klupfel, P., Reinhard, P. G., Burvenich, T. J., and Maruhn, J. A. (2009). Variations on a theme by Skyrme: A systematic study of adjustments of model parameters. *Physical Review C*, 79:034310.

Klüpfel, P., Reinhard, P.-G., Bürvenich, T. J., and Maruhn, J. A. (2009). Variations on a theme by Skyrme: A systematic study of adjustments of model parameters. *Physical Review C*, 79(3):034310.

Kortelainen, M., Lesinski, T., Moré, J. J., Nazarewicz, W., Sarich, J., Schunck, N., Stoitsov, M. V., and Wild, S. M. (2010a). Nuclear energy density optimization. *Physical Review C*, 82(2):024313.

Kortelainen, M., Lesinski, T., Moré, J., Nazarewicz, W., Sarich, J., Schunck, N., Stoitsov, M. V., and Wild, S. (2010b). Nuclear energy density optimization. *Physical Review C*, 82:024313.

Kortelainen, M., McDonnell, J., Nazarewicz, W., Olsen, E., Reinhard, P.-G., Sarich, J., Schunck, N., Wild, S. M., Davesne, D., Erler, J., and Pastore, A. (2014). Nuclear energy density optimization: Shell structure. *Physical Review C*, 89:054314.

Kortelainen, M., McDonnell, J., Nazarewicz, W., Reinhard, P.-G., Sarich, J., Schunck, N., Stoitsov, M. V., and Wild, S. M. (2012). Nuclear energy density optimization: large deformations. *Physical Review C*, 85:024304.

Krane, K. (1987). *Introductory Nuclear Physics*. Wiley.

Lawrence, E., Heitmann, K., White, M., Higdon, D., Wagner, C., Habib, S., and Williams, B. (2010). The Coyote Universe III: simulation suite and precision emulator for the nonlinear matter power spectrum. *The Astrophysical Journal*, 713(2):1322–1331.

Lynch, P. (2008). The origins of computer weather prediction and climate modeling. *Journal of Computational Physics*, 227(7):3431 – 3444. Predicting weather, climate and extreme events.

Madigan, D., Gavrin, J., and Raftery, A. E. (1995). Eliciting prior information to enhance the predictive performance of bayesian graphical models. *Communications in Statistics - Theory and Methods*, 24(9):2271–2292.

Martino, L., Laparra, V., and Camps-Valls, G. (2017). Probabilistic cross-validation estimators for gaussian process regression. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 823–827.

McDonnell, J. D., Schunck, N., Higdon, D., Sarich, J., Wild, S. M., and Nazarewicz, W. (2015). Uncertainty quantification for nuclear density functional theory and information content of new measurements. *Physical Review Letters*, 114(12):122501.

Morris, M. D. and Mitchell, T. J. (1995). Exploratory designs for computational experiments. *Journal of Statistical Planning and Inference*, 43(3):381 – 402.

Myers, W. D. and Swiatecki, W. J. (1966). Nuclear masses and deformations. *Nuclear Physics*, 81(2):1 – 60.

Nazarewicz, W. (2016). Challenges in nuclear structure theory. *Journal of Physics G: Nuclear and Particle Physics*, 43:044002.

Nazarewicz, W. (2018). The limits of nuclear mass and charge. *Nature Physics*, 14(6):537–541.

Neiswanger, W., Wang, C., and Xing, E. P. (2014). Asymptotically exact, embarrassingly parallel mcmc. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, UAI'14, pages 623–632, Arlington, VA. AUAI Press.

Neufcourt, L., Cao, Y., Giuliani, S., Nazarewicz, W., Olsen, E., and Tarasov, O. B. (2020a). Beyond the proton drip line: Bayesian analysis of proton-emitting nuclei. *Physical Review C*, 101:014319.

Neufcourt, L., Cao, Y., Giuliani, S. A., Nazarewicz, W., Olsen, E., and Tarasov, O. B. (2020b). Quantified limits of the nuclear landscape. *Physical Review C*, 101:044307.

Neufcourt, L., Cao, Y., Nazarewicz, W., Olsen, E., and Viens, F. (2019). Neutron drip line in the Ca region from Bayesian Model Averaging. *Physical Review Letters*, 122:062502.

Park, I. and Grandhi, R. V. (2012). Quantification of model-form and parametric uncertainty using evidence theory. *Structural Safety*, 39:44—-51.

Parkinson, D. and Liddle, A. R. (2013). Bayesian model averaging in astrophysics: A review. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 6(1):3–14.

Pastore, A. (2019). An introduction to bootstrap for nuclear physics. *Journal of Physics G: Nuclear and Particle Physics*, 46(5):052001.

Peterson, C. and Anderson, J. R. (1987). A mean field theory learning algorithm for neural networks. *Complex Systems*, 1:995–1019.

Plumlee, M. (2017). Bayesian calibration of inexact computer models. *Journal of the American Statistical Association*, 112:1274–1285.

Plumlee, M. (2019). Computer model calibration with confidence and consistency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(3):519–545.

Plumlee, M., Joseph, V. R., and Yang, H. (2016). Calibrating functional parameters in the ion channel models of cardiac cells. *Journal of the American Statistical Association*, 111:500–509.

Pollard, D., Chang, W., Haran, M., Applegate, P., and DeConto, R. (2016). Large ensemble modeling of the last deglacial retreat of the West Antarctic Ice Sheet: comparison of simple and advanced statistical techniques. *Geoscientific Model Development*, 9(5):1697–1723.

Quiñonero-Candela, J. and Rasmussen, C. E. (2005). A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, pages 1939–1959.

Radaideh, M. I., Borowiec, K., and Kozlowski, T. (2019). Integrated framework for model assessment and advanced uncertainty quantification of nuclear computer codes under Bayesian statistics. *Reliability Engineering & System Safety*, 189:357–377.

Ranganath, R., Gerrish, S., and Blei, D. (2014). Black box variational inference. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 814–822. PMLR.

Ranganath, R., Tran, D., and Blei, D. M. (2016). Hierarchical variational models. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning – Volume 48*, ICML'16, pages 2568–2577. JMLR.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press.

Reinhard, P.-G., Bender, M., Nazarewicz, W., and Vertse, T. (2006). From finite nuclei to the nuclear liquid drop: Leptodermous expansion based on self-consistent mean-field theory. *Physical Review C*, 73:014309.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics*, 22(3):400–407.

Robert, C. and Casella, G. (2005). *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer New York.

Ross, S. M. (2006). *Simulation*. Academic Press, Inc., Orlando, FL, fourth edition.

Ruiz, F. J. R., Titsias, M. K., and Blei, D. M. (2016). Overdispersed black-box variational inference. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'16, page 647–656, Arlington, Virginia, USA. AUAI Press.

Schaeffer, D. G. and Cain, J. W. (2016). *Nonlinear Systems: Local Theory*, pages 79–109. Springer New York, New York, NY.

Schorning, K., Bornkamp, B., Bretz, F., and Dette, H. (2016). Model selection versus model averaging in dose finding studies. *Statistics in Medicine*, 35:4021–4040.

Sexton, D. M. H., Murphy, J. M., Collins, M., and Webb, M. J. (2012). Multivariate probabilistic projections using imperfect climate models Part i: outline of methodology. *Climate Dynamics*, 38(11):2513–2542.

Shelley, M., Becker, P., Gration, A., and Pastore, A. (2014). Advanced statistical methods to fit nuclear models. *Acta Physica Polonica B*, 12:649.

Shiryaev, A. N. (1996). *Convergence of Probability Measures. Central Limit Theorem*, pages 308–378. Springer New York, New York, NY.

Silvestro, D., Schnitzler, J., Liow, L. H., Antonelli, A., and Salamin, N. (2014). Bayesian estimation of speciation and extinction from incomplete fossil occurrence data. *Systematic Biology*, 63:349–367.

Skilling, J. (2006). Nested sampling for general bayesian computation. *Bayesian Analysis*, 1(4):833–860.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8:229–231.

Smith, M. S., Loaiza-Maya, R., and Nott, D. J. (2020). High-dimensional copula variational approximation through transformation. *Journal of Computational and Graphical Statistics*, pages 1–35.

Sundararajan, S. and Keerthi, S. S. (2001). Predictive approaches for choosing hyperparameters in gaussian processes. *Neural Computation*, 13(5):1103–1118.

Tieleman, T. and Hinton, G. (2012). Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning.

Titsias, M. (2009). Variational learning of inducing variables in sparse Gaussian processes. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, volume 5, pages 567–574. PMLR.

Toivanen, J., Dobaczewski, J., Kortelainen, M., and Mizuyama, K. (2008). Error analysis of nuclear mass fits. *Physical Review C*, 78:034306.

Tokdar, S. T. and Ghosh, J. K. (2007). Posterior consistency of logistic gaussian process priors in density estimation. *Journal of Statistical Planning and Inference*, 137(1):34 – 42.

Tran, D., Blei, D. M., and Airoldi, E. M. (2015). Copula variational inference. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NeurIPS'15, pages 3564–3572, Cambridge, MA. MIT Press.

Tran, D., Ranganath, R., and Blei, D. M. (2017). Hierarchical implicit models and likelihood-free variational inference. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NeurIPS'17, pages 5529–5539.

Trotta, R. (2008). Bayes in the sky: Bayesian inference and model selection in cosmology. *Contemporary Physics*, 49:71–104.

Tuo, R. and Wu, C. F. J. (2015). Efficient calibration for imperfect computer models. *Annals of Statistics*, 43:2331–2352.

Tuo, R. and Wu, C. F. J. (2016). A theoretical framework for calibration in computer models: parametrization, estimation and convergence properties. *SIAM/ASA Journal on Uncertainty Quantification*, 4:767–795.

Utama, R., Piekarewicz, J., and Prosper, H. B. (2016). Nuclear mass predictions for the crustal composition of neutron stars: A Bayesian neural network approach. *Physical Review C*, 93:014311.

van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer.

Wahba, G. (1990). *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305.

Wang, M., Audi, G., Kondev, F. G., Huang, W. J., Naimi, S., and Xu, X. (2017). The AME2016 atomic mass evaluation (II). tables, graphs and references. *Chinese Physics C*, 41:030003.

Wang, Y. and Blei, D. M. (2019). Frequentist consistency of variational bayes. *Journal of the American Statistical Association*, 114(527):1147–1161.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Physics*, 44:92 – 107.

Wei, W., Visweswaran, S., and Cooper, G. F. (2011). The application of naive Bayes model averaging to predict Alzheimer's disease from genome-wide data. *Journal of the American Medical Informatics Association*, 18:370–375.

Weizsäcker, C. F. v. (1935). Zur theorie der kernmassen. *Z. Phys.*, 96(7):431–458.

Wen, X. (2015). Bayesian model comparison in genetic association analysis: linear mixed modeling and SNP set testing. *Biostatistics*, 16:701–712.

Williams, B., Higdon, D., Gattiker, J., Moore, L., McKay, M., and Keller-McNulty, S. (2006). Combining experimental data and computer simulations, with an application to flyer plate experiments. *Bayesian Analysis*, 1(4):765–792.

Xie, F. and Xu, Y. (2020). Bayesian projected calibration of computer models. *Journal of the American Statistical Association*, pages 1–47.

Yuan, C. (2016). Uncertainty decomposition method and its application to the liquid drop model. *Physical Review C*, 93:034310.

Zeiler, M. D. (2012). Adadelta: An adaptive learning rate method. *ArXiv*, 1212.5701.

Zhang, H. F., Wang, L. H., Yin, J. P., Chen, P. H., and Zhang, H. F. (2017). Performance of the Levenberg-Marquardt neural network approach in nuclear mass prediction. *Journal of Physics G: Nuclear and Particle Physics*, 44(4):045110.

Zhang, L., Jiang, Z., Choi, J., Lim, C.-Y., Maiti, T., and Baek, S. (2019). Patient-specific prediction of abdominal aortic aneurysm expansion using Bayesian calibration. *IEE Journal of Biomedical and Health Informatics*.