Gaussian Process for Ground-Motion Prediction and Emulation of Systems of Computer Models

Deyu Ming

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Statistical Science

University College London

October 26, 2020

I, Deyu Ming, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

to *Grandpa*

Abstract

In this thesis, several challenges in both ground-motion modelling and the surrogate modelling, are addressed by developing methods based on Gaussian processes (GP). The first chapter contains an overview of the GP and summarises the key findings of the rest of the thesis.

In the second chapter, an estimation algorithm, called the Scoring estimation approach, is developed to train GP-based ground-motion models with spatial correlation. The Scoring estimation approach is introduced theoretically and numerically, and it is proven to have desirable properties on convergence and computation. It is a statistically robust method, producing consistent and statistically efficient estimators of spatial correlation parameters. The predictive performance of the estimated ground-motion model is assessed by a simulation-based application, which gives important implications on the seismic risk assessment.

In the third chapter, a GP-based surrogate model, called the integrated emulator, is introduced to emulate a system of multiple computer models. It generalises the state-of-the-art linked emulator for a system of two computer models and considers a variety of kernels (exponential, squared exponential, and two key Matérn kernels) that are essential in advanced applications. By learning the system structure, the integrated emulator outperforms the composite emulator, which emulates the entire system using only global inputs and outputs. Furthermore, its analytic expressions allow a fast and efficient design algorithm that could yield significant computational and predictive gains by allocating

different runs to individual computer models based on their heterogeneous functional complexity. The benefits of the integrated emulator are demonstrated in a series of synthetic experiments and a feed-back coupled fire-detection satellite model.

Finally, the developed method underlying the integrated emulator is used to construct a non-stationary Gaussian process model based on deep Gaussian hierarchy.

Impact Statement

The method presented in Chapter 2 of this thesis provides a statistically robust way to construct ground-motion models with spatial correlation, which have the potential to generate shake intensity maps with higher accuracy and better uncertainty measurements. The method would provide researchers with better understanding about the uncertainties of ground-motion intensities. In addition, it would also benefit the government and (re)insurance companies on assessing their exposures to seismic risks accurately as high-quality ground-motion models play key roles in catastrophe models.

The integrated emulator introduced in Chapter 3 of this thesis opens the door for the uncertainty quantification of many complex systems of computer models, which are computationally expensive to run and thus prohibitive to implement further analysis. The fast formulae and adaptive design of the integrated emulator allow natural scientists, biologists, meteorologists to build efficient surrogate models for their sophisticated systems of simulators, and therefore any subsequent inferences such as sensitivity analysis, uncertainty propagation and model calibration become feasible.

The non-stationary Gaussian process (GP) model proposed in Chapter 4 of this thesis addresses the non-stationarity and heteroscedasticity inherent in the datasets that cannot be handled by the conventional stationary GP model. The flexibility of the non-stationary GP model is powered by the state-of-the-art deep learning technique. In addition, our non-stationary model could greatly expand the capacity of the integrated emulator because many real-world data and scientific simulators are by nature non-stationary.

Acknowledgements

I would like to convey my sincere gratitude to my supervisor Prof. Serge Guillas for his countless support and advice. I am grateful for his encouragement when I was at the low point of my PhD study. I appreciate him giving me lots of opportunities to present my work and develop my academic career.

I would also like to thank Prof. Gareth W. Peters and Dr. Carmine Galasso for their help in the first two years of my PhD study. I thank Chen Huang for her help during our collaboration. My thanks also go to Prof. Mario Wüthrich for offering me opportunities to present my research work at ETH Zurich. Many thanks to Ayao Ehara, Devaraj Gopinathan and Dimitra Salmanidou for their help when I struggled on my research.

Finally, I would like to thank my wife Qiaolu, my parents, my grandparents and my families for their undying support when I am aboard. Without them, I would have never been able to come this far.

Contents

Chapter 1

Introduction

1.1 The Basics of Gaussian Process

Gaussian process (GP) has gained its popularity in recent decades due to its successful applications in the machine learning community, e.g., Williams and Rasmussen (2006); Damianou and Lawrence (2013); Cutajar et al. (2019). However, Gaussian process is itself not a new concept and has a long history in statistics. At its early phase, Gaussian process is actively used for spatial analysis, e.g., Mardia and Marshall (1984). It is then utilised for computer experiments, e.g., Santner et al. (2003) and more recently in the area of uncertainty quantification, e.g., Bilionis and Zabaras (2016).

Definition 1.1 (Gaussian process) *A real-valued stochastic process* $(Y_i)_{i\in\mathbb{N}}$ *is called a Gaussian process if the random vector* $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ *for* $n \in \mathbb{N}$ *follows the multivariate Gaussian distribution with mean* $\boldsymbol{\mu} \in \mathbb{R}^n$ *and covariance matrix* $\boldsymbol{\Sigma} \in \mathbb{R}^{n\times n}$*, denoted by*

$$\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{1.1}$$

where the i*-th element of* $\boldsymbol{\mu}$ *is given by* $\mu_i = \mathbb{E}(Y_i)$ *and the* ij*-th element of* $\boldsymbol{\Sigma}$ *is given by* $\Sigma_{ij} = \mathrm{cov}(Y_i, Y_j)$.

In this thesis, the Gaussian process is mainly used for regression (i.e., supervised learning) where for each Y_i there is a corresponding input vector $\mathbf{x}_i$ that

represents its covariates (or features). In such an case, the elements of mean and covariance matrix of the Gaussian process can be specified as follow:

$$\begin{aligned}\mu_i =& m(\mathbf{x}_i)\\ \Sigma_{ij} =& \sigma^2 k(\mathbf{x}_i, \mathbf{x}_j),\end{aligned}$$

where $m(\cdot)$ is the mean function, and $k(\cdot,\cdot)$ is the kernel function that is symmetric and positive semi-definite. When $k(\mathbf{x}_i, \mathbf{x}_j) = k(\|\mathbf{x}_i - \mathbf{x}_j\|_2)$, it is called stationary and isotropic. Examples of this class of kernel function include

- **Exponential:**
$$k(\cdot,\cdot) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2}{h}\right),$$
with a positive range parameter (or length-scale) h, which indicates the value of $\|\mathbf{x}_i - \mathbf{x}_j\|_2$ at which the correlation is around 0.37, i.e., when $\|\mathbf{x}_i - \mathbf{x}_j\|_2 = h$ the correlation ρ_{ij} is given by
$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{h}{h}\right) = \exp(-1) \approx 0.37.$$

- **Squared Exponential:**
$$k(\cdot,\cdot) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2h^2}\right).$$
This type of correlation function is sometime called Gaussian ;

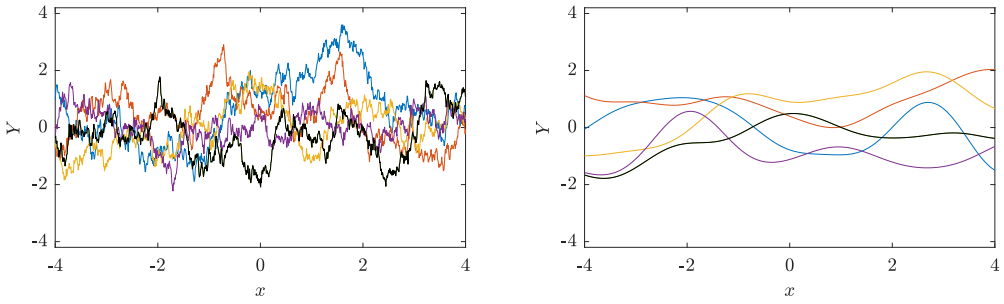
- **Periodic:**
$$k(\cdot,\cdot) = \exp\left(-\frac{2\sin^2(\pi\|\mathbf{x}_i - \mathbf{x}_j\|_2/p)}{h^2}\right),$$
where p determines the distance between repetitions of the function;

- **Matérn:**
$$k(\cdot,\cdot) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}\|\mathbf{x}_i - \mathbf{x}_j\|_2}{h}\right)^{\nu} K_\nu\left(\frac{\sqrt{2\nu}\|\mathbf{x}_i - \mathbf{x}_j\|_2}{h}\right)$$
with positive parameters ν and d , where $\Gamma(\cdot)$ is the gamma function and $K_\nu(\cdot)$ is the modified Bessel function of the second kind. The Matérn kernel can be simplified to exponential and square exponential kernels by setting $\nu = 1/2$ and $\nu \to \infty$ respectively.

Anisotropic kernels can be constructed from stationary and isotropic ones by replacing the Euclidean distance $\|\mathbf{x}_i - \mathbf{x}_j\|_2$ by the Mahalanobis distance:

$$\sqrt{(\mathbf{x}_i - \mathbf{x}_j)^\top \boldsymbol{\Gamma}^{-1} (\mathbf{x}_i - \mathbf{x}_j)}\,,$$

where $\boldsymbol{\Gamma}$ is an arbitrary positive definite matrix and when $\boldsymbol{\Gamma} = \mathbf{I}$ the Mahalanobis distance becomes the Euclidean distance. Examples of other types of kernel functions are illustrated in Williams and Rasmussen (2006). The Gaussian process for regression tries to represent the relations between Y_i and $\mathbf{x}_i$ via the chosen kernel function. For example, Figure 1.1 presents some random paths generated from a zero-mean Gaussian process with one dimensional feature (i.e., $\mathbf{x}_i \in \mathbb{R}$) and covariance matrix specified by the exponential and squared exponential kernel functions.



(a) Exponential **(b)** Squared Exponential

Figure 1.1: Random sample paths between Y and $x \in [-4, 4]$ generated from a zero-mean Gaussian process with one dimensional input feature and covariance matrix specified by the exponential and squared exponential kernel functions where $h = 1$ and $\sigma^2 = 1$.

Given a set of N observations $\{\mathbf{x}^{\mathcal{D}}, \mathbf{y}^{\mathcal{D}}\}$ with $\mathbf{x}^{\mathcal{D}} = (\mathbf{x}_1^{\mathcal{D}}, \ldots, \mathbf{x}_N^{\mathcal{D}})^\top$ and $\mathbf{y}^{\mathcal{D}} = (y_1^{\mathcal{D}}, \ldots, y_N^{\mathcal{D}})^\top$, the Gaussian process predictive distribution of Y^* at an unobserved input position $\mathbf{x}^*$ is then given by

$$Y^* \sim \mathcal{N}\left(m(\mathbf{x}^*) + \mathbf{K}^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y}^{\mathcal{D}} - \boldsymbol{\mu}),\, \sigma^2 - \mathbf{K}^\top \boldsymbol{\Sigma}^{-1} \mathbf{K}\right), \tag{1.2}$$

where $\mathbf{K} = \sigma^2 \left(k(\mathbf{x}^*, \mathbf{x}_1^{\mathcal{D}}), \ldots, k(\mathbf{x}^*, \mathbf{x}_N^{\mathcal{D}})\right)^\top$; $\boldsymbol{\mu} = \left(m(\mathbf{x}_1^{\mathcal{D}}), \ldots, m(\mathbf{x}_N^{\mathcal{D}})\right)^\top$; and the ij-th element of $\boldsymbol{\Sigma}$ is given by $\Sigma_{ij} = \sigma^2 k(\mathbf{x}_i^{\mathcal{D}}, \mathbf{x}_j^{\mathcal{D}})$. Note that the predictive distribution (1.2) interpolates the observations $\{\mathbf{x}^{\mathcal{D}}, \mathbf{y}^{\mathcal{D}}\}$ because when $\mathbf{x}^* = \mathbf{x}_i^{\mathcal{D}}$

we have $\mathbf{K}^\top \boldsymbol{\Sigma}^{-1}$ being an unit vector with the i-th element equal to one.

The real-world data often have noise and such feature can be incorporated into the Gaussian process model (1.1) by adding a nugget term $\tau > 0$ to the diagonal elements of $\boldsymbol{\Sigma}$. In this case, the Gaussian process predictive distribution (1.2) becomes to

$$Y^* \sim \mathcal{N}\left(m(\mathbf{x}^*) + \mathbf{K}^\top (\boldsymbol{\Sigma} + \tau \mathbf{I})^{-1} (\mathbf{y}^{\mathcal{D}} - \boldsymbol{\mu}),\, \sigma^2 + \tau - \mathbf{K}^\top (\boldsymbol{\Sigma} + \tau \mathbf{I})^{-1} \mathbf{K}\right). \tag{1.3}$$

However, by introducing the nugget term τ the predictive distribution (1.3) no longer interpolates the observations. Instead, as τ increases the mean and variance of the Gaussian process predictive distribution (1.3) tend to $m(\mathbf{x}^*)$ and $\sigma^2 + \tau$ (Mohammadi et al., 2016). If the nugget τ is set to a sufficiently small value, then it can be treated as a regularisation term that enhances the condition number of $\boldsymbol{\Sigma}$ to prevent from the ill-conditioning problems (Andrianakis and Challenor, 2012). In fact, Gramacy and Lee (2012) suggest to always include the nugget in the Gaussian process models, even the underlying process is deterministic, to retain statistical robustness in cases such as the data are sparse and non-stationary.

We note that this section only gives a general overview of the most basic form of the Gaussian process model for regression. In each chapter of the thesis, we will present and detail different modifications of this base form to address challenges in different contexts.

1.2 Scope of the Thesis

This thesis explores methodological developments in three research fields, namely the ground-motion modelling, the surrogate modelling and non-stationary modelling, on the basis of Gaussian processes.

Ground-motion models, also known as ground-motion prediction equations (GMPEs) and attenuation relationships, are empirical models widely used in probabilistic seismic hazard analysis (PSHA), to predict ground-motion inten-

sity measures (IMs) occurring at sites due to a nearby earthquake of a certain magnitude. Ground-motion models require robust estimation techniques. The accuracy of the estimated ground-motion models is important for assessing earthquake risk and resilience of engineered systems. In the last decade, the increasing interest in assessing earthquake risk and resilience of spatially distributed portfolios of buildings and infrastructures has motivated the modelling of ground-motion spatial correlation. This introduces further challenges for researchers to incorporate spatial correlation into the ground-motion models and develop statistically rigorous and computationally efficient algorithms to perform the estimation of the models. To this aim, in Chapter 2, repeated Gaussian processes are used to represent ground-motion models with spatial correlation and a one-stage estimation algorithm, called the Scoring estimation approach, is introduced to fit the constructed models. By comparing, both theoretically and numerically, to the state-of-the-art estimation algorithm proposed by Jayaram and Baker (2010), we find that the proposed Scoring estimation approach presents comparable or higher accuracy in estimating ground-motion model parameters, especially when the spatial correlation becomes smoother. The approach is also capable of quantifying the uncertainties in spatial correlation. The statistical robustness of the estimation approach further allows us to investigate the impact of spatial correlation on ground-motion predictions.

Gaussian process-based surrogate models (also known as Gaussian process emulators) have been used to emulate systems of computer models in many fields including environmental science, biology and geophysics because of their attractive statistical properties. However, their construction often neglects system structures and thus requires additional computational costs (which may become unaffordable for some expensive systems) to achieve a satisfactory accuracy. To address this issue, in Chapter 3, we generalise the linked emulator (proposed by Kyzyurova et al. (2018)) for a system of two computer models to an integrated emulator for any feed-forward system of multiple computer models. The integrated emulator combines Gaussian process emulators of

individual computer models, and implicitly takes the system structure into account. Comparing to the composite emulator, which is a GP emulator of the entire system built with only global inputs and outputs, the integrated emulator can achieve orders of magnitude prediction improvement for moderate-size designs. Thanks to the analytic expressions, the predictive performance of the integrated emulator can be further enhanced by an adaptive designing strategy that only refines the GP emulators with insufficient accuracy. The skills of the integrated emulator are shown in several synthetic experiments and a multi-disciplinary satellite model.

Conventional Gaussian process models assume stationarity, which is often insufficient in real-world data. In Chapter 4, we introduce a new type of non-stationary Gaussian process model which utilises the state-of-the-art deep learning technique and thus is demonstrated to be flexible to learn the non-stationarity and heteroscedasticity embed in the data in an automatic manner.

In Chapter 5, key findings of the thesis are summarised. Some future research directions and associated challenges are discussed. We note that the work presented in Chapter 2, 3 and 4 are self-contained and the results in Chapter 2 have been published in Ming et al. (2019).

Chapter 2

Ground-Motion Modelling with Spatial Correlation

2.1 Background to Earthquakes

In this chapter a few key concepts in seismology are frequently referred and thus are briefly described in this section.

2.1.1 Intensity measures

Intensity measures (IM) are engineering characteristics of earthquake ground-motion records that are used to estimate structural damages and loss. The IM are simplified representations of ground-motion time histories and are key component in ground-motion models. Some examples of IM are given below:

- **Peak ground acceleration** (PGA): PGA (in g or cm/s^2) is the maximum absolute value of ground-motion acceleration time history. The ground-motion acceleration time history is the processed ground-motion records obtained from accelerographs;

- **Peak ground velocity** (PGV): PGV (in cm/s) is the maximum absolute value of ground-motion velocity time history. The ground-motion velocity time history is the integration of the acceleration time history;

- **Peak ground displacement** (PGD): PGD (in cm) is the maximum absolute value of ground-motion displacement time history. The ground-motion displacement time history is the double integration of acceleration time history;

- **Elastic response spectral acceleration** ($S_a(T)$): $S_a(T)$ (in g or cm/s^2) is the maximum absolute value of structural response acceleration time history with 5% critical damping at structural period T (in s).

Among the different IM, PGA and $S_a(T)$ are the most popular measures that are widely used in published ground-motion models (Douglas and Edwards, 2016) and the seismic design in worldwide building codes.

2.1.2 Magnitude

Magnitude represents the size (i.e. the energy released) of an earthquake. The following list outlines four magnitude scales (unitless) that are commonly used by international seismic networks:

- **Local magnitude** (M_L): M_L, also known as "Richter magnitude" (Richter, 1935), is determined as the logarithm of maximum amplitude of the ground shaking:

$$M_L = \log_{10} A - \log_{10} \sigma(R)\,,$$

  where A is the maximum amplitude of ground shaking (in micrometres) and $\sigma(R)$ is an empirical function of epicentre distance (in kilometres), R;

- **Surface-wave magnitude** (M_S): M_S is derived from measuring the magnitude of Rayleigh surface waves, a type of seismic waves that travel primarily along the Earth's surface, and was the standard magnitude scale in China from 1999 till 2017:

$$M_S = \log_{10}\left(\frac{A}{T}\right) + 1.66\log_{10} R + 3.3\,,$$

where A is the maximum amplitude (in micrometres) of the Rayleigh waves, T is the corresponding period (in seconds) and R is the epicentre distance (in degrees);

- **Body-wave magnitude** (m_b): m_b is computed according to the amplitude of the P-wave, a type of seismic waves that travels through the interior of the Earth and reaches seismograph stations first. The body-wave magnitude formula is defined by

$$m_b = \log_{10}\left(\frac{A}{T}\right) + \sigma(R,\, h)\,,$$

where A is the maximum amplitude (in micrometres) of the P-waves, T is the corresponding period (in seconds) and $\sigma(R,\, h)$ is a function of epicentre distance (in degrees), R and focal depth (in kilometres), h;

- **Moment magnitude** (M_W): M_W is a measure of the seismic moment introduced by Hanks and Kanamori (1979) and is given by

$$M_W = \frac{3}{2}\log_{10} M_0 - 6.06\,,$$

where M_0 is the seismic moment (in Newton metres) defined by

$$M_0 = \mu A D$$

where μ is the shear strength of the rocks involved in the earthquake (in N/m^2), A is the area of the fault rupture plane (in m^2), and D is the average displacement on the fault rupture plane (in m).

Unlike other scales that measure the sizes of earthquakes via amplitude of waves produced at a certain distance and frequency band, the moment magnitude relates the magnitude to the physical properties of earthquakes. Besides, the moment magnitude scale has no saturation point for magnitude, meaning that there are no upper limits to the possible measurable magnitudes. However, moment magnitude scale requires more seismology knowledge than other scales and thus is more difficult to compute.

The advantages of the moment magnitude scale have accelerated its popularity

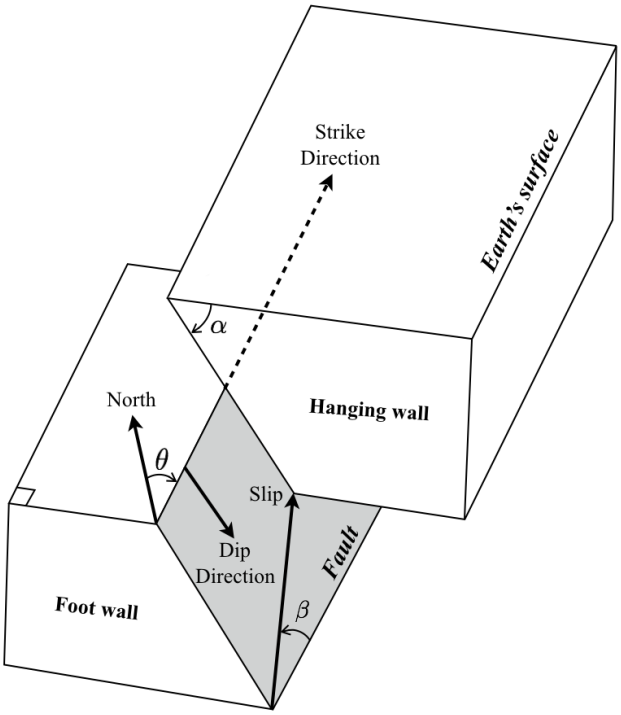


Figure 2.1: The geometry of a faulting: The grey area is the fault surface; α is the dip; θ is the strike; β is the rake.

in the seismic hazard community (Di Giacomo et al., 2015) and it has been the scale used by the United States Geological Survey (USGS) to report the magnitudes of all modern large earthquakes since January, 2002 (The USGS earthquake magnitude working group, 2002). However, many older earthquakes are still measured by other magnitude scales. Thus, there are many research (e.g., Das et al. (2011, 2012); Di Giacomo et al. (2015)) being carried out to find empirical relations between the moment magnitude and other magnitude scales.

2.1.3 Faulting geometry

The geometry of an earthquake faulting (Figure 2.1) can be described by three angular measurements (strike, dip and rake) and the magnitude of the slip.

Dip and dip direction

Dip (α in Figure 2.1) is the angle used to describe the steepness of a fault surface. The angle is between 0° and 90° and is measured from the Earth's surface, or a tangent plane parallel to the Earth's surface, to the fault surface. The **dip direction** is the direction towards which the fault surface is inclined. A fault with a dip of 0° is called a horizontal fault while a fault with a dip of 90° is called a vertical fault.

Foot wall and hanging wall

For non-vertical faults, the **foot wall** is the lower fault block beneath the Earth's surface and the fault surface (grey area in Figure 2.1), while the **hanging wall** is the upper fault block that is beneath the Earth's surface and above the fault surface. For vertical fault, the foot wall is assumed to be on the left of an observer looking in the strike direction.

Strike and strike direction

The **strike** (θ in Figure 2.1) is the angle between 0° and 360° used to specify the orientation of a fault. To determine the strike, strike direction needs to be decided first. The **strike direction** is the direction in which an observer looks along the fault line (i.e., the intersection of the Earth's surface and the fault surface) when they stand on the Earth's surface with the foot wall on their left and the hanging wall on the right. The strike is then measured clockwise from North direction to the strike direction.

Slip and rake

The **slip** is a parameter used to describe the motion of a fault. The slip is a vector, meaning that it has a magnitude and direction. The **magnitude of a slip** is simply the distance that a hanging wall moves relative to the foot wall. The **direction of slip** is the direction that the hanging wall moves relative to the foot wall. The **rake** (β in Figure 2.1) is the angle between 0°

and 360° measured anticlockwise from the strike direction to the slip direction. According to the slip direction, the faulting can be classified into two types termed *strike-slip* and *dip-slip*. The faulting is strike-slip if the slip direction is parallel to the strike direction (i.e., the rake $\beta = 0°$ or $\beta = 180°$); the fault is dip-slip if the slip direction is perpendicular to the strike direction (i.e., the rake $\beta = 90°$ or $\beta = 270°$). The strike-slip and dip-slip faulting can be further categorised:

- **Strike-slip**: If an observer, standing on one side of a fault, finds that the adjunct side moves to the left, then the faulting is *left-lateral* strike-slip (i.e., slip has the same direction with the strike direction or the rake $\beta = 0°$). If the adjunct side moves to the right, then the faulting is *right-lateral* strike-slip (i.e., slip has the opposite direction with the strike direction or the rake $\beta = 180°$).

- **Dip-slip**: If the hanging wall moves upward relative to the foot wall (i.e., $\beta = 90°$), the faulting is termed *reverse*, whereas when the hanging wall moves downward relative to the foot wall (i.e., $\beta = 270°$), the faulting is called *normal*.

Figure 2.2 illustrates the faulting types explained. There are some unusual faulting types such as tensile faulting that not only include strike- and dip-slips but also have expansion and compression of faults.

2.1.4 Source-to-site distance

The metrics of the source-to-site distance varies with different definitions. The following list summarises four types of source-to-site distance (illustrated in Figure 2.3):

- **Epicentre distance** (R_{epi}): the Euclidean distance between a site and epicentre;

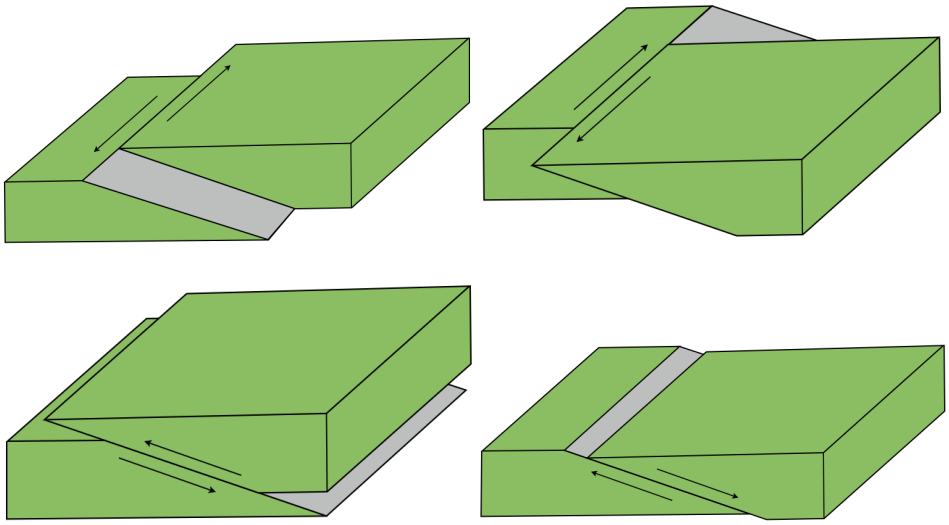


Figure 2.2: (*Top-left*) The left-lateral strike-slip faulting. (*Top-right*) The right-lateral strike-slip faulting. (*Bottom-left*) The reverse dip-slip faulting. (*Bottom-right*) The normal dip-slip faulting.

- **Hypocentre distance** (R_{hyp}): the Euclidean distance between a site and hypo-centre;

- **Rupture distance** (R_{rup}): the shortest Euclidean distance from a site to the rupture surface;

- **Joyner-Boore distance** (R_{JB}): the shortest Euclidean distance from a site to the surface projection of the rupture surface.

The choice of distance metrics depends on the magnitude of the earthquake and the availability of information about the rupture fault. For example, if an earthquake with a small-to-moderate magnitude occurs, a point source is typically assumed because the geometry of the fault plane is negligible compared to epicentre distances. Thus, the epicentre or hypo-centre distances are preferred. When an earthquake with a large magnitude (e.g., moment magnitude higher than 7) happens, the geometry of the fault plane is often assumed to be non-ignorable compared to the epicentre distances. Therefore, R_{JB} or R_{rup} are used accordingly.

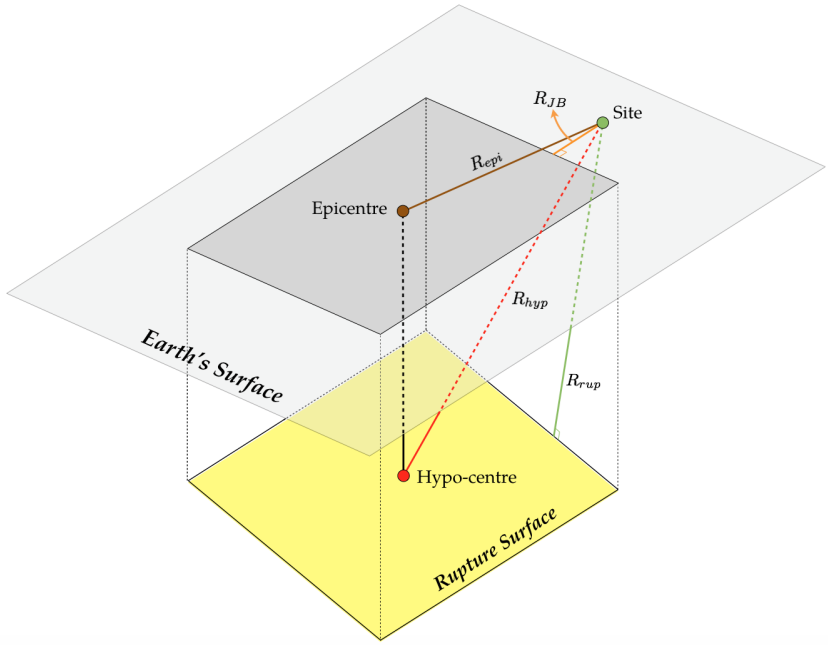


Figure 2.3: An illustrative example of different source-to-site distances. R_{epi} (brown line) is the epicentre distance; R_{hyp} (red line) is the hypo-centre distance, R_{rup} (green line) is the rupture distance; R_{JB} (orange line) is the Joyner-Boore distance. The yellow plane denotes the rupture surface. The grey plane is the surface projection of the rupture surface.

2.1.5 Soil property

The properties of near-surface soil at the sites of interest play an important role in filtering the ground-motion signals. The soil may amplify or de-amplify the ground-motion amplitude, change the frequency content and influence the earthquake duration, ultimately resulting in different degrees of damage to structures at the sites. The soil property at a site is often characterised by either a discrete or continuous fashion. In the discrete fashion, the soil profile at a site is classified into several catalogues such as soft soil, stiff soil and rock based on soil description and opinions of experts (Trifunac and Brady, 1975, 1976). The ground shaking tends to be stronger at sites with softer soil types because the seismic waves travel more slowly. Therefore, the wave amplification increases as the soil type shifts from rock to soft soil. The discrete characterisation is often used when detailed survey at sites is unavailable. However, when such a survey is available, the continuous characterisation is preferred and the average

shear-wave velocity (in m/s) in the upper 30 meters of the ground, known as V_S30 , is often used as the prime indicator of the site soil property (CEN, 2005).

2.2 Introduction

Initial ground-motion models were formulated as fixed-effects models without considering variations across different events. To further characterise the aleatory variability in ground shaking intensities, the uncertainties are separated into the inter-event and the intra-event components, where the inter-event components were introduced as random effects to the ground-motion model (Brillinger and Preisler, 1984). The modern ground-motion model is thus constructed as a mixed-effects model in the following form,

$$Y_{ij} = f(\mathbf{X}_{ij},\, \mathbf{b}) + \eta_i + \varepsilon_{ij}\,, \quad i = 1, \ldots, N,\ j = 1, \ldots, n_i\,, \tag{2.1}$$

where $Y_{ij} = \log \mathrm{IM}_{ij}$ is the logarithm of the IM of interest (e.g., peak ground acceleration (PGA), peak ground velocity (PGV), etc.) at site j during earthquake i ; $f(\mathbf{X}_{ij},\, \mathbf{b})$ is the ground-motion prediction function of $\mathbf{b}$, a vector of unknown parameters, and $\mathbf{X}_{ij}$, a vector of predictors (e.g., magnitude, source-to-site distance, soil type at site, etc.) for site j during event i ; η_i and ε_{ij} are the inter-event error and the intra-event error respectively; N is the total number of earthquakes and n_i is the number of recording sites during the i-th earthquake.

Traditionally, the ground-motion model in equation (2.1) is treated without spatial correlation by assuming the intra-event errors are spatially independent of each other, and is primarily estimated by algorithms proposed by Abrahamson and Youngs (1992) and Joyner and Boore (1993). However, it is well known that the intra-event errors are spatially correlated due to the common source and wave travelling paths and to similar site conditions (Goda and Hong, 2008; Jayaram and Baker, 2009). Hong et al. (2009) investigated the effects of spatial correlation on ground-motion model estimation and observed that the estimates of variances for inter-event and intra-event errors change significantly when

spatial correlation is considered. Jayaram and Baker (2010) confirmed the results and also demonstrated that the changes in variances of inter-event and intra-event errors have important implications for the seismic risk assessment of spatially distributed systems. Hence, we argue that it is crucial to develop an efficient and accurate estimation method for ground-motion models with spatial correlation.

Indeed, the consideration of spatial correlation complicates the estimation of ground-motion models. In particular, Hong et al. (2009) illustrated how to incorporate the spatial correlation into a ground-motion model and performed estimation using the method under the framework proposed by Joyner and Boore (1993). However, the estimation method proposed by Hong et al. (2009) uses the linearisation of the ground-motion prediction function, an inefficient technique that can add bias due to model misspecifications and was subsequently criticised by Draper and Smith (2014) for its slow convergence, wide oscillation and possibility of divergence. Based on the framework of Abrahamson and Youngs (1992), Jayaram and Baker (2010) introduced a multi-stage algorithm to account for the spatial correlation by adopting the idea of the classical geostatistical analysis (Zimmerman and Stein, 2010). However, this algorithm may not be statistically optimal and can result in inefficient parameter estimation, poor conclusions on model structure and variable selection, which in turn affects predictions of spatially distributed ground-motion intensities and, eventually, reliability of the seismic risk assessment and loss estimation for portfolios of spatially distributed buildings and lifelines.

In addition to the bespoke algorithms mentioned above, there is also a more generic existing computer package, namely `nlme` in R, available to fit ground-motion models with or without spatial correlation. However, this package is based on the method proposed by Lindstrom and Bates (1990) for mixed-effects models with nonlinear random effects and thus introduces excessive computational expenses during its implementation. Besides, the package may experience numerical instabilities when spatial correlation is considered even

though the estimation is performed on a small number of events. Jayaram and Baker (2010) also reported the numerical instability of the package. We argue that the failure of the package is due to the numerical issues that can arise when working with the Hessian matrices during its implementation of the Newton-Raphson algorithm. Furthermore, the package only considers limited types of spatial correlation structures (Pinheiro and Bates, 2000).

In this chapter, we first specify a ground-motion model as repeated Gaussian processes to incorporate spatial correlation. The multi-stage algorithm introduced by Jayaram and Baker (2010) is then reviewed and its limitations are highlighted. The new training method, referred to as the Scoring estimation approach, will then be formally introduced. The method is based on the method of Scoring (Fisher, 1925) as a specialised alternative procedure for fitting ground-motion models with spatial correlation. Numerical considerations for the Scoring estimation approach are also discussed. A simulation study is followed to measure the performances of the Scoring estimation approach by comparing against those of the multi-stage algorithm. Finally, we discuss the performance of the Scoring estimation approach when spatial correlation structure is neglected in the ground-motion model.

2.3 The Ground-Motion Model

The ground-motion model is expressed as the vector form of equation (2.1):

$$\mathbf{Y}_i = \mathbf{f}(\mathbf{X}_i, \mathbf{b}) + \boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i\,, \quad i = 1, \ldots, N\,, \tag{2.2}$$

where

- $\mathbf{Y}_i = \log \mathbf{IM}_i = (\log \mathrm{IM}_{i1}, \ldots, \log \mathrm{IM}_{ij}, \ldots, \log \mathrm{IM}_{in_i})^\top$ is an $n_i \times 1$ vector of logarithmic IMs of interest at all sites $j \in \{1, \ldots, n_i\}$ during earthquake i;

- $\mathbf{f}(\mathbf{X}_i, \mathbf{b}) = (f(\mathbf{X}_{i1}, \mathbf{b}), \ldots, f(\mathbf{X}_{in_i}, \mathbf{b}))^\top$ is an $n_i \times 1$ vector of ground-motion prediction functions $f(\mathbf{X}_{ij}, \mathbf{b})$ at all sites $j \in \{1, \ldots, n_i\}$ during

earthquake i;

- $\mathbf{X}_{ij}$ represents a vector of covariates (e.g., magnitude, source-to-site distance, soil type at site, etc.) for site j during earthquake i;
- $\mathbf{b} \in \mathbb{R}^{p_1}$ is a vector of unknown model parameters;
- $\boldsymbol{\eta}_i = \eta_i \mathbf{1}_{n_i}$ for all $i \in \{1, \ldots, N\}$, where $\mathbf{1}_{n_i}$ is an $n_i \times 1$ vector of ones and $(\eta_i)_{i=1,\ldots,N}$ are independent and identically distributed inter-event errors with the Gaussian distribution, $\mathcal{N}(0, \tau^2)$;
- $(\boldsymbol{\varepsilon}_i)_{i=1,\ldots,N}$ are independent intra-event error vectors of size $n_i \times 1$ with the multivariate Gaussian distribution, $\mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{\Omega}_i(\boldsymbol{\omega}))$, where $\boldsymbol{\Omega}_i(\boldsymbol{\omega})$ is the correlation matrix corresponding to earthquake i with $\boldsymbol{\omega}$, a vector of unknown parameters;
- $(\eta_i)_{i=1,\ldots,N}$ and $(\boldsymbol{\varepsilon}_i)_{i=1,\ldots,N}$ are mutually independent.

It can be seen that the ground-motion model (2.2) specifies N repeated Gaussian processes as it can be written as N independent random vectors following the multivariate Gaussian distribution:

$$\mathbf{Y}_i | (\mathbf{X}_i, \mathbf{S}_i, n_i) \overset{ind}{\sim} \mathcal{N}\left(\mathbf{f}(\mathbf{X}_i, \mathbf{b}), \tau^2 \mathbf{1}_{n_i \times n_i} + \sigma^2 \boldsymbol{\Omega}_i(\boldsymbol{\omega})\right), \tag{2.3}$$

where $\mathbf{S}_i = \{\mathbf{s}_{ij}\}_{j=1,\ldots,n_i}$ is a set of spatial locations (e.g., longitude and latitude) of the sites in earthquake i. To take the spatial correlation into account, the jj'-th entry, $\boldsymbol{\Omega}_{i,jj'}(\boldsymbol{\omega})$, of $\boldsymbol{\Omega}_i(\boldsymbol{\omega})$ is specified as

$$\boldsymbol{\Omega}_{i,jj'}(\boldsymbol{\omega}) = k(\mathbf{s}_{ij}, \mathbf{s}_{ij'})$$

for all $i \in \{1, \ldots, N\}$ and $j, j' \in \{1, \ldots, n_i\}$, where $k(\mathbf{s}_{ij}, \mathbf{s}_{ij'})$ is the kernel that gives the correlation $\rho(\varepsilon_{ij}, \varepsilon_{ij'})$ between ε_{ij} and $\varepsilon_{ij'}$ at locations $\mathbf{s}_{ij}$ and $\mathbf{s}_{ij'}$ of sites j and j' during earthquake i:

$$k(\mathbf{s}_{ij}, \mathbf{s}_{ij'}) = \rho(\varepsilon_{ij}, \varepsilon_{ij'}).$$

In this chapter, we mainly consider stationary and isotropic kernels, meaning that $\rho(\varepsilon_{ij}, \varepsilon_{ij'})$ only depends on $d_{i,jj'} = \|\mathbf{s}_{ij} - \mathbf{s}_{ij'}\|_2$, the Euclidean distance

between sites j and j' during earthquake i. Thus,

$$k(\mathbf{s}_{ij}, \mathbf{s}_{ij'}) = k(d_{i,jj'}).$$

Note that if no spatial correlation is incorporated, we have

$$k(\mathbf{s}_{ij}, \mathbf{s}_{ij'}) = 0 \tag{2.4}$$

for all sites j and j' during earthquake i. It is also worth noting that the covariance matrix of the Gaussian process model specified in (2.3) is in fact determined by the kernel function

$$k^*(d_{i,jj'}) = \frac{\tau^2}{\sigma^2} + k(d_{i,jj'}),$$

which is still a valid kernel function.

In the rest of this chapter, we denote $\boldsymbol{\alpha} = (\mathbf{b}^\top, \boldsymbol{\theta}^\top)^\top \in \mathbb{R}^p$ as the complete vector of model parameters, in which $\boldsymbol{\theta} = (\tau^2, \sigma^2, \boldsymbol{\omega}^\top)^\top \in \mathbb{R}^{p_2}$ with $\boldsymbol{\omega}$ being a vector of the parameters contained in the kernel function $k(\mathbf{s}_{ij}, \mathbf{s}_{ij'})$.

2.4 Jayaram and Baker's Multi-Stage Algorithm

In this section, we review the multi-stage algorithm proposed by Jayaram and Baker (2010) to estimate ground-motion models with spatial correlation. This algorithm will serve as the current best benchmark procedure for our new proposed method, so it is important to discuss its properties and compare its approach to our proposed Scoring estimation approach. The algorithm consists of three stages (see Figure 2.4) and follows the framework of the classical geostatistical method (Zimmerman and Stein, 2010). In the preliminary stage, the algorithm provisionally estimates the model parameters ignoring the spatial correlation. In the second stage, the residuals from the estimated provisional ground-motion prediction function are used to estimate the parameters in the kernel function by fitting a parametric semivariogram model to the empirical semivariogram. In the final stage, the preliminary estimates of model parameters

from the first stage are updated given the spatial correlation structure fitted in the second stage. We proceed to outline each stage in detail below.

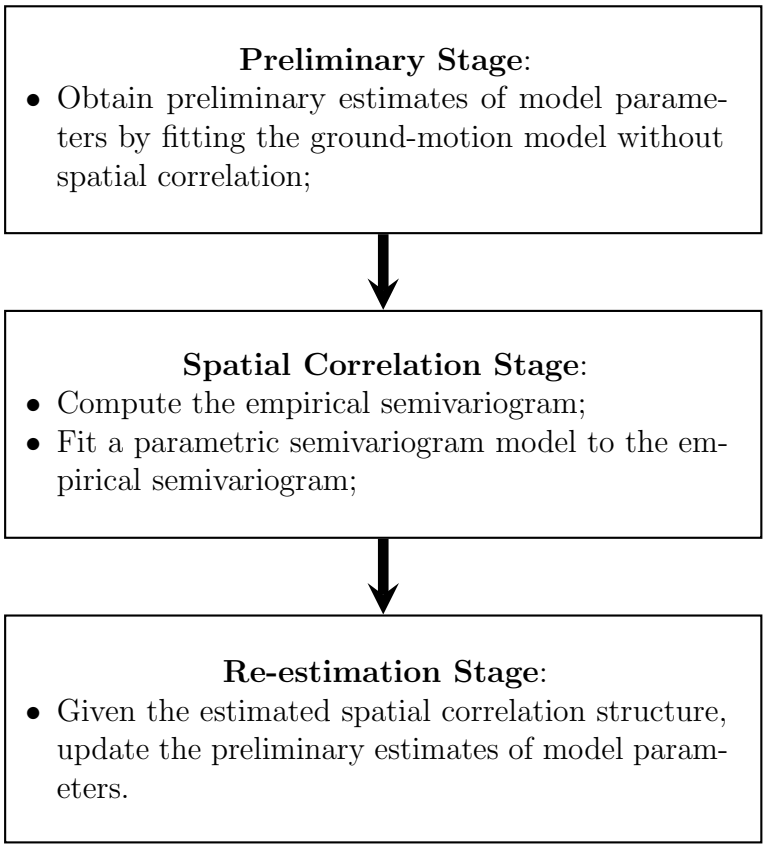


Figure 2.4: Flowchart of the multi-stage algorithm proposed by Jayaram and Baker (2010).

2.4.1 The preliminary stage

The preliminary stage of the algorithm aims at estimating ground-motion models requiring no knowledge about the spatial correlation. Because the spatial correlation is being ignored at this stage, authors such as Goda and Hong (2008); Goda and Atkinson (2009, 2010); Sokolov et al. (2010) adopted estimation methods introduced by Abrahamson and Youngs (1992) or Joyner and Boore (1993) to obtain the estimates of unknown model parameters $\mathbf{b}$, τ^2, and σ^2. Other authors such as Wang and Takada (2005); Jayaram and Baker (2009); Esposito and Iervolino (2011, 2012) obtained the estimates of $\mathbf{b}$, τ^2, and σ^2 by simply adopting existing ground-motion models developed without

consideration of spatial correlation.

2.4.2 The spatial correlation stage

The spatial correlation stage is designed to estimate $\boldsymbol{\omega}$, a vector of unknown parameters in the kernel function, from the total residuals

$$e_{ij}^{(t)} = Y_{ij} - f(\mathbf{X}_{ij}, \widehat{\mathbf{b}}),$$

in which $\widehat{\mathbf{b}}$ is the estimate of $\mathbf{b}$ given by the preliminary stage. Because the total error term

$$\varepsilon_{ij}^{(t)} = \varepsilon_{ij} + \eta_i$$

consists of intra-event errors ε_{ij} and inter-event errors η_i, the total residuals can be represented by intra-event residuals $\widehat{\varepsilon}_{ij}$ and inter-event residuals $\widehat{\eta}_i$:

$$e_{ij}^{(t)} = \widehat{\varepsilon}_{ij} + \widehat{\eta}_i .$$

Then one defines a random process of the standardised intra-event errors

$$\widetilde{\boldsymbol{\varepsilon}} = \frac{\boldsymbol{\varepsilon}}{\sigma}$$

with $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1^\top, \ldots, \boldsymbol{\varepsilon}_N^\top)^\top$ and $\widetilde{\boldsymbol{\varepsilon}} = (\widetilde{\boldsymbol{\varepsilon}}_1^\top, \ldots, \widetilde{\boldsymbol{\varepsilon}}_N^\top)^\top$. Assuming that the process of intra-event errors is second-order stationary and isotropic, Jayaram and Baker (2009) constructed for each earthquake i the empirical semivariogram $\widehat{\gamma}_i(d)$, a moment-based estimator defined by Cressie (1993), of $\widetilde{\boldsymbol{\varepsilon}}_i$ from the scaled intra-event residuals:

$$\widehat{\widetilde{\varepsilon}}_{ij} = \frac{\widehat{\varepsilon}_{ij}}{\widehat{\sigma}}.$$

The empirical semivariogram $\widehat{\gamma}_i(d)$ is calculated by

$$\begin{aligned}
\widehat{\gamma}_i(d) &= \frac{1}{2|N_{i,\delta}(d)|} \sum_{N_{i,\delta}(d)} \left(\widehat{\widetilde{\varepsilon}}_{ij} - \widehat{\widetilde{\varepsilon}}_{ij'} \right)^2 \\
&= \frac{1}{2|N_{i,\delta}(d)|} \sum_{N_{i,\delta}(d)} \left(\frac{e_{ij}^{(t)} - \widehat{\eta}_i}{\widehat{\sigma}} - \frac{e_{ij'}^{(t)} - \widehat{\eta}_i}{\widehat{\sigma}} \right)^2 \\
&= \frac{1}{2|N_{i,\delta}(d)|} \sum_{N_{i,\delta}(d)} \left(\frac{e_{ij}^{(t)} - e_{ij'}^{(t)}}{\widehat{\sigma}} \right)^2 ,
\end{aligned} \tag{2.5}$$

in which $N_{i,\delta}(d)$ is a δ-neighbourhood set consisting of all site pairs $(j,\, j')$ such that

$$d - \delta < \|\mathbf{s}_{ij} - \mathbf{s}_{ij'}\|_2 < d + \delta$$

during earthquake i and $|N_{i,\delta}(d)|$ is the number of distinct pairs in $N_{i,\delta}(d)$.

Each empirical seimivariogram $\widehat{\gamma}_i(d)$ is then fitted by a common parametric semivariogram model $\gamma(d)$ constructed from a stationary and isotropic kernel function $k(d)$ according to the relationship given by

$$\gamma(d) = 1 - k(d), \tag{2.6}$$

whose proof is available in Section A.1 of Appendix A. One can then obtain the estimate $\widehat{\boldsymbol{\omega}}_i$ of $\boldsymbol{\omega}$ for each earthquake i by fitting $\gamma(d)$ to the sample estimator given by $\widehat{\gamma}_i(d)$ via estimation methods such as least-squares and trial-and-error methods (i.e., a manual fitting method focusing on fitting the empirical semivariogram at short separation distances d). Jayaram and Baker (2009) then computed the estimates $\widehat{\boldsymbol{\omega}}_{i=1,\ldots,N}$ for spectral accelerations at different structural periods and built linear regression models to obtain the estimate of $\boldsymbol{\omega}$ for a given structural period.

Unlike Jayaram and Baker (2009) who estimated $\boldsymbol{\omega}$ by constructing empirical semivariogram for each earthquake i, Esposito and Iervolino (2011, 2012) built a pooled empirical semivariogram $\widehat{\gamma}(d)$ given by

$$\widehat{\gamma}(d) = \frac{1}{2|N_\delta(d)|} \sum_{N_\delta(d)} \left(\frac{e_{ij}^{(t)} - e_{ij'}^{(t)}}{\widehat{\sigma}} \right)^2 ,$$

in which $N_\delta(d)$ is a δ-neighbourhood set consisting of all site pairs $(j,\, j')$ such that

$$d - \delta < \|\mathbf{s}_{ij} - \mathbf{s}_{ij'}\|_2 < d + \delta$$

across all earthquakes $i \in \{1, \ldots, N\}$. The estimate of $\boldsymbol{\omega}$ is then obtained by fitting a parametric semivariogram model $\gamma(d)$ to $\widehat{\gamma}(d)$ via least-squares and trial-and-error methods.

Jayaram and Baker (2009) discussed the method of least squares and the trial-

and-error method and suggested that the trial-and-error method is a better choice because of its simplicity and better fit at separation "distances that are of practical interest" (Jayaram and Baker, 2009).

2.4.3 The re-estimation stage

The objective of the re-estimation stage is to update the estimates $\widehat{\mathbf{b}}$, $\widehat{\sigma^2}$ and $\widehat{\tau^2}$ obtained in the preliminary stage by considering the spatial correlation structure established in the spatial correlation stage. Algorithm 1 illustrates the re-estimation procedure proposed by Jayaram and Baker (2010). However, Jayaram and Baker (2010) did not report any convergence properties of the procedure. In Section A.2 of Appendix A, we demonstrate that the re-estimation procedure can be alternatively constructed based on the idea of the Expectation-Maximisation (EM) algorithm (Laird and Ware, 1982; Brillinger and Preisler, 1985; Laird et al., 1987). Therefore, the re-estimation procedure is a non-decreasing algorithm (i.e., $l(\sigma^2, \tau^2, \mathbf{b}|\boldsymbol{\omega} = \widehat{\boldsymbol{\omega}})$ is increased at each iteration) as long as the fixed-effects regression algorithm (step 4 of the Algorithm 1) solves the following generalised least squares problem with respect to $\mathbf{b}$:

$$\widehat{\mathbf{b}}^{(k+1)} = \arg\min \sum_{i=1}^{N} [\mathbf{Y}_i - \mathbf{f}(\mathbf{X}_i,\, \mathbf{b}) - \widehat{\eta}_i \mathbf{1}_{n_i}]^\top \boldsymbol{\Omega}_i^{-1}(\widehat{\boldsymbol{\omega}})[\mathbf{Y}_i - \mathbf{f}(\mathbf{X}_i,\, \mathbf{b}) - \widehat{\eta}_i \mathbf{1}_{n_i}]\,. \tag{2.7}$$

2.4.4 Problems of the multi-stage algorithm

Although the multi-stage algorithm is feasible in practice and may be numerically stable by estimating the spatial kernel function in separate steps (i.e., the preliminary and spatial correlation stages), it is not optimal in various aspects from a statistical estimation perspective.

First, the least squares estimator of $\boldsymbol{\omega}$ produced by the first two stages of the algorithm is inconsistent (i.e., $\widehat{\boldsymbol{\omega}}$ does not converge in probability to the true value of $\boldsymbol{\omega}$). Lahiri et al. (2002) and Kerby (2016) discussed the conditions for the consistency of the least squares estimator of $\boldsymbol{\omega}$. To have a consistent

Algorithm 1 The re-estimation procedure (Jayaram and Baker, 2010)

Require: 1) $\mathbf{Y}_i$, $\mathbf{X}_{ij}$ and $\mathbf{s}_{ij}$ for $i \in \{1, \ldots, N\}$ and $j \in \{1, \ldots, n_i\}$;
2) Estimate $\widehat{\boldsymbol{\omega}}$ of $\boldsymbol{\omega}$ obtained in the spatial correlation stage.

Ensure: Updated estimates of $\mathbf{b}$, σ^2 and τ^2.

1: **Initialisation:**

1) obtain the initial estimate $\widehat{\mathbf{b}}^{(1)}$ of $\mathbf{b}$ by a fixed-effects regression algorithm setting $\eta_{i=1,\ldots,N} = 0$;

2) obtain the initial estimates $\widehat{\sigma^2}^{(1)}$ and $\widehat{\tau^2}^{(1)}$ by maximising the log-likelihood function:

$$l\left(\sigma^2,\, \tau^2 \middle| \mathbf{b} = \widehat{\mathbf{b}}^{(1)},\, \boldsymbol{\omega} = \widehat{\boldsymbol{\omega}}\right) = -\frac{\sum_{i=1}^{N} n_i}{2} \ln(2\pi) - \frac{1}{2} \sum_{i=1}^{N} \ln \left| \tau^2 \mathbf{1}_{n_i \times n_i} + \sigma^2 \boldsymbol{\Omega}_i(\widehat{\boldsymbol{\omega}}) \right| - \frac{1}{2} \sum_{i=1}^{N} [\mathbf{Y}_i - \mathbf{f}(\mathbf{X}_i,\, \widehat{\mathbf{b}}^{(1)})]^\top \left(\tau^2 \mathbf{1}_{n_i \times n_i} + \sigma^2 \boldsymbol{\Omega}_i(\widehat{\boldsymbol{\omega}}) \right)^{-1} [\mathbf{Y}_i - \mathbf{f}(\mathbf{X}_i,\, \widehat{\mathbf{b}}^{(1)})] ;$$

2: **repeat**

3: Given $\widehat{\mathbf{b}}^{(k)}$, $\widehat{\sigma^2}^{(k)}$, $\widehat{\tau^2}^{(k)}$ and $\widehat{\boldsymbol{\omega}}$, obtain $\widehat{\eta}_{i=1,\ldots,N}$ from

$$\widehat{\eta}_i = \frac{\frac{1}{\widehat{\sigma^2}^{(k)}} \mathbf{1}_{n_i}^\top \boldsymbol{\Omega}_i^{-1}(\widehat{\boldsymbol{\omega}}) \left[\mathbf{Y}_i - \mathbf{f}(\mathbf{X}_i,\, \widehat{\mathbf{b}}^{(k)})\right]}{\frac{1}{\widehat{\tau^2}^{(k)}} + \frac{1}{\widehat{\sigma^2}^{(k)}} \mathbf{1}_{n_i}^\top \boldsymbol{\Omega}_i^{-1}(\widehat{\boldsymbol{\omega}})\, \mathbf{1}_{n_i}} ; \tag{2.8}$$

4: Given $\widehat{\eta}_{i=1,\ldots,N}$, obtain $\widehat{\mathbf{b}}^{(k+1)}$, the estimate of $\mathbf{b}$ at iteration $k+1$, using a fixed-effects regression algorithm by setting $\eta_i = \widehat{\eta}_i$ for all $i \in \{1, \ldots, N\}$;

5: Given $\widehat{\mathbf{b}}^{(k+1)}$ and $\widehat{\boldsymbol{\omega}}$, obtain $\widehat{\sigma^2}^{(k+1)}$ and $\widehat{\tau^2}^{(k+1)}$ by maximising the log-likelihood function $l(\sigma^2,\, \tau^2 | \mathbf{b} = \widehat{\mathbf{b}}^{(k+1)},\, \boldsymbol{\omega} = \widehat{\boldsymbol{\omega}})$;

6: **until** $l(\sigma^2, \tau^2, \mathbf{b} | \boldsymbol{\omega} = \widehat{\boldsymbol{\omega}})$ is maximised and parameter estimates converge.

least squares estimator of $\boldsymbol{\omega}$, we need the empirical semivariogram $\widehat{\gamma}(d)$ to be a consistent estimator of $\gamma(d)$. However, this consistency typically requires very restrictive asymptotic conditions in which "not only the number of locations increases but the distance between them decreases" (Kerby, 2016). Furthermore, Kerby (2016) showed that observation locations must not be heavily clustered (which is the case in reality where the recording sites are indeed clustered, especially at near-fault locations) and the bandwidth δ need to be carefully chosen so that the consistency of the empirical semivariogram $\widehat{\gamma}(d)$ is ensured. In addition, the consistency of the empirical semivariogram $\widehat{\gamma}(d)$

requires the estimators of $\mathbf{b}$ and σ^2 obtained from the preliminary stage to be consistent. However, although the estimator of $\mathbf{b}$ obtained in the preliminary stage is consistent (due to asymptotic independence between $\widehat{\mathbf{b}}$ and $\widehat{\boldsymbol{\omega}}$), the estimator of σ^2 is not (as $\widehat{\sigma^2}$ and $\widehat{\boldsymbol{\omega}}$ are asymptotically dependent). Consequently, the least squares estimator of $\boldsymbol{\omega}$ obtained at the spatial correlation stage is not consistent. Finally, the least squares estimator of $\boldsymbol{\omega}$ can be statistically inefficient (Lahiri et al., 2002), and naively using the formula of asymptotic standard error estimate produced by software packages based on ordinary least squares can cause incorrect confidence interval on $\boldsymbol{\omega}$.

With regard to the trial-and-error method, although it fits the parametric semivariogram model to the empirical semivariograms better than the least squares at short separation distances, Stein (1999) illustrated in a simulation study that this eyeball procedure leads to substantial prediction errors, especially when the spatial correlation structure is misspecified. Besides, this manual fitting procedure makes it impossible to evaluate the asymptotic properties of the estimator of $\boldsymbol{\omega}$. Therefore, such a heuristic procedure should not become the standard.

Moreover, the first two stages are only capable of estimations of isotropic and stationary correlation structures and inflexible in considering more advanced (e.g., non-stationary) spatial kernel functions.

In addition, the re-estimation procedure maximises the conditional log-likelihood function $l(\sigma^2, \tau^2, \mathbf{b}|\boldsymbol{\omega} = \widehat{\boldsymbol{\omega}})$ given the pre-computed estimate $\widehat{\boldsymbol{\omega}}$. Because the least squares estimator of $\boldsymbol{\omega}$ is inconsistent, the resulting estimators of $\mathbf{b}$ (although consistent) are statistically inefficient, and estimators of τ^2 and σ^2 are both inconsistent and statistically inefficient.

Additionally, because the re-estimation procedure can be interpreted via the idea of the EM algorithm, it suffers from the "hopelessly slow linear convergence" (Couvreur, 1997) and is very sensitive to the initial parameter values (Gao and Wang, 2013).

Furthermore, unlike the Scoring estimation approach introduce in Section 2.5, the multi-stage algorithm does not produce asymptotic standard error estimates of model parameters as by-products. As a consequence, the multi-stage algorithm requires extra computations and complexities in its implementation when asymptotic standard error estimates are desired. Finally, it is worth noting that the equations provided by Jayaram and Baker (2010) for asymptotic standard error estimates of τ^2 and σ^2 are only valid when estimators of τ^2 and σ^2 are asymptotically independent. However, $\widehat{\tau^2}$ and $\widehat{\sigma^2}$ are not asymptotically independent, thus, their asymptotic variance estimates should be obtained by taking the first and the second diagonal entry of

$$2\begin{bmatrix} \operatorname{tr}\left\{\left(\mathbf{C}(\boldsymbol{\theta})^{-1}\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial(\tau^2)}\right)^2\right\} & \operatorname{tr}\left\{\mathbf{C}(\boldsymbol{\theta})^{-1}\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial(\tau^2)}\mathbf{C}(\boldsymbol{\theta})^{-1}\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial(\sigma^2)}\right\} \\ \operatorname{tr}\left\{\mathbf{C}(\boldsymbol{\theta})^{-1}\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial(\sigma^2)}\mathbf{C}(\boldsymbol{\theta})^{-1}\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial(\tau^2)}\right\} & \operatorname{tr}\left\{\left(\mathbf{C}(\boldsymbol{\theta})^{-1}\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial(\sigma^2)}\right)^2\right\} \end{bmatrix}^{-1}_{\boldsymbol{\theta}=(\widehat{\tau^2},\widehat{\sigma^2},\widehat{\boldsymbol{\omega}}^\top)^\top}, \tag{2.9}$$

in which

$$\mathbf{C}(\boldsymbol{\theta}) = \begin{bmatrix} \tau^2\mathbf{1}_{n_1\times n_1} + \sigma^2\boldsymbol{\Omega}_1(\boldsymbol{\omega}) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \tau^2\mathbf{1}_{n_2\times n_2} + \sigma^2\boldsymbol{\Omega}_2(\boldsymbol{\omega}) & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \tau^2\mathbf{1}_{n_N\times n_N} + \sigma^2\boldsymbol{\Omega}_N(\boldsymbol{\omega}) \end{bmatrix}.$$

However, even matrix (2.9) may not give the correct asymptotic standard error estimates of $\widehat{\tau^2}$ and $\widehat{\sigma^2}$ because the least squares estimator of $\boldsymbol{\omega}$ is inconsistent and asymptotic variances of $\widehat{\tau^2}$ and $\widehat{\sigma^2}$ depend on that of $\widehat{\boldsymbol{\omega}}$.

To avoid the above complications and statistical deficiencies inherent in the Jayaram and Baker (2010) multi-stage estimation procedure, we introduce the Scoring estimation approach, a method based on maximum likelihood estimation framework. The proposed Scoring estimation approach produces model parameter estimators consistently in a single stage algorithm, which admits any parametric class of kernel functions and associated spatial correlation properties, including anisotropic or non-stationary choices.

2.5 A One-Stage Algorithm: the Scoring Estimation Approach

The one-stage estimation approach we propose here aims at obtaining the maximum likelihood estimate of $\boldsymbol{\alpha}$ by maximising the following log-likelihood function:

$$\begin{aligned} l(\boldsymbol{\alpha}) &= \ln L(\boldsymbol{\alpha}) \\ &= -\frac{\sum_{i=1}^{N} n_i}{2} \ln(2\pi) - \frac{\ln\left|\mathbf{C}(\boldsymbol{\theta})\right|}{2} - \frac{1}{2}[\mathbf{Y} - \mathbf{f}(\mathbf{X},\, \mathbf{b})]^\top \mathbf{C}^{-1}(\boldsymbol{\theta})[\mathbf{Y} - \mathbf{f}(\mathbf{X},\, \mathbf{b})]\,, \end{aligned} \tag{2.10}$$

where $L(\boldsymbol{\alpha}|\mathbf{Y})$ is the likelihood function, $\mathbf{f}(\mathbf{X},\, \mathbf{b}) = \left(\mathbf{f}(\mathbf{X}_1,\, \mathbf{b})^\top, \ldots, \mathbf{f}(\mathbf{X}_N,\, \mathbf{b})^\top\right)^\top$ and $\mathbf{Y} = (\mathbf{Y}_1^\top, \ldots, \mathbf{Y}_N^\top)^\top$.

The classic statistical method to maximise the log-likelihood function (2.10) is via the Newton-Raphson algorithm. The Newton-Raphson algorithm finds the estimate of $\boldsymbol{\alpha}$ that maximises the log-likelihood function (2.10) via the updating equation:

$$\widehat{\boldsymbol{\alpha}}^{(k+1)} = \widehat{\boldsymbol{\alpha}}^{(k)} - \mathbf{H}^{-1}(\widehat{\boldsymbol{\alpha}}^{(k)})\mathbf{S}(\widehat{\boldsymbol{\alpha}}^{(k)})\,, \tag{2.11}$$

in which $\widehat{\boldsymbol{\alpha}}^{(k)}$ denotes the estimate of $\boldsymbol{\alpha}$ at iteration step k, and

$$\mathbf{S}(\boldsymbol{\alpha}) = \frac{\partial l(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} \quad \text{and} \quad \mathbf{H}(\boldsymbol{\alpha}) = \frac{\partial^2 l(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^\top}$$

represent the gradient and Hessian matrix of $l(\boldsymbol{\alpha})$, respectively. In general, however, the Newton-Raphson algorithm may not be a robust maximisation algorithm when applied directly to applications such as the one in this study. There are numerous reasons for this. First, even though the Hessian matrix is negative definite at the local maximum, the Hessian matrix may not be negative definite at every iteration. Thus, the algorithm does not guarantee an ascent direction of the log-likelihood function and may converge to a local minimum if positive definite Hessian matrices are encountered during the updates. Second, the Hessian matrix can sometimes have poor sparsity and thus can be computationally expensive to evaluate at each iteration. Finally,

the Hessian matrix can be indefinite or even singular (Seber and Wild, 2003), causing numerical instabilities in the Newton-Raphson algorithm.

To overcome these issues, the Scoring estimation approach is proposed to obtain the maximum likelihood estimate of $\boldsymbol{\alpha}$. The Scoring estimation approach is based on the method of Scoring introduced by Fisher (1925), which is a modified version of the Newton-Raphson algorithm. The updating equation for the Scoring estimation approach is obtained by replacing the negative Hessian matrix, $-\mathbf{H}(\boldsymbol{\alpha})$, by the expected (or Fisher) information matrix, $\mathbf{I}(\boldsymbol{\alpha})$:

$$\widehat{\boldsymbol{\alpha}}^{(k+1)} = \widehat{\boldsymbol{\alpha}}^{(k)} + \mathbf{I}^{-1}(\widehat{\boldsymbol{\alpha}}^{(k)})\mathbf{S}(\widehat{\boldsymbol{\alpha}}^{(k)}) \tag{2.12}$$

with

$$\mathbf{I}(\boldsymbol{\alpha}) = -\mathbb{E}\left[\mathbf{H}(\boldsymbol{\alpha})\right] = -\mathbb{E}\left[\frac{\partial^2 l(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^\top}\right].$$

Let $\boldsymbol{\alpha}_0$ be the true parameter value of $\boldsymbol{\alpha}$ and assume that $L(\boldsymbol{\alpha})$ and its first derivatives with respect to $\boldsymbol{\alpha}$ are continuous in the domains of $\boldsymbol{\alpha}$ and $\mathbf{Y}$. Then it can be shown (Wooldridge, 2010) that

$$\mathbf{I}(\boldsymbol{\alpha}_0) = \mathbf{A}(\boldsymbol{\alpha}_0) \tag{2.13}$$

with

$$\mathbf{A}(\boldsymbol{\alpha}) = \mathbb{E}\left[\frac{\partial l(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} \frac{\partial l(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}^\top}\right],$$

which is positive-definite. This result states that the expected information matrix $\mathbf{I}(\boldsymbol{\alpha}_0)$ is always positive-definite, meaning that if we replace $\boldsymbol{\alpha}_0$ in $\mathbf{I}(\boldsymbol{\alpha}_0)$ by $\widehat{\boldsymbol{\alpha}}^{(k)}$, then each iteration of the approach will lead the log-likelihood function in an uphill direction. Therefore, the Scoring estimation approach is more numerically stable than the Newton-Raphson algorithm. Furthermore, equation (2.13) states that only the gradient of $l(\boldsymbol{\alpha})$ is required for the calculation of the expected information matrix $\mathbf{I}(\boldsymbol{\alpha})$, implying that computation in each iteration of the approach is usually quicker than that of Newton-Raphson.

Denote the gradient $\mathbf{S}(\boldsymbol{\alpha})$ and expected information matrix $\mathbf{I}(\boldsymbol{\alpha})$ of $l(\boldsymbol{\alpha})$ by

the partitions

$$\mathbf{S}(\boldsymbol{\alpha}) = \begin{bmatrix} \mathbf{S}_{\mathbf{b}}(\boldsymbol{\alpha}) \\ \mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\alpha}) \end{bmatrix} \quad \text{and} \quad \mathbf{I}(\boldsymbol{\alpha}) = \begin{bmatrix} \mathbf{I}_{\mathbf{bb}}(\boldsymbol{\alpha}) & \mathbf{I}_{\mathbf{b}\boldsymbol{\theta}}(\boldsymbol{\alpha}) \\ \mathbf{I}_{\boldsymbol{\theta}\mathbf{b}}(\boldsymbol{\alpha}) & \mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\alpha}) \end{bmatrix} .$$

Then, the Scoring estimation approach obtains the maximum likelihood estimate of $\boldsymbol{\alpha}$ by the updating equations in Theorem 2.1.

Theorem 2.1 *The updating equations for the Scoring estimation approach are given by*

$$\widehat{\mathbf{b}}^{(k+1)} = \widehat{\mathbf{b}}^{(k)} + \mathbf{I}^{-1}_{\mathbf{bb}}(\widehat{\boldsymbol{\alpha}}^{(k)})\, \mathbf{S}_{\mathbf{b}}(\widehat{\boldsymbol{\alpha}}^{(k)})\,, \tag{2.14}$$

$$\widehat{\boldsymbol{\theta}}^{(k+1)} = \widehat{\boldsymbol{\theta}}^{(k)} + \mathbf{I}^{-1}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\widehat{\boldsymbol{\alpha}}^{(k)})\, \mathbf{S}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\alpha}}^{(k)})\,, \tag{2.15}$$

in which

- *the i-th element of $\mathbf{S}_{\mathbf{b}}(\boldsymbol{\alpha})$ is given by*

$$[\mathbf{S}_{\mathbf{b}}(\boldsymbol{\alpha})]_i = \left[\frac{\partial \mathbf{f}(\mathbf{X},\, \mathbf{b})}{\partial \mathbf{b}_i}\right]^{\top} \mathbf{C}^{-1}(\boldsymbol{\theta})[\mathbf{Y} - \mathbf{f}(\mathbf{X},\, \mathbf{b})]\,;$$

- *the i-th element of $\mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\alpha})$ is given by*

$$\begin{aligned}[\mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\alpha})]_i = &-\frac{1}{2}\mathrm{tr}\left\{\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i}\right\} \\ &+ \frac{1}{2}[\mathbf{Y} - \mathbf{f}(\mathbf{X},\, \mathbf{b})]^{\top}\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i}\mathbf{C}^{-1}(\boldsymbol{\theta})[\mathbf{Y} - \mathbf{f}(\mathbf{X},\, \mathbf{b})]\,;\end{aligned}$$

- *the ij-th element of $\mathbf{I}_{\mathbf{bb}}(\boldsymbol{\alpha})$ is given by*

$$[\mathbf{I}_{\mathbf{bb}}(\boldsymbol{\alpha})]_{ij} = \left[\frac{\partial \mathbf{f}(\mathbf{X},\, \mathbf{b})}{\partial \mathbf{b}_i}\right]^{\top} \mathbf{C}^{-1}(\boldsymbol{\theta})\, \frac{\partial \mathbf{f}(\mathbf{X},\, \mathbf{b})}{\partial \mathbf{b}_j}\,;$$

- *the ij-th element of $\mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\alpha})$ is given by*

$$[\mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\alpha})]_{ij} = \frac{1}{2}\mathrm{tr}\left\{\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i}\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j}\right\} .$$

Proof The proof is given in Section A.3 of Appendix A. □

It can be seen from the updating equations (2.14) and (2.15) that the Scoring estimation approach is able to update the estimates of $\mathbf{b}$ and $\boldsymbol{\theta}$ by separate equations. This separation has two advantages. For the Newton-Raphson

update equation (2.11), it requires at each iteration the complexity (i.e., a concept in computer sciences describing the amount of time required for running an algorithm) of $\mathcal{O}(p^3)$ dominated by the inversion of the Hessian matrix $\mathbf{H}(\widehat{\boldsymbol{\alpha}}^{(k)})$. However, thanks to the separation, the Scoring estimation approach only requires at each iteration the complexity of $\mathcal{O}(p_1^3 + p_2^3)$ dominated by inversions of

$$\mathbf{I}_{\mathbf{bb}}(\widehat{\boldsymbol{\alpha}}^{(k)}) \in \mathbb{R}^{p_1 \times p_1} \quad \text{and} \quad \mathbf{I}_{\boldsymbol{\theta\theta}}(\widehat{\boldsymbol{\alpha}}^{(k)}) \in \mathbb{R}^{p_2 \times p_2},$$

in which $p_1 + p_2 = p$ and p_1 and p_2 are dimensions of $\mathbf{b}$ and $\boldsymbol{\theta}$, respectively. Therefore, the separate updating equations in the Scoring estimation approach reduce computational expenses. In addition, equations (2.14) and (2.15) indicate that the Scoring estimation approach only requires inversions of $\mathbf{I}_{\mathbf{bb}}(\widehat{\boldsymbol{\alpha}}^{(k)})$ and $\mathbf{I}_{\boldsymbol{\theta\theta}}(\widehat{\boldsymbol{\alpha}}^{(k)})$, each of which has a smaller size than the Hessian matrix $\mathbf{H}(\widehat{\boldsymbol{\alpha}}^{(k)})$ in the Newton-Raphson algorithm. Pyzara et al. (2011) showed that the size of a matrix is positively connected to its condition number, and the condition number of an ill-conditioned matrix (e.g., a Hilbert matrix) can grow at a remarkably higher rate than that of a well-conditioned matrix as its size increases. Thus, inversions of matrices of smaller sizes in the Scoring estimation approach mitigate the risk of developing large condition numbers, which reduces the effects of round-off error and thus improves the computational stability.

2.5.1 Asymptotic properties of the maximum likelihood estimator

Applying the asymptotic results of M-estimator (Wooldridge, 2010; Demidenko, 2013), we have that the maximum likelihood estimator $\widehat{\boldsymbol{\alpha}}$ is consistent, asymptotically normal, and statistically efficient when $N \to \infty$. The asymptotic standard error estimate $\widehat{\text{se}}(\widehat{\boldsymbol{\alpha}})$ of $\widehat{\boldsymbol{\alpha}} = (\widehat{\mathbf{b}}^\top, \widehat{\boldsymbol{\theta}}^\top)^\top$ can be obtained by

$$\widehat{\text{se}}(\widehat{\mathbf{b}}) = \sqrt{\text{diag}\left[\mathbf{I}_{\mathbf{bb}}^{-1}(\widehat{\boldsymbol{\alpha}}^{(K)})\right]} \tag{2.16}$$

and

$$\widehat{\text{se}}(\widehat{\boldsymbol{\theta}}) = \sqrt{\text{diag}\left[\mathbf{I}_{\boldsymbol{\theta\theta}}^{-1}(\widehat{\boldsymbol{\alpha}}^{(K)})\right]}, \tag{2.17}$$

in which $\widehat{\boldsymbol{\alpha}}^{(K)}$ is the final estimate of $\boldsymbol{\alpha}$ (i.e., the estimate of $\boldsymbol{\alpha}$ given by the Scoring estimation approach at iteration K where the convergence is reached).

Because $\mathbf{I}_{\mathbf{bb}}^{-1}(\widehat{\boldsymbol{\alpha}}^{(k)})$ and $\mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}(\widehat{\boldsymbol{\alpha}}^{(k)})$ are involved in the updating equations of the Scoring estimation approach, the asymptotic standard error estimates are by-products of the approach and can be obtained easily after the final iteration K.

2.5.2 Implementing the Scoring estimation approach

Algorithm 2 illustrates the implementation procedure of the Scoring estimation approach. The convergence criterion can be defined either as absolute distance or relative distance between estimate $\widehat{\boldsymbol{\alpha}}^{(k+1)}$ and $\widehat{\boldsymbol{\alpha}}^{(k)}$. According to Golub and Van Loan (2012), the absolute convergence criterion in q-norm can be defined as

$$\kappa_{\text{abs}} = \|\widehat{\boldsymbol{\alpha}}^{(k+1)} - \widehat{\boldsymbol{\alpha}}^{(k)}\|_q \,.$$

However, when magnitudes of model parameters in $\boldsymbol{\alpha}$ differ widely, a sufficient low tolerance level is required to achieve a satisfactory accuracy at the cost of speed. In such a case and if $\widehat{\boldsymbol{\alpha}}^{(k)} \neq \mathbf{0}$, the relative convergence criterion in q-norm defined by

$$\kappa_{\text{rel}} = \frac{\|\widehat{\boldsymbol{\alpha}}^{(k+1)} - \widehat{\boldsymbol{\alpha}}^{(k)}\|_q}{\|\widehat{\boldsymbol{\alpha}}^{(k)}\|_q}$$

is preferred. The choice of tolerance levels for κ_{abs} and κ_{rel} depends on problems under consideration and trade-offs between accuracy and speed.

Algorithm 2 Scoring estimation approach

Require: $\mathbf{Y}_i$, $\mathbf{X}_{ij}$ and $\mathbf{s}_{ij}$ for $i \in \{1, \ldots, N\}$ and $j \in \{1, \ldots, n_i\}$.
Ensure: Estimates of $\mathbf{b}$ and $\boldsymbol{\theta}$ with corresponding asymptotic standard error estimates.
1: Initialisation: choose values for $\widehat{\mathbf{b}}^{(1)}$ and $\widehat{\boldsymbol{\theta}}^{(1)}$;
2: **repeat**
3: Update the estimate of $\boldsymbol{\alpha} = (\mathbf{b}^\top, \boldsymbol{\theta}^\top)^\top$ by equations (2.14) and (2.15);
4: **until** the convergence criterion is met;
5: Obtain estimates of asymptotic standard errors of $\widehat{\mathbf{b}}$ and $\widehat{\boldsymbol{\theta}}$ by equations (2.16) and (2.17).

2.5.3 Numerical considerations

Many ground-motion prediction functions contain both linear and nonlinear parameters in $\mathbf{b}$. When the dimension of $\mathbf{b}$ is large, it can be more computationally effective to separate linear and nonlinear parameters and update their estimates separately to make the Scoring estimation approach better-conditioned and faster to maximise the log-likelihood function. This can be achieved in many families of ground-motion prediction functions, which contain combinations of linear and nonlinear components in the parameters.

To carry out updates for the linear and nonlinear parameter estimates separately in the Scoring estimation approach (named separable Scoring estimation approach thereafter), the ground-motion prediction function $\mathbf{f}(\mathbf{X}_i, \mathbf{b})$ is decomposed as

$$\mathbf{f}(\mathbf{X}_i, \mathbf{b}) = \mathbf{g}(\mathbf{X}_i, \boldsymbol{\gamma})\boldsymbol{\beta}\,, \tag{2.18}$$

in which $\boldsymbol{\beta} \in \mathbb{R}^{p_{11}}$ represents a vector of linear parameters in $\mathbf{b}$ with its design matrix $\mathbf{g}(\mathbf{X}_i, \boldsymbol{\gamma})$ and $\boldsymbol{\gamma} \in \mathbb{R}^{p_{12}}$ is a vector of the nonlinear parameters in $\mathbf{b}$.

Let $\boldsymbol{\alpha} = (\boldsymbol{\gamma}^\top, \boldsymbol{\beta}^\top, \boldsymbol{\theta}^\top)^\top$ and denote the gradient $\mathbf{S}(\boldsymbol{\alpha})$ and expected information matrix $\mathbf{I}(\boldsymbol{\alpha})$ of $l(\boldsymbol{\alpha})$ by the partitions

$$\mathbf{S}(\boldsymbol{\alpha}) = \begin{bmatrix} \mathbf{S}_{\boldsymbol{\gamma}}(\boldsymbol{\alpha}) \\ \mathbf{S}_{\boldsymbol{\beta}}(\boldsymbol{\alpha}) \\ \mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\alpha}) \end{bmatrix} \quad \text{and} \quad \mathbf{I}(\boldsymbol{\alpha}) = \begin{bmatrix} \mathbf{I}_{\boldsymbol{\gamma\gamma}}(\boldsymbol{\alpha}) & \mathbf{I}_{\boldsymbol{\gamma\beta}}(\boldsymbol{\alpha}) & \mathbf{I}_{\boldsymbol{\gamma\theta}}(\boldsymbol{\alpha}) \\ \mathbf{I}_{\boldsymbol{\beta\gamma}}(\boldsymbol{\alpha}) & \mathbf{I}_{\boldsymbol{\beta\beta}}(\boldsymbol{\alpha}) & \mathbf{I}_{\boldsymbol{\beta\theta}}(\boldsymbol{\alpha}) \\ \mathbf{I}_{\boldsymbol{\theta\gamma}}(\boldsymbol{\alpha}) & \mathbf{I}_{\boldsymbol{\theta\beta}}(\boldsymbol{\alpha}) & \mathbf{I}_{\boldsymbol{\theta\theta}}(\boldsymbol{\alpha}) \end{bmatrix}.$$

Then, the updating equations are given by Theorem 2.2.

Theorem 2.2 *The updating equations for the separable Scoring estimation approach are given by*

$$\widehat{\boldsymbol{\gamma}}^{(k+1)} = \widehat{\boldsymbol{\gamma}}^{(k)} + \left(\mathbf{I}_{\boldsymbol{\gamma\gamma}}(\widehat{\boldsymbol{\alpha}}^{(k)}) - \mathbf{I}_{\boldsymbol{\gamma\beta}}(\widehat{\boldsymbol{\alpha}}^{(k)})\,\mathbf{I}^{-1}_{\boldsymbol{\beta\beta}}(\widehat{\boldsymbol{\alpha}}^{(k)})\,\mathbf{I}_{\boldsymbol{\beta\gamma}}(\widehat{\boldsymbol{\alpha}}^{(k)})\right)^{-1}\mathbf{S}_{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\alpha}}^{(k)})\,, \tag{2.19}$$

$$\widehat{\boldsymbol{\theta}}^{(k+1)} = \widehat{\boldsymbol{\theta}}^{(k)} + \mathbf{I}^{-1}_{\boldsymbol{\theta\theta}}(\widehat{\boldsymbol{\alpha}}^{(k)})\,\mathbf{S}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\alpha}}^{(k)})\,, \tag{2.20}$$

$$\widehat{\boldsymbol{\beta}}^{(k+1)} = \mathbf{I}^{-1}_{\boldsymbol{\beta\beta}}(\widehat{\boldsymbol{\gamma}}^{(k+1)}, \widehat{\boldsymbol{\theta}}^{(k+1)})\left[\mathbf{g}^\top(\mathbf{X}, \widehat{\boldsymbol{\gamma}}^{(k+1)})\mathbf{C}^{-1}\left(\widehat{\boldsymbol{\theta}}^{(k+1)}\right)\mathbf{Y}\right], \tag{2.21}$$

in which $\mathbf{g}(\mathbf{X}, \boldsymbol{\gamma}) = \left(\mathbf{g}(\mathbf{X}_1, \boldsymbol{\gamma})^\top, \ldots, \mathbf{g}(\mathbf{X}_N, \boldsymbol{\gamma})^\top\right)^\top$ *and*

- *the i-th element of* $\mathbf{S}_{\boldsymbol{\gamma}}(\boldsymbol{\alpha})$ *is given by*

$$[\mathbf{S}_{\boldsymbol{\gamma}}(\boldsymbol{\alpha})]_i = \left[\frac{\partial \mathbf{g}(\mathbf{X}, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}_i}\boldsymbol{\beta}\right]^\top \mathbf{C}^{-1}(\boldsymbol{\theta})[\mathbf{Y} - \mathbf{g}(\mathbf{X}, \boldsymbol{\gamma})\boldsymbol{\beta}]\,;$$

- *The i-th element of* $\mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\alpha})$ *is given by*

$$\begin{aligned}[\mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\alpha})]_i =& -\frac{1}{2}\mathrm{tr}\left\{\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i}\right\} \\ &+ \frac{1}{2}[\mathbf{Y} - \mathbf{g}(\mathbf{X}, \boldsymbol{\gamma})\boldsymbol{\beta}]^\top \mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i}\mathbf{C}^{-1}(\boldsymbol{\theta})[\mathbf{Y} - \mathbf{g}(\mathbf{X}, \boldsymbol{\gamma})\boldsymbol{\beta}]\,;\end{aligned}$$

- $\mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\alpha})$ *is given by*

$$\mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\alpha}) = \mathbf{g}(\mathbf{X}, \boldsymbol{\gamma})^\top \mathbf{C}^{-1}(\boldsymbol{\theta})\mathbf{g}(\mathbf{X}, \boldsymbol{\gamma})\,;$$

- *the ij-th element of* $\mathbf{I}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}(\boldsymbol{\alpha})$ *is given by*

$$[\mathbf{I}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}(\boldsymbol{\alpha})]_{ij} = \left[\frac{\partial \mathbf{g}(\mathbf{X}, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}_i}\boldsymbol{\beta}\right]^\top \mathbf{C}^{-1}(\boldsymbol{\theta})\,\frac{\partial \mathbf{g}(\mathbf{X}, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}_j}\boldsymbol{\beta}\,;$$

- *the ij-th element of* $\mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\alpha})$ *is given by*

$$[\mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\alpha})]_{ij} = \frac{1}{2}\mathrm{tr}\left\{\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i}\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j}\right\}\,;$$

- *the i-th row of* $\mathbf{I}_{\boldsymbol{\gamma}\boldsymbol{\beta}}(\boldsymbol{\alpha})$ $\left(\textit{or the i-th column of } \mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\gamma}}(\boldsymbol{\alpha})\right)$ *is given by*

$$[\mathbf{I}_{\boldsymbol{\gamma}\boldsymbol{\beta}}(\boldsymbol{\alpha})]_{i*} = [\mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\gamma}}(\boldsymbol{\alpha})]_{*i}^\top = \left[\frac{\partial \mathbf{g}(\mathbf{X}, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}_i}\boldsymbol{\beta}\right]^\top \mathbf{C}^{-1}(\boldsymbol{\theta})\mathbf{g}(\mathbf{X}, \boldsymbol{\gamma})\,.$$

Proof The proof is given in Section A.4 of Appendix A. □

It can be seen from equation (2.19)-(2.21), that after separating the linear and nonlinear parameters in ground-motion prediction functions via decomposition in equation (2.18), the Scoring estimation approach amounts to three updating equations in each iteration. The updating equation (2.21) for $\boldsymbol{\beta}$ has an analytical form given the estimates of $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ obtained from updating equations (2.19) and (2.20). The further separation of the update scheme caused by the isolation between linear and nonlinear parameters reduces the complexity of each iteration from $\mathcal{O}(p_1^3 + p_2^3)$ (in the ordinary Scoring estimation approach) to $\mathcal{O}(p_{11}^3 + p_{12}^3 +$

$p_2^3 + p_{11}^2 p_{12} + p_{12}^2 p_{11})$, in which $p_{11} + p_{12} = p_1$ and p_{11} and p_{12} are dimensions of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, respectively. Another advantage of the separable Scoring estimation approach is that the conditioning of the algorithm is improved because of the further separation. Finally, the separable Scoring estimation approach only requires initial values of $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ to be set because the initial value $\widehat{\boldsymbol{\beta}}^{(1)}$ of $\boldsymbol{\beta}$ can be obtained by (2.21) using $\widehat{\boldsymbol{\gamma}}^{(1)}$ and $\widehat{\boldsymbol{\theta}}^{(1)}$. Consequently, the convergence criterion is only required for $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$, implying that the convergence may be achieved with fewer iterations.

Define

$$\mathbf{I}_{-\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\alpha}) = \mathbf{I}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}(\boldsymbol{\alpha}) - \mathbf{I}_{\boldsymbol{\gamma}\boldsymbol{\beta}}(\boldsymbol{\alpha})\,\mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1}(\boldsymbol{\alpha})\,\mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\gamma}}(\boldsymbol{\alpha})$$

and apply block matrix inversion on equation (2.16), the asymptotic standard error estimates of $\widehat{\boldsymbol{\gamma}}$, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\theta}}$ are then given by

$$\widehat{\mathrm{se}}(\widehat{\boldsymbol{\gamma}}) = \sqrt{\mathrm{diag}\left[\mathbf{I}_{-\boldsymbol{\beta}\boldsymbol{\beta}}^{-1}(\widehat{\boldsymbol{\alpha}}^{(K)})\right]}, \tag{2.22}$$

$$\widehat{\mathrm{se}}(\widehat{\boldsymbol{\beta}}) = \sqrt{\mathrm{diag}\left[\mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1}(\widehat{\boldsymbol{\alpha}}^{(K)}) + \mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1}(\widehat{\boldsymbol{\alpha}}^{(K)})\,\mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\gamma}}(\widehat{\boldsymbol{\alpha}}^{(K)})\mathbf{I}_{-\boldsymbol{\beta}\boldsymbol{\beta}}^{-1}(\widehat{\boldsymbol{\alpha}}^{(K)})\,\mathbf{I}_{\boldsymbol{\gamma}\boldsymbol{\beta}}(\widehat{\boldsymbol{\alpha}}^{(K)})\,\mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1}(\widehat{\boldsymbol{\alpha}}^{(K)})\right]}, \tag{2.23}$$

$$\widehat{\mathrm{se}}(\widehat{\boldsymbol{\theta}}) = \sqrt{\mathrm{diag}\left[\mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}(\widehat{\boldsymbol{\alpha}}^{(K)})\right]}. \tag{2.24}$$

The Algorithm 3 outlines the implementation procedure for the separable Scoring estimation approach.

Algorithm 3 Separable Scoring estimation approach

Require: $\mathbf{Y}_i$, $\mathbf{X}_{ij}$ and $\mathbf{s}_{ij}$ for $i \in \{1, \ldots, N\}$ and $j \in \{1, \ldots, n_i\}$.
Ensure: Estimates of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ with corresponding asymptotic standard error estimates.
1: Initialisation:
   1) choose values for $\widehat{\boldsymbol{\gamma}}^{(1)}$ and $\widehat{\boldsymbol{\theta}}^{(1)}$;
   2) compute the value of $\widehat{\boldsymbol{\beta}}^{(1)}$ by equation (2.21);
2: **repeat**
3: Update the estimates of $\boldsymbol{\alpha} = (\boldsymbol{\gamma}^\top, \boldsymbol{\beta}^\top, \boldsymbol{\theta}^\top)^\top$ by equation (2.19) to (2.21);
4: **until** the convergence criterion is met;
5: Obtain the asymptotic standard error estimate of $\widehat{\boldsymbol{\alpha}}$ by equation (2.22) to (2.24).

Although the separable Scoring estimation approach is generally fast to converge

and numerically stable, it can be improved to further speed up the computation and reduce the chances of numerical errors. For example, we can perform inexact line search to promote the convergence by adding a step length $\varphi^{(k)}$ to the updating equation (2.12) of the Scoring estimation approach:

$$\widehat{\boldsymbol{\alpha}}^{(k+1)} = \widehat{\boldsymbol{\alpha}}^{(k)} + \varphi^{(k)} \mathbf{I}^{-1}(\widehat{\boldsymbol{\alpha}}^{(k)}) \mathbf{S}(\widehat{\boldsymbol{\alpha}}^{(k)})$$

and identify an appropriate value of $\varphi^{(k)}$ at each iteration k such that the log-likelihood function value is increased adequately at minimum cost. Desirable values for step lengths can be searched by algorithms that terminate upon certain conditions, such as the Wolfe conditions (Wolfe, 1969, 1971). For details of the inexact line search, its implementation algorithms as well as other optimisation techniques that may be applied to improve the numerical performances of the Scoring estimation approach, readers can refer to Gill et al. (1981) and Nocedal and Wright (2006).

2.6 Simulation Study

The purpose of this section is to quantify and compare the performances of the multi-stage algorithm and the Scoring estimation approach. The performance of an estimation method can be measured by the accuracy of the obtained model parameter estimates and the resulting predictions. However, this requires knowledge about the true underlying model that is unknown in reality, causing the evaluation of an estimation method difficult in terms of its true performance. To resolve this issue, simulation studies can be implemented. Simulation studies are synthetic experiments conducted on computers under planned conditions, meaning that the generator of the ground-motion data (i.e., the true underlying ground-motion model and its parameter values) is chosen by experimenters and thus fully informative. As a result, the performance of an estimation method can be tested. Simulation studies have been used previously in earthquake modelling in work such as Chen and Tsai (2002); Arroyo and Ordaz (2010); Worden et al. (2018).

2.6.1 Generator settings

The first step of the simulation study is to specify the underlying generator (i.e., the true ground-motion model) of the considered IM. Specifically, in this simulation study, PGA is used as the considered ground-motion IM. To eliminate the effects of model misspecification, the true ground-motion model is chosen to have the same model representation as the hypothetical ground-motion model specified in Section 2.3 with the ground-motion prediction function (proposed by Akkar and Bommer (2010)):

$$f(\mathbf{X}_{ij},\, \mathbf{b}) = b_1 + b_2\, M_i + b_3\, M_i^2 + (b_4 + b_5\, M_i) \log_{10} \sqrt{R_{ij}^2 + b_6^2} + b_7\, S_{S,ij} + b_8\, S_{A,ij} + b_9\, F_{N,i} + b_{10}\, F_{R,i}\,, \quad (2.25)$$

in which

- M_i is the moment magnitude (M_W) of earthquake i;
- R_{ij} is the Joyner-Boore distance (R_{JB}) (i.e., the closest distance to the surface projection of the rupture plane) in kilometres of site j in earthquake i;
- $S_{S,ij}$ and $S_{A,ij}$ are dummy variables determining the soil type at site j during earthquake i according to

$$(S_{S,ij},\, S_{A,ij}) = \begin{cases} (1,\, 0)\,, & \text{soft soil,} \\ (0,\, 1)\,, & \text{stiff soil,} \\ (0,\, 0)\,, & \text{rock;} \end{cases}$$

- $F_{N,i}$ and $F_{R,i}$ are dummy variables indicating the faulting type of earthquake i according to

$$(F_{N,i},\, F_{R,i}) = \begin{cases} (1,\, 0)\,, & \text{normal fault,} \\ (0,\, 1)\,, & \text{reverse fault,} \\ (0,\, 0)\,, & \text{strike-slip fault.} \end{cases}$$

Two kernel functions are selected for illustrative purposes:

$$k_1(d) = \exp\left(-\frac{d}{h}\right) \tag{2.26}$$

and

$$k_2(d) = \left(1 + \frac{\sqrt{3}d}{h}\right)\exp\left(-\frac{\sqrt{3}d}{h}\right), \tag{2.27}$$

which are special cases of Matérn kernel with $\nu = 0.5$ and $\nu = 1.5$, respectively. The first kernel function (2.26) (i.e., exponential kernel function) represents a type of spatial correlation structure that is commonly used in works such as Jayaram and Baker (2009, 2010); Esposito and Iervolino (2011, 2012) and allows for an instructive comparison between the two estimation methods. The second kernel function (2.27) is smoother than the kernel function (2.26) and admits the comparison between the two estimation approaches when the logarithmic PGA field is smooth.

The parameter values in the true ground-motion model are outlined in Table 2.1. The values for $b_1, \ldots, b_{10}$, τ^2 and σ^2 are chosen based on the regression results given by Akkar and Bommer (2010) for the ground-motion model of PGA. The value of the range parameter h in the kernel function (2.26) is set arbitrarily to 11.5 km. This value of h corresponds to $d = 34.45$ km when $\rho = 0.05$ with the kernel function (2.26). To get the same ρ value at the same distance $d = 34.45$ km, it is found that $h = 12.58$ km for the kernel function (2.27).

2.6.2 Choice for covariates

Before synthetic PGA datasets can be generated, the information of covariates needs to be known. The information of covariates includes the number of earthquakes N, the number of recording sites n_i during each event (i.e., earthquake) as well as their locations $\mathbf{s}_{ij}$, and the values of predictors

$$\mathbf{X}_{ij} = (M_i,\, R_{ij},\, S_{S,ij},\, S_{A,ij},\, F_{N,i},\, F_{R,i}).$$

In this simulation study, the information of covariates is extracted from a historical ground-motion database, the European Strong-Motion (ESM) database,

Table 2.1: Parameter values chosen for the assumed true ground-motion model

Parameter	Value	Parameter	Value
b_1	1.0416	b_8	0.0153
b_2	0.9133	b_9	-0.0419
b_3	-0.0814	b_{10}	0.0802
b_4	-2.9273	τ^2	0.0099
b_5	0.2812	σ^2	0.0681
b_6	7.8664	$h\,(\nu = 0.5)^*$	11.50 km
b_7	0.0875	$h\,(\nu = 1.5)^\dagger$	12.58 km

* The range parameter h in the kernel function (2.26) (i.e., Matérn with $\nu = 0.5$).
† The range parameter h in the kernel function (2.27) (i.e., Matérn with $\nu = 1.5$).

which ensures the generation of realistic scenarios for comparison of the two estimation methods. In using this database, we apply to the database the selection criteria detailed below so that the proposed simulation study can be independently verified and reproduced:

- retain events occurred within Italy;
- retain events with moment magnitude $M_W \geq 5$, removing events without M_W information;
- remove events without information of fault types;
- retain recording sites with epicentral distance $R_{epi} \leq 250$ km;
- remove recording sites without information of V_{S30} , the average shear-wave velocity (in m/s) in the upper 30 meters of the soil;
- remove recording sites that are not free-field;
- remove recording sites with redundant site information (e.g., co-located recording sites) in a single event; and
- retain events with at least two recording sites.

After the implementation of the above selection criteria, the resulting catalogue used in this simulation study consists of 2150 entries of recording sites (in which the same recording site may appear in different earthquakes) from 62 earthquakes of $5 \leq M_W \leq 6.9$ in Italy from 1976 to 2016. The geographical distribution of the 62 earthquakes with their moment magnitudes, and the distribution of inter-site distance in each earthquake are shown in Figure 2.5.

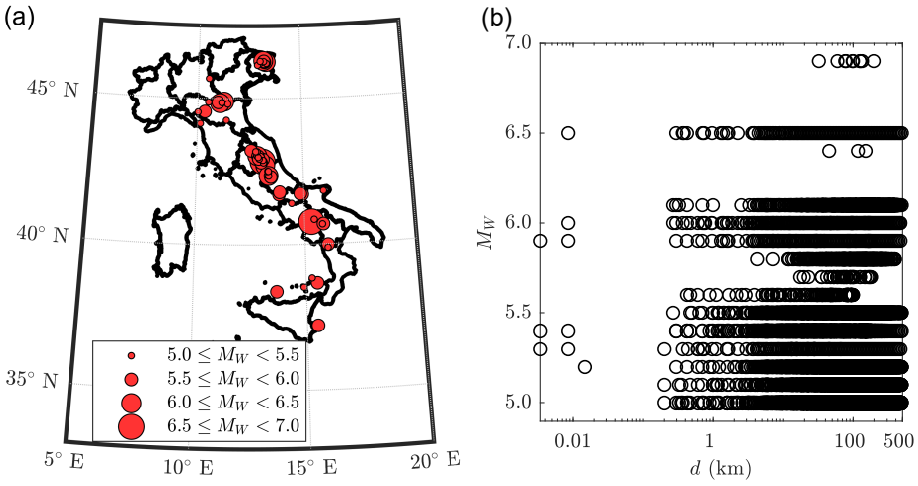


Figure 2.5: (a) The geographical distribution of 62 earthquakes of $5 \leq M_W \leq 6.9$ in Italy from 1976 to 2016. The epicentre of each event is labelled by a filled circle (∘), whose size is scaled by the moment magnitude (M_W) of the event. (b) The distribution of inter-site distance in each earthquake (represented by its corresponding moment magnitude) on a log scale.

The R_{JB} of each recording site in each earthquake is calculated based on the corresponding fault geometry (e.g., strike angle, dip angle, rake angle, length, and width), if information of the finite-fault model is available. Otherwise, R_{JB} is estimated by the empirical relationship between R_{epi} and R_{JB} (Stucchi et al., 2011) if the corresponding earthquake is with $M_W > 5.5$ and is set to be R_{epi} if the corresponding earthquake is with $M_W \leq 5.5$. The obtained R_{JB} for each recording site of each earthquake in the resulting catalogue for this simulation study is less than 250 km. The site classification of each recording site in each earthquake is obtained based on the information of V_{S30} from the ESM database. In ESM database, V_{S30} is either obtained from in-situ experiments or inferred from the topographic slope according to Wald and Allen (2007). It is preferable to use V_{S30} from the experimental measurements, and if that is not

available, the inferred V_{S30} is used instead. The soil type of each recording site of each earthquake in the catalogue for this simulation study is then classified (according to Akkar and Bommer (2010)) as soft soil if $V_{S30} < 360$ m/s, stiff soil if 360 m/s $\leq V_{S30} \leq 750$ m/s, and rock if $V_{S30} > 750$ m/s.

2.6.3 PGA data generation

Given the true ground-motion model and information of covariates, we can simulate synthetic datasets of logarithmic PGAs through Algorithm 4.

Algorithm 4 Synthetic logarithmic PGA dataset generation

Require: Specified true ground-motion model and information of covariates.
Ensure: A synthetic dataset of logarithmic PGAs (denoted by $\mathbf{y}$).
1: Compute the covariance matrix $\mathbf{C}(\boldsymbol{\theta})$ where $\boldsymbol{\theta} = (\tau^2, \sigma^2, h)^\top$;
2: Compute the Cholesky factor $\mathbf{L}$ such that $\mathbf{L}\mathbf{L}^\top = \mathbf{C}(\boldsymbol{\theta})$;
3: Compute the value of $\mathbf{f}(\mathbf{X}, \mathbf{b})$;
4: Generate independently $G = \sum_{i=1}^{N} n_i$ standard normal random numbers $\mathbf{v} = (v_1, \ldots, v_G)^\top$;
5: Return a synthetic dataset of logarithmic PGAs by $\mathbf{y} = \mathbf{f}(\mathbf{X}, \mathbf{b}) + \mathbf{L}\mathbf{v}$.

2.6.4 Evaluation of the estimation performance

In this section, estimation performances of the multi-stage algorithm and the Scoring estimation approach are evaluated and compared. We first generate $\mathcal{T} = 1000$ synthetic datasets of logarithmic PGAs via Algorithm 4. Then for each of the synthetic dataset, the multi-stage algorithm and the Scoring estimation approach are implemented. Let $\widehat{\alpha}_t$ and $\widehat{\text{se}}(\widehat{\alpha}_t)$ represent, respectively, the estimate and the asymptotic standard error estimate of a model parameter $\alpha \in \{\mathbf{b}, \tau^2, \sigma^2, h\}$ produced by one of the two estimation methods on some synthetic dataset $t \in \{1, \ldots, \mathcal{T}\}$. The estimation performance of either method then can be evaluated by computing the following criteria:

- **root mean squared error (RMSE)**, computed by

$$\text{RMSE} = \sqrt{\frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} (\widehat{\alpha}_t - \alpha_0)^2},$$

in which α_0 is the true parameter value (given in Table 2.1) of α;

- **coverage rate (CR)**, defined by the percentage of $\mathcal{T}$ synthetic datasets in which the true parameter value α_0 falls into the 95% confidence interval constructed from $\widehat{\alpha}_t$ and $\widehat{\text{se}}(\widehat{\alpha}_t)$.

Table 2.2 illustrates the estimation criteria of the parameter estimators produced by the multi-stage algorithm and the Scoring estimation approach under the kernel functions (2.26) and (2.27). It can be observed that the RMSEs of all parameter estimators from the Scoring estimation approach are less than those from the multi-stage algorithm under both types of kernel functions. Although the RMSEs of estimators of $b_1, \ldots, b_{10}$ produced by the multi-stage algorithm are not significantly higher than those produced by the Scoring estimation approach, the RMSEs of $\widehat{\tau^2}$, $\widehat{\sigma^2}$ and $\widehat{h}$ are noticeably different between the two methods. For $\widehat{\tau^2}$, the multi-stage algorithm produces 50% higher RMSE than the Scoring estimation approach under the kernel function (2.26) and two times larger RMSE than the Scoring estimation approach under the kernel function (2.27). With regard to $\widehat{\sigma^2}$, the RMSE from the multi-stage algorithm is around eight times larger than that from the Scoring estimation approach under the kernel function (2.26) and more than 30 times larger than that from the Scoring estimation approach under the kernel function (2.27). Similar observations can be seen regarding the estimator of h, whose RMSE from the multi-stage algorithm is 12 times higher than that from the Scoring estimation approach under the kernel function (2.26) and about 26 times larger than that from the Scoring estimation approach under the kernel function (2.27). These findings imply that the estimators, particularly the estimators of τ^2, σ^2, and h, given by the Scoring estimation approach are more robust.

Finally, it can be found that the CRs under the Scoring estimation approach are relatively stable across different model parameters, the CRs for τ^2, σ^2, and h under the multi-stage algorithm are remarkably lower than the expected 95% confidence level, indicating that the constructed confidence interval from

Table 2.2: Comparison of the estimation performance between the multi-stage algorithm and the Scoring estimation approach

	Multi-Stage Algorithm*				Scoring Estimation Approach			
	$\nu = 0.5^{\dagger}$		$\nu = 1.5^{\ddagger}$		$\nu = 0.5$		$\nu = 1.5$	
	RMSE[§]	CR[‖]	RMSE	CR	RMSE	CR	RMSE	CR
b_1	2.5540	94.8	2.8160	95.7	2.5156	94.4	2.6551	92.8
b_2	0.8875	94.0	0.9795	94.8	0.8749	94.0	0.9234	92.8
b_3	0.0780	93.6	0.0862	94.0	0.0769	93.6	0.0811	92.3
b_4	0.3184	98.3	0.3529	99.9	0.3013	94.4	0.3071	95.0
b_5	0.0573	98.4	0.0634	99.8	0.0541	93.9	0.0551	94.7
b_6	0.8631	96.6	0.9250	89.7	0.8438	95.9	0.8092	94.3
b_7	0.0158	93.3	0.0055	80.3	0.0154	95.3	0.0054	94.5
b_8	0.0087	91.6	0.0017	83.3	0.0085	94.3	0.0016	96.2
b_9	0.0661	92.4	0.0723	92.9	0.0649	92.4	0.0651	92.8
b_{10}	0.0712	91.2	0.0740	92.9	0.0701	91.0	0.0683	92.7
τ^2	0.0052	51.3	0.0076	26.5	0.0034	88.9	0.0035	89.2
σ^2	0.0197	1.6	0.0790	0.0	0.0025	94.2	0.0026	94.9
h	8.6122	0.2	9.8763	0.0	0.7582	93.7	0.3773	94.3

* Jayaram and Baker (2010).
† Corresponding to the kernel function (2.26) (i.e., Matérn type with $\nu = 0.5$) with $h = 11.50$ km.
‡ Corresponding to the kernel function (2.27) (i.e., Matérn type with $\nu = 1.5$) with $h = 12.58$ km.
§ Root mean squared error of the corresponding parameter estimator.
‖ Coverage rate (in percentage and rounded to one decimal place) of the corresponding parameter.

the multi-stage algorithm is biased in a non-conservative manner, that is, too narrow on average, and there exist risks of wrong decisions on hypothesis tests relating to model structure for the resulting GMPE under such an estimation procedure. The low CRs of τ^2 and σ^2 are partly due to the non-optimal formulas of asymptotic standard error estimates given by Jayaram and Baker (2010) and partly due to the separate estimation of h and the inconsistency of $\widehat{h}$. The low CR of h is because of the naive use of the asymptotic standard error formula for ordinary least squares and the inconsistency of $\widehat{\sigma^2}$ produced from the preliminary stage.

To examine how the estimation performances of the multi-stage algorithm

and the Scoring estimation approach change, when the sample (i.e., event) size N varies, we extract two sub-catalogues from the full catalogue described in Section 2.6.2. One sub-catalogue has the size of $N = 46$, which includes the events occurred by the end of the year 2010. Another sub-catalogue has the size of $N = 29$, which includes the events occurred by the end of the year 2000. We then generate 1000 synthetic datasets of logarithmic PGAs for both sub-catalogues and implement the multi-stage algorithm and the Scoring estimation approach, which provides 1000 sets of estimates for each sub-catalogue under each estimation method. Figure 2.6 and 2.7 present the sampling distributions of $\widehat{b}_1, \ldots, \widehat{b}_{10}$ under kernel function (2.26) and (2.27), respectively. As we expected in Section 2.4.4, both the multi-stage algorithm and the Scoring estimation approach produce consistent estimators of $b_1, \ldots, b_{10}$ (i.e., the sampling distributions of $\widehat{b}_1, \ldots, \widehat{b}_{10}$ converge to the true parameter values as N increases).

We emphasise in Section 2.4.4 that $\widehat{\tau^2}$, $\widehat{\sigma^2}$, and $\widehat{h}$ produced by the multi-stage algorithm are inconsistent, meaning that the sampling distribution of $\widehat{\tau^2}$, $\widehat{\sigma^2}$, and $\widehat{h}$ from the multi-stage algorithm will not converge to the true parameter values as N grows. This statement is illustrated in Figure 2.8. Under both the kernel function (2.26) and (2.27), the sampling distributions of $\widehat{\tau^2}$, $\widehat{\sigma^2}$, and $\widehat{h}$ produced by the Scoring estimation approach converge to the true parameter values as N increases. In contrast, the sampling distributions of $\widehat{\tau^2}$, $\widehat{\sigma^2}$, and $\widehat{h}$ produced by the multi-stage algorithm are biased. Moreover, the sampling distributions of $\widehat{\tau^2}$ and $\widehat{\sigma^2}$ produced by the multi-stage algorithm under the kernel function (2.27) behave worse than those under the kernel function (2.26) because increasing sampling variances and a larger number of outliers are observed.

2.6.5 Evaluation of the predictive performance

The estimated ground-motion models allow one to perform ground-motion predictions at locations where recording sites are unavailable (e.g., generate

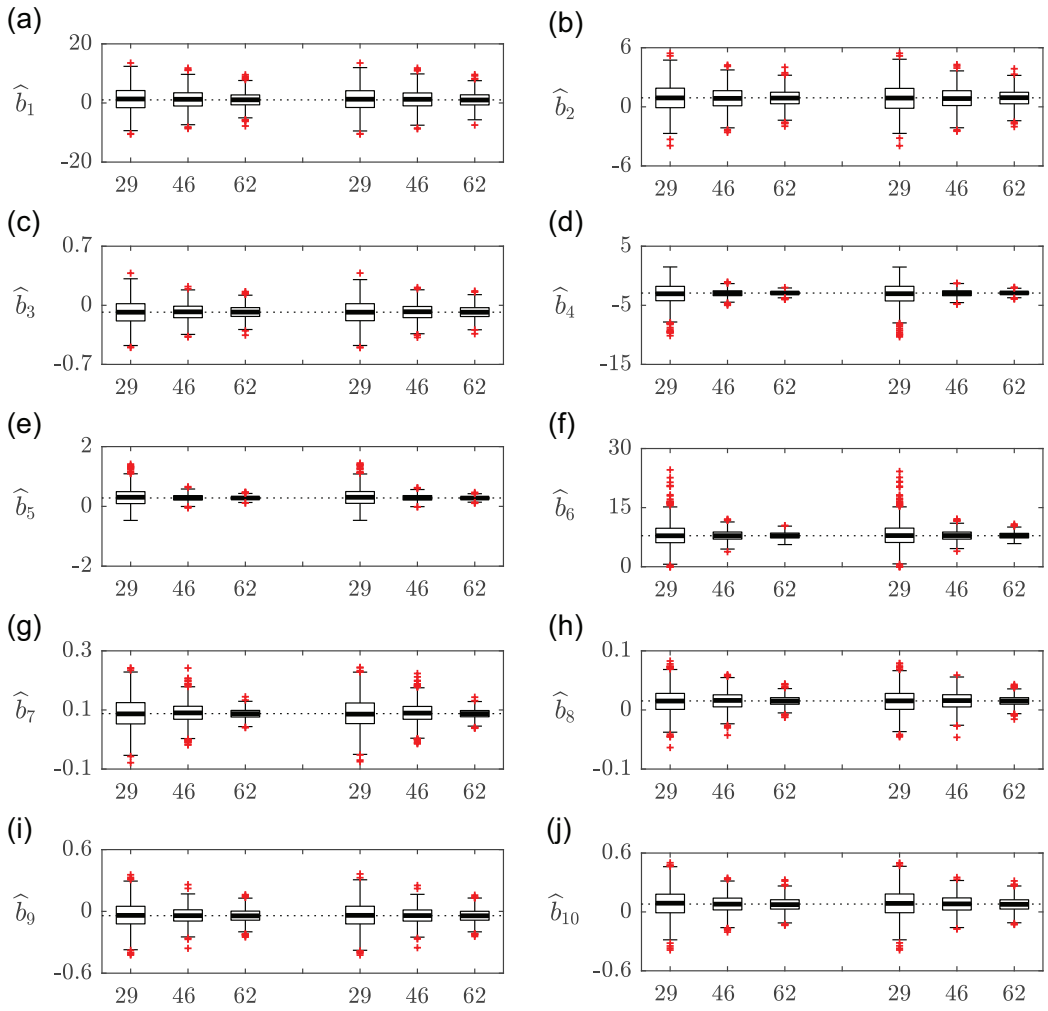


Figure 2.6: Sampling distributions for estimators $\widehat{b}_1, \ldots, \widehat{b}_{10}$ under the kernel function (2.26) with $h = 11.50$ km. The left three boxplots (reading from left to right) in each panel correspond to event sizes of $N = 29$, 46, and 62 under the multi-stage algorithm, respectively; the right three boxplots (reading from left to right) in each panel correspond to event sizes of $N = 29$, 46, and 62 under the Scoring estimation approach, respectively; the three event sizes correspond to events by the end of the year 2000, 2010, and 2016, respectively. The dashed line in each panel represents the true parameter value.

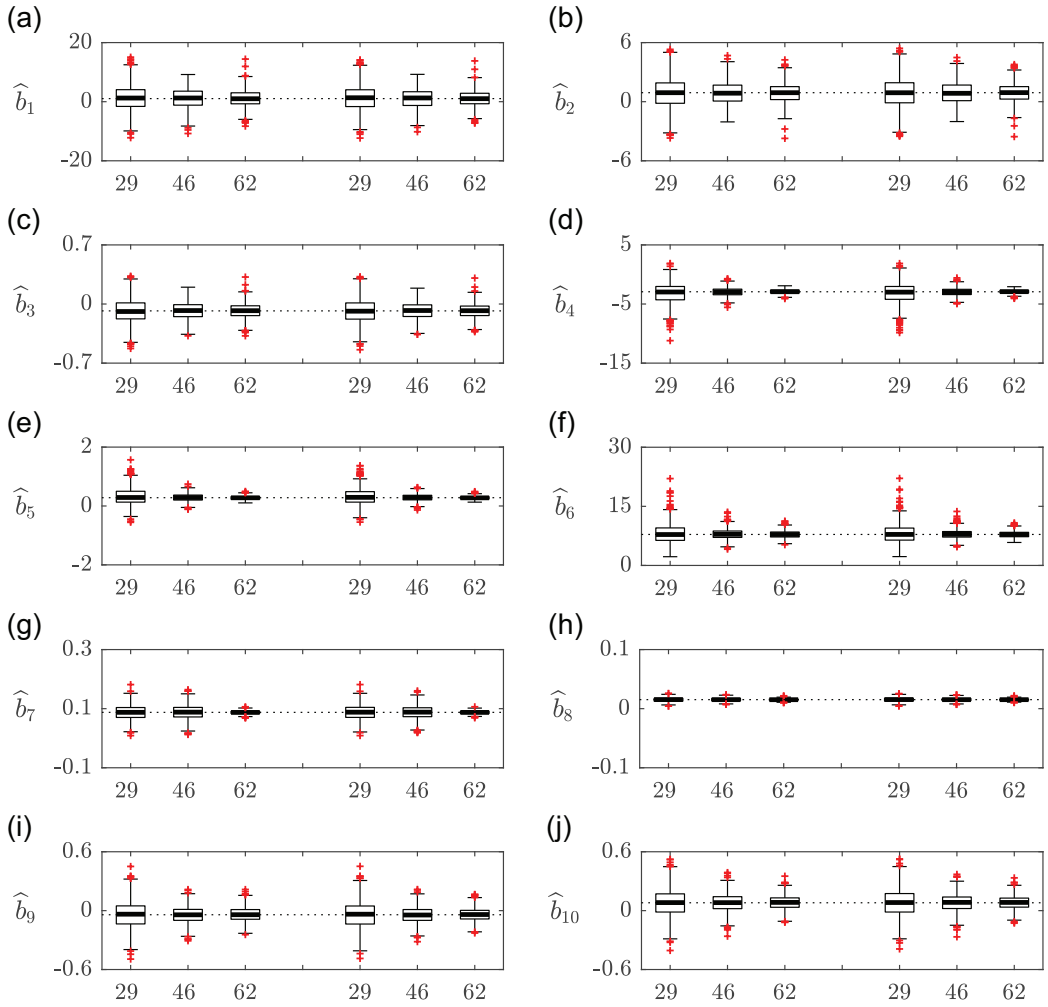


Figure 2.7: Sampling distributions for estimators $\widehat{b}_1, \ldots, \widehat{b}_{10}$ under the kernel function (2.27) with $h = 12.58$ km. The left three boxplots (reading from left to right) in each panel correspond to event sizes of $N = 29$, 46, and 62 under the multi-stage algorithm, respectively; the right three boxplots (reading from left to right) in each panel correspond to event sizes of $N = 29$, 46, and 62 under the Scoring estimation approach, respectively; the three event sizes correspond to events by the end of the year 2000, 2010, and 2016, respectively. The dashed line in each panel represents the true parameter value.

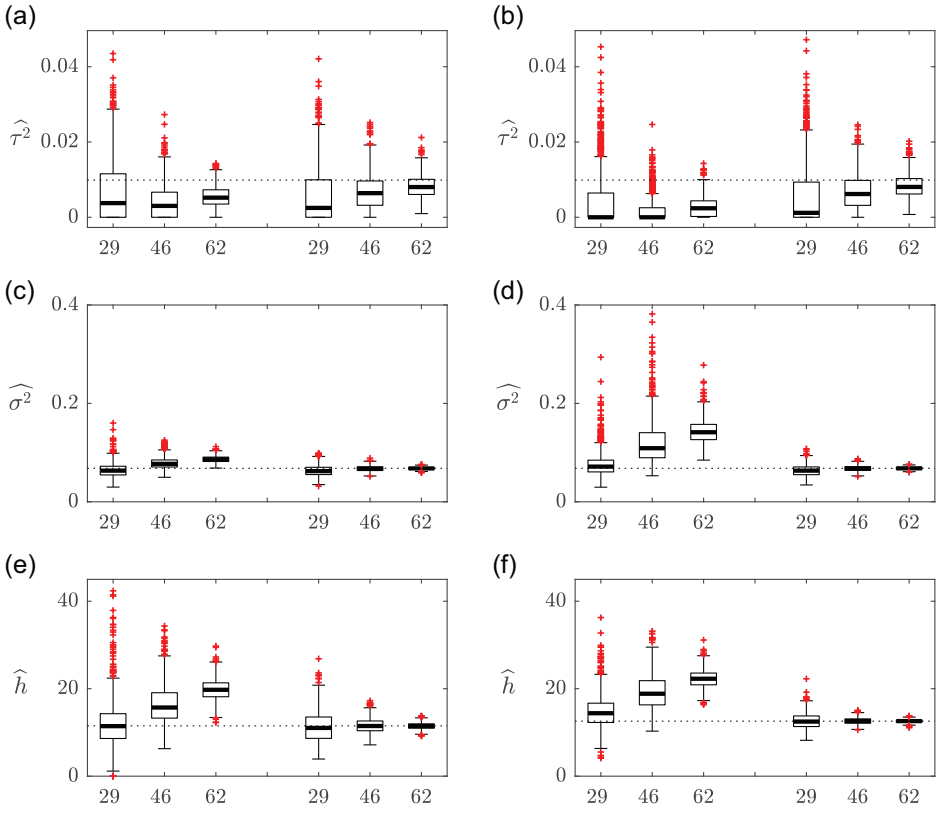


Figure 2.8: Sampling distributions for estimators $\widehat{\tau^2}$, $\widehat{\sigma^2}$, and $\widehat{h}$: (a), (c) and (e) correspond to the kernel function (2.26) with $h = 11.50$ km; (b), (d) and (f) correspond to the kernel function (2.27) with $h = 12.58$ km. The left three boxplots (reading from left to right) in each panel correspond to event sizes of $N = 29$, 46, and 62 under the multi-stage algorithm, respectively; the right three boxplots (reading from left to right) in each panel correspond to event sizes of $N = 29$, 46, and 62 under the Scoring estimation approach, respectively; the three event sizes correspond to events by the end of the year 2000, 2010, and 2016, respectively. The dashed line in each panel represents the true parameter value.

a ground-motion shaking intensity map). Therefore, it is vital to assess the predictive performances of the ground-motion models estimated by the multi-stage algorithm and the Scoring estimation approach. To this goal, we examine the prediction accuracy for a selected event with ID 'IT-1997-0137', which corresponds to the earthquake with $M_W = 5.6$ occurred in the regions of Umbria and Marche in 1997 and has $n_e = 15$ recording sites. This particular event is selected because it is included in both the full catalogue (events by the end of the year 2016) and the two sub-catalogues (events by the end of the

year 2000 and 2010) described in Section 2.6.4. This allows us to examine how the predictive performance of an estimation method changes as the number of events used for estimation varies. The prediction region of the event is set to be within a distance of 250 km from the epicentre (see Figure 2.9). The ground-motion models used for predictions are those estimated from the full catalogue and the two sub-catalogues in Section 2.6.4.

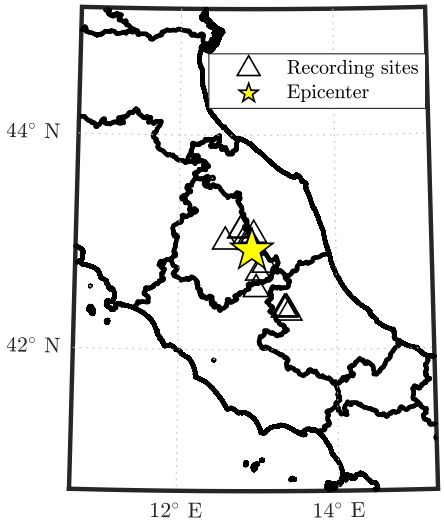


Figure 2.9: The region (within a distance of 250 km from the epicentre) of the selected event with ID 'IT-1997-0137'. The epicentre of the event is labelled by a filled star (☆); triangles (△) represent the recording sites whose logarithmic peak ground acceleration (PGA) records (generated in Section 2.6.4) are observed and used for predictions.

We first discretise the prediction region of the event by fine square grids with mesh size $\Delta = 5$ km and treat the resulting $K = 5228$ grid points as prediction locations. Then, for each estimation method and each catalogue (i.e., the full catalogue and the two sub-catalogues) we proceed with the following steps:

1. For each synthetic dataset t, compute the predictions $\widehat{\mathbf{z}}_t = (\widehat{z}_{1,t}, \ldots, \widehat{z}_{K,t})$ on all grid points $k \in \{1, \ldots, K\}$ by the plug-in predictor (Stein, 1999)

$$\widehat{\mathbf{z}}_t = \mathbf{f}(\mathbf{W}, \widehat{\mathbf{b}}_t) + \boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}}_t)\mathbf{c}^{-1}(\widehat{\boldsymbol{\theta}}_t)\left(\mathbf{y}_t - \mathbf{f}(\mathbf{X}_e, \widehat{\mathbf{b}}_t)\right), \tag{2.28}$$

   in which

- $\widehat{\mathbf{b}}_t$ and $\widehat{\boldsymbol{\theta}}_t = (\widehat{\tau^2}_t, \widehat{\sigma^2}_t, \widehat{h}_t)$ are parameter estimates obtained from synthetic dataset t;
- $\mathbf{f}(\mathbf{W}, \widehat{\mathbf{b}}_t) = (\mathbf{f}(\mathbf{W}_1, \widehat{\mathbf{b}}_t), \ldots, \mathbf{f}(\mathbf{W}_K, \widehat{\mathbf{b}}_t))^\top$ is a $K \times 1$ vector of mean logarithmic PGAs with $\mathbf{W}_k$ being a vector of predictors at grid point k. The soil types at grid points are obtained from the U.S. Geological Survey global V_{S30} database;
- $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \text{cov}(\mathbf{Z}, \mathbf{Y})$ and $\mathbf{c}(\boldsymbol{\theta}) = \text{var}(\mathbf{Y})$ with $\mathbf{Z}$ and $\mathbf{Y}$ representing vectors of logarithmic PGAs at grid points and recording sites, respectively;
- $\mathbf{y}_t$ is an $n_e \times 1$ vector of logarithmic PGAs at recording sites and is obtained from the the t-th synthetic dataset of logarithmic PGAs simulated in Section 2.6.4;
- $\mathbf{f}(\mathbf{X}_e, \widehat{\mathbf{b}}_t) = (\mathbf{f}(\mathbf{X}_{e,1}, \widehat{\mathbf{b}}_t), \ldots, \mathbf{f}(\mathbf{X}_{e,n_e}, \widehat{\mathbf{b}}_t))^\top$ is an $n_e \times 1$ vector of mean PGAs with $\mathbf{X}_{e,j}$ being a vector of predictors at the recording site $j \in \{1, \ldots, n_e\}$ of the event.

In this step, a ground-motion shaking intensity map can be generated from the obtained $\widehat{\mathbf{z}}_t$, which represent the logarithmic PGAs on grid points predicted by the estimated ground-motion model given the synthetic observations $\mathbf{y}_t$;

2. For each $\mathbf{y}_t$, generate a synthetic logarithmic PGA dataset $\mathbf{z}_t = (z_{1,t}, \ldots, z_{K,t})$ on all grid points $k \in \{1, \ldots, K\}$ from the multivariate Gaussian distribution with mean

$$\mathbf{f}(\mathbf{W}, \mathbf{b}_0) + \boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\mathbf{c}^{-1}(\boldsymbol{\theta}_0)\left(\mathbf{y}_t - \mathbf{f}(\mathbf{X}_e, \mathbf{b}_0)\right),$$

and covariance matrix

$$\boldsymbol{\Psi}(\boldsymbol{\theta}_0) - \boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\mathbf{c}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{\Sigma}^\top(\boldsymbol{\theta}_0)$$

in which $\boldsymbol{\Psi}(\boldsymbol{\theta}) = \text{var}(\mathbf{Z})$, and $\mathbf{b}_0$ and $\boldsymbol{\theta}_0$ are true parameter values chosen for $\mathbf{b}$ and $\boldsymbol{\theta}$ in Section 2.6.1. To assess the quality of the ground-motion shaking intensity map (i.e., the accuracy of the predictions $\widehat{\mathbf{z}}_t$)

produced by the estimated ground-motion model in the last step, this step generates the benchmark logarithmic PGAs (i.e., $\mathbf{z}_t$) on grid points using the underlying true ground-motion model given the synthetic observations $\mathbf{y}_t$;

3. At each grind point k , compute the root mean squared error of predictions (RMSEP) by

$$\text{RMSEP}_k = \sqrt{\frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}(\widehat{z}_{k,t} - z_{k,t})^2}\,,$$

which measures the predictive accuracy of the estimated ground-motion model at each grid point k .

In Figure 2.10, we plot at each grid point the percentage increase in RMSEP from the multi-stage algorithm relative to that from the Scoring estimation approach under three sample sizes of $N = 29$, 46 and 62 (corresponding to events by the end of the year 2000, 2010 and 2016) with kernel function (2.26) and (2.27). It can be seen that for both kernel function (2.26) and (2.27), as N increases, the region where the RMSEP from the multi-stage algorithm is greater than that from the Scoring estimation approach expands. When the kernel function (2.26) is considered, we find that the RMSEP from the Scoring estimation approach are smaller than those from the multi-stage algorithm, especially around the recording sites (triangles in Figure 2.10). This is because the spatial correlation structure in the ground-motion model is estimated with higher accuracy by the Scoring estimation approach. Because recording sites are often concentrated in the near-fault regions, the difference between the RMSEP from the Scoring estimation approach and that from the multi-stage algorithm becomes more distinct within the near-field (the region bounded by the dashed circle in Figure 2.10). This observation becomes remarkable when the kernel function (2.27) is considered, in which the RMSEP from the multi-stage algorithm can exceed that from the Scoring estimation approach by more than 10% near the recording sites. Furthermore, Figure 2.10 also

indicates that the Scoring estimation approach is less sensitive to the overfitting problem than the multi-stage algorithm. As we can observe from (a) and (b) in Figure 2.10, even the number of events is scare (i.e., $N = 29$), the predictive performance of the Scoring estimation approach is still comparable or better than that of the multi-stage algorithm over the region, especially when the underlying spatial correlation follows the kernel function (2.27).

2.7 Impacts of Ignoring the Spatial Correlation

We have demonstrated that the Scoring estimation approach outperforms the multi-stage algorithm in terms of estimation and prediction. However, if the spatial correlation structure is neglected from the ground-motion model while the spatial correlation is significant in the ground-motion data, we could obtained very biased estimates of model parameters, give misleading interpretation on the contributions of covariates. In addition, the predictive performance of the estimated ground-motion model may be degraded. Because most of the existing ground-motion models (e.g., Akkar and Bommer (2010); Abrahamson et al. (2014); Bindi et al. (2014); Boore et al. (2014); Campbell and Bozorgnia (2014); Chiou and Youngs (2014); Idriss (2014)) are proposed without any form of spatial correlation structure, we investigate in this section how the ignorance of spatial correlation influences the model parameter estimates and the predictive performance of the estimated ground-motion model.

2.7.1 Impact on parameter estimation

To assess how the parameter estimates could be influenced by the ignorance of spatial correlation in the ground-motion model, 1000 synthetic datasets of logarithmic PGAs, which form a training set, are generated using the kernel function (2.26) with $h = 11.50$ km. The Scoring estimation approach is then applied to estimate, respectively, the ground-motion model with well-specified

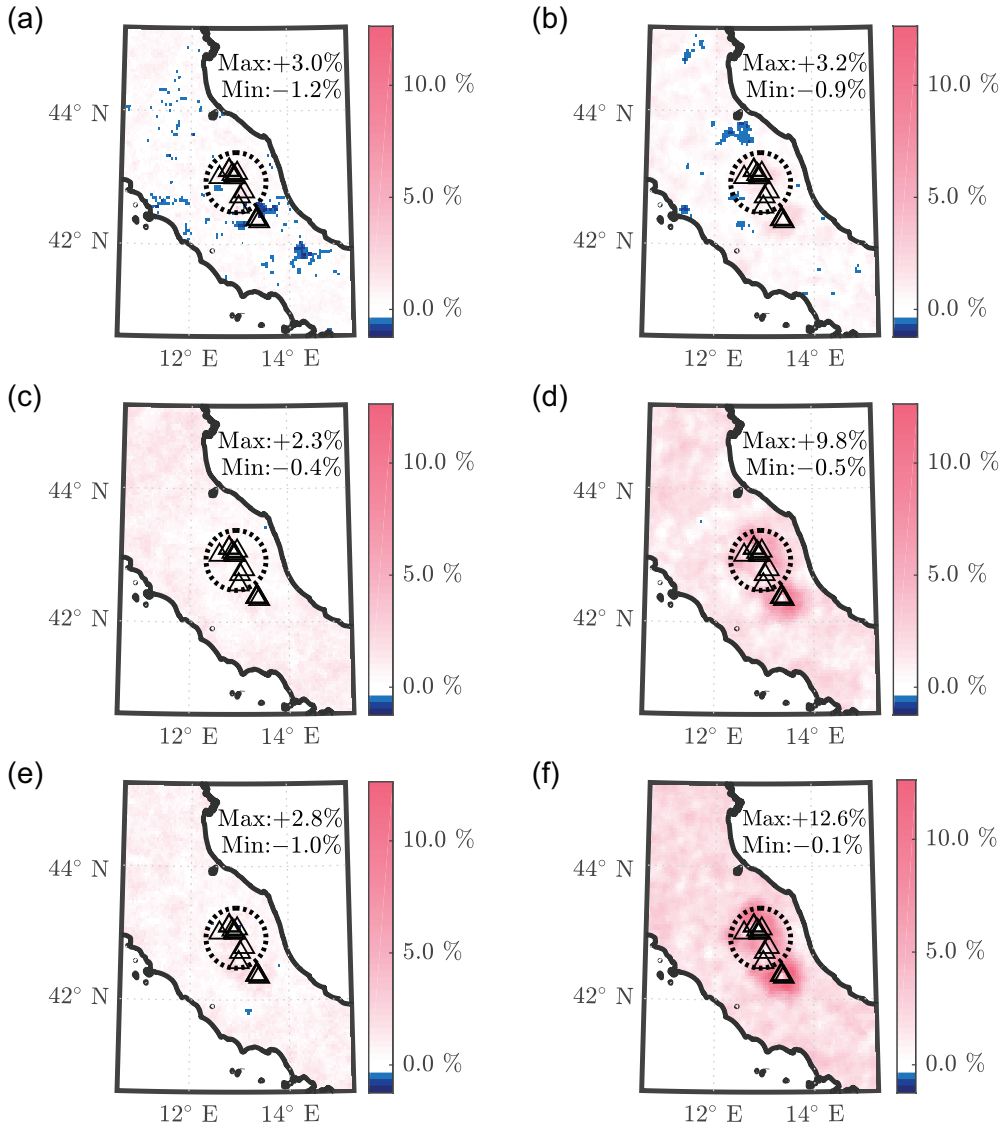


Figure 2.10: Maps of percentage increases in root mean squared error of predictions (RMSEP) from the multi-stage algorithm relative to those from the Scoring estimation approach at grid points: (a), (c) and (e) correspond to the kernel function (2.26) with $h = 11.50$ km when $N = 29$, 46, and 62, reading from top to bottom; (b), (d) and (f) correspond to the kernel function (2.27) with $h = 12.58$ km when $N = 29$, 46, and 62, reading from top to bottom. Triangles ($\triangle$) are recording sites and the dashed circle defines the border of the near-field (within 50 km from the epicenter).

spatial correlation structure (i.e., with the kernel function (2.26)) and the ground-motion model without spatial correlation structure (i.e., with the kernel function (2.4)). The sampling distributions for $\widehat{b}_1, \ldots, \widehat{b}_{10}$ obtained under the two ground-motion models are shown in Figure 2.11. It can be seen that although the estimators of $b_1, \ldots, b_{10}$ produced by the Scoring estimation approach are generally unbiased for both models, estimators such as $\widehat{b}_5, \ldots, \widehat{b}_8$ exhibit larger variances when the spatial correlation structure is ignored in the ground-motion model. Comparisons between the sampling distributions for $\widehat{\tau^2}$ and $\widehat{\sigma^2}$ under the two models are presented in Figure 2.12. We observe that when the training set is generated by the kernel function (2.26) with $h = 11.50$ km, the estimates of the inter-event variance τ^2 from the ground-motion model without spatial correlation structure are overestimated, but the estimates of the intra-event variance σ^2 are underestimated. For the ground-motion model with well-specified spatial correlation structure, however, the estimates of τ^2 and σ^2 produced essentially match their true values. To further investigate such overestimation on τ^2 and underestimation on σ^2 when spatial correlation is ignored from the ground-motion model, we refit the two ground-motion models to two additional training sets, each of which consists of 1000 synthetic datasets of logarithmic PGAs, generated using the kernel function (2.26) with $h = 30.00$ and 60.00 km, respectively. From Figure 2.12, it can be seen that as the value of h increases (i.e., the spatial correlation implied by the training data becomes stronger), the overestimation on τ^2 and underestimation on σ^2 due to the ignorance of spatial correlation are amplified. On the contrary, the estimates of τ^2 and σ^2 from the ground-motion model with well-specified spatial correlation structure are still concentrated around the true parameter values.

We repeated the above procedure using the training sets generated by the kernel function (2.27). The sampling distributions for $\widehat{b}_1, \ldots, \widehat{b}_{10}$, $\widehat{\tau^2}$, and $\widehat{\sigma^2}$ under the two completing ground-motion models are visualised in Figure 2.13 and 2.14. Figure 2.13 indicates that the loss of statistical efficiency on the

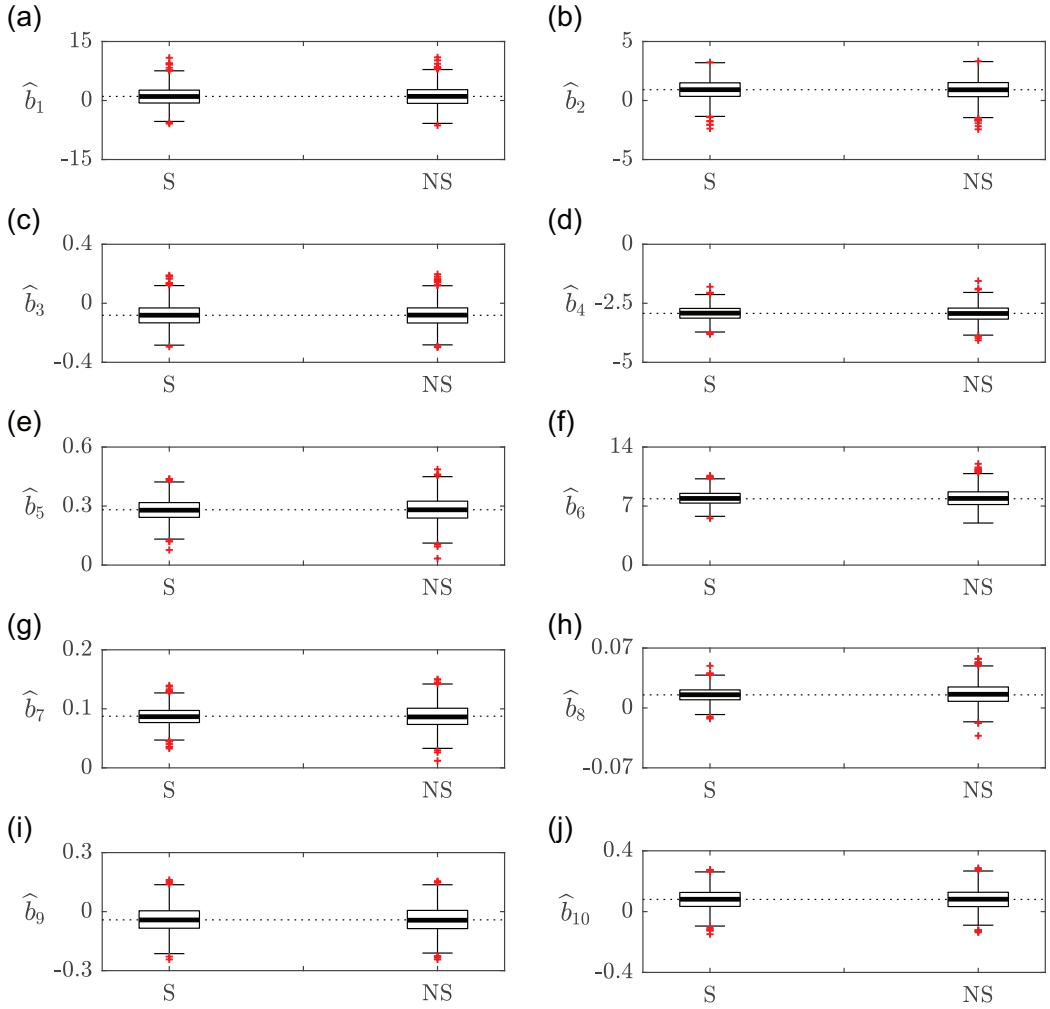


Figure 2.11: Sampling distributions for $\widehat{b}_1, \ldots, \widehat{b}_{10}$ of ground-motion models with (S) and without (NS) spatial correlation structure. The estimates are obtained from 1000 synthetic datasets generated under the kernel function (2.26) with $h = 11.50$ km. The left boxplot in each panel corresponds to the ground-motion model with spatial correlation structure; the right boxplot in each panel corresponds to the ground-motion model without spatial correlation structure. The dashed line in each panel represents the true parameter value.

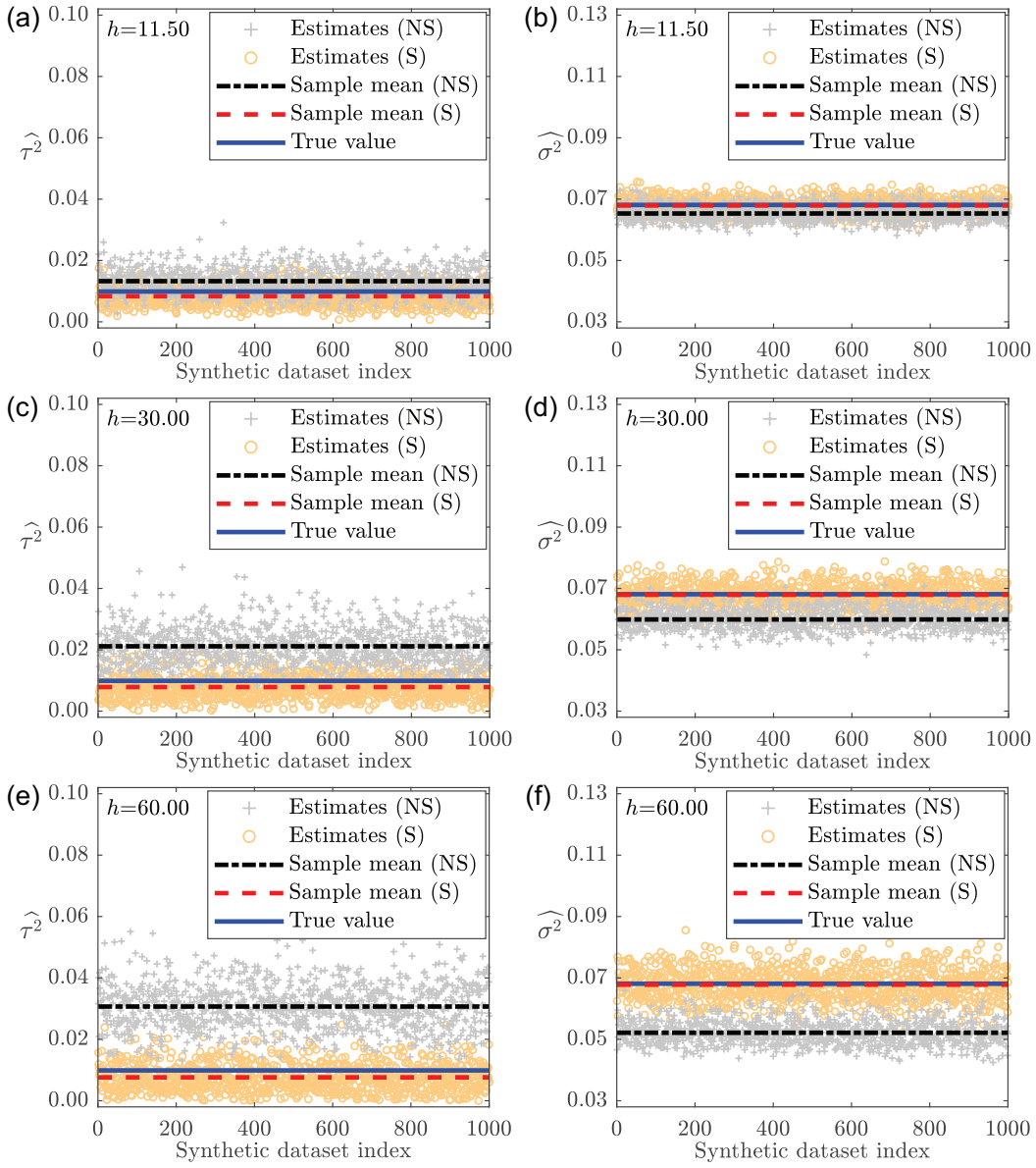


Figure 2.12: Sampling distributions for $\widehat{\tau^2}$ and $\widehat{\sigma^2}$ of ground-motion models with (S) and without (NS) spatial correlation structure (specified by the kernel function (2.26)). The estimates are obtained from 1000 synthetic datasets generated under the kernel function (2.26) with h = 11.50, 30.00, and 60.00 km, respectively. (a), (c) and (e) correspond to the estimates of τ^2; (b), (d) and (f) correspond to the estimates of σ^2.

estimator of $\mathbf{b}$ becomes more apparent when the ground-motion model without spatial correlation structure is fitted to the training data with smoother spatial correlation. From Figure 2.14, we find that fitting the ground-motion model without spatial correlation structure to the training data with smoother spatial correlations will cause severer overestimation on τ^2 and underestimation on σ^2. In contrast, the changed smoothness of the spatial correlation in the training data does not influence the accuracy of estimating τ^2 and σ^2 in the ground-motion model with well-specified spatial correlation structure.

2.7.2 Impact on predictive performance

In this section, we consider the predictive performance of the estimated (via the Scoring estimation approach) ground-motion model without spatial correlation structure for the event selected in Section 2.6.5. To investigate the predictive performance when observations are available in the far-field, 15 artificial recording sites are added to the event (see Figure 2.15). The addition of the 15 artificial recording sites increases the entries of recording sites in the catalogue, which is described in Section 2.6.2, from 2150 to 2165.

On the basis of the updated catalogue, we then generate six training sets, each of which includes 1000 synthetic datasets of logarithmic PGAs, using the generator specified in Section 2.6.1 with $h = 11.50$, 30.00 and 60.00 km for the kernel function (2.26) and with $h = 12.58$, 32.81 and 65.63 km for the kernel function (2.27). For each training set, we estimate the ground-motion model with well-specified spatial correlation structure (i.e., with the same kernel function as the underlying generator) and the ground-motion model with no spatial correlation structure by the Scoring estimation approach. The predictive performances of the estimated ground-motion models are subsequently assessed by the RMSEP obtained via the procedure detailed in Section 2.6.5. The RMSEPs produced by the estimated ground-motion models with and without spatial correlation are plotted in Figure 2.16 and 2.17. These figures show that when the spatial correlation structure is ignored from the ground-motion

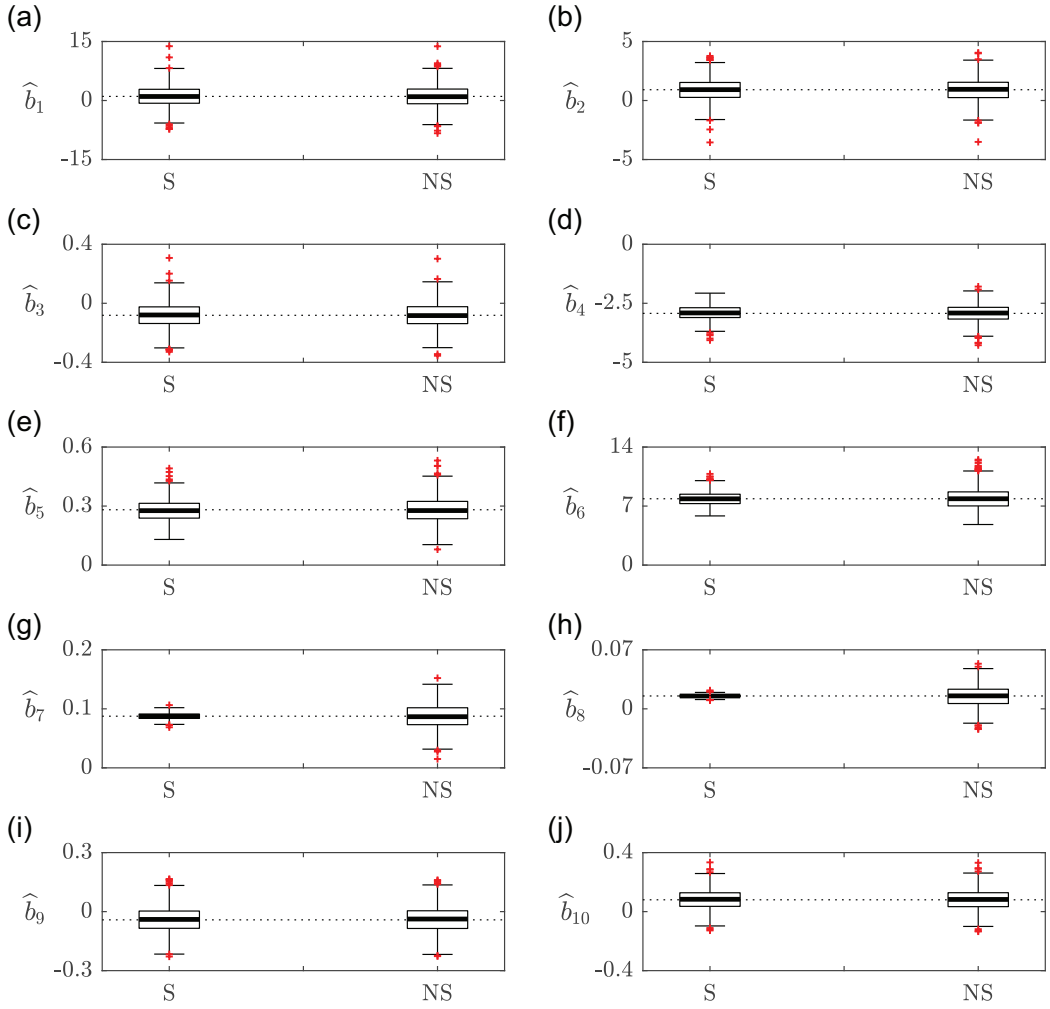


Figure 2.13: Sampling distributions for $\widehat{b}_1, \ldots, \widehat{b}_{10}$ of ground-motion models with (S) and without (NS) spatial correlation structure. The estimates are obtained from 1000 synthetic datasets generated under the kernel function (2.27) with $h = 12.58$ km. The left boxplot in each panel corresponds to the ground-motion model with spatial correlation structure; the right boxplot in each panel corresponds to the ground-motion model without spatial correlation structure. The dashed line in each panel represents the true parameter value.

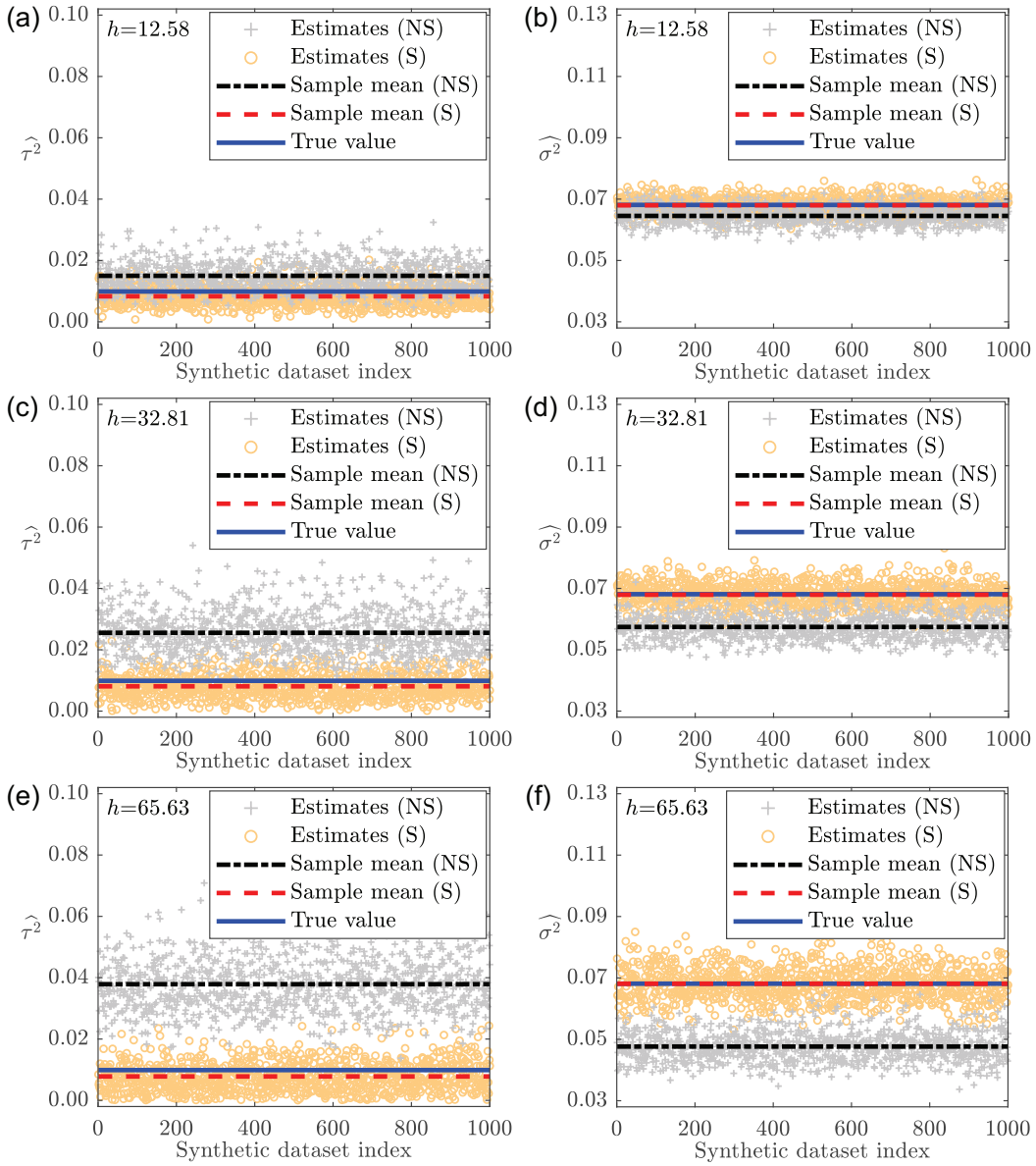


Figure 2.14: Sampling distributions for $\widehat{\tau^2}$ and $\widehat{\sigma^2}$ of ground-motion models with (S) and without (NS) spatial correlation structure (specified by the kernel function (2.27)). The estimates are obtained from 1000 synthetic datasets generated under the kernel function (2.27) with h = 12.58, 32.81, and 65.63 km, respectively. (a), (c) and (e) correspond to the estimates of τ^2; (b), (d) and (f) correspond to the estimates of σ^2.

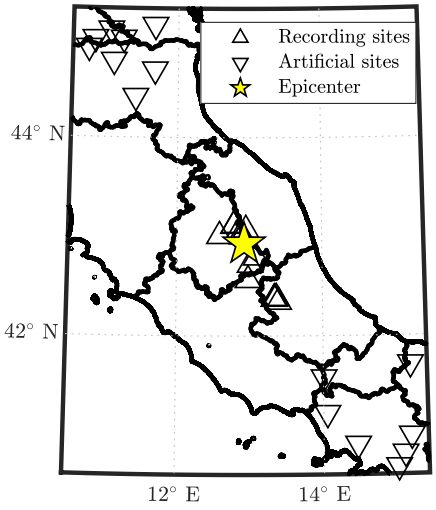


Figure 2.15: The region (within a distance of 250 km from the epicentre) of the selected event with ID 'IT-1997-0137', to which artificial recording sites are added. The epicentre of the event is labelled by a filled star (☆); triangles (△) represent the historical recording sites of the selected event described in Section 2.6.5. Inverted triangles (▽) represent the artificial recording sites that are added to the selected event. The observations at both historical and artificial recording sites are used for prediction.

model, the resulting predictions are poor across the study region regardless of the strength (i.e., the magnitude of h) and the smoothness (i.e., the choice between the kernel function (2.26) and (2.27)) of the spatial correlation implied by the training data. In addition, we find that whereas the RMSEP around the recording sites are only weakly improved when the spatial correlation is ignored from the ground-motion model, the RMSEP near the recording sites are significantly reduced when the spatial correlation is well-specified in the ground-motion model. For example, when the spatial correlation implied by the training data are characterised by the kernel function (2.26) with $h = 60$ km, little reductions in RMSEP can be observed around the recording sites if the data are fitted by the ground-motion model without spatial correlation structure (see (f) in Figure 2.16). However, the improvement of predictions near the recording sites is obvious when the spatial correlation structure is well-specified

in the ground-motion model (see (e) in Figure 2.16). Furthermore, it is found that the reductions of RMSEP caused by the availability of recording sites are consistent in near-field and far-field, when the ground-motion model with well-specified spatial correlation structure is considered. However, under the ground-motion model without spatial correlation structure, the improvement of predictions caused by the proximity to the recording sites is clearer in the near-field than in the far-field, which suffers high RMSEP in all considered scenarios.

2.8 Conclusion

In this chapter, we construct ground-motion models with repeated Gaussian processes and introduce a one-stage training algorithm, namely the Scoring estimation approach. The estimators produced by the approach have good statistical properties such as consistency, statistical efficiency and asymptotic normality. In addition, to yield consistent, statistically efficient and asymptotically normal estimators, the approach requires only a large number of events (that can be assumed to be independent) even with a small number of records per event, something that is historically relevant to earthquake records. The simulation study demonstrates that the Scoring estimation approach generally outperforms the multi-stage algorithm proposed by Jayaram and Baker (2010) in terms of estimation and prediction. With regard to estimation, the Scoring estimation approach produces parameter estimators in an accurate and stable manner under both smooth (e.g., kernel function (2.27)) and less smooth (e.g., kernel function (2.26)) kernel functions. Regarding the predictive performance, the simulation study indicates that the ground-motion model with spatial correlation estimated via the Scoring estimation approach produces smaller prediction errors than the multi-stage algorithm does, especially at locations around the recording sites and when the spatial correlation is smooth. Because the estimation of ground-motion models with spatial correlation is a key ingredient in developing GMPEs for use in PHSA, the Scoring estimation approach

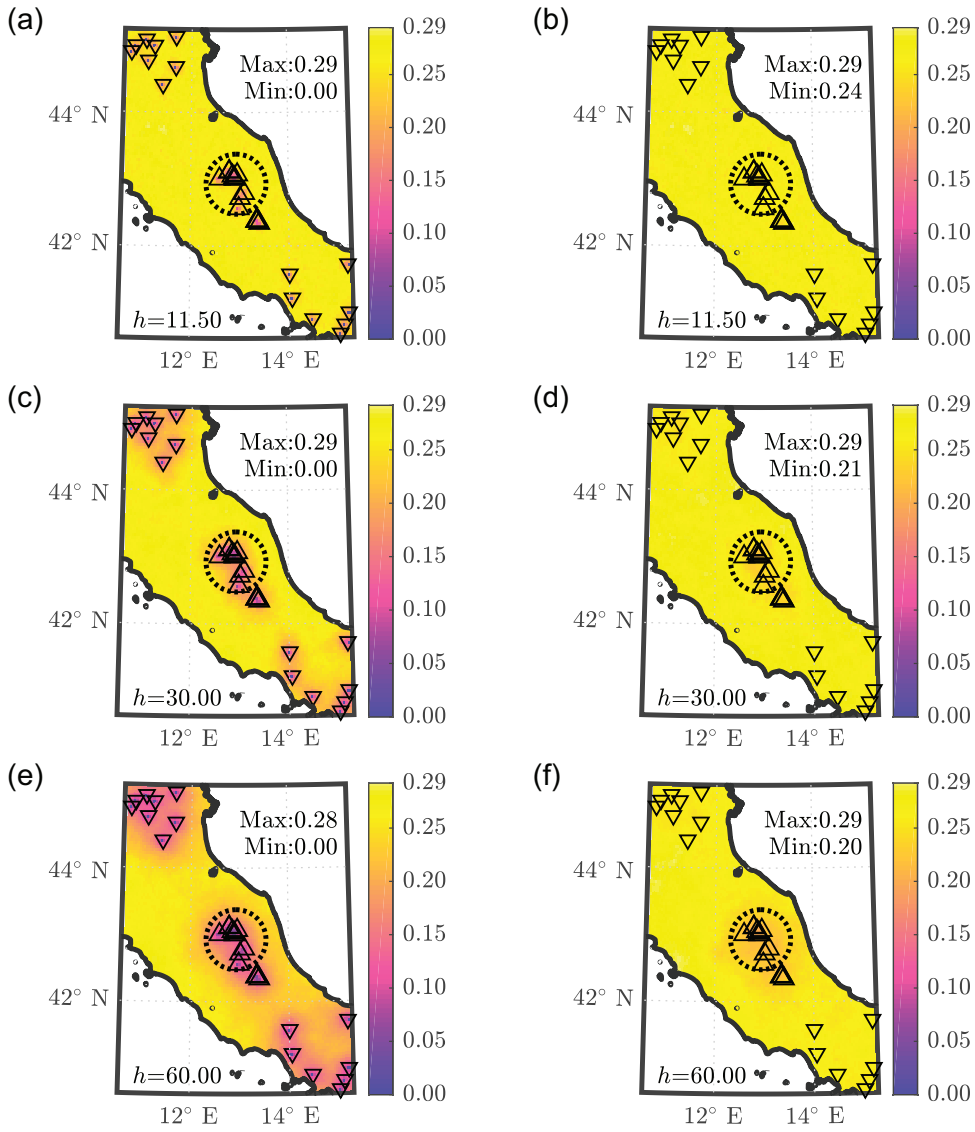


Figure 2.16: Maps of RMSEP from ground-motion models with and without spatial correlation structure (specified by the kernel function (2.26)). Ground-motion models are fitted to synthetic datasets generated under the kernel function (2.26) with $h = 11.50$, 30.00, and 60.00 km. (a), (c) and (e) correspond to the ground-motion model with spatial correlation structure; (b), (d) and (f) correspond to the ground-motion model without spatial correlation structure. Triangles ($\triangle$) and inverted triangles ($\triangledown$) are historical and artificial recording sites, respectively. The dashed circle defines the border of the near-field (within 50 km from the epicentre).

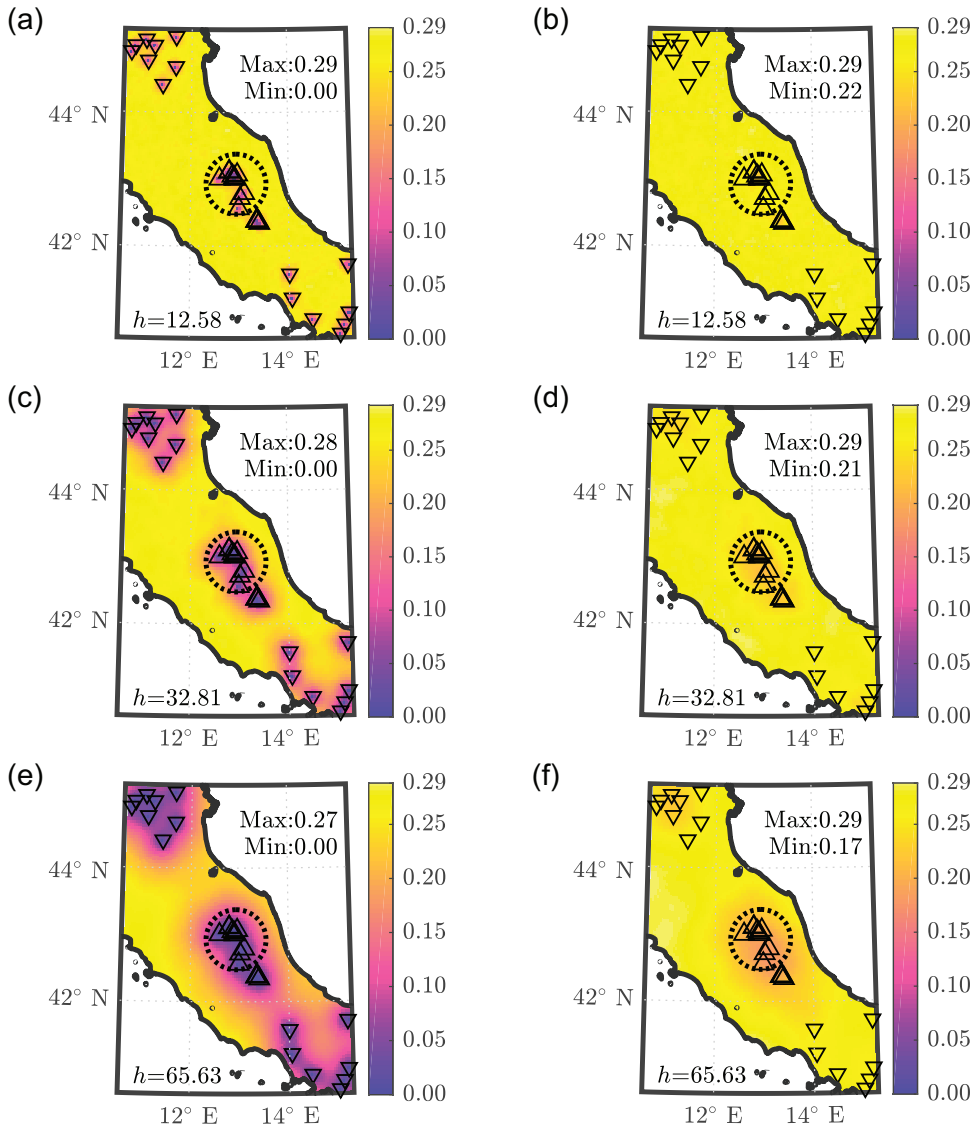


Figure 2.17: Maps of RMSEP from ground-motion models with and without spatial correlation structure (specified by the kernel function (2.27)). Ground-motion models are fitted to synthetic datasets generated under the kernel function (2.27) with $h = 12.58$, 32.81, and 65.63 km. (a), (c) and (e) correspond to the ground-motion model with spatial correlation structure; (b), (d) and (f) correspond to the ground-motion model without spatial correlation structure. Triangles ($\triangle$) and inverted triangles ($\triangledown$) are historical and artificial recording sites, respectively. The dashed circle defines the border of the near-field (within 50 km from the epicentre).

provides a statistically robust way that increases the estimation accuracy in ground-motion model construction and has the potential to reduce prediction errors in ground-motion shaking intensity maps, which in turn can improve the earthquake-induced loss assessment process.

The Scoring estimation approach is then used to assess the accuracy of model parameter estimates and subsequent prediction under the condition that spatial correlation structure is ignored in ground-motion models. It is demonstrated that neglecting spatial correlation structure in ground-motion models can cause inconsistent and statistically inefficient estimators, and inaccurate predictions.

Finally, because the Scoring estimation approach provides a relatively accurate estimation of the spatial correlation parameters (e.g., h in the exponential kernel function), as a by-product of the ground-motion model estimation, this approach could be applied to areas that do not have well-recorded events, giving the opportunity to provide a first estimate of a spatial correlation model.

2.8.1 Practicalities

In this last section we discuss several aspects of the practicalities of the Scoring estimation approach, aiming to address two key numerical issues of the approach and the importance of the asymptotic information produced by the approach when it is applied to real ground-motion datasets.

Local and global maxima

The maximum likelihood estimation framework used in the Scoring estimation approach requires the global maximum to be found so that the asymptotic properties can be established. However, it is generally not guaranteed that the global maximum of the likelihood function can be located by the Scoring estimation approach. If the likelihood surface is multimodal, the Scoring estimation approach may be trapped at local maxima, where the corresponding optimised parameter estimates can be unrealistic, e.g., the estimate of range parameter is very large indicating that the IMs at all sites are perfectly corre-

lated. Nevertheless, the issue of local maxima can be checked and mitigated to some extent by using some practical methods. The simplest way would be initiating the Scoring estimation algorithm with different initial values of the model parameters and checking whether the optimised model parameters give higher likelihood. Since the multimodality (i.e., the nonconcavity of the likelihood function) is often in respect of kernel parameter $\boldsymbol{\omega}$ and intra-event variance σ^2, another approach is to plot the likelihood surface between $\boldsymbol{\omega}$ and σ^2 given the optimised values of other model parameters and check visually if the global maximum is indeed reached. This approach is generally feasible and not very computationally expensive since $\boldsymbol{\omega}$ is one-dimensional in kernel functions such as (2.26) and (2.27).

It is worth noting that even we have the global maximum, the optimised model parameter estimates may not give sensible interpretations. For example, the range parameter can be very small or large at the global maximum. In such situations, one may need to increase the data size in the hope that the likelihood function becomes well-behaved. Otherwise, one may have to retreat to Jayaram and Baker's multi-stage algorithm assuming that the estimate of h given by the spatial correlation stage is the true value of the range parameter, and thus lose the capacity to measure the uncertainty of spatial correlation.

Ill-conditioned covariance matrix

An ill-conditioned covariance matrix is associated with a large condition number, which can cause numerical instabilities (e.g., accumulation of rounding errors) of the Scoring estimation approach and thus deteriorate the subsequent predictions. There are two main sources of an ill-conditioned covariance matrix. One source is the dataset. The locations of some recording sites included in the dataset may overlap or are very close to each other, causing singularity (condition number being infinite) or near-singularity (condition number being very large) of the constructed covariance matrix. The ill-conditioning issue due to this reason can often be alleviated by only including one recording site at every location

or by adding a very small jitter (e.g., 10^{-5}) to the diagonal elements of the covariance matrix. The latter approach effectively adds a small noise term to the ground-motion models. Changing kernel functions is another way to reduce the condition number of a covariance matrix. Exponential kernel function (2.26) is usually less prone to the ill-conditioned problem than the Matérn kernel (2.27) with $\nu = 1.5$, which decays faster to zero as the distance between two sites increases. Therefore, using the exponential kernel function for the ground-motion model is often robust in terms of the numerical stability. Another source of an ill-conditioned covariance matrix is the poor estimate of the range parameter. When the Scoring estimation approach traps at local maxima or the global maximum of the likelihood function is not well-behaved, the estimate of the range parameter can be very large, causing a large condition number of the covariance matrix. In such situations, one need resort to techniques discussed in the last section to reduce the condition number.

Why asymptotic properties are useful?

Given that the global maximum is found and the model parameter estimates are sensible, the asymptotic properties established under the Scoring estimation approach allow one to conduct various analysis of the ground-motion model. For example, one can construct hypothesis tests to conduct variable selection and model comparison. One can also build confidence intervals to check the uncertainties of the estimated spatial correlation, intra- and inter-event variances. Therefore, the asymptotic information produced by the Scoring estimation approach provides a useful toolbox for the practitioners to design new ground-motion models that appreciate the underlying data and to test the goodness-of-fit of the designed models.

Chapter 3

Integrated Emulators for Systems of Computer Models

3.1 Introduction

Systems of computer models constitute the new frontier of many scientific and engineering simulations. These can be multi-physics systems of computer simulators such as coupled tsunami simulators with earthquake and landslide sources (Salmanidou et al., 2017; Ulrich et al., 2019), coupled multi-physics model of the human heart (Santiago et al., 2018), and multi-disciplinary systems such as automotive and aerospace systems (Kodiyalam et al., 2004; Fazeley et al., 2016; Zhao et al., 2018). Other examples include climate models where climate variability arises from atmospheric, oceanic, land, and cryospheric processes and their coupled interactions (Kay et al., 2015; Hawkins et al., 2016), or highly multi-disciplinary future biodiversity models (Thuiller et al., 2019) using combinations of species distribution models, dispersal strategies, climate models, and representative concentration pathways. The number and complexity of computer models involved can hinder the analysis of such systems. For instance, the engineering design optimisation of an aerospace system typically requires hundreds of thousands of system evaluations. When the system has feed-backs across computer models, the number of simulations becomes computationally

prohibitive (Chaudhuri et al., 2018). Therefore, building and using a surrogate model is crucial: the system outputs can be predicted at little computational cost, and subsequent sensitivity analysis, uncertainty propagation or inverse modelling can be conducted in a computationally efficient manner.

Gaussian process emulators have gained popularity as surrogate models of systems of computer models. However, many studies (Jandarov et al., 2014; Johnstone et al., 2016; Salmanidou et al., 2017; Simpson et al., 2001; Tagade et al., 2013) construct global GP emulators (named as composite emulators hereinafter) of such systems based on a single GP model trained by global inputs and outputs without consideration of system structures. One major drawback of such a structural ignorance is that designing experiments can be expensive because system structures may induce high non-linearity between global inputs and outputs (Sanson et al., 2019). Furthermore, runs of the whole system are required to produce new training points, even though the overall functional complexity between global inputs and outputs originates from a few computer models. This pitfall is particularly undesirable because modern engineering and physical systems can include multiple computer models.

To overcome the disadvantages of the composite emulator, we propose a structure-informed emulator, called integrated emulator, as the surrogate for a system of computer models by integrating GP emulators of individual computer models. The idea of integrating GP emulators has been explored by Sanson et al. (2019) in a feed-forward system, but only using the Monte Carlo simulation to approximate the predictive mean and variance of the system output. The Monte Carlo method suffers from a low convergence rate and heavy computational cost, especially when the number of layers in a system is high (Rainforth et al., 2018) and the number of new input positions to be evaluated is large, making it prohibitive for complex systems. Recently, two studies by Kyzyurova et al. (2018) and Marque-Pucheu et al. (2019) have derived an emulator, called linked emulator (Kyzyurova et al., 2018), for a feed-forward system of two computer models in analytical form under the assumption that every computer model in

the system is represented by the GP with a product of squared exponential kernels over different input dimensions.

Inspired by the linked emulator, our integrated emulator provides analytical expressions for mean and variance of the predicted output of any feed-forward system at an unexplored input position. Furthermore, our analytical formulas for the integrated emulator are derived under a general and flexible framework that allows different computer models to be modelled by different GPs with a wide range of kernel choices, such as the Matérn kernel with smoothness parameter of 2.5. Indeed, the squared exponential kernel has been criticised for its over-smoothness (Stein, 1999) and associated ill-conditioned problem (Dalbey, 2013; Gu et al., 2018). Particularly, the integrated emulator is more prone to the latter issue than the composite emulator because the design (e.g., the Latin hypercube design) of the global input can produce poor designs for GP emulators of internal computer models. Thus, the generalisation of the kernel assumption is necessary and several of our examples below require it. Our framework can also be readily extended to systems with feed-back-coupled computer models as such systems can be converted to feed-forward ones by applying decoupling procedures such as the optimal approximations of coupling (Baptista et al., 2018) or the surrogate-based approximation of coupling variables (Chaudhuri et al., 2018).

The remainder of the chapter is organised as follows. In Section 3.2, we detail the procedure and the theoretical method to construct the integrated emulator. Synthetic experiments are provided in Section 3.3 to compare the training cost and predictive performances of the integrated and composite emulators. A feed-back coupled fire-detection satellite example is demonstrated in Section 3.4. An adaptive designing strategy allowed by the integrated emulation is discussed in Section 3.5. We conclude in Section 3.7. Key closed form expressions for the integrated emulator and proofs of results are contained in the Appendix B and Appendix C, respectively.

3.2 Model and Method

We consider a system of computer models with a feed-forward hierarchy. In such a hierarchy, the outputs of lower-layer computer models act as the inputs of higher-layer computer models. An illustrative example of this type of hierarchy is shown in Figure 3.1.

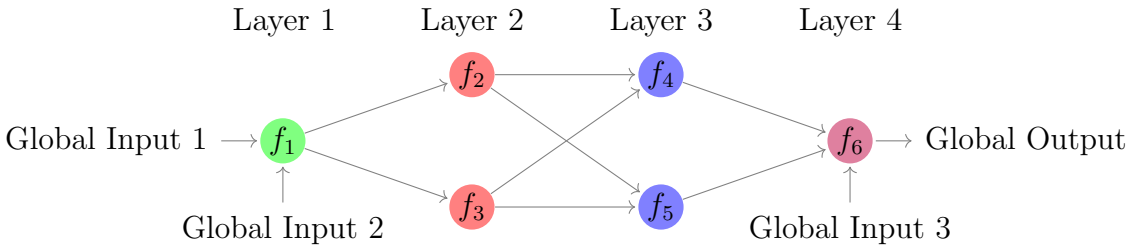


Figure 3.1: An example of a four-layered feed-forward system of six computer models.

3.2.1 GP emulators for individual computer models

The first step to construct the integrated emulator of a feed-forward system of computer models is to build GP emulators for individual computer models. The GP emulator of a computer model is itself a collection of GP emulators, approximating the functional dependence between the inputs of the computer model and its one-dimensional outputs. Each 1-D output emulator is constructed independently without the consideration of cross-output dependence, as in Gu and Berger (2016) and Kyzyurova et al. (2018).

Let $\mathbf{X} \in \mathbb{R}^p$ be a p-dimensional vector of inputs of a computer model and $Y(\mathbf{X})$ be the corresponding scalar-valued output. Then, given m sets of inputs $\{\mathbf{X}_1, \ldots, \mathbf{X}_m\}$, the GP model is defined by

$$Y(\mathbf{X}_i) = t(\mathbf{X}_i, \mathbf{b}) + \varepsilon_i, \quad i = 1, \ldots, m$$

where $t(\mathbf{X}_i, \mathbf{b}) = \mathbf{h}(\mathbf{X}_i)^\top \mathbf{b}$ is the trend with q basis functions $\mathbf{h}(\mathbf{X}_i) = [h_1(\mathbf{X}_i), \ldots, h_q(\mathbf{X}_i)]^\top$ and coefficients $\mathbf{b} = [b_1, \ldots, b_q]^\top$; $(\varepsilon_1, \ldots, \varepsilon_m)^\top \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{R})$ with ij-th element of the correlation matrix $\mathbf{R}$ given by $R_{ij} = c(\mathbf{X}_i, \mathbf{X}_j) + \eta \mathbb{1}_{\{\mathbf{X}_i = \mathbf{X}_j\}}$, where $c(\cdot, \cdot)$ is a given kernel function; η is the nugget

term; and $\mathbb{1}_{\{\cdot\}}$ is the indicator function.

The specification of the kernel function $c(\cdot, \cdot)$ plays an important role in GP emulation as it characterises the sample paths of a GP model (Stein, 1999). In this study we consider the kernel function with the following multiplicative form:

$$c(\mathbf{X}_i, \mathbf{X}_j) = \prod_{k=1}^{p} c_k(X_{ik}, X_{jk}),$$

where $c_k(\cdot, \cdot)$ is a one-dimensional kernel function for the k-th input dimension. Popular candidates for $c_k(\cdot, \cdot)$ are summarised in Table 3.1. In Section 3.2.2, we will show that the integrated emulator is applicable to all these aforementioned choices. In Appendix C, we also derive the integrated emulator under the additive form of $c(\cdot, \cdot)$.

Table 3.1: Choices of $c_k(\cdot, \cdot)$. $\gamma_k > 0$ is the range parameter for the k-th input dimension.

Exponential	$c_k(\cdot, \cdot) = \exp\left\{-\frac{\lvert X_{ik} - X_{jk}\rvert}{\gamma_k}\right\}$
Squared Exponential	$c_k(\cdot, \cdot) = \exp\left\{-\frac{(X_{ik} - X_{jk})^2}{\gamma_k^2}\right\}$
Matérn-1.5	$c_k(\cdot, \cdot) = \left(1 + \frac{\sqrt{3}\lvert X_{ik} - X_{jk}\rvert}{\gamma_k}\right) \exp\left\{-\frac{\sqrt{3}\lvert X_{ik} - X_{jk}\rvert}{\gamma_k}\right\}$
Matérn-2.5	$c_k(\cdot, \cdot) = \left(1 + \frac{\sqrt{5}\lvert X_{ik} - X_{jk}\rvert}{\gamma_k} + \frac{5(X_{ik} - X_{jk})^2}{3\gamma_k^2}\right) \exp\left\{-\frac{\sqrt{5}\lvert X_{ik} - X_{jk}\rvert}{\gamma_k}\right\}$

Assume that the GP model parameters σ^2, η and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^\top$ are known but $\mathbf{b}$ is a random vector that has a Gaussian distribution with mean $\mathbf{b}_0$ and variance $\tau^2\mathbf{V}_0$. Then, given m inputs $\mathbf{x}^{\mathcal{T}} = (\mathbf{x}_1^{\mathcal{T}}, \ldots, \mathbf{x}_m^{\mathcal{T}})^\top$ and the corresponding outputs $\mathbf{y}^{\mathcal{T}} = (y_1^{\mathcal{T}}, \ldots, y_m^{\mathcal{T}})^\top$, the GP emulator of the computer model is defined by the predictive distribution of $Y(\mathbf{x}_0)$ (i.e., conditional distribution of $Y(\mathbf{x}_0)$ given $\mathbf{y}^{\mathcal{T}}$) at a new input position $\mathbf{x}_0$ (Santner et al., 2003), which is

$$Y(\mathbf{x}_0)|\mathbf{y}^{\mathcal{T}} \sim \mathcal{N}(\mu_0(\mathbf{x}_0),\, \sigma_0^2(\mathbf{x}_0)) \tag{3.1}$$

with

$$\mu_0(\mathbf{x}_0) = \mathbf{h}(\mathbf{x}_0)^\top \widehat{\mathbf{b}} + \mathbf{r}(\mathbf{x}_0)^\top \mathbf{R}^{-1} \left(\mathbf{y}^{\mathcal{T}} - \mathbf{H}(\mathbf{x}^{\mathcal{T}}) \widehat{\mathbf{b}} \right) \tag{3.2}$$

$$\begin{aligned} \sigma_0^2(\mathbf{x}_0) = \sigma^2 \Big[1 + \eta - \mathbf{r}(\mathbf{x}_0)^\top \mathbf{R}^{-1} \mathbf{r}(\mathbf{x}_0) + \left(\mathbf{h}(\mathbf{x}_0) - \mathbf{H}(\mathbf{x}^{\mathcal{T}})^\top \mathbf{R}^{-1} \mathbf{r}(\mathbf{x}_0) \right)^\top \\ \times \left(\mathbf{H}(\mathbf{x}^{\mathcal{T}})^\top \mathbf{R}^{-1} \mathbf{H}(\mathbf{x}^{\mathcal{T}}) + \frac{\sigma^2}{\tau^2} \mathbf{V}_0^{-1} \right)^{-1} \left(\mathbf{h}(\mathbf{x}_0) - \mathbf{H}(\mathbf{x}^{\mathcal{T}})^\top \mathbf{R}^{-1} \mathbf{r}(\mathbf{x}_0) \right) \Big], \end{aligned} \tag{3.3}$$

where $\mathbf{r}(\mathbf{x}_0) = [c(\mathbf{x}_0, \mathbf{x}_1^{\mathcal{T}}), \ldots, c(\mathbf{x}_0, \mathbf{x}_m^{\mathcal{T}})]^\top$, $\mathbf{H}(\mathbf{x}^{\mathcal{T}}) = [\mathbf{h}(\mathbf{x}_1^{\mathcal{T}}), \ldots, \mathbf{h}(\mathbf{x}_m^{\mathcal{T}})]^\top$ and

$$\widehat{\mathbf{b}} \overset{\text{def}}{=\!=} \left(\mathbf{H}(\mathbf{x}^{\mathcal{T}})^\top \mathbf{R}^{-1} \mathbf{H}(\mathbf{x}^{\mathcal{T}}) + \frac{\sigma^2}{\tau^2} \mathbf{V}_0^{-1} \right)^{-1} \left(\mathbf{H}(\mathbf{x}^{\mathcal{T}})^\top \mathbf{R}^{-1} \mathbf{y}^{\mathcal{T}} + \frac{\sigma^2}{\tau^2} \mathbf{V}_0^{-1} \mathbf{b}_0 \right).$$

Let $\tau^2 \to \infty$ (i.e., the Gaussian distribution of $\mathbf{b}$ gets more and more non-informative), then all terms associated with $\mathbf{b}_0$ and $\mathbf{V}_0$ in equation (3.2) and (3.3) become increasingly insignificant and thus we obtain the GP emulator defined by the predictive distribution of $Y(\mathbf{x}_0)$ with its mean and variance given by

$$\mu_0(\mathbf{x}_0) = \mathbf{h}(\mathbf{x}_0)^\top \widehat{\mathbf{b}} + \mathbf{r}(\mathbf{x}_0)^\top \mathbf{R}^{-1} \left(\mathbf{y}^{\mathcal{T}} - \mathbf{H}(\mathbf{x}^{\mathcal{T}}) \widehat{\mathbf{b}} \right) \tag{3.4}$$

$$\begin{aligned} \sigma_0^2(\mathbf{x}_0) = \sigma^2 \Big[1 + \eta - \mathbf{r}(\mathbf{x}_0)^\top \mathbf{R}^{-1} \mathbf{r}(\mathbf{x}_0) + \left(\mathbf{h}(\mathbf{x}_0) - \mathbf{H}(\mathbf{x}^{\mathcal{T}})^\top \mathbf{R}^{-1} \mathbf{r}(\mathbf{x}_0) \right)^\top \\ \times \left(\mathbf{H}(\mathbf{x}^{\mathcal{T}})^\top \mathbf{R}^{-1} \mathbf{H}(\mathbf{x}^{\mathcal{T}}) \right)^{-1} \left(\mathbf{h}(\mathbf{x}_0) - \mathbf{H}(\mathbf{x}^{\mathcal{T}})^\top \mathbf{R}^{-1} \mathbf{r}(\mathbf{x}_0) \right) \Big] \end{aligned} \tag{3.5}$$

with $\widehat{\mathbf{b}} \overset{\text{def}}{=\!=} \left[\mathbf{H}(\mathbf{x}^{\mathcal{T}})^\top \mathbf{R}^{-1} \mathbf{H}(\mathbf{x}^{\mathcal{T}}) \right]^{-1} \mathbf{H}(\mathbf{x}^{\mathcal{T}})^\top \mathbf{R}^{-1} \mathbf{y}^{\mathcal{T}}$, where $\mu_0(\mathbf{x}_0)$ and $\sigma_0^2(\mathbf{x}_0)$ match the best linear unbiased predictor (BLUP) of $Y(\mathbf{x}_0)$ and its mean squared error (Stein, 1999). In the remainder of the study we use the predictive distribution with mean and variance given in equation (3.4) and (3.5) as the GP emulator of a computer model. Note that the GP model parameters σ^2, η and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^\top$ in equation (3.4) and (3.5) are typically unknown and need to be estimated. One may estimate these parameters by solving the objective function

$$(\widehat{\eta}, \widehat{\boldsymbol{\gamma}}) = \underset{\eta, \boldsymbol{\gamma}}{\operatorname{argmax}} \, \mathcal{L}(\widehat{\sigma^2}, \eta, \boldsymbol{\gamma}),$$

where

$$\mathcal{L}(\widehat{\sigma^2},\, \eta,\, \boldsymbol{\gamma}) = \frac{|\mathbf{R}|^{-\frac{1}{2}}|\mathbf{H}(\mathbf{x}^{\mathcal{T}})^{\top}\mathbf{R}^{-1}\mathbf{H}(\mathbf{x}^{\mathcal{T}})|^{-\frac{1}{2}}}{(2\pi\widehat{\sigma^2})^{\frac{m-q}{2}}} \times \exp\left\{-\frac{1}{2\widehat{\sigma^2}}\left(\mathbf{y}^{\mathcal{T}} - \mathbf{H}(\mathbf{x}^{\mathcal{T}})\widehat{\mathbf{b}}\right)^{\top}\mathbf{R}^{-1}\left(\mathbf{y}^{\mathcal{T}} - \mathbf{H}(\mathbf{x}^{\mathcal{T}})\widehat{\mathbf{b}}\right)\right\},$$

is the marginal likelihood obtained by integrating out $\mathbf{b}$ from the full likelihood function $\mathcal{L}(\mathbf{b},\, \sigma^2,\, \eta,\, \boldsymbol{\gamma})$ and have σ^2 replaced by its maximum likelihood estimator

$$\widehat{\sigma^2} = \frac{1}{m-q}\left(\mathbf{y}^{\mathcal{T}} - \mathbf{H}(\mathbf{x}^{\mathcal{T}})\widehat{\mathbf{b}}\right)^{\top}\mathbf{R}^{-1}\left(\mathbf{y}^{\mathcal{T}} - \mathbf{H}(\mathbf{x}^{\mathcal{T}})\widehat{\mathbf{b}}\right) \tag{3.6}$$

with $\widehat{\mathbf{b}} \overset{\text{def}}{=\!=} \left[\mathbf{H}(\mathbf{x}^{\mathcal{T}})^{\top}\mathbf{R}^{-1}\mathbf{H}(\mathbf{x}^{\mathcal{T}})\right]^{-1}\mathbf{H}(\mathbf{x}^{\mathcal{T}})^{\top}\mathbf{R}^{-1}\mathbf{y}^{\mathcal{T}}$. Alternatively, the maximum a posterior (MAP) method is a more robust estimation technique (Gu et al., 2018). It maximises the marginal posterior mode with respect to the objective function

$$(\widehat{\eta},\, \widehat{\boldsymbol{\gamma}}) = \underset{\eta,\boldsymbol{\gamma}}{\operatorname{argmax}}\, \mathcal{L}(\widehat{\sigma^2},\, \eta,\, \boldsymbol{\gamma})\pi(\eta,\, \boldsymbol{\gamma}), \tag{3.7}$$

where $\pi(\eta,\, \boldsymbol{\gamma})$ is the reference prior, see Gu et al. (2018) for different choices and parameterisations.

After the estimates of σ^2, η and $\boldsymbol{\gamma}$ are obtained, they are plugged into the predictive distribution mean (3.4) and variance (3.5), forming the empirical GP emulator of a computer model. In the remainder of the study, all GP models of individual computer models are estimated using the MAP method via the `R` package `RobustGaSP`. Note that `RobustGaSP` in fact estimates η and $\boldsymbol{\gamma}$ with the marginal likelihood obtained by integrating out both $\mathbf{b}$ and σ^2. However, as demonstrated in Andrianakis and Challenor (2009) the estimates of η and $\boldsymbol{\gamma}$ are not influenced by the integration of σ^2. As a result, we can implement `RobustGaSP` to obtain the estimates of η and $\boldsymbol{\gamma}$ produced by the discussed MAP method and then have them plugged in equation (3.6) to obtain the estimate of σ^2.

3.2.2 Integration of GP emulators

Integrating GP emulators of individual computer models in a complex feed-forward system is a challenging analytical work because it requires the integration of predictive distributions across a large number of layers. To reduce the analytical efforts, we propose an iterative approach that collapses a complex system into a sequence of two-layered computer systems so that at each iteration we only need to integrate emulators across two layers.

Consider a general feed-forward system of computer models, denoted by $e_{1\to L}$, with L layers. The iterative method constructs its emulator by successively building integrated emulators of $e_{1\to(i+1)}$ for $i = 1, \ldots, L-1$. For example, the system in Figure 3.1 can be decomposed into three recursive systems shown in Figure 3.2. The iterative approach then takes three iterations to produce the integrated emulator of $e_{1\to 4}$.

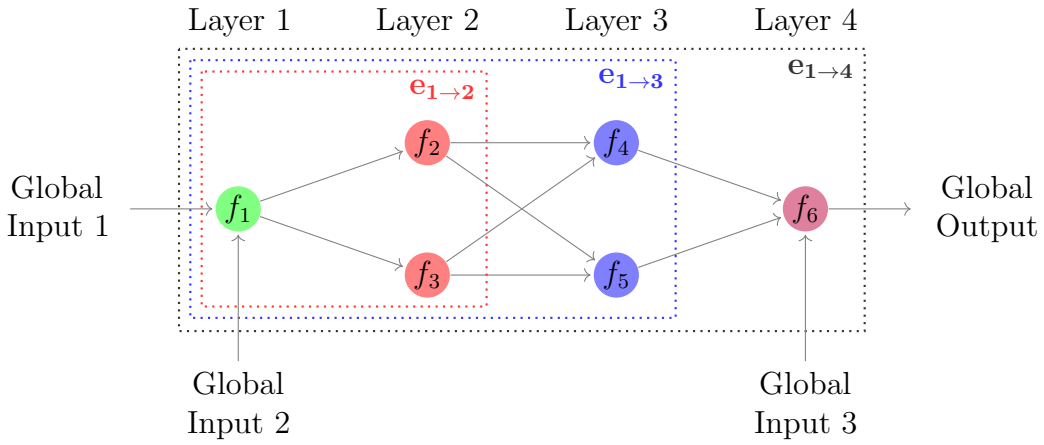


Figure 3.2: The recursive systems $e_{1\to 2}$, $e_{1\to 3}$ and $e_{1\to 4}$ of the computer system in Figure 3.1.

Without loss of generality, we consider the i-th iteration of the iterative approach to emulate $e_{1\to(i+1)}$ with respect to its one scalar-valued output y. At this iteration, we effectively have a two-layered computer system with $e_{1\to i}$ in the first layer and a computer model g (belonging to the system e_{i+1} in layer $i+1$) that produces y in the second layer. Assume that $e_{1\to i}$ have a d-dimensional output and is approximated by a collection of d one-dimensional emulators

$\widehat{f}_1, \ldots, \widehat{f}_d$, which are GP emulators when $i = 1$. Otherwise, they are integrated emulators. Let $\widehat{g}$ be the GP emulator of g with respect to y. Then, the connections between these emulators are visualised in Figure 3.3.

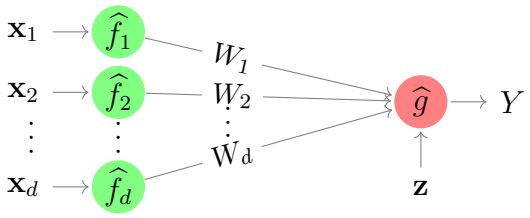


Figure 3.3: The connections of emulators to be integrated at the i-th iteration of the iterative approach for emulating a general feed-forward computer system $e_{1\to L}$ with L layers. $\widehat{f}_1$, $\widehat{f}_2 \ldots, \widehat{f}_d$ are one-dimensional emulators approximating the computer system $e_{1\to i}$; $\widehat{g}$ is a one-dimensional GP emulator of the computer model g (belonging to the system e_{i+1} in layer $i+1$) with respect to the scalar-valued output y.

The integrated emulator of $e_{1\to(i+1)}$ with respect to the one-dimensional output y is defined as the predictive distribution of $Y(\mathbf{x}_1, \ldots, \mathbf{x}_d, \mathbf{z})$, given the global inputs $\mathbf{x}_1, \ldots, \mathbf{x}_d$ and $\mathbf{z}$. This predictive distribution is naturally given by the probability density function

$$p(y|\mathbf{x}_1, \ldots, \mathbf{x}_d, \mathbf{z}) = \int_{\mathbf{w}} p(y|\mathbf{w}, \mathbf{z})\, p(\mathbf{w}|\mathbf{x}_1, \ldots, \mathbf{x}_d)\, \mathrm{d}\mathbf{w}, \tag{3.8}$$

where $\mathbf{w} = (w_1, \ldots, w_d)^\top$. However, $p(y|\mathbf{x}_1, \ldots, \mathbf{x}_d, \mathbf{z})$ often has no closed form expression and the resulting predictive distribution is not Gaussian in general. One might employ methods such as Monte Carlo simulation to compute the integral in equation (3.8) numerically at each given input position and use the resulting sampled density as the predictive distribution. However, such an approach is computationally expensive and the resulting integrated emulator is analytically intractable. To obtain the integrated emulator analytically, in the following, we demonstrate that under Assumption 1 and 2 below, the mean and variance of the predictive distribution of $Y(\mathbf{x}_1, \ldots, \mathbf{x}_d, \mathbf{z})$ can be calculated in closed form, subject to the choice of the 1-D kernel functions in GP emulator $\widehat{g}$.

Let $Y(\mathbf{W}, \mathbf{z})$ be the output of the GP emulator $\widehat{g}$ at inputs

$$\mathbf{W} = [W_1(\mathbf{x}_1), \ldots, W_d(\mathbf{x}_d)]^\top \quad \text{and} \quad \mathbf{z} = (z_1, \ldots, z_p)^\top,$$

where $W_1(\mathbf{x}_1), \ldots, W_d(\mathbf{x}_d)$ are outputs of (GP or integrated) emulators $\widehat{f}_1, \ldots, \widehat{f}_d$ at the input positions $\mathbf{x}_1, \ldots, \mathbf{x}_d$. Assume that the GP emulator $\widehat{g}$ is built with m training points $\mathbf{w}^{\mathcal{T}} = (\mathbf{w}_1^{\mathcal{T}}, \ldots, \mathbf{w}_m^{\mathcal{T}})^\top$, $\mathbf{z}^{\mathcal{T}} = (\mathbf{z}_1^{\mathcal{T}}, \ldots, \mathbf{z}_m^{\mathcal{T}})^\top$ and $\mathbf{y}^{\mathcal{T}} = (y_1^{\mathcal{T}}, \ldots, y_m^{\mathcal{T}})^\top$, where $\mathbf{w}_i^{\mathcal{T}} = (w_{i1}^{\mathcal{T}}, \ldots, w_{id}^{\mathcal{T}})^\top$ and $\mathbf{z}_i^{\mathcal{T}} = (z_{i1}^{\mathcal{T}}, \ldots, z_{ip}^{\mathcal{T}})^\top$ for all $i = 1, \ldots, m$. We make the following assumptions:

Assumption 1 *The trend function $t(\mathbf{W}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\beta})$ in the GP model for the computer model g is specified by $t(\mathbf{W}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\beta}) = \mathbf{W}^\top\boldsymbol{\theta} + \mathbf{h}(\mathbf{z})^\top\boldsymbol{\beta}$, where*

- *$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^\top$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)^\top$;*
- *$\mathbf{h}(\mathbf{z}) = [h_1(\mathbf{z}), \ldots, h_q(\mathbf{z})]^\top$ are basis functions of $\mathbf{z}$;*

Assumption 2 *$W_k(\mathbf{x}_k) \overset{ind}{\sim} \mathcal{N}(\mu_k(\mathbf{x}_k), \sigma_k^2(\mathbf{x}_k))$ for $k = 1, \ldots, d$.*

Theorem 3.1 *Under Assumption 1 and 2, the output $Y(\mathbf{x}_1, \ldots, \mathbf{x}_d, \mathbf{z})$ of the computer system $e_{1\to(i+1)}$ predicted at the input positions $\mathbf{x}_1, \ldots, \mathbf{x}_d$ and $\mathbf{z}$ has analytical mean μ_I and variance σ_I^2 given by*

$$\mu_I = \boldsymbol{\mu}^\top\widehat{\boldsymbol{\theta}} + \mathbf{h}(\mathbf{z})^\top\widehat{\boldsymbol{\beta}} + \mathbf{I}^\top\mathbf{A}, \tag{3.9}$$

$$\sigma_I^2 = \underbrace{\mathbf{A}^\top\left(\mathbf{J} - \mathbf{I}\mathbf{I}^\top\right)\mathbf{A} + 2\widehat{\boldsymbol{\theta}}^\top\left(\mathbf{B} - \boldsymbol{\mu}\mathbf{I}^\top\right)\mathbf{A} + \operatorname{tr}\left\{\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^\top\boldsymbol{\Omega}\right\}}_{V_1} + \underbrace{\sigma^2\left(1 + \eta + \operatorname{tr}\{\mathbf{Q}\mathbf{J}\} + \mathbf{G}^\top\mathbf{C}\mathbf{G} + \operatorname{tr}\left\{\mathbf{C}\mathbf{P} - 2\mathbf{C}\widetilde{\mathbf{H}}^\top\mathbf{R}^{-1}\mathbf{K}\right\}\right)}_{V_2}, \tag{3.10}$$

where

- *$\boldsymbol{\mu} = [\mu_1(\mathbf{x}_1), \ldots, \mu_d(\mathbf{x}_d)]^\top$ and $\left[\widehat{\boldsymbol{\theta}}^\top, \widehat{\boldsymbol{\beta}}^\top\right]^\top \overset{\text{def}}{=\!=} \left(\widetilde{\mathbf{H}}^\top\mathbf{R}^{-1}\widetilde{\mathbf{H}}\right)^{-1}\widetilde{\mathbf{H}}^\top\mathbf{R}^{-1}\mathbf{y}^{\mathcal{T}}$;*
- *$\boldsymbol{\Omega} = \operatorname{diag}(\sigma_1^2(\mathbf{x}_1), \ldots, \sigma_d^2(\mathbf{x}_d))$ and $\mathbf{P} = \operatorname{blkdiag}(\boldsymbol{\Omega}, \mathbf{0})$;*
- *$\mathbf{A} = \mathbf{R}^{-1}\left(\mathbf{y}^{\mathcal{T}} - \mathbf{w}^{\mathcal{T}}\widehat{\boldsymbol{\theta}} - \mathbf{H}(\mathbf{z}^{\mathcal{T}})\widehat{\boldsymbol{\beta}}\right)$ with $\mathbf{H}(\mathbf{z}^{\mathcal{T}}) = [\mathbf{h}(\mathbf{z}_1^{\mathcal{T}}), \ldots, \mathbf{h}(\mathbf{z}_m^{\mathcal{T}})]^\top$;*
- *$\mathbf{Q} = \mathbf{R}^{-1}\widetilde{\mathbf{H}}\left(\widetilde{\mathbf{H}}^\top\mathbf{R}^{-1}\widetilde{\mathbf{H}}\right)^{-1}\widetilde{\mathbf{H}}^\top\mathbf{R}^{-1} - \mathbf{R}^{-1}$ with $\widetilde{\mathbf{H}} = \left[\mathbf{w}^{\mathcal{T}}, \mathbf{H}(\mathbf{z}^{\mathcal{T}})\right]$;*
- *$\mathbf{G} = [\boldsymbol{\mu}^\top, \mathbf{h}(\mathbf{z})^\top]^\top$, $\mathbf{C} = \left(\widetilde{\mathbf{H}}^\top\mathbf{R}^{-1}\widetilde{\mathbf{H}}\right)^{-1}$ and $\mathbf{K} = \left[\mathbf{B}^\top, \mathbf{I}\mathbf{h}(\mathbf{z})^\top\right]$;*

- $\mathbf{I}$ *is a* $m \times 1$ *column vector with the* i*-th element given by*

$$I_i = \prod_{k=1}^{p} c_k(z_k, z_{ik}^{\mathcal{T}}) \prod_{k=1}^{d} \xi_{ik},$$

where $\xi_{ik} \stackrel{\text{def}}{=\!=} \mathbb{E}\left[c_k(W_k(\mathbf{x}_k),\, w_{ik}^{\mathcal{T}})\right]$*;*

- $\mathbf{J}$ *is a* $m \times m$ *matrix with the* ij*-th element given by*

$$J_{ij} = \prod_{k=1}^{p} c_k(z_k,\, z_{ik}^{\mathcal{T}})\, c_k(z_k,\, z_{jk}^{\mathcal{T}}) \prod_{k=1}^{d} \zeta_{ijk},$$

where $\zeta_{ijk} \stackrel{\text{def}}{=\!=} \mathbb{E}\left[c_k(W_k(\mathbf{x}_k),\, w_{ik}^{\mathcal{T}})\, c_k(W_k(\mathbf{x}_k),\, w_{jk}^{\mathcal{T}})\right]$*;*

- $\mathbf{B}$ *is a* $d \times m$ *matrix with the* lj*-th element given by*

$$B_{lj} = \psi_{jl} \prod_{\substack{k=1 \\ k \neq l}}^{d} \xi_{jk} \prod_{k=1}^{p} c_k(z_k, z_{jk}^{\mathcal{T}}),$$

where $\psi_{jl} \stackrel{\text{def}}{=\!=} \mathbb{E}\left[W_l(\mathbf{x}_l)\, c_l(W_l(\mathbf{x}_l),\, w_{jl}^{\mathcal{T}})\right]$*.*

Proof The proof is in Section C.1 of Appendix C. □

Note that V_1 and V_2 in formula (3.10) give a closed form expression for $\text{Var}\left(\mu_g(\mathbf{W}, \mathbf{z})\right)$ and $\mathbb{E}\left[\sigma_g^2(\mathbf{W}, \mathbf{z})\right]$ respectively with $\mu_g(\mathbf{W}, \mathbf{z})$ and $\sigma_g^2(\mathbf{W}, \mathbf{z})$ being the mean and variance of $\widehat{g}$ (see Section C.1 of Appendix C). If we define V_2 as the contribution of $\widehat{g}$ to the variance σ_I^2, V_1 then represents the overall contribution of emulators $\widehat{f}_1, \ldots, \widehat{f}_d$ to the variance σ_I^2. One can also define

$$V_1(\mathbb{S}) \stackrel{\text{def}}{=\!=} \text{Var}_{W_{k \in \mathbb{S}}}\left(\mathbb{E}_{W_{k \in \mathbb{S}^{\mathsf{c}}}}\left[\mu_g(\mathbf{W}, \mathbf{z})\right]\right) \tag{3.11}$$

where $\mathbb{S} \subseteq \{1, \ldots, d\}$ and $\mathbb{S}^{\mathsf{c}}$ is the complement of $\mathbb{S}$, as the contribution of emulators $\widehat{f}_{k \in \mathbb{S}}$ to the variance σ_I^2.

Proposition 3.2 $V_1(\mathbb{S})$ *defined in equation* (3.11) *has the closed form expression given by*

$$V_1(\mathbb{S}) = \mathbf{A}^\top \left(\widetilde{\mathbf{J}} - \mathbf{I}\mathbf{I}^\top\right) \mathbf{A} + 2\widehat{\boldsymbol{\theta}}^\top \left(\widetilde{\mathbf{B}} - \boldsymbol{\mu}\mathbf{I}^\top\right) \mathbf{A} + \text{tr}\left\{\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^\top \widetilde{\boldsymbol{\Omega}}\right\},$$

where

- $\widetilde{\mathbf{\Omega}}$ *is a* $d \times d$ *diagonal matrix with its* k*-th diagonal element given by* $\sigma_k^2(\mathbf{x}_k)\mathbb{1}_{\{k\in\mathbb{S}\}}$;

- $\widetilde{\mathbf{J}}$ *is a* $m \times m$ *matrix with the* ij*-th element given by*

$$\widetilde{J}_{ij} = \prod_{k\in\mathbb{S}} \zeta_{ijk} \prod_{k\in\mathbb{S}^c} \xi_{ik}\xi_{jk} \prod_{k=1}^{p} c_k(z_k,\, z_{ik}^{\mathcal{T}})\, c_k(z_k,\, z_{jk}^{\mathcal{T}});$$

- $\widetilde{\mathbf{B}}$ *is a* $d \times m$ *matrix with the* lj*-th element given by*

$$\widetilde{B}_{lj} = \begin{cases} \psi_{jl} \prod\limits_{\substack{k=1 \\ k\neq l}}^{d} \xi_{jk} \prod\limits_{k=1}^{p} c_k(z_k, z_{jk}^{\mathcal{T}}), & l \in \mathbb{S}, \\ \mu_l \prod\limits_{k=1}^{d} \xi_{jk} \prod\limits_{k=1}^{p} c_k(z_k, z_{jk}^{\mathcal{T}}), & l \in \mathbb{S}^c. \end{cases}$$

Proof The proof is in Section C.2 of Appendix C. □

Equation (3.10) together with Proposition 3.2 thus provides a fast way to evaluate the uncertainty contributions of emulators from different layers, and will be utilised to improve designs of GP emulators across layers in Section 3.5.

Proposition 3.3 *The three expectations* ξ_{ik}, ζ_{ijk} *and* ψ_{jl} *defined in Theorem 3.1 have closed form expressions for all 1-D kernel functions in Table 3.1.*

Proof The derivations under exponential case, squared exponential case and more challenging cases of Matérn-1.5 and Matérn-2.5 are given in Section C.3 of Appendix C. The final closed form expressions for the three expectations are summarised in Appendix B. □

Note that the closed form expressions of μ_I and σ_I^2 in Theorem 3.1 are established under Assumption 2 where the emulators $\widehat{f}_1, \ldots, \widehat{f}_d$ (i.e., the predictive distributions of $W_1(\mathbf{x}_1), \ldots, W_d(\mathbf{x}_d)$) need to be Gaussian. However, $\widehat{f}_1, \ldots, \widehat{f}_d$ may not be Gaussian when the iterative approach reaches the second step ($i = 2$) because the integrated emulators built in the first iteration ($i = 1$) are not Gaussian in general. Therefore, to ensure that the integrated emulator of the computer system $e_{1\to L}$ can be constructed by the iterative approach analytically,

we employ the Gaussian distribution $\mathcal{N}(\mu_I,\, \sigma_I^2)$ with its mean μ_I and variance σ_I^2 matching those given by Theorem 3.1 at each given iteration i. Although the Gaussian distribution with matching mean and variance may not be a good approximation of the actual predictive distribution of $Y(\mathbf{x}_1, \ldots, \mathbf{x}_d,\, \mathbf{z})$ when $i \geq 2$, it minimises the Kullback–Leibler (KL) divergence between the actual predictive density $p(y|\mathbf{x}_1, \ldots, \mathbf{x}_d,\, \mathbf{z})$ and a Gaussian density $\mathcal{N}(\mu, \sigma^2)$ (Minka, 2013):

$$(\mu_I,\, \sigma_I^2) = \underset{\mu,\, \sigma^2}{\operatorname{argmin}}\, \mathcal{KL}\left(p(y|\mathbf{x}_1, \ldots, \mathbf{x}_d,\, \mathbf{z})||\mathcal{N}(\mu, \sigma^2)\right).$$

Thus, the utilisation of Gaussian approximation with matching moments (i.e., mean and variance) at each iteration is justified in the sense of minimised information loss. Once the integrated emulator is constructed by the iterative approach, its empirical version is obtained by plugging the estimates of parameters of individual GP models into the mean and variance of the integrated emulator.

In the remainder of the chapter, the Matérn-2.5 kernel will be used as the default 1-D kernel function for integrated emulation, unless otherwise stated. We choose Matérn-2.5 because we found that it can often prevent from the ill-conditioned correlation matrices (with condition number close to the machine precision) created by the large training size or the poor design (i.e., very closed training points) under the squared exponential kernel. In addition, the Matérn-2.5 kernel still retains most of the smoothness induced by the squared exponential kernel (Gu et al., 2018). As we will demonstrate in Section 3.3, we sometimes need to switch to a Matérn-1.5 kernel when the design becomes extremely poor due to a higher density of training points under large training sizes, a situation where the Matérn-1.5 kernel provides both satisfactory mean predictions and predictive uncertainties. Meanwhile, it provides sufficient smoothness, compared to a very rough exponential kernel. Nevertheless, our integrated emulator can function with all kernels presented in Table 3.1, and different kernels can be used in the GP emulators of different computer models.

3.3 Synthetic Experiments

In this section, we compare the training cost and predictive performance of the integrated emulator with those of the composite emulator in two synthetic computer systems with a different feed-forward structure.

3.3.1 Experiment 1

The first experiment is a system with three computer models composed sequentially (see Figure 3.4). The individual computer models f_1, f_2 and f_3 with scalar-valued output w_1, w_2 and y respectively are defined by the following analytical expressions:

$$f_1 = \sin(\pi x), \quad f_2 = \cos(5w_1) \quad \text{and} \quad f_3 = \sin(w_2^2),$$

where the range of interest for the global input x is between -1 and 1.

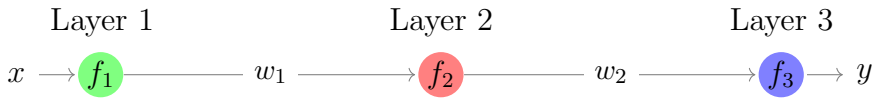


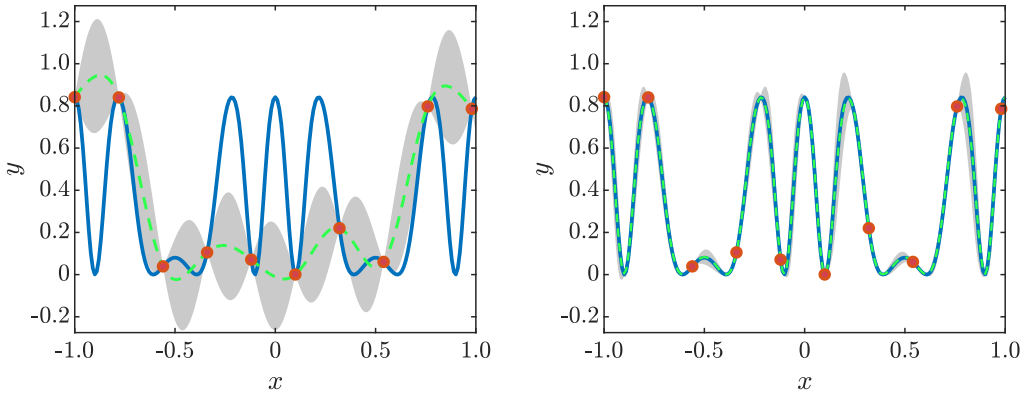
Figure 3.4: Computer system in experiment 1 where f_1, f_2 and f_3 have 1-D input and output.

The constructed composite and integrated emulators with the same ten equally spaced training points are shown in Figure 3.5(a) and 3.5(b) respectively. The comparison demonstrates that the integrated emulator drastically outperforms the composite one, with excellent mean predictions and small predictive variances under identical information.

To compare the training cost between the composite and integrated emulators, we compute at seven different training set sizes (i.e., 5, 10, 15, 20, 30, 40 and 50) the normalised root mean squared error of prediction (NRMSEP) that is defined by

$$\text{NRMSEP} = \frac{\sqrt{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(y(\mathbf{x}_i) - \mu_Y^t(\mathbf{x}_i))^2}}{\max\{y(\mathbf{x}_i)_{i=1,\ldots,n}\} - \min\{y(\mathbf{x}_i)_{i=1,\ldots,n}\}}, \tag{3.12}$$

where $y(\mathbf{x}_i)$ denotes the true global output of the system evaluated at the



(a) Composite Emulator **(b)** Integrated Emulator

Figure 3.5: Composite and integrated emulators of the computer system in experiment 1. The solid line is the true functional form between the global input and output of the system; the dashed line is the mean prediction; the shaded area represents 95% prediction interval; the filled circles are training points used to construct the emulators.

testing input position $\mathbf{x}_i$ for $i = 1, \ldots, n$; $\mu_Y^t(\mathbf{x}_i)$ is the mean prediction of the respective (integrated or composite) emulator built with the t-th training set of total T training sets, each of which has the same size of training points.

At each training set size, the corresponding NRMSEP is evaluated at $n = 100$ testing positions equally spaced over $[-1, 1]$ and $T = 100$ randomly generated training sets from the maximin Latin hypercube sampling. For the training set size of 40 and 50, we use Matérn-1.5 instead the default Matérn-2.5 kernel for the GP emulator of f_2. This is because when training size is large Latin hypercube designs on x can produce poor designs on w_1 (i.e., very closed training positions), causing ill-conditioned correlation matrix (i.e., large condition number exceeding 10^{12}) for the GP model of f_2 with Matérn-2.5 kernel and thus inaccurate mean predictions from the resulting integrated emulator. The comparison in Figure 3.6 provides two implications. Firstly, the integrated emulator effectively reduces to almost zero NRMSEP with a small number of training points (i.e., around 15). In contrast, the composite emulator slowly reaches to a negligible NRMSEP with 50 training points. Secondly, at a given training set size (e.g., 15), the integrated emulator can achieve significantly more reductions in predictive error

than the composite emulator.

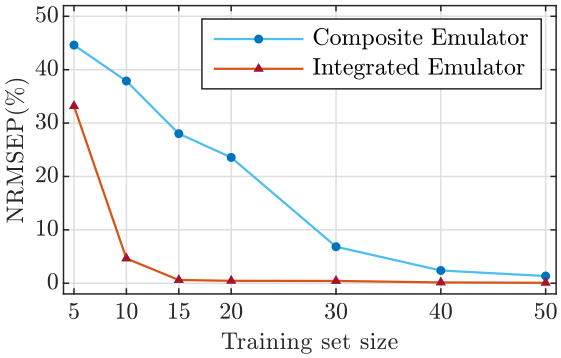


Figure 3.6: NRMSEP of composite and integrated emulators in experiment 1.

3.3.2 Experiment 2

In this experiment, we explore the predictive performance of the integrated emulator in the computer system shown in Figure 3.7. The three computer models in the system have the following analytical functional forms:

$$f_1 = 30 + 5x_1 \sin(5x_1), \quad f_2 = 4 + \exp(-5x_2) \quad \text{and} \quad f_3 = (w_1 w_2 - 100)/6$$

with $x_1 \in [0, 2]$ and $x_2 \in [0, 2]$.

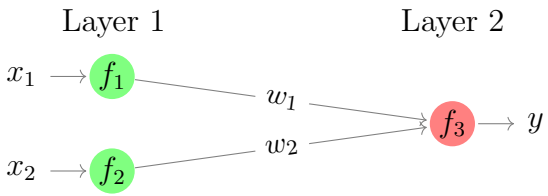
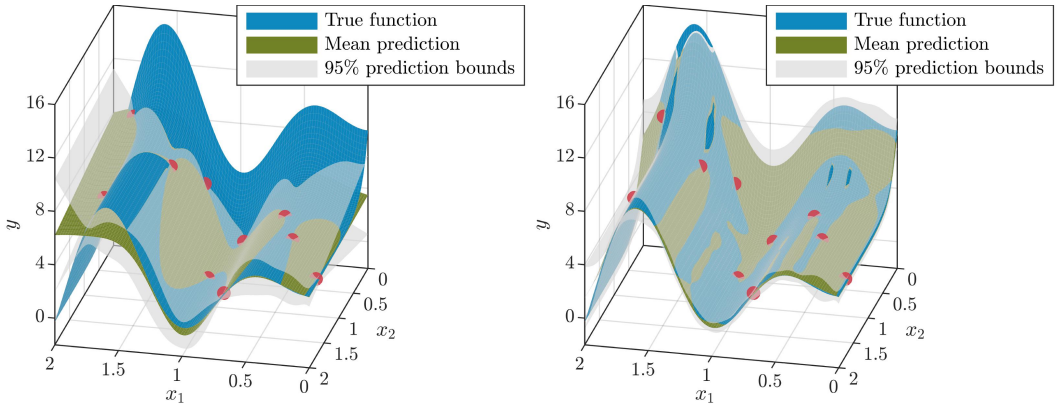


Figure 3.7: The computer system in experiment 2 where f_1 and f_2 are two computer models with one-dimensional input and output, and f_3 is a computer model with two-dimensional input and one-dimensional output.

The composite (Figure 3.8(a)) and integrated (Figure 3.8(b)) emulators of the system are constructed with ten training points generated by the maximin Latin hypercube sampling. For the integrated emulator, a Matérn-1.5 kernel with a nugget term is chosen for the GP emulator of f_3. This is because under a Matérn-2.5 kernel (even with a nugget term), the estimated correlation matrix is ill-conditioned (with condition number around 10^{15}) due to the relatively

large estimates of range parameters. Such an ill-conditioned matrix causes significant round-off errors in double precision arithmetic, and thus severely degrades the predictive accuracy of the integrated emulator. Figure 3.8 shows that the integrated emulator outperforms the composite emulator in terms of both mean predictions and prediction bounds. While the composite emulator fails to mimic the true system function in areas where the training points are scarce, the integrated emulator matches the true function well even over regions (e.g, the peak and ridge) far away from the training points.



(a) Composite Emulator **(b)** Integrated Emulator

Figure 3.8: The composite and integrated emulators of the system in experiment 2. The filled circles are training points used to construct the emulators.

The predictive performances of the composite and integrated emulators are further compared by computing the NRMSEP at 12 training set sizes (i.e., 5, 10, 15, 20, 30, . . . , 100). At each selected training set size, NRMSEP of both composite and integrated emulators are calculated based on $n = 10000$ testing position equally spaced over the global input domain $[0, 2] \times [0, 2]$ and $T = 100$ Latin hypercube samples. Figure 3.9 shows that the NRMSEP of the integrated emulator quickly drops to values close to zero with only 20 training points. In contrast, the NRMSEP of the composite emulator slowly decays to a negligible level at a training set size around 60. This corroborates the superiority of the integrated emulator for a computer system with multiple computer models in a layer.

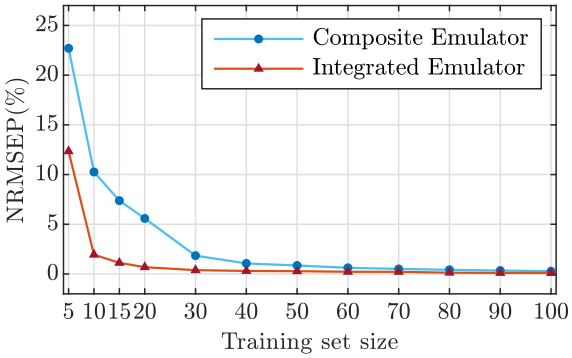


Figure 3.9: NRMSEP of composite and integrated emulators in Experiment 2.

From both this experiment and the experiment 1, we note that Matérn-2.5 and Matérn-1.5 kernels are essential to build integrated emulators of feed-forward computer systems because they offer reasonable choices on smoothness while at the same time efficiently alleviate the issue of ill-conditioned correlation matrices caused by sources such large range parameter estimates and poor designs (especially when sample size is large). Furthermore, in Section 3.5 we will discuss a smart designing strategy that can further mitigate such numerical issues caused by the poor designs of individual computer models.

3.4 Integrated Emulator for a Feed-Back Coupled Satellite Model

In this section, we construct the integrated emulator of the fire-detection satellite model studied in Sankararaman and Mahadevan (2012). This satellite is designed to conduct near-real-time detection, identification and monitoring of forest fires. The satellite system consists of three sub-models, namely the orbit analysis, the attitude control and power analysis. The satellite system is shown in Figure 3.10. It can be seen from Figure 3.10 that there are nine global input variables H, F_s, θ, L_{sp}, q, R_D, L_a, C_d, P_{other} and three global output variables of interest τ_{tot}, P_{tot}, A_{sa}. The coupling variables are Δt_{orbit}, $\Delta t_{eclipse}$, ν, θ_{slew}, P_{ACS}, I_{max} and I_{min}. Since Δt_{orbit} is the input to both power analysis and attitude control, there are total eight coupling variables.

Note that the system has feed-back coupling because the coupling variables P_{ACS}, I_{max} and I_{min} form an internal loop between power analysis and attitude control. Therefore, to implement the integrated emulation framework on the global output variables, the system is converted to a feed-forward one by applying the decoupling algorithm proposed in Baptista et al. (2018). The decoupling algorithm identifies four weakly coupled variables Δt_{orbit} (between orbit analysis and attitude control), θ_{slew}, I_{max} and I_{min}. Since the weakly coupled variables have insignificant impact on the accuracy of global outputs, they are neglected from the interaction terms between sub-models, producing a feed-forward system (see Figure 3.10 without the dashed arrows). Table 3.2 gives the domains of global inputs considered for the emulation.

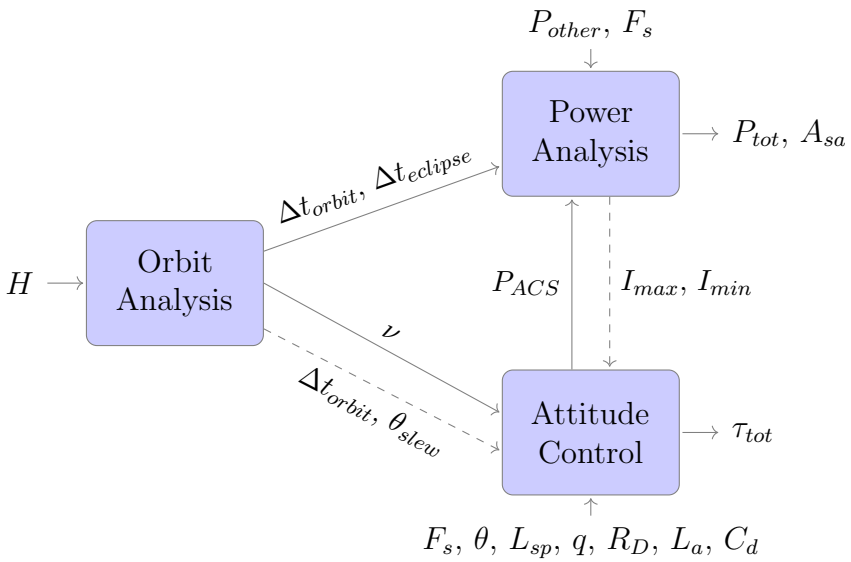


Figure 3.10: Fire-detection satellite model from Sankararaman and Mahadevan (2012), where H is altitude; Δt_{orbit} is orbit period; $\Delta t_{eclipse}$ is eclipse period; ν is satellite velocity; θ_{slew} is maximum slewing angel; P_{other} represents other sources of power; P_{ACS} is power of attitude control system; I_{max}, I_{min} are maximum and minimum moment of inertia respectively; F_s, θ, L_{sp}, q, R_D, L_a, C_d represent average solar flux, deviation of moment axis from vertical, moment arm for the solar radiation torque, reflectance factor, residual dipole, moment arm for aerodynamic torque, and drag coefficient respectively; P_{tot} is total power; A_{sa} is area of solar array; and τ_{tot} is total torque. The dashed arrows indicate the connections that can be decoupled between sub-models, according to the decoupling algorithm from Baptista et al. (2018).

Table 3.2: Domains of the nine global input variables to be considered for the emulation.

Global input variable (unit)	Symbol	Domain
Altitude (m)	H	$[1.50 \times 10^{17}, 2.10 \times 10^{17}]$
Other sources of power (W)	P_{other}	$[8.50 \times 10^{2}, 1.15 \times 10^{3}]$
Average solar flux (W/m^2)	F_s	$[1.34 \times 10^{3}, 1.46 \times 10^{3}]$
Deviation of moment axis from vertical ($^\circ$)	θ	[12.00, 18.00]
Moment arm for the solar radiation torque (m)	L_{sp}	[0.80, 3.20]
Reflectance factor	q	[0, 1]
Residual dipole ($A \cdot m^2$)	R_D	[2.00, 8.00]
Moment arm for aerodynamic torque (m)	L_a	[0.80, 3.20]
Drag coefficient	C_d	[0.10, 1, 90]

Maximin Latin hypercube sampling is then used to generate inputs positions for seven training sets, with sizes of 10, 15, 20, 25, 30, 35 and 40 respectively. The corresponding output positions are consequently obtained by running the satellite model. For each of the seven training set and each of the three global output variables, we build the composite and integrated emulators. Leave-one-out cross-validation is utilised for assessing the predictive performance of the emulators. For example, in case of the composite emulation of the output variable P_{tot} with training set size of 10, we build ten composite emulators, each based on nine training points by dropping one training point out of the set. The dropped training point is then serves as the testing point to assess the associated composite emulator. The performance of the emulator (composite or integrated) of a global output variable given a certain training set is ultimately summarised by

$$\text{NRMSEP} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(f(\mathbf{x}_i) - \mu^{-i}(\mathbf{x}_i))^2}}{\max\{f(\mathbf{x}_i)_{i=1,\ldots,n}\} - \min\{f(\mathbf{x}_i)_{i=1,\ldots,n}\}},$$

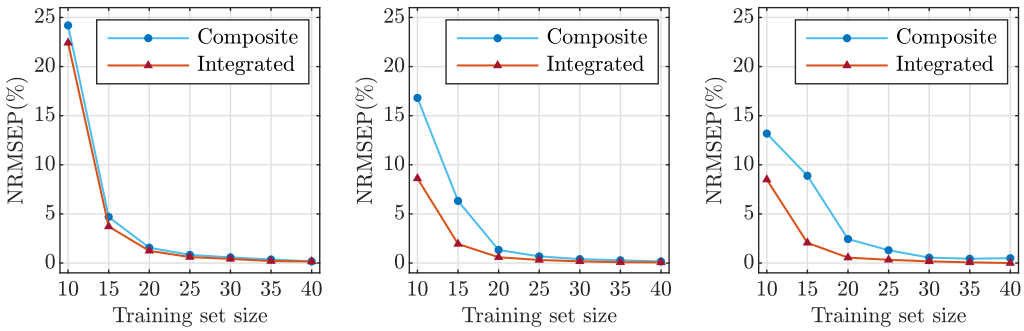
where $\mathbf{x}_i$ is the i-th input position of a training set with size n; $f(\mathbf{x}_i)$ is the value of the output variable of interest produced by the satellite model at the input $\mathbf{x}_i$; the mean prediction $\mu^{-i}(\mathbf{x}_i)$ at input $\mathbf{x}_i$ is provided by the corresponding (composite or integrated) emulator constructed using all n training points

except for $\mathbf{x}_i$.

The NRMSEP of the composite and integrated emulators of the three global output variables τ_{tot}, P_{tot} and A_{sa} against seven different training sizes are presented in Figure 3.11. It can be seen that for the output variable τ_{tot}, the integrated emulator is only marginally better than the composite one. This is because the functional complexity between the global inputs and the output τ_{tot} is dominated by the sub-model attitude control, and thus the integrated emulator shows no obvious superiority over the composite emulator. This explanation can be inferred from Figure 3.12(a) and 3.12(b), where the GP emulator of the attitude control with respect to τ_{tot} requires more training points than that of the orbit analysis with respect to ν to reach a low NRMSEP. For the output variables P_{tot} and A_{sa}, the integrated emulators present better predictive performance than the composite ones at training set size ranging from 10 to 20, while show little superiority after the training set size increases over 20. The better predictive performance of the integrated emulators at small training sizes can be explained by noting that P_{tot} and A_{sa} are produced not only by the orbit analysis and attitude control, but also by the power analysis. Although the attitude control still dominates the functional complexity between the global inputs and P_{tot} and A_{sa} (see Figure 3.12), the power analysis has higher input dimensions than the orbit analysis, causing the composite emulators slow to learn the functional dependence of P_{tot} and A_{sa} to the global inputs with a small number of training points.

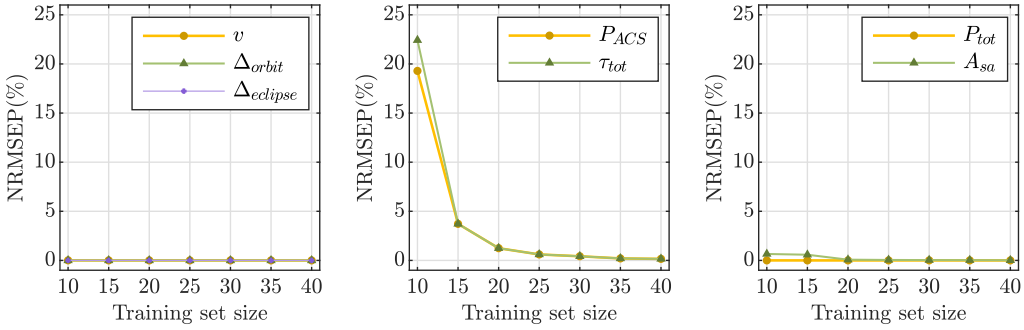
3.5 Towards a Smart Design for Integrated Emulation

We have so far demonstrated that the integrated emulator outperforms the composite emulator in general, while in cases where the functional complexity of the whole system is dominated by a single computer model, the integrated emulator naturally provides comparable predictive performance to the composite



(a) τ_{tot} **(b)** P_{tot} **(c)** A_{sa}

Figure 3.11: The NRMSEP of the composite and integrated emulators of the three global output variables τ_{tot}, P_{tot} and A_{sa} against different training set sizes.



(a) Orbit Analysis **(b)** Attitude Control **(c)** Power Analysis

Figure 3.12: The NRMSEP of the GP emulators of outputs produced by the three subsystems: orbit analysis, attitude control and power analysis.

one. Nevertheless, even in this situation, the design for the integrated emulation can be improved, with potentially large gains. In this section, we discuss an adaptive designing strategy for the integrated emulator. Before exploring the strategy under a general system structure, we first present the design in a simple feed-forward system of two computer models f_1 and f_2 (producing scalar-valued output w and y respectively) with the following analytical functional forms:

$$f_1 = \frac{2}{1+\exp(-2x)} \quad \text{and} \quad f_2 = \cos(2\pi w), \quad x \in [-4, 4].$$

3.5.1 Latin hypercube design

The space-filling Latin hypercube design (LHD) (Santner et al., 2003) has been used to construct the integrated emulators of all examples illustrated so far. For the computer system under the consideration, the LHD first samples the training positions of global input x via the maximin Latin hypercube method to determine the GP emulator $\widehat{f_1}$. The design for the GP emulator $\widehat{f_2}$ (i.e., the training positions of w) is then specified by evaluating f_1 at the training positions of x. However, such design may not be optimal for the integrated emulation.

In Figure 3.13(a), 3.13(b) and 3.13(c), showcasing our example, GP emulators $\widehat{f_1}$ and $\widehat{f_2}$, and the corresponding integrated emulator $\widehat{f_2} \circ \widehat{f_1}$ constructed by the LHD are presented respectively. Although the ten training points drawn from the LHD produce a well-behaved GP emulator of computer model f_1, the GP emulator of computer model f_2 presents unsatisfactory predictive performance between 0.5 and 1.5. Such predictive deficiency in GP emulator of f_2 propagates to the integrated emulator, which fails to capture the peak shape of $f_2 \circ f_1$ around 0. The reason for the unsatisfactory predictive performance of the resulting integrated emulator is that f_1 exhibits a steep rise as x increases from -1 to 1, causing few training points to be sampled by the LHD over this range. Consequently, the design for the GP emulator of f_2 is poorly spaced with insufficient information over [0.5, 1.5]. Another issue with the LHD is that the design for $\widehat{f_2}$ consists of excessive training points at its boundary. These dense points are created by the flat wings of f_1 and may cause numerical challenges for GP model fitting and prediction, especially when the size of the training set is large. Therefore, a better designing strategy is needed to improve the LHD by smartly choosing designs for individual computer models, especially for f_2.

3.5.2 An adaptive design for integrated emulation

Note that the variance of the integrated emulator can be decomposed into contributions V_1 and V_2 from GP emulator $\widehat{f_1}$ and $\widehat{f_2}$ respectively (see the

discussion following Theorem 3.1). By utilising this fact, an adaptive strategy is developed in Algorithm 5 to smartly enrich the existing designs for f_1 and f_2 and update their corresponding GP emulators.

Algorithm 5 Adaptive design for emulating a system of two computer models

1: Choose K number of enrichment to the existing design.
2: **for** $k = 1, \ldots, K$ **do**
3: Find x_0 and l_0 such that

$$(x_0,\, l_0) = \underset{x,\, l \in \{1,\, 2\}}{\operatorname{argmax}} V_l(x),$$

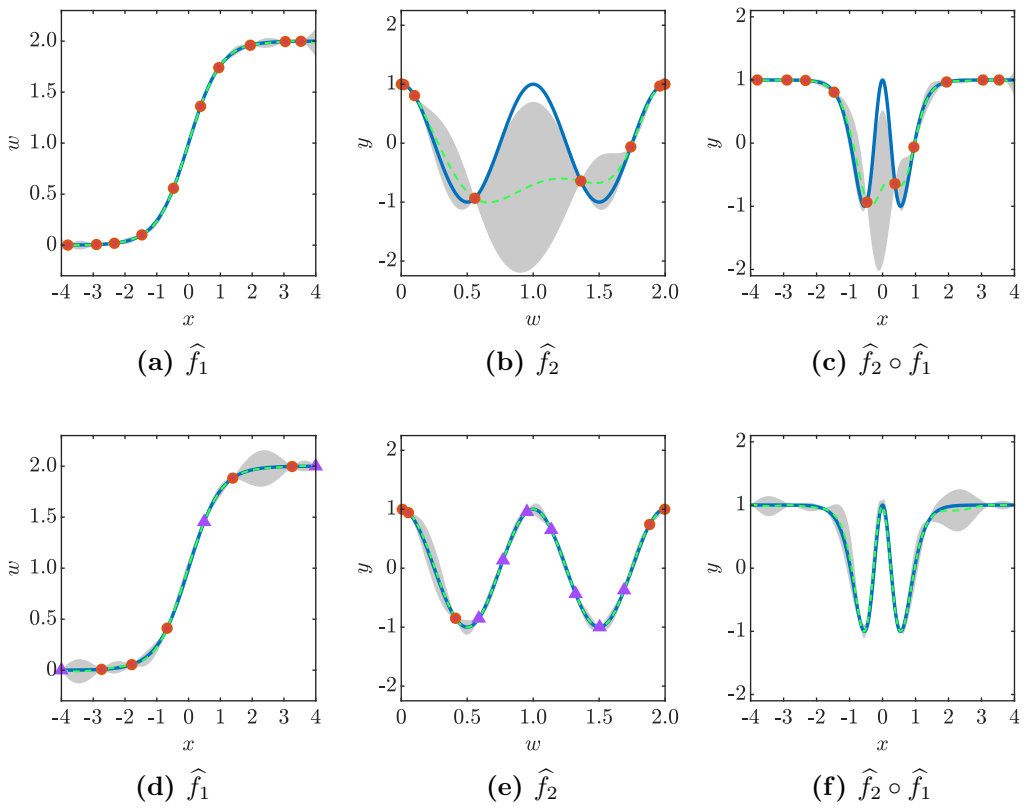
where $V_l(x)$ is the contribution of $\widehat{f}_l$ to the variance of the integrated emulator;
4: **if** $l_0 = 1$ **then**
5: Enrich the training points for $\widehat{f}_1$ by evaluating f_1 at the input position x_0;
6: **else**
7: Enrich the training points for $\widehat{f}_2$ by evaluating f_2 at the input position $\mu_1(x_0)$, obtained by evaluating the predictive mean μ_1 of $\widehat{f}_1$ at the input position x_0;
8: **end if**
9: Update the GP emulator $\widehat{f}_l$ with the added training point.
10: **end for**

A similar training strategy to Algorithm 5 is discussed by Sanson et al. (2019). However, they compute V_1 and V_2 numerically, resulting inaccurate and slow evaluation of the maximisation problem on line 3 of Algorithm 5. Thanks to the analytical framework of the integrated emulation, V_1 and V_2 can be expressed in closed form using the formula (3.10), and therefore Algorithm 5 can be implemented faster and more accurately.

To demonstrate the performance of this design, we construct the initial designs for f_1 and f_2 with five training points generated by the maximin Latin hypercube sampling. The adaptive design is then applied to enrich the designs of f_1 and f_2 with $K = 10$. The resulting GP emulators for f_1 and f_2 and the corresponding integrated emulator are shown in Figure 3.13(d), 3.13(e) and 3.13(f) respectively. It can be observed that the adaptive designing strategy smartly enriches the initial design for f_2 by choosing positions of w that correspond to the steep

segment of f_1. As a result, the final integrated emulator provides a better predictive performance than that constructed by the LHD.

Furthermore, at each iteration the adaptive design only requires the evaluation of a single computer model without running the whole system. In this case, the adaptive design asks for three evaluations of f_1 while seven evaluations of f_2. This property of the adaptive design can be particularly useful when the system contains computer models with heterogeneous functional complexity (i.e., non-linearity) because it allows different computer models with different functional complexities to be trained with different training costs.



(a) $\widehat{f}_1$ **(b)** $\widehat{f}_2$ **(c)** $\widehat{f}_2 \circ \widehat{f}_1$

(d) $\widehat{f}_1$ **(e)** $\widehat{f}_2$ **(f)** $\widehat{f}_2 \circ \widehat{f}_1$

Figure 3.13: The GP emulators $\widehat{f}_1$, $\widehat{f}_2$ and the integrated emulator $\widehat{f}_2 \circ \widehat{f}_1$ trained with the LHD *(first row)* and adaptive design *(second row)*. The filled circles are training points for LHD or the initial design for the adaptive design; the filled triangles are training points created by the adaptive design; the solid line is the underlying true function; the dashed line is the mean prediction; the shaded area represents 95% prediction interval.

3.5.3 Design comparison

In this section, we compare the LHD and the adaptive design in terms of the predictive performance of the resulting integrated emulator and the associated training cost. For the LHD, ten integrated emulators, each based on a different sample from the maximin Latin hypercube method, are constructed at nine training set sizes (i.e., 5, 6, 8, 10, . . . , 18, 20). These training set sizes correspond to the total number of computer model evaluations that are 10, 12, 16, 20, . . . , 36, 40 respectively (double due to two computer models). For the adaptive design, ten random samples with five training points (i.e., ten computer model runs) are generated by the maximin Latin hypercube method as the initial designs and each initial design is enriched by 30 training points (i.e., 30 computer model runs). The NRMSEP defined by equation (3.12) is used for both designs. From the left plot in Figure 3.14 we see that the integrated emulator under the adaptive design provides better predictive performance than the one under the LHD with the same number of computer model runs. Given the same overall number of computer model evaluations, the adaptive design allocates more runs to the computer model f_2 than to f_1, which is less functionally complex. Whereas, the LHD allocates runs equally to f_1 and f_2 without appreciating the difference of functional complexity between the two computer models (see the right plot in Figure 3.14).

The left plot in Figure 3.14 also indicates that to achieve a similar accuracy (in terms of NRMSEP) the integrated emulator trained with the adaptive design requires significantly smaller amount of evaluations of computer models. To see how this saving of evaluations on computer models translates to the reduction of system run time for the integrated emulation, we consider three scenarios where the computational time for running computer model f_2 is 100, 1 and 0.01 times that for running computer model f_1, respectively.

The first scenario represents the cases where the computer models with more complex functional forms are also more expensive to run, while the third

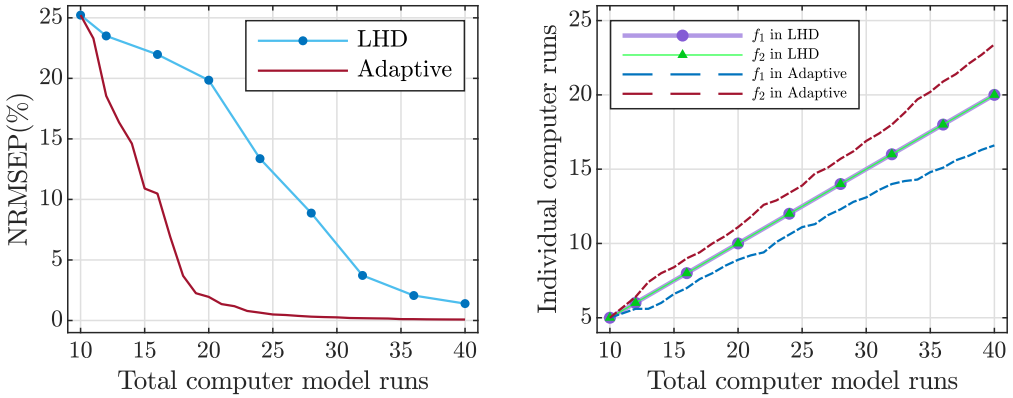


Figure 3.14: *(Left)* The NRMSEP of the integrated emulators constructed under the LHD and the adaptive design at various number of computer model runs. *(Right)* The number of evaluations of computer models f_1 and f_2 under the LHD and the adaptive design.

scenario represents the situations where the computational cost is expensive for computer models with simple functional forms. The reductions on the system run time due to the use of the adaptive design for the integrated emulation at different levels of NRMSEP are illustrated in Figure 3.15.

For all three scenarios, the adaptive design reduces the run time used by the LHD for integrated emulation, and such reduction becomes more remarkable when a higher accuracy of the integrated emulator is targeted. In scenario 2 the adaptive design saves more than 40% of the time spent by the LHD to construct the integrated emulator with a moderate-to-low NRMSEP. This reduction goes around 50% and above in scenario 3. Even for scenario 1, the adaptive design can save more than 30% of total run time for a relatively well performed integrated emulator.

In addition to the run time reduction, the adaptive design also reduces the risk of numerical issues related to the integrated emulation. Since the adaptive design only updates the GP emulators that contribute most to the variance of the final integrated emulator (i.e., GP emulators who contribute less are not retrained at each enrichment), numerical issues, such as the increased computational time for inverting the correlation matrices with larger training

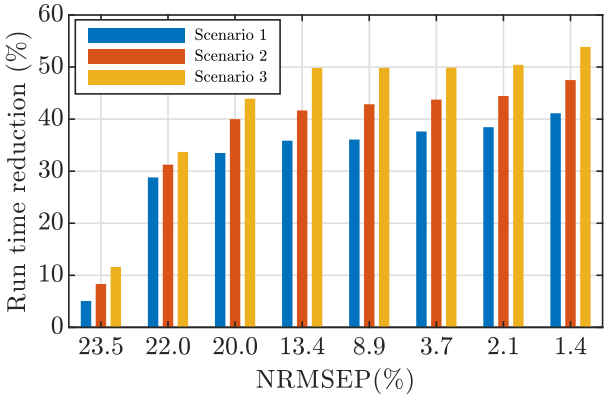


Figure 3.15: The run time reduction for the integrated emulation by the adaptive design under three different hypothetical scenarios. In scenario 1, the computer model f_2 is 100 times more expensive than the computer model f_1 to run; in scenario 2, computer model f_1 and f_2 are equally expensive to run; in scenario 3, the computer model f_1 is 100 times more expensive than the computer model f_2 to run.

sizes and the ill-conditioned correlation matrices due to the poorly spaced training points, can be mitigated to some extent.

3.5.4 Generalisation of the adaptive design

By utilising Theorem 3.1 with Proposition 3.2, one can generalise the adaptive design to the integrated emulation of any feed-forward computer system. In this section, we demonstrate such generalisation by considering the two synthetic experiments discussed in Section 3.3. Algorithm 6 and 7 present the adaptive design strategies for the two experiments, respectively.

The training of the integrated emulators of the two systems by the adaptive designs in Algorithm 6 and 7 are shown in Figure 3.16(a) and 3.16(b), respectively. It can be seen from Figure 3.16(a) that for the computer system in experiment 1, the integrated emulator trained by the adaptive design can achieve a low NRMSEP with smaller number of computer model runs than that built by the gridded design (i.e., equally spaced design points). Similar observation can be seen for the integrated emulator of the computer system in experiment 2 from Figure 3.16(b). However, unlike experiment 1, in experiment 2 the integrated emulator by the adaptive design can achieve lower NRMSEP than that by

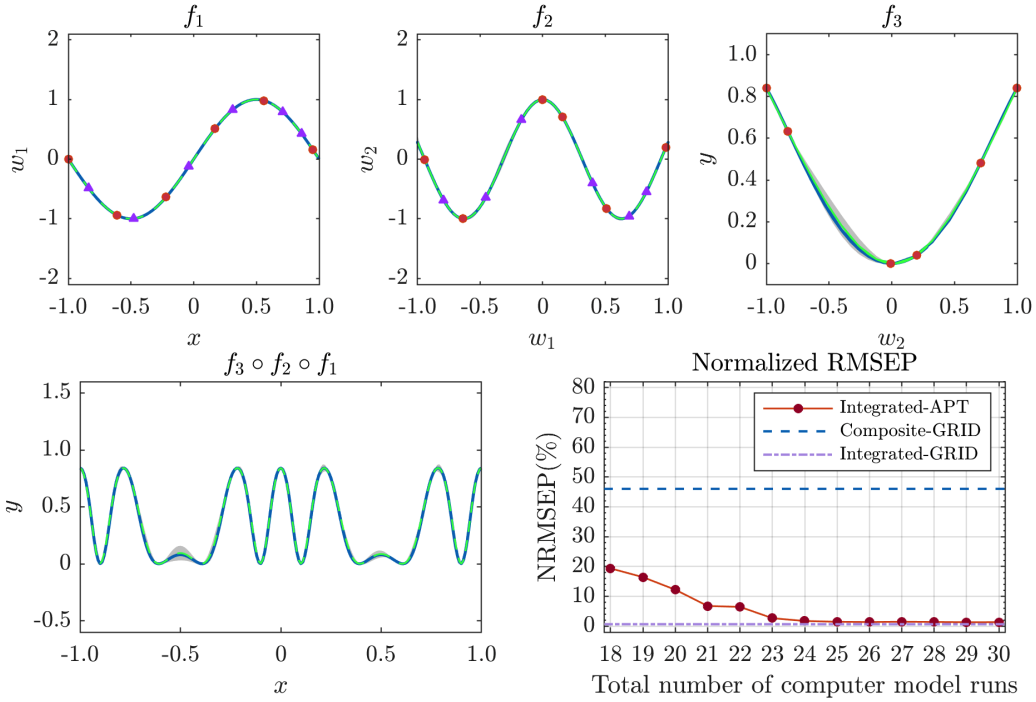
Algorithm 6 Adaptive design for emulating the computer system in experiment 1 of Section 3.3

Choose K number of enrichment to the existing design.
for $k = 1, \ldots, K$ **do**
  Find x_0 and l_0 such that

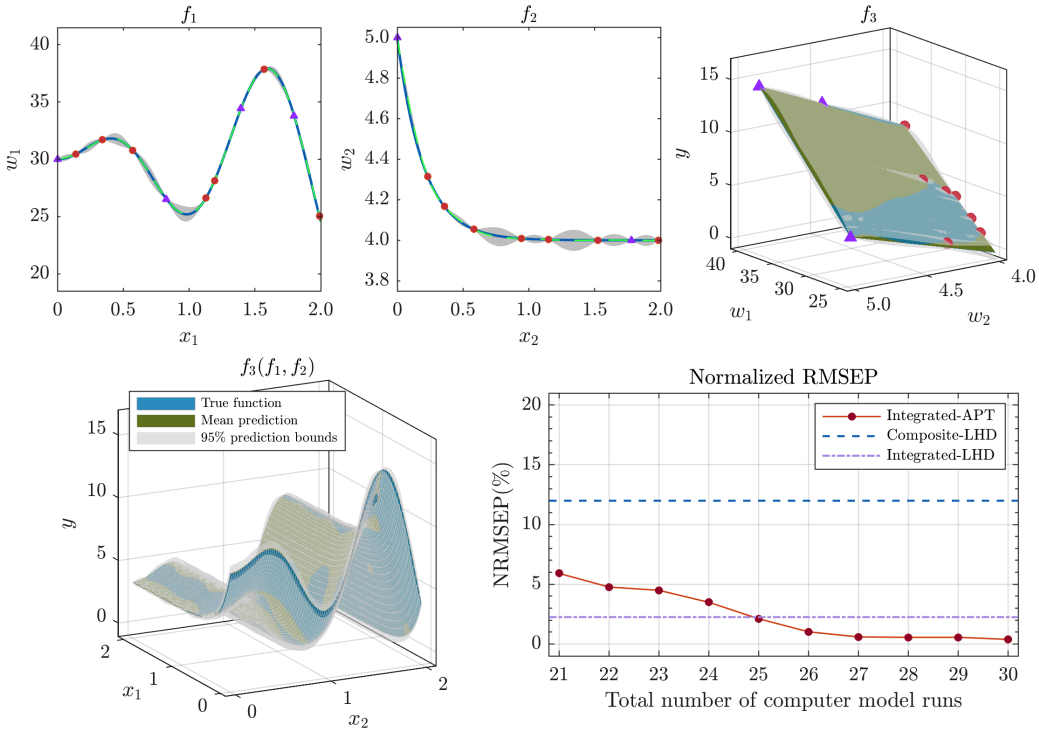
$$(x_0,\, l_0) = \underset{x,\, l \in \{1,\, 2\}}{\operatorname{argmax}} V_l^{1\to 3}(x),$$

  where $V_1^{1\to 3}(x)$ and $V_2^{1\to 3}(x)$ respectively are contributions of $\widehat{e}_{1\to 2}$ (i.e., the integrated emulator of system $e_{1\to 2}$ consisting of f_1 and f_2) and $\widehat{f}_3$ (i.e., the GP emulator of f_3) to the variance of integrated emulator $\widehat{e}_{1\to 3}$;
  if $l_0 = 1$ **then**
    Compute $V_k^{1\to 2}(x_0)$ for $k \in \{1,\, 2\}$ according to Theorem 3.1, where $V_k^{1\to 2}(x_0)$ is the contribution of $\widehat{f}_k$ to the variance of integrated emulator $\widehat{e}_{1\to 2}$;
    if $V_1^{1\to 2}(x_0) > V_2^{1\to 2}(x_0)$ **then**
      Enrich the training points for $\widehat{f}_1$ by evaluating f_1 at the input position x_0;
    else
      Enrich the training points for $\widehat{f}_2$ by evaluating f_2 at the input position $\mu_1(x_0)$, obtained by evaluating the predictive mean μ_1 of $\widehat{f}_1$ at the input position x_0;
    end if
  else
    Enrich the training points for $\widehat{f}_3$ by evaluating f_3 at the input position $\mu_I(x_0)$, obtained by evaluating the predictive mean μ_I of $\widehat{e}_{1\to 2}$ at the input position x_0;
  end if
  Update the GP emulator $\widehat{f}_1$, $\widehat{f}_2$ or $\widehat{f}_3$ with the added training point.
end for

the space-filling design. This is because in experiment 1 the gridded design points for the global input x create designs for w_1 and w_2 that are relatively well-spaced, producing the integrated emulator with comparable performance to that trained by the adaptive design. On the contrary, in experiment 2 the LHD of the global inputs x_1 and x_2 produce poor designs for f_3 (i.e., designing points are concentrated along one boundary of the input space), causing the resulting integrated emulator with higher NRMSEP.



(a) Experiment 1: 30 computer runs with the first 18 runs for the initial design



(b) Experiment 2: 30 computer runs with the first 21 runs for the initial design

Figure 3.16: The adaptive designs for the two synthetic experiments in Section 3.3. In each sub-panel ((a) or (b)): *(Top)* the GP emulators of f_1, f_2 and f_3 trained after each enrichment (i.e., computer model run); *(Bottom-left)* integrated emulator trained after each enrichment; *(Bottom-right)* NRMSEP of the integrated emulator after each enrichment: the dashed and dash-dot lines represent the NRMSEP of the composite and integrated emulators trained with 10 equally spaced (for experiment 1) and maximin Latin hypercube (for experiment 2) designing points.

Algorithm 7 Adaptive design for emulating the computer system in experiment 2 of Section 3.3

Choose K number of enrichment to the existing design.
for $k = 1, \ldots, K$ **do**
  Find $\mathbf{x}_0$ and l_0 such that

$$(\mathbf{x}_0,\, l_0) = \underset{\mathbf{x},\, l \in \{1,\, 2\}}{\operatorname{argmax}} V_l(\mathbf{x}),$$

  where $\mathbf{x} = (x_1,\, x_2)$, $\mathbf{x}_0 = (x_{01},\, x_{02})$, and $V_1(\mathbf{x})$ and $V_2(\mathbf{x})$ respectively are contributions of $\widehat{e}_1$ (i.e., GP emulators $\widehat{f}_1$ and $\widehat{f}_2$ in the first layer) and $\widehat{f}_3$ to the variance of integrated emulator;
  if $l_0 = 1$ **then**
    Compute $V_{1k}(\mathbf{x}_0)$ for $k \in \{1,\, 2\}$ according to Proposition 3.2, where $V_{1k}(\mathbf{x}_0)$ is the contribution of $\widehat{f}_k$ to the variance of integrated emulator;
    if $V_{11}(\mathbf{x}_0) > V_{12}(\mathbf{x}_0)$ **then**
      Enrich the training points for $\widehat{f}_1$ by evaluating f_1 at the input position x_{01};
    else
      Enrich the training points for $\widehat{f}_2$ by evaluating f_2 at the input position x_{02};
    end if
  else
    Enrich the training points for $\widehat{f}_3$ by evaluating f_3 at the input position $(\mu_1(x_{01}),\, \mu_2(x_{02}))$, obtained by evaluating the predictive mean μ_1 and μ_2 of $\widehat{f}_1$ and $\widehat{f}_2$ at the input position x_{01} and x_{02}, respectively;
  end if
  Update the GP emulator $\widehat{f}_1$, $\widehat{f}_2$ or $\widehat{f}_3$ with the added training point.
end for

3.6 Discussion

The development of integrated emulators depends on Assumption 1 and 2 presented in Section 3.2.2. The first assumption requires the trend function of individual Gaussian process emulators in layer 2 and above have linear forms of their respective inputs. This assumption generally is not an issue because one can set constant trends for Gaussian process emulators and have the underlying function forms to be explained by the chosen kernel functions (i.e., covariance matrices). This constant trend specification is used in all examples illustrated in this study to construct Gaussian process emulators. However, it is worth

noting that well-specified trend functions can help reduce design points needed to build Gaussian process emulators (e.g., a function with sine components can be much easier to learnt by the Gaussian process with a sine trend), while it is a challenging work to specify the trend function that is a good description of the underlying black-box computer model. Therefore, there is a trade-off between the modelling flexibility of Gaussian process emulators and the required cost of designs (i.e., the number of computer model evaluations).

The second assumption asks for both independence and normality of input variables of individual Gaussian process emulators. The independence assumption helps reduce analytical efforts in deriving the closed form mean and variance of the integrated emulator. In addition, the consideration of dependence between input variables requires specification of their dependence structures, which is a difficult task as this requires careful dependence modelling and extra computational cost for Gaussian process model estimation and prediction. Nevertheless, ignoring the dependence structure between input variables can cause biased mean and variance of integrated emulators constructed by Theorem 3.1. To show how such biases can be quantified, in Proposition 3.4 we present the mean and variance of integrated emulators under the squared exponential kernel when the dependence between inputs are considered.

Proposition 3.4 *Assume that* $\mathbf{W} \sim \mathcal{MN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*, where* $\boldsymbol{\Sigma}$ *is the covariance matrix of* $\mathbf{W}$ *with diagonal elements being* $\sigma_1^2(\mathbf{x}_1), \ldots, \sigma_d^2(\mathbf{x}_d)$*. The mean and variance of the integrated emulator under the squared exponential kernel is given by those from Theorem 3.1 with* $\boldsymbol{\Omega} = \boldsymbol{\Sigma}$ *and*

- *the i-th element of* $\mathbf{I}$*:*

$$I_i = \widetilde{\xi}_i \prod_{k=1}^{p} c_k(z_k, z_{ik}^{\mathcal{T}}),$$

where

$$\widetilde{\xi}_i = \frac{1}{\sqrt{|(\boldsymbol{\Lambda} + \boldsymbol{\Sigma})\boldsymbol{\Lambda}^{-1}|}} \exp\left\{-\frac{1}{2}(\boldsymbol{\omega}_i^{\mathcal{T}} - \boldsymbol{\mu})^{\top}(\boldsymbol{\Lambda} + \boldsymbol{\Sigma})^{-1}(\boldsymbol{\omega}_i^{\mathcal{T}} - \boldsymbol{\mu})\right\}$$

with $\boldsymbol{\Lambda} = \mathrm{diag}(\frac{\gamma_1^2}{2}, \ldots, \frac{\gamma_d^2}{2})$*;*

- *the ij-th element of* $\mathbf{J}$*:*

$$J_{ij} = \widetilde{\zeta}_{ij} \prod_{k=1}^{p} c_k(z_k,\, z_{ik}^{\mathcal{T}})\, c_k(z_k,\, z_{jk}^{\mathcal{T}}),$$

where

$$\widetilde{\zeta}_{ij} = \frac{1}{\sqrt{|(\boldsymbol{\Gamma}+\boldsymbol{\Sigma})\boldsymbol{\Gamma}^{-1}|}} \exp\left\{-\frac{1}{8}(\boldsymbol{\omega}_i^{\mathcal{T}} - \boldsymbol{\omega}_j^{\mathcal{T}})^{\top}\boldsymbol{\Gamma}^{-1}(\boldsymbol{\omega}_i^{\mathcal{T}} - \boldsymbol{\omega}_j^{\mathcal{T}})\right\}$$
$$\times \exp\left\{-\frac{1}{2}\left(\frac{\boldsymbol{\omega}_i^{\mathcal{T}} + \boldsymbol{\omega}_j^{\mathcal{T}}}{2} - \boldsymbol{\mu}\right)^{\top}(\boldsymbol{\Gamma}+\boldsymbol{\Sigma})^{-1}\left(\frac{\boldsymbol{\omega}_i^{\mathcal{T}} + \boldsymbol{\omega}_j^{\mathcal{T}}}{2} - \boldsymbol{\mu}\right)\right\}$$

with $\boldsymbol{\Gamma} = \mathrm{diag}(\frac{\gamma_1^2}{4}, \ldots, \frac{\gamma_d^2}{4})$*;*

- *the lj-th elemen of* $\mathbf{B}$*:*

$$B_{lj} = \widetilde{\psi}_{jl} \prod_{k=1}^{p} c_k(z_k, z_{jk}^{\mathcal{T}}),$$

where

$$\widetilde{\psi}_{jl} = \mathbf{e}_l[\boldsymbol{\Lambda}(\boldsymbol{\Lambda}+\boldsymbol{\Sigma})^{-1}\boldsymbol{\mu} + \boldsymbol{\Sigma}(\boldsymbol{\Lambda}+\boldsymbol{\Sigma})^{-1}\boldsymbol{\omega}_j^{\mathcal{T}}]\,\widetilde{\xi}_j.$$

Proof The proof is in Section C.4 of Appendix C. □

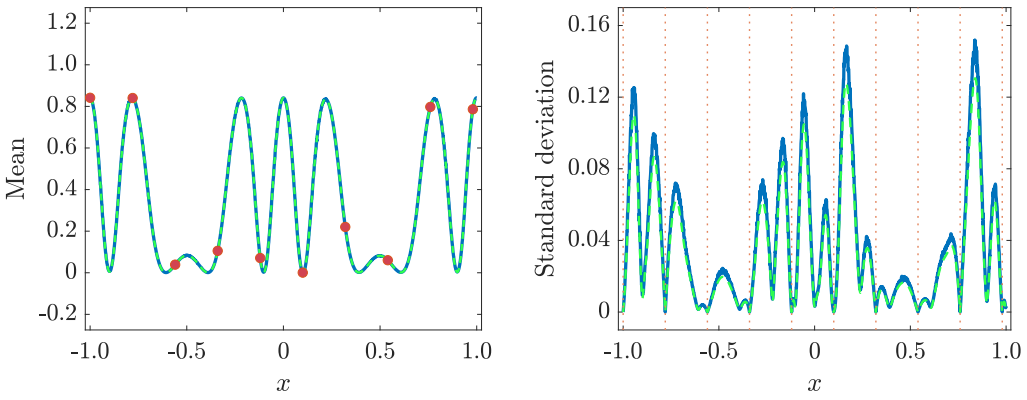
It can be seen from Proposition 3.4 that the covariance matrix $\boldsymbol{\Sigma}$ appears in the forms of inversions and determinants of $\boldsymbol{\Lambda}+\boldsymbol{\Sigma}$ and $\boldsymbol{\Gamma}+\boldsymbol{\Sigma}$ in most cases and appears only in these two forms (i.e., inversion and determinant) when the trend function is set to a constant (i.e., $\mathbf{B}$ has no effects on the mean and variance of integrated emulators). Thus, how much influence the correlations (i.e., the off-diagonal elements of $\boldsymbol{\Sigma}$) between input variables $\mathbf{W}$ have on integrated emulators depends on the magnitudes of $\gamma_1^2, \ldots, \gamma_d^2$. When the magnitudes of $\gamma_1^2, \ldots, \gamma_d^2$ are sufficiently large such that $\boldsymbol{\Lambda}+\boldsymbol{\Sigma}$ and $\boldsymbol{\Gamma}+\boldsymbol{\Sigma}$ become diagonally dominant, the inversions and determinants of $\boldsymbol{\Lambda}+\boldsymbol{\Sigma}$ and $\boldsymbol{\Gamma}+\boldsymbol{\Sigma}$ can be well approximated by those of $\boldsymbol{\Lambda}+\mathrm{diag}(\boldsymbol{\Sigma})$ and $\boldsymbol{\Gamma}+\mathrm{diag}(\boldsymbol{\Sigma})$ (Demmel, 1992; Ipsen and Lee, 2011). Thus, in practice one may first construct integrated emulators by assuming independence and check the ratio of γ_i^2 to σ_i^2 for all $i = 1, \ldots, d$ to determine whether the dependence between input variables are non-negligible.

If the dependence is determined to be significant to the integrated emula-

tion, one shall consider to build multivariate Gaussian process emulators to incorporate the correlation structures between emulator outputs, which consequently serve as the inputs to the linked Gaussian process emulators in the context of integrated emulation. However, all these existing literature, such as Rougier et al. (2009); Fricker et al. (2013); Zhang et al. (2015), only consider dependence between outputs from a single emulator, while in the framework of integrated emulation a Gaussian process emulator can have its inputs fed by the outputs from different linked emulators. As a result, further studies need to done to explore how dependence between outputs from different Gaussian process emulators can be addressed during the integration of Gaussian process emulators.

In Section 3.2.2, we have discussed some intuitions on why the Gaussian assumption is important to have analytical expressions for integrated emulators and justified the Gaussian approximation in terms of the KL divergence. It is noted that one can use other distributions other than the Gaussian distribution to approximate the true predictive distribution (named as exact integrated emulator hereinafter) by minimising the KL divergence. However, whether these distributions can be determined analytically and allow closed form expressions for integrated emulators are not clear. In particular, the accuracy of Gaussian approximation is not essential because the full probabilistic description is considered as non-critical in the integrated emulation of deterministic computer models. Instead, mean predictions and associated variances are treated as the primitive quantities, which carries the information of individual computer models and their structural relations by combining the mean predictions and variances of individual GP emulators. For example, the smart designs and some calibration methods such as the history matching (Vernon et al., 2014) only require information of mean and variance of an emulator. Since the mean and variance are key quantities for integrated emulation, it could be useful in practice to conduct diagnosis on how well the analytical mean and variance of the integrated emulator using the Gaussian approximation represent those

of the exact integrated emulator. Figure 3.17 compares means and standard deviations of the integrated emulator (for experiment 1 described in Section 3.3) using the Gaussian approximation and those of the exact integrated emulator. It can be observed that the mean and standard deviation of the exact integrated emulator is well approximated by those of the integrated emulator using the Gaussian approximation, even through there are slight underestimations of large standard deviations with the approximation. For simple computer models with low input dimensions such as the one in experiment 1, the diagnosis of the mean-variance approximation can be implemented by generating samples from the exact integrated emulator at a large amount of testing input locations with a moderate computational cost. However, for computer models with a high-dimensional input space, it becomes impractical to check the goodness of the mean-variance approximation with a high number of testing input positions. Therefore, one may only use a small number of testing positions, that are space-filling conditional on the training positions, for the diagnosis.



(a) Mean **(b)** Standard Deviation

Figure 3.17: Comparison of means and standard deviations between the integrated emulator using the Gaussian approximation and the exact integrated emulator in experiment 1. The solid lines in (a) and (b) are the mean and standard deviation of the exact integrated emulator, respectively; the dashed lines in (a) and (b) are the mean and standard deviation of the integrated emulator using the Gaussian approximation, respectively; the filled circles are training points used to construct the integrated emulators; the dashed vertical lines indicate the locations of the training inputs.

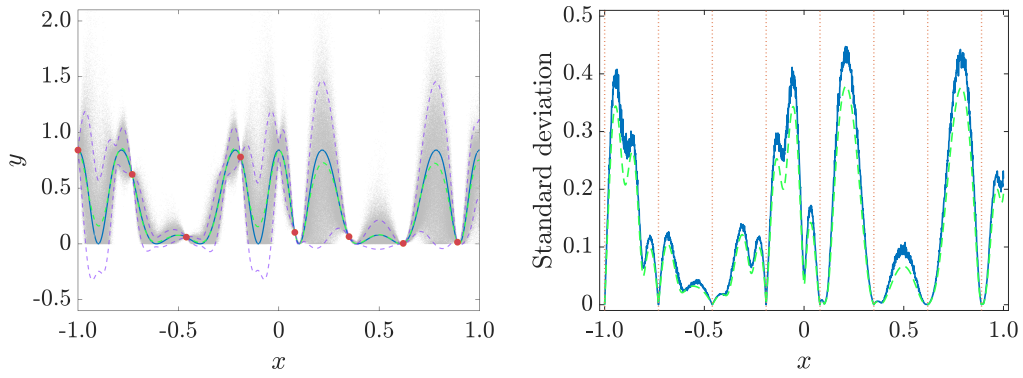
Even though we argue that it is not essential if the probability density of the exact integrated emulator is not well captured by the Gaussian distribution (as long as the mean and variance are sufficiently approximated) when the underlying computer models are deterministic, such distributional discrepancy can be problematic if the underlying computer models are stochastic and one assumes that the outputs of individual computer models are Gaussian distributed. For example, Figure 3.18(a) compares the probability densities of the exact integrated emulator and the integrated emulator using the Gaussian approximation. Both integrated emulators are constructed on the basis of eight training points, which are assumed here as random realisations of the underlying stochastic computer model given their input locations. It can be clearly observed that the exact integrated emulator is not Gaussian distributed. The exact integrated emulator has almost all of its probability density above zero. This is because the Gaussian process emulator $\widehat{f_3}$ of computer model f_3 has a very low predictive uncertainty over its input space (see top-right plot in Figure 3.16(a) for an instance where the Gaussian process emulator of f_3 is very accurate even with six training points), and thus nearly all the probability densities generated by the Gaussian process emulators of f_1 and f_2 are mapped above zero by $\widehat{f_3}$. Because of the Gaussian assumption, the integrated emulator using the Gaussian approximation still puts some probability masses below zero even though its variances capture those of the exact integrated emulator adequately (see Figure 3.18(b)). Therefore, if the probability distribution is of concern to the emulation (e.g., the computer model outputs are stochastic) the integrated emulator using the Gaussian approximation could make wrong predictions. For example, the integrated emulator using the Gaussian approximation in Figure 3.18(a) can predict negative values with a high probability in some input spaces of the underlying computer model.

Nevertheless, if the target system of computer models is deterministic, we could safely ignore the distributional inaccuracy caused by the Gaussian approximation since the mean of the integrated emulator (using the Gaussian

approximation) serves as the surrogate of the underlying deterministic computer model while the corresponding variance serves as a predictive uncertainty measurement from the emulator at unrealised input positions. If the eight training points in Figure 3.18(a) are from a deterministic system, it can be observed that the integrated emulator using the Gaussian approximation emulates the underlying model well with its mean close to the true functional form, which also fall within the prediction bounds (i.e., 5-th and 95-th percentiles). It is noted that there are some discrepancies of the prediction bounds between the approximated and exact integrated emulators. These discrepancies exist because the Gaussian approximation in the approximated integrated emulator distributes the probability masses differently from the exact integrated emulator. However, as the distributional assumption is not critical for emulating deterministic models, such discrepancies would not degrade subsequent analysis based on the integrated emulator using the Gaussian approximation.

3.7 Conclusion

In this chapter, we generalise the linked emulator to the integrated emulator for any feed-forward system of computer models. It explicitly exploits the internal system structures to produce better predictive performance than the composite emulator, which only learns the systems from the global inputs and outputs. The integrated emulator is defined by employing a Gaussian approximation with explicit mean and variance derived analytically under a variety of kernel functions, offering a flexible and computationally efficient way to emulate computer systems. The ability to use two key Matérn kernels is essential to the success of the framework. It mitigates the numerical issues while maintaining sufficient smoothness. The integrated emulation can also be applied to systems with internal loops by utilising decoupling techniques. In our experiment 1 and 2 above, significant reductions in predictive errors can be gained by the integrated emulator with moderate-size designs. Compared to the composite emulator, the integrated emulator can alternatively achieve



(a) Probability Density Comparison **(b)** Standard Deviation Comparison

Figure 3.18: *(a)*: Comparison of probability densities of the exact integrated emulator and the integrated emulator using the Gaussian approximation. The grey scattered particles are random samples generated from the exact integrated emulator; the blue solid line is the true functional form between the global input and output of the system; the dashed green line is the mean prediction of the integrated emulator using the Gaussian approximation; the dashed purple lines represents 5-th and 95-th percentiles of the integrated emulator using the Gaussian approximation; the filled circles are training points used to construct the integrated emulators. *(b)*: Comparison of standard deviations between the exact integrated emulator and the integrated emulator using the Gaussian approximation. The blue solid line is the standard deviation of the exact integrated emulator and the dashed line is the standard deviation of the integrated emulator using the Gaussian approximation; the dashed vertical lines indicate the locations of the training inputs.

similar error levels with reduced computational costs.

The integrated emulator also allows a smart adaptive designing strategy that can further reduce the predictive errors (or computational cost) remarkably by recognising the heterogeneous functional complexity of different computer models. Although the adaptive design is only illustrated via a few synthetic examples, we anticipate that the integrated emulator enhanced by this design can achieve multiple orders of magnitude reductions in predictive errors with moderate training cost in real systems, compared to the composite emulator.

Furthermore, since the integrated emulator may not show significant predictive improvement with respective to the composite emulator when a single computer

model dominates the functional complexity of the whole system, decomposition of a sophisticated system into a number of small computer models with similar functional complexity could take the advantages of the skills of the integrated emulator. This opens the door to potentially very fruitful new multi-physics approaches that split processes to facilitate surrogate modelling.

Finally, the integrated emulator offers a framework to unify and couple (simple with sophisticated, statistical-based with physics-based) models and simulators from distinct fields, creating opportunities to tackle challenges on integrating different expertise involved in cross-disciplinary studies.

Chapter 4

Non-Stationary Gaussian Processes using Deep Gaussian Hierarchy

4.1 Introduction

The Gaussian process (GP) models often assume stationarity, which in practice may not be adequate to capture non-stationary behaviours, i.e., changing smoothness over input space, of underlying function. Many work have been done on non-stationary GP models. For example, the Bayesian treed GP model (TGP) proposed by Gramacy and Lee (2008) splits the input space into several partitions and uses independent stationary GPs to each sub-region. Ba et al. (2012) apply the composition of two stationary GPs to model both global and local details of a non-stationary function. Other studies such as Montagna and Tokdar (2016) use augmented kernel function and Volodina and Williamson (2020) utilise mixtures of stationary GP processes to capture the non-stationarity. However, most of the current work may not be extended easily to more general non-stationary behaviours and require tailored kernel functions. In this chapter, we present some preliminary work on a new type of non-stationary GP model that is inspired by the deep Gaussian architectures (Duvenaud et al., 2014) and

aims to automatically learn the non-stationarity exhibited by the target dataset. Rather than adopting the conventional deep Gaussian process (DGP) models used in studies such as Damianou and Lawrence (2013); Bui et al. (2016); Salimbeni and Deisenroth (2017); Havasi et al. (2018), our model embeds the deep Gaussian architecture into the mean function and variance components of a Gaussian distribution analytically and we show that the resulting GP is able to reproduce non-stationary features, such as jump discontinuity, and heteroscedasticity, even only with a shallow hierarchy.

The remainder of the chapter is organised as follow. In Section 4.2, we describe the construction of our non-stationary GP model and visualise the functional behaviours it can produce. Then we show how to train the model via two synthetic examples in Section 4.3. Section 4.4 concludes the chapter.

4.2 Model Specification

The work of Duvenaud et al. (2014) reveals that the paths of deep Gaussian architecture exhibit non-stationary behaviours, thus our non-stationary GP model is built by integrating L Gaussian process models $\mathcal{GP}_1, \ldots, \mathcal{GP}_L$ connected in a feed-forward hierarchy, as shown in Figure 4.1.

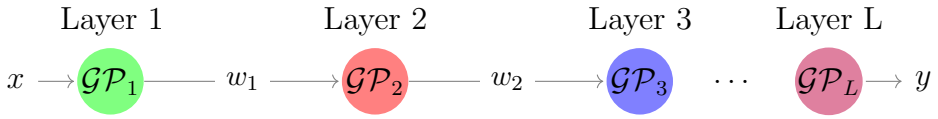


Figure 4.1: The hierarchy used to construct the non-stationary GP model.

Assume that we have N data points $\{x_i, y_i\}_{i=1,\ldots,N}$, the l-th GP model is specified as

$$\mathbf{w}_l|\mathbf{w}_{l-1} \sim \mathcal{GP}_l(\mathbf{w}_{l-1}) = \mathcal{N}\left(\boldsymbol{\mu}_l(\mathbf{w}_{l-1}),\, \sigma_l^2\left[\boldsymbol{\Sigma}_l(\mathbf{w}_{l-1}) + \eta_l\mathbf{I}\right]\right), \quad l = 1, \ldots, L$$

where $\boldsymbol{\mu}_l$ is a column vector of size N and $\boldsymbol{\Sigma}_l$ is a $N \times N$ diagonal matrix; $\mathbf{w}_{l-1} = (w_{l-1,1}, \ldots, w_{l-1,N})^\top$ with $\mathbf{w}_0 = \mathbf{x}$ and $\mathbf{w}_L = \mathbf{y}$; σ_l^2 and η_l are variance and nugget at layer l.

To introduce non-stationarity into our model, we adopt the formulation of the

sparse GP model proposed in Snelson and Ghahramani (2006) and specify the i-th element of $\boldsymbol{\mu}_l(\mathbf{w}_{l-1})$ by

$$\boldsymbol{\mu}_{l,i}(\mathbf{w}_{l-1}) = m_l(w_{l-1,i}) = \mathbf{r}_l(w_{l-1,i})^\top (\mathbf{R}_l(\mathbf{w}_{l-1}^p) + \eta_l \mathbf{I})^{-1} \mathbf{w}_l^p \tag{4.1}$$

and the i-th diagonal element of $\boldsymbol{\Sigma}_l(\mathbf{w}_{l-1})$ by

$$\boldsymbol{\Sigma}_{l,i}(\mathbf{w}_{l-1}) = v_l(w_{l-1,i}) = 1 + \mathbf{r}_l(w_{l-1,i})^\top (\mathbf{R}_l(\mathbf{w}_{l-1}^p) + \eta_l \mathbf{I})^{-1} \mathbf{r}_l(w_{l-1,i}), \tag{4.2}$$

where

$$\mathbf{w}_l^p | \mathbf{w}_{l-1}^p \sim \mathcal{N}\left(\mathbf{0}, \sigma_l^2 \left[\mathbf{R}(\mathbf{w}_{l-1}^p) + \eta_l \mathbf{I}\right]\right). \tag{4.3}$$

The above specification follows the GP predictive distribution but is conditional on a set of unknown pseudo input $\mathbf{w}_{l-1}^p = \left(w_{l-1,1}^p, \ldots, w_{l-1,M}^p\right)^\top$ and output $\mathbf{w}_l^p = \left(w_{l,1}^p, \ldots, w_{l,M}^p\right)^\top$ for each layer l. As emphasised in Snelson and Ghahramani (2006), the pseudo points induce extra flexibility to the model so that the non-stationarity (i.e., heteroscedasticity in particular) could be achieved.

In equation (4.1) and (4.2), $\mathbf{r}_l(w_{l-1,i}) \in \mathbb{R}^M$ is the kernel vector with its j-th element given by $c_l(w_{l-1,i}, w_{l-1,j}^p)$, and $\mathbf{R}_l(\mathbf{w}_{l-1}^p) \in \mathbb{R}^{M \times M}$ is the kernel matrix with its ij-th element given by $c_l(w_{l-1,i}^p, w_{l-1,j}^p)$. We employ in this study the squared exponential kernel

$$c_l(w, w') = \exp\left\{-\frac{(w - w')^2}{\gamma_l^2}\right\}$$

with γ_l being the length-scale for all layers $l = 1, \ldots, L$.

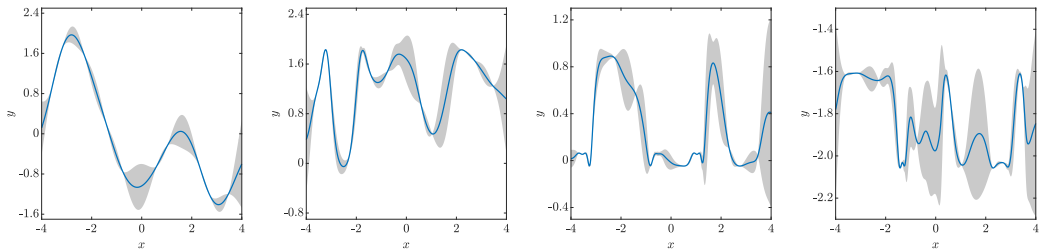
Note that in our non-stationary model the pseudo points $\mathbf{w}_{l-1}^p$ and $\mathbf{w}_l^p$ are hidden values that need to be estimated. Therefore, for each layer l (i.e., each $\mathcal{GP}_l$), we have a collection $\boldsymbol{\Phi}_l$ of unknown model parameters,

$$\boldsymbol{\Phi}_l = \left\{\mathbf{w}_{l-1}^p,\, \mathbf{w}_l^p,\, \sigma_l^2,\, \eta_l,\, \gamma_l\right\}.$$

Assume that $\boldsymbol{\Phi}_l$ are known for each layer, then the mean and variance of the Gaussian approximation $\mathcal{GP}_{1 \to L}(\mathbf{x})$ to the distribution associated with the probability density

$$p(\mathbf{y}|\mathbf{x}) = \int \prod_{l=1}^{L} p(\mathbf{w}_l|\mathbf{w}_{l-1})\, \mathrm{d}\mathbf{w}_l,$$

where $p(\cdot)$ denotes the probability density function (PDF), can be achieved using the moment matching approach presented in Chapter 3. Particularly, because $\mathbf{\Sigma}_l$ are diagonal, the Theorem 3.1 and Proposition 3.3 can be applied data point-wisely, i.e., $\mathcal{GP}_{1\to L}(\mathbf{x})$ can be obtained by constructing $\mathcal{GP}_{1\to L}(x_i)$ for each i independently. Figure 4.2 shows the mean and variance of the Gaussian approximation $\mathcal{GP}_{1\to L}(x)$ with different number of layers, each of which is generated with a realisation of a sequence of pseudo points sampled from (4.3) over $x \in (-4, 4)$. It can be seen that with a single layer we effectively have the sparse GP model introduced in Snelson and Ghahramani (2006) that is demonstrated to be able to capture heteroscedasticity. As the number of layers increases, we observe that the Gaussian approximation exhibits more non-stationary behaviours in its mean and heteroscedasticity in its variance. This indicates that with the deep Gaussian hierarchy, $\mathcal{GP}_{1\to L}(x)$, which formally defines our ultimate form of non-stationary GP model, could capture more complex behaviours of a target dataset.



(a) 1 Layer **(b)** 2 Layers **(c)** 3 Layers **(d)** 4 Layers

Figure 4.2: Non-stationary Gaussian process $\mathcal{GP}_{1\to L}(x)$ for four different number of layers (i.e., $L = 1, 2, 3, 4$). The number of pseudo points at each layer is chosen to 10. Length scale γ_l, variance σ_l^2 and nugget η_l are set to $\frac{2}{\sqrt{\pi}}$, 1 and 10^{-8} uniformly for all layers. The solid line is the mean and the shaded area represents uncertainty bound that equals to 2 standard deviations away from the mean.

Note that the flexibility of our model is induced by both the deep hierarchy and the locations of pseudo points. One may choose a large number of layers for the

model, while in our study we fix the total number of layers to be three for several reasons. Firstly, with three layers we already observe sufficient non-stationary features as implied in Figure 4.2. In addition, both Duvenaud et al. (2014) and Dunlop et al. (2018) argue that a moderate number of layers would be sufficient for inference and significant depth may even cause pathological issues. Finally, incorporating deeper Gaussian architecture will lead greater number of pseudo points and other model parameters, and thus may risk over-fitting during the inference especially when the dataset is small, e.g., in the context of surrogate modelling where only limited number of training data points are available. In Section 4.3, we show how to adjust the locations of pseudo points by means of optimisation, which provides a way to achieve automatic learning of the non-stationarity inherent in the underlying process.

Since $\mathcal{GP}_{1\to L}(\mathbf{x})$ has diagonal covariance matrix (due to the diagonal covariance matrix of $\mathcal{GP}_l(\cdot)$ for all l), after we obtain the estimates of model parameters $\mathbf{\Phi}_{i=1,\dots,L}$ the predictive distribution of our non-stationary GP model realised at a new input position x^* is given by $\mathcal{GP}_{1\to L}(x^*)$.

Although we introduce our non-stationary GP model under one-dimensional setting (i.e., $w_{l,i}$ being scalar-valued), the extension to higher dimensions is straightforward because the formulae provided by Theorem 3.1 can assemble stationary GP models with any number of dimensions in different layers.

4.3 Model Inferences and Examples

Assume that we have a set of data $\{\mathbf{y}^{\mathcal{D}}, \mathbf{x}^{\mathcal{D}}\}$ from an unknown process, then the automatic learning of the pattern of the process (i.e., the inference of model parameters) can be accomplished by the following maximum a posterior (MAP) problem:

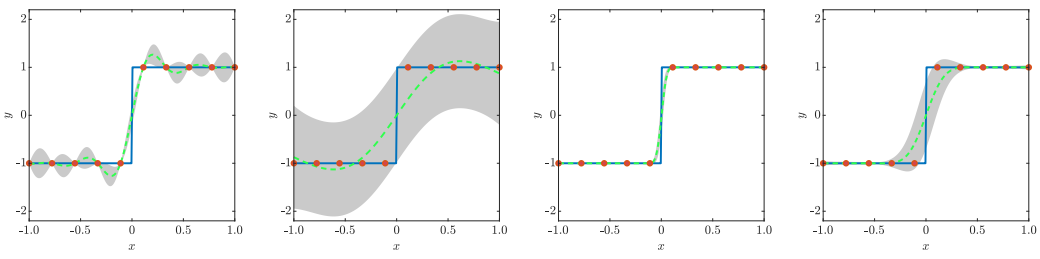
$$\widehat{\mathbf{\Phi}}_{l=1,\dots,L} = \underset{\mathbf{\Phi}_{l=1,\dots,L}}{\operatorname{argmax}} \log q_L(\mathbf{y}^{\mathcal{D}}|\mathbf{x}^{\mathcal{D}}) + \lambda \sum_{l=1}^{L} \log p(\mathbf{w}_l^p|\mathbf{w}_{l-1}^p)\,, \tag{4.4}$$

where $q_L(\mathbf{y}|\mathbf{x})$ denotes the PDF of $\mathbf{y} \sim \mathcal{GP}_{1\to L}(\mathbf{x})$; the second term serves as a regulariser on the pseudo points to guard against the over-fitting; and $\lambda \geq 0$ is the regularisation parameter that controls the weight of the regularisation.

To examine the performance of our non-stationary GP model and the inference method given by (4.4), we conduct a synthetic experiment on a step function,

$$f(x) = \begin{cases} -1, & x \leq 0 \\ 1, & x > 0 \end{cases}$$

for $x \in [-1, 1]$, which has been extensively used by many studies (Vafa, 2016; Montagna and Tokdar, 2016; Dunlop et al., 2018) to test non-stationary models. We train our non-stationary GP model with 10 equally spaced data points sampled from the step function. Figure 4.3 compares the performance of our non-stationary GP model to the stationary GP model (with squared exponential kernel) trained by the MAP method via the `R` package `RobustGaSP` (Gu et al., 2018) with and without the nugget. It can be seen that our non-stationary GP models outperform the conventional GPs both in terms of the mean prediction and the uncertainty bound. It is also demonstrated that by adjusting the value of regularisation parameter λ we are able to improve the fitting of our non-stationary GP model to the data.

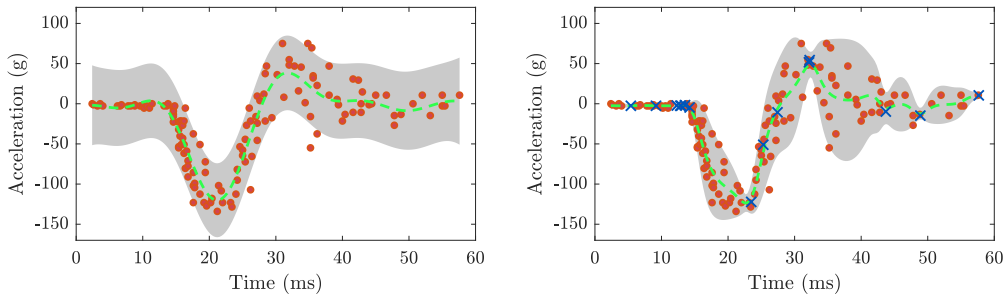


(a) GP without nugget **(b)** GP with nugget **(c)** NstGP ($\lambda = 0.3$) **(d)** NstGP ($\lambda = 1.0$)

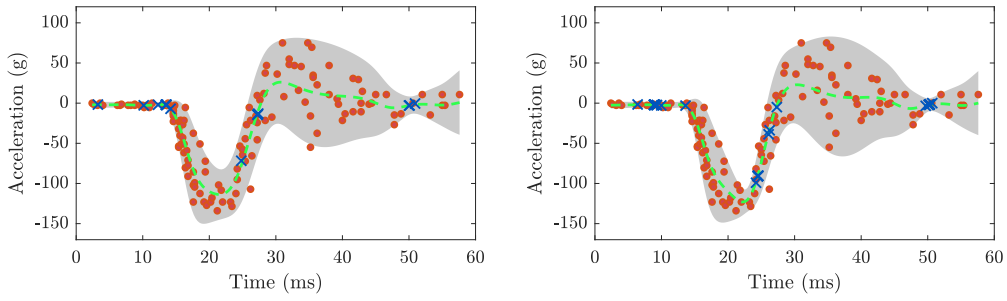
Figure 4.3: Comparison of our non-stationary (Nst) GP models (trained by optimisation problem (4.4)) to the stationary GP model with and without the nugget. For our non-stationary GP models, we choose $L = 3$ and $M = 8$. The solid line is the true data generating process; the dashed line is the mean prediction; the shaded area represents 95% prediction interval; the filled circles are sampled training points used to construct the models.

To further investigate the inference method (4.4), we train our non-stationary GP model to the motorcycle accident dataset (see the filled circles in Figure 4.4) used by Silverman (1985) and recent studies (Rasmussen and Ghahramani, 2002; Gramacy and Lee, 2008) to demonstrate the successful non-stationary modelling. The dataset gives the measurements of the motorcycle helmet acceleration over time after an accident. It can be observed that the dataset roughly has three regimes: a flat low-noise phase over $(0\,ms, 15\,ms)$, a curved noisy section between $(15\,ms, 35\,ms)$, and a smooth region after $35\,ms$ with moderate noises. From Figure 4.4(a), we see that the stationary GP is unable to capture the first linear section with low noises and the final smooth region with moderate noises. All the non-stationary GP models in Figure 4.4 learn the first region of data well, while it seems that the model in Figure 4.4(b) experiences over-fitting when $\lambda = 0.3$. The over-fitting emerges because the pseudo points are located close to the training data that are also near the predictive mean, causing the variance of the model realised at positions around the pseudo points shrinking to zero. The issue could be partially mitigated by increasing the value of λ, as shown in Figure 4.4(c) when $\lambda = 1.0$. However, we found that some pseudo points are placed on top of others. This clumping effect, as discussed by Bauer et al. (2016) in the context of sparse GP inference, can be explained by the mechanism of the inference method (4.4) that tries to learn better the heteroscedastic noise in data by avoiding to spread out the pseudo points. We argue that this is not a pleasant feature of the inference method because there is a waste of the resources (e.g., the model flexibility offered by additional pseudo points and computational efforts required to accommodate extra pseudo points) and increasing the number of pseudo points may not improve the modelling performance. The latter argument is showcased by the scenario in Figure 4.4(d), where the added pseudo points are clustered and do not help the model better represent the up-and-down behaviours of acceleration over $(30\,ms, 35\,ms)$.

To address the issue encountered to the inference method describe above, we



(a) Stationary GP **(b)** NstGP ($\lambda = 0.3$, $M = 15$)



(c) NstGP ($\lambda = 1.0$, $M = 15$) **(d)** NstGP ($\lambda = 1.0$, $M = 30$)

Figure 4.4: Comparison of the stationary GP model to our non-stationary (Nst) GP models (trained by the optimisation problem (4.4)) using two different values of the regularisation parameter λ and number of pseudo points M. For our non-stationary GP models, we choose $L = 3$. The dashed line is the mean prediction; the shaded area represents 95% prediction interval; the filled circles are training points used to construct the models; the cross markers are pseudo points represented by $\{\mathbf{w}_0^p, \mathbf{w}_3^p\}$.

reformulate the optimisation problem (4.4) to the following:

$$\widehat{\boldsymbol{\Phi}}_{l=1,\dots,L} = \underset{\boldsymbol{\Phi}_{l=1,\dots,L}}{\operatorname{argmax}} \log g_L(\mathbf{y}^{\mathcal{D}}|\mathbf{x}^{\mathcal{D}}) - \sum_{l=1}^{L} \frac{1}{2\eta_l} \operatorname{tr}\left(\mathbb{E}_{q_{l-1}(\mathbf{w}_{l-1}|\mathbf{x}^{\mathcal{D}})}\left[\boldsymbol{\Sigma}_l(\mathbf{w}_{l-1})\right]\right) + \lambda \sum_{l=1}^{L} \log p(\mathbf{w}_l^p|\mathbf{w}_{l-1}^p)\,, \quad (4.5)$$

where $g_L(\mathbf{y}|\mathbf{x})$ denotes the PDF of $\mathbf{y} \sim \mathcal{GP}_{1\to L}(\mathbf{x})$ that approximates the PDF

$$p(\mathbf{y}|\mathbf{x}) = \int \prod_{l=1}^{L} p(\mathbf{w}_l|\mathbf{w}_{l-1})\, \mathrm{d}\mathbf{w}_l$$

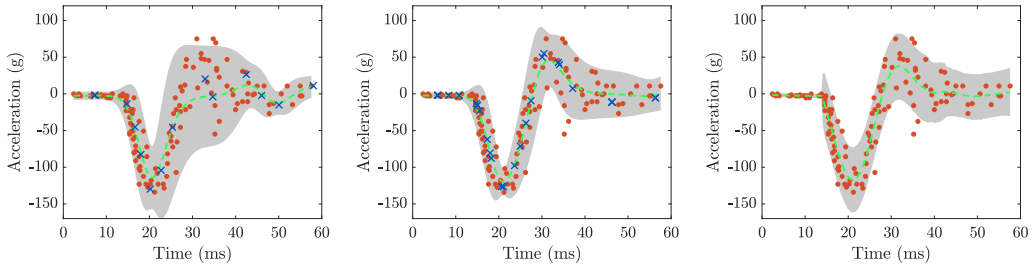
with $\mathbf{w}_l|\mathbf{w}_{l-1} \sim \mathcal{GP}_l(\mathbf{w}_{l-1}) = \mathcal{N}\left(\boldsymbol{\mu}_l(\mathbf{w}_{l-1}),\, \sigma_l^2\eta_l\mathbf{I}\right)$. The formulation of objective function in (4.5) stems from the work of Titsias (2009); Bauer et al. (2016) and the first two terms can be seen as an optimisation problem in the sense

of minimising the mean squared error between our non-stationary model and the underlying data. Since $g_L(\mathbf{y}|\mathbf{x})$ no longer propagates $\mathbf{\Sigma}_l$ over layers and the maximisation of the second term in (4.5) means minimising the conditional variances of data on the pseudo points at each layer, the inference method given by (4.5) will spread out the pseudo points for such objective.

To test the performance of the inference method (4.5), we re-train our non-stationary GP model to the motorcycle accident dataset. Figure 4.5 illustrates the trained non-stationary models with two different choices of the number of pseudo points. As it is expected, with the inference method (4.5) the pseudo points are scattered evenly across the input space of the training data. In addition, the predictive performance of the model is improved when the number of pseudo points is increased. From Figure 4.5(b), we can clearly observe from our model the three different regions indicated by the dataset. Since the treed GP (Gramacy and Lee, 2008) is one of the most popular tools to model non-stationarity, the treed GP model (see Figure 4.5(c)) is trained using the `R` package `tgp` to the same motorcycle accident dataset. From Figure 4.5, we observe that our model offers comparable modelling performance to that of the treed GP model. Besides, our model provides tighter prediction interval than the treed GP over the second (e.g., notice the spike of the predictive bound at the beginning of the second region by the treed GP model) and third regions because our model has built-in power to appreciate the heteroscedasticity. Furthermore, due to the partition used in the treed GP, the predictive bound of the treed GP model shows obvious tracks of the segmentation, while our model provides smooth transitions across different regimes.

4.3.1 Implementation notes

We build our non-stationary GP model and the inference algorithms using TensorFlow (Abadi et al., 2015) in Python and the optimisation for inference problems (4.4) and (4.5) is implemented by the gradient descent-based method Adam (Kingma and Ba, 2015). All the examples in the chapter are produced



(a) NstGP ($M = 15$) **(b)** NstGP ($M = 30$) **(c)** Treed GP

Figure 4.5: Comparison of the treed GP to our non-stationary (Nst) GP models (trained by the optimisation problem (4.5)) using two different number of pseudo points M. We choose $L = 3$ and $\lambda = 1.0$. The dashed line is the mean prediction; the shaded area represents 95% prediction interval; the filled circles are training points used to construct the models; the cross markers are pseudo points represented by $\{\mathbf{w}_0^p, \mathbf{w}_3^p\}$. Some pseudo points are out of the range of training data and thus not plotted.

after $1,000$ iterations of Adam by first (for the first 300 iterations) using learning rate of 0.01 for quicker increase of the objective function and then (for the rest 700 iterations) 0.001 for fine-tuning. We initialise the nugget η_l and variance σ_l^2 to 0.01 and 1 for all layers. As suggested in Vafa (2016), the initial length-scale γ_l for each layer is set to the mode of the pair-wise distances between the elements of $\mathbf{x}^{\mathcal{D}}$. The number of pseudo points M is often chosen much smaller than the total number of data points if the size of dataset is large, which achieves sparse approximation and thus reduction of computational expenses. Otherwise, its value can be set equal to the size of dataset. Once M is chosen, we initialise the pseudo points $\{\mathbf{w}_0^p, \mathbf{w}_3^p\}$ by selecting an evenly-spaced subset of $\{\mathbf{y}^{\mathcal{D}}, \mathbf{x}^{\mathcal{D}}\}$ and set $\mathbf{w}_2^p = \mathbf{w}_1^p = \mathbf{w}_0^p$.

4.4 Conclusion

In this chapter, we introduce a non-stationary GP model based on deep Gaussian hierarchy. Due to its hierarchical construction and introduction of pseudo points, the non-stationary GP model has the flexibility to learn the underlying data features automatically with its mean function representing non-stationarity and its variance components characterising heteroscedasticity. Since the model

has its roots to the sparse GP (Snelson and Ghahramani, 2006), it can be naturally scaled to big data size by reducing the complexity of conventional GP-based models from $\mathcal{O}(N^3)$ to $\mathcal{O}(LNM^2)$ when the number of pseudo points $M \ll N$. In addition, the independence of data points specified in the model would allow the inference to be handled by many state-of-the-art optimisers, such as stochastic gradient descent (SGD), mini-batch gradient descent and Adam, that could further reduce the complexity to $\mathcal{O}(LM^3)$. This sparse approximation incorporated in our model makes it more competitive than other non-stationary models when the data size is huge. Our non-stationary GP model is examined in a jump discontinuity function and a real dataset. The resulting good modelling performance indicates that the proposed model could be a promising candidate to model even high-dimensional datasets of large sizes.

Chapter 5

Conclusions and Future Directions

In this thesis, we explore the methodological development when Gaussian processes are used in ground-motion prediction, emulation and non-stationary modelling. In the first case, Gaussian processes are utilised to model the ground-motion intensities with spatial correlation taken into account. A statistically robust estimation algorithm, called Scoring estimation approach, is then proposed to train the model in comparison to the state-of-the-art technique that has bunch of statistical and numerical deficiencies. The approach is demonstrated to be accurate on the estimation of model parameters and able to capture the corresponding uncertainties, especially those of the spatial correlation parameters, in terms of asymptotic standard deviations. Such uncertainty measurement is not available for existing methods for ground-motion modelling. Since the proposed Scoring estimation approach is consistent and statistically efficient when model is well-specified, it allows us to examine the impacts of the ignorance of spatial correlation on model parameter estimates and the resulting ground-motion predictions.

The investigation in Chapter 2 provides two important implications for seismic risk assessment. Firstly, as any estimation technique, the Scoring estimation approach is only as good as the proposed ground-motion model. Therefore, a

rigorous assessment of spatial correlation in the ground-motion data should be addressed during the GMPE construction such that the resulting ground-motion model is a good representation of the underlying data. In return, the Scoring estimation approach can serve as a competitive method for accurate ground-motion model estimation and shaking intensity map generation. Secondly, we show that ignoring spatial correlation in ground-motion models can result in overestimation of the inter-event variance and underestimation of the intra-event variance, and such biases increase when the spatial correlation implied by the underlying data becomes stronger and smoother. These results generalise the findings of Jayaram and Baker (2010) and further emphasise the importance to accurately estimate the inter-event and intra-event variances as their changes "have implications for risk assessments of spatially-distributed systems" (Jayaram and Baker, 2010).

There are several aspects of the work in Chapter 2 that can be further addressed in future work. Firstly, we could consider non-stationary and anisotropic kernels in modelling spatial structure as the correlation of the intensity measures between two sites are also related to their soil conditions and distances to the epicentre. This is particularly sensible because the ground-motion intensities of two sites, even though they are far away, may strongly correlated if their distances to the fault are similar. Another aspect that is worth exploring is the assumption of independence between earthquakes. This assumption is crucial to the validation of consistency and statistical efficiency offered by the Scoring estimation approach. However, if the earthquake events in the dataset are correlated, e.g., there are main-shocks and after-shocks, one needs to shown theoretically under what extra conditions such statistical properties would still hold. Finally, since only mean predictions are considered in the work, the predictive uncertainties should be incorporated in the future as they can enhance the predictive capacity of shaking intensity maps, which in turn provide additional information for decision-makers.

In Chapter 3, we generalise the linked emulator to the integrated emulator

for any feed-forward systems of computer models. This integrated emulator combines individual Gaussian process emulators analytically, providing a flexible and fast way for its construction. Furthermore, the integrated emulation and its bespoke designing strategies could reduce orders of magnitude computational costs and predictive errors. This methodological development enables a more cost-efficient construction of surrogate models for systems of computer models, such as climate models, in many research areas, and opens several potential future research directions.

The first research direction is how to incorporate dimension reduction into the framework. Many computer systems have high-dimensional data (e.g., maps or time series) transferred from one computer model to another. For example, in the earthquake-induced tsunami model the outputs from the earthquake simulator, e.g., SPECFEM3D (Komatitsch and Tromp, 2002*a*,*b*) is a time series of high-dimensional displacement fields, which consequently serve as the input to the tsunami simulator, e.g., VOLNA (Reguly et al., 2018). This high-dimensional output from the earthquake simulator poses a challenging issue for training the integrated emulator. One option to resolve this issue could be reducing the dimensions of the output from the earthquake model by dimension reduction techniques and then building the integrated emulator on GP emulators of earthquake and tsunami simulators with the low-dimensional manifold. However, this solution treats dimension reduction as a pre-step and how such procedure would affect the accuracy of the final emulation is worth a further investigation. In addition, one may try to come up with a method to integrate the dimension reduction naturally into the integrated emulation.

Another research direction is how to apply the integrated emulation to systems with feed-back coupled computer models in a natural way. In Chapter 3 we address the issue by fixing values of weakly coupled variables in a feed-back coupled system so that it is transformed into a feed-forward system. However, the detection of weakly coupled variables can be computationally expensive and in some complex systems such procedure may soon become infeasible.

Therefore, complex feed-back coupled systems should be investigated more thoroughly. One possible solution would be using GP emulators to represent the functional relations of computer model inputs and the stabilised values of feed-back coupled variables. In this way, any feed-back coupled system can be emulated naturally by our framework.

Note that Algorithm 6 and 7 may not be the only applicable adaptive design strategies for the integrated emulation of the two synthetic systems in Chapter 3. Thus, one may investigate further on other variants of the proposed adaptive design depending on the system structures at hand by using the key results from Theorem 3.1 and Proposition 3.2. For example, in a computer system with large number of layers one may apply adaptive designs at each iteration of the iterative procedure for the integrated emulation. It is also worth investigating further on how to efficiently and effectively conduct the optimisation of the maximisation problems involved in these algorithms when the global input space is high dimensional.

It would be also interesting to extend the integrated emulator to accommodate sparse Gaussian processes (Quiñonero-Candela and Rasmussen, 2005; Liu et al., 2020). It may seem that it is not necessary to utilise sparse GP in the context of emulation as we often have small number of computer runs (i.e., training data points). However, many modern computer models have high-dimensional outputs (e.g., time series and maps). These high-dimensional outputs are often tackled by dimension reduction such as principle component analysis (PCA) (Salter et al., 2019) or by construction of emulators for individual output dimensions. Nevertheless, the former approach may not always work if the output is highly nonlinear or the lower-dimensional manifold does not exist. The latter approach may soon become computationally infeasible if the output dimension is too high and cannot provide emulators at output locations that are not generated by the computer model. Alternatively, the sparse Gaussian processes could be used to address the issues by including the coordinates (for maps) or time stamps (for times series) into the model inputs, and the resulting

GP model could naturally give uncertainty predictions at positions that are not produced by the simulator under different model settings.

The non-stationary GP model proposed in Chapter 4 has many potentials to be explored in the future. For example, its performance needs to be investigated in datasets with more than one dimensional inputs. One possible candidate would be the synthetic earthquake surface deformation field (Okada, 1985), which has discontinuity along the fault trace (i.e., the intersection of the fault plane with the ground surface). Besides, since the non-stationary GP model is built from the moment matching technique, it is straightforward to incorporate the model into the framework of integrated emulation. Therefore, how the predictive performance of integrated emulator improves with the incorporation of non-stationary GP model would be an interesting future research direction. The nature of the non-stationary GP model indicates that it may not interpolate the training data points, and thus challenges, such as how to implement efficient sequential design strategies when it is used for the emulation of deterministic functions, need to be tackled. The current inference method for the model is based on the MAP, while it would be worth exploring fully Bayesian approach so that the over-fitting issue could be further addressed and the uncertainties of pseudo points are taken into account.

Appendix A

Proofs in Chapter 2

A.1 Proof of Equation (2.6)

The semivariogram of $\widetilde{\boldsymbol{\varepsilon}}$ is defined by

$$\gamma(\widetilde{\varepsilon}_{ij},\, \widetilde{\varepsilon}_{ij'}) = \frac{1}{2}\mathrm{var}(\widetilde{\varepsilon}_{ij} - \widetilde{\varepsilon}_{ij'}).$$

Then, we have

$$\begin{aligned}
\gamma(\widetilde{\varepsilon}_{ij},\, \widetilde{\varepsilon}_{ij'}) =& \frac{1}{2}\mathbb{E}\left[\left(\frac{\varepsilon_{ij}}{\sigma} - \frac{\varepsilon_{ij'}}{\sigma}\right)^2\right] \\
=& \frac{1}{2\sigma^2}\mathbb{E}\left[(\varepsilon_{ij} - \varepsilon_{ij'})^2\right] \\
=& \frac{1}{2\sigma^2}\left(\mathbb{E}[\varepsilon_{ij}^2] + \mathbb{E}[\varepsilon_{ij'}^2] - 2\mathbb{E}[\varepsilon_{ij}\varepsilon_{ij'}]\right) \\
=& \frac{1}{2\sigma^2}\mathrm{var}(\varepsilon_{ij}) + \frac{1}{2\sigma^2}\mathrm{var}(\varepsilon_{ij'}) - \frac{1}{\sigma^2}\mathrm{cov}(\varepsilon_{ij},\, \varepsilon_{ij'}) \\
=& 1 - k(\mathbf{s}_{ij},\, \mathbf{s}_{ij'}).
\end{aligned}$$

Since the kernel function is stationary and isotropic, we have

$$k(\mathbf{s}_{ij},\, \mathbf{s}_{ij'}) = k(d_{i,jj'})$$

with $d_{i,jj'} = \|\mathbf{s}_{ij} - \mathbf{s}_{ij'}\|_2$. Thus, the semivariogram of $\widetilde{\boldsymbol{\varepsilon}}$ is a function of $d_{i,jj'}$:

$$\gamma(\widetilde{\varepsilon}_{ij},\, \widetilde{\varepsilon}_{ij'}) = \gamma(d_{i,jj'}) = 1 - k(d_{i,jj'}).$$

Then, for all site pairs $(j,\, j')$ such that $d_{i,jj'} = d$ we have

$$\gamma(d) = 1 - k(d).$$

A.2 Alternative Construction of the Re-Estimation Procedure

In this section, we show how to reconstruct the re-estimation procedure based on the idea of the EM algorithm.

Treating the random effects $\eta_{i=1,\ldots,N}$ as unobservable, at iteration $k+1$ we first increase $l(\widehat{\sigma^2}^{(k)}, \widehat{\tau^2}^{(k)}, \mathbf{b}|\boldsymbol{\omega}=\widehat{\boldsymbol{\omega}})$ with respect to $\mathbf{b}$ via one Expectation-Maximisation (EM) step, which consists of an E-step and a M-step:

- **E-step:** find the expected log-likelihood function

$$Q(\widehat{\sigma^2}^{(k)}, \widehat{\tau^2}^{(k)}, \mathbf{b}|\boldsymbol{\omega}=\widehat{\boldsymbol{\omega}}) = \sum_{i=1}^{N} \mathbb{E}\left[l_i^F(\widehat{\sigma^2}^{(k)}, \widehat{\tau^2}^{(k)}, \mathbf{b}|\boldsymbol{\omega}=\widehat{\boldsymbol{\omega}})\right],$$

  where the expectation is taken with respect to $\eta_{i=1,\ldots,N}$ conditional on $\mathbf{Y}_{i=1,\ldots,N}$ and estimates $\widehat{\sigma^2}^{(k)}$, $\widehat{\tau^2}^{(k)}$ and $\widehat{\mathbf{b}}^{(k)}$; and

$$\begin{aligned}
&l_i^F(\widehat{\sigma^2}^{(k)}, \widehat{\tau^2}^{(k)}, \mathbf{b}|\boldsymbol{\omega}=\widehat{\boldsymbol{\omega}}) \\
&= \ln f(\mathbf{Y}_i|\eta_i) f(\eta_i)\big|_{\sigma^2=\widehat{\sigma^2}^{(k)}, \tau^2=\widehat{\tau^2}^{(k)}, \boldsymbol{\omega}=\widehat{\boldsymbol{\omega}}} \\
&\propto -\frac{1}{2}\ln \widehat{\tau^2}^{(k)} - \frac{1}{2}\ln|\widehat{\sigma^2}^{(k)}\boldsymbol{\Omega}_i(\widehat{\boldsymbol{\omega}})| - \frac{1}{2\widehat{\tau^2}^{(k)}}\eta_i^2 \\
&\quad - \frac{1}{2\widehat{\sigma^2}^{(k)}}[\mathbf{Y}_i - \mathbf{f}(\mathbf{X}_i, \mathbf{b}) - \eta_i \mathbf{1}_{n_i}]^\top \boldsymbol{\Omega}_i^{-1}(\widehat{\boldsymbol{\omega}})[\mathbf{Y}_i - \mathbf{f}(\mathbf{X}_i, \mathbf{b}) - \eta_i \mathbf{1}_{n_i}];
\end{aligned}$$

- **M-step:** obtain the estimate $\widehat{\mathbf{b}}^{(k+1)}$ such that

$$Q(\widehat{\sigma^2}^{(k)}, \widehat{\tau^2}^{(k)}, \widehat{\mathbf{b}}^{(k+1)}|\boldsymbol{\omega}=\widehat{\boldsymbol{\omega}}) \geq Q(\widehat{\sigma^2}^{(k)}, \widehat{\tau^2}^{(k)}, \widehat{\mathbf{b}}^{(k)}|\boldsymbol{\omega}=\widehat{\boldsymbol{\omega}}).$$

Up to a constant, the expected log-likelihood function can be written as

$$\begin{aligned}
&Q(\widehat{\sigma^2}^{(k)}, \widehat{\tau^2}^{(k)}, \mathbf{b}|\boldsymbol{\omega}=\widehat{\boldsymbol{\omega}}) \\
&\propto -\frac{N}{2}\ln \widehat{\tau^2}^{(k)} - \frac{1}{2}\sum_{i=1}^{N}\ln|\widehat{\sigma^2}^{(k)}\boldsymbol{\Omega}_i(\widehat{\boldsymbol{\omega}})| \\
&\quad - \frac{1}{2\widehat{\tau^2}^{(k)}}\sum_{i=1}^{N}\widehat{\eta}_i^2 - \frac{1}{2\widehat{\sigma^2}^{(k)}}\sum_{i=1}^{N}\mathrm{tr}\left\{\boldsymbol{\Omega}_i^{-1}(\widehat{\boldsymbol{\omega}})\mathbf{V}_i\right\} \\
&\quad - \frac{1}{2\widehat{\sigma^2}^{(k)}}\sum_{i=1}^{N}[\mathbf{Y}_i - \mathbf{f}(\mathbf{X}_i, \mathbf{b}) - \widehat{\eta}_i \mathbf{1}_{n_i}]^\top \boldsymbol{\Omega}_i^{-1}(\widehat{\boldsymbol{\omega}})[\mathbf{Y}_i - \mathbf{f}(\mathbf{X}_i, \mathbf{b}) - \widehat{\eta}_i \mathbf{1}_{n_i}],
\end{aligned}$$

where

$$\mathbf{V}_i = \text{var}(\eta_i|\mathbf{Y}_i, \widehat{\sigma^2}^{(k)}, \widehat{\tau^2}^{(k)}, \widehat{\mathbf{b}}^{(k)}, \boldsymbol{\omega} = \widehat{\boldsymbol{\omega}})\mathbf{1}_{n_i \times n_i},$$

$$\widehat{\eta_i^2} = \mathbb{E}[\eta_i^2|\mathbf{Y}_i, \widehat{\sigma^2}^{(k)}, \widehat{\tau^2}^{(k)}, \widehat{\mathbf{b}}^{(k)}, \boldsymbol{\omega} = \widehat{\boldsymbol{\omega}}]$$

and

$$\widehat{\eta_i} = \mathbb{E}[\eta_i|\mathbf{Y}_i, \widehat{\sigma^2}^{(k)}, \widehat{\tau^2}^{(k)}, \widehat{\mathbf{b}}^{(k)}, \boldsymbol{\omega} = \widehat{\boldsymbol{\omega}}].$$

Note that

$$\begin{aligned}\widehat{\eta_i} =& \mathbb{E}[\eta_i|\mathbf{Y}_i, \widehat{\sigma^2}^{(k)}, \widehat{\tau^2}^{(k)}, \widehat{\mathbf{b}}^{(k)}, \boldsymbol{\omega} = \widehat{\boldsymbol{\omega}}] \\ =& \widehat{\tau^2}^{(k)}\mathbf{1}_{n_i}^\top \left(\widehat{\tau^2}^{(k)}\mathbf{1}_{n_i \times n_i} + \widehat{\sigma^2}^{(k)}\boldsymbol{\Omega}_i(\widehat{\boldsymbol{\omega}})\right)^{-1} [\mathbf{Y}_i - \mathbf{f}(\mathbf{X}_i, \widehat{\mathbf{b}}^{(k)})], \end{aligned} \tag{A.1}$$

where the second equality is given by the formula for the expectation of the conditional multivariate normal distribution (Flury, 2013). Also note that

$$\begin{aligned}&\left(\widehat{\tau^2}^{(k)}\mathbf{1}_{n_i \times n_i} + \widehat{\sigma^2}^{(k)}\boldsymbol{\Omega}_i(\widehat{\boldsymbol{\omega}})\right)^{-1} \\ =& \left(\widehat{\sigma^2}^{(k)}\boldsymbol{\Omega}_i(\widehat{\boldsymbol{\omega}})\right)^{-1} \\ &- \left(\widehat{\sigma^2}^{(k)}\boldsymbol{\Omega}_i(\widehat{\boldsymbol{\omega}})\right)^{-1}\mathbf{1}_{n_i}\left(\frac{1}{\widehat{\tau^2}^{(k)}} + \mathbf{1}_{n_i}^\top\left(\widehat{\sigma^2}^{(k)}\boldsymbol{\Omega}_i(\widehat{\boldsymbol{\omega}})\right)^{-1}\mathbf{1}_{n_i}\right)^{-1}\mathbf{1}_{n_i}^\top\left(\widehat{\sigma^2}^{(k)}\boldsymbol{\Omega}_i(\widehat{\boldsymbol{\omega}})\right)^{-1} \\ =& \left(\widehat{\sigma^2}^{(k)}\boldsymbol{\Omega}_i(\widehat{\boldsymbol{\omega}})\right)^{-1} - \frac{\left(\widehat{\sigma^2}^{(k)}\boldsymbol{\Omega}_i(\widehat{\boldsymbol{\omega}})\right)^{-1}\mathbf{1}_{n_i}\mathbf{1}_{n_i}^\top\left(\widehat{\sigma^2}^{(k)}\boldsymbol{\Omega}_i(\widehat{\boldsymbol{\omega}})\right)^{-1}}{\frac{1}{\widehat{\tau^2}^{(k)}} + \mathbf{1}_{n_i}^\top\left(\widehat{\sigma^2}^{(k)}\boldsymbol{\Omega}_i(\widehat{\boldsymbol{\omega}})\right)^{-1}\mathbf{1}_{n_i}}, \end{aligned} \tag{A.2}$$

where the first step uses the Woodbury identity (Petersen and Pedersen, 2012). Plugging equation (A.2) into (A.1), we have

$$\begin{aligned}\widehat{\eta_i} &= \frac{\mathbf{1}_{n_i}^\top\left(\widehat{\sigma^2}^{(k)}\boldsymbol{\Omega}_i(\widehat{\boldsymbol{\omega}})\right)^{-1}}{\frac{1}{\widehat{\tau^2}^{(k)}} + \mathbf{1}_{n_i}^\top\left(\widehat{\sigma^2}^{(k)}\boldsymbol{\Omega}_i(\widehat{\boldsymbol{\omega}})\right)^{-1}\mathbf{1}_{n_i}}[\mathbf{Y}_i - \mathbf{f}(\mathbf{X}_i, \widehat{\mathbf{b}}^{(k)})] \\ &= \frac{\frac{1}{\widehat{\sigma^2}^{(k)}}\mathbf{1}_{n_i}^\top\boldsymbol{\Omega}_i^{-1}(\widehat{\boldsymbol{\omega}})\,[\mathbf{Y}_i - \mathbf{f}(\mathbf{X}_i, \widehat{\mathbf{b}}^{(k)})]}{\frac{1}{\widehat{\tau^2}^{(k)}} + \frac{1}{\widehat{\sigma^2}^{(k)}}\mathbf{1}_{n_i}^\top\boldsymbol{\Omega}_i^{-1}(\widehat{\boldsymbol{\omega}})\,\mathbf{1}_{n_i}}, \end{aligned}$$

which equals to equation (2.8) in step 3 of Algorithm 1 (re-estimation procedure).

In M-step (corresponding to step 4 in the Algorithm 1) we obtain the estimate

$\mathbf{b}^{(k+1)}$ by solving the generalised least squares problem:

$$\widehat{\mathbf{b}}^{(k+1)} = \arg\min \sum_{i=1}^{N} [\mathbf{Y}_i - \mathbf{f}(\mathbf{X}_i, \mathbf{b}) - \widehat{\eta}_i \mathbf{1}_{n_i}]^\top \boldsymbol{\Omega}_i^{-1}(\widehat{\boldsymbol{\omega}})[\mathbf{Y}_i - \mathbf{f}(\mathbf{X}_i, \mathbf{b}) - \widehat{\eta}_i \mathbf{1}_{n_i}].$$

Then we have

$$Q(\widehat{\sigma^2}^{(k)}, \widehat{\tau^2}^{(k)}, \widehat{\mathbf{b}}^{(k+1)} | \boldsymbol{\omega} = \widehat{\boldsymbol{\omega}}) \geq Q(\widehat{\sigma^2}^{(k)}, \widehat{\tau^2}^{(k)}, \widehat{\mathbf{b}}^{(k)} | \boldsymbol{\omega} = \widehat{\boldsymbol{\omega}})$$

and subsequently, by monotonicity one obtains that

$$l(\widehat{\sigma^2}^{(k)}, \widehat{\tau^2}^{(k)}, \widehat{\mathbf{b}}^{(k+1)} | \boldsymbol{\omega} = \widehat{\boldsymbol{\omega}}) \geq l(\widehat{\sigma^2}^{(k)}, \widehat{\tau^2}^{(k)}, \widehat{\mathbf{b}}^{(k)} | \boldsymbol{\omega} = \widehat{\boldsymbol{\omega}}). \tag{A.3}$$

Finally, we obtain estimates $\widehat{\sigma^2}^{(k+1)}$ and $\widehat{\tau^2}^{(k+1)}$ by solving

$$(\widehat{\sigma^2}^{(k+1)}, \widehat{\tau^2}^{(k+1)}) = \arg\max l\left(\sigma^2, \tau^2 \middle| \mathbf{b} = \widehat{\mathbf{b}}^{(k+1)}, \boldsymbol{\omega} = \widehat{\boldsymbol{\omega}}\right),$$

which is the step 5 in the Algorithm 1 and implies that

$$l(\widehat{\sigma^2}^{(k+1)}, \widehat{\tau^2}^{(k+1)}, \widehat{\mathbf{b}}^{(k+1)} | \boldsymbol{\omega} = \widehat{\boldsymbol{\omega}}) \geq l(\widehat{\sigma^2}^{(k)}, \widehat{\tau^2}^{(k)}, \widehat{\mathbf{b}}^{(k+1)} | \boldsymbol{\omega} = \widehat{\boldsymbol{\omega}}),$$

and by inequality (A.3)

$$l(\widehat{\sigma^2}^{(k+1)}, \widehat{\tau^2}^{(k+1)}, \widehat{\mathbf{b}}^{(k+1)} | \boldsymbol{\omega} = \widehat{\boldsymbol{\omega}}) \geq l(\widehat{\sigma^2}^{(k)}, \widehat{\tau^2}^{(k)}, \widehat{\mathbf{b}}^{(k)} | \boldsymbol{\omega} = \widehat{\boldsymbol{\omega}}).$$

A.3 Proof of Theorem 2.1

The elements of gradient $\mathbf{S}(\boldsymbol{\alpha})$ and expected information matrix $\mathbf{I}(\boldsymbol{\alpha})$ can be calculated as follow:

- the i-th element of $\mathbf{S}_{\mathbf{b}}(\boldsymbol{\alpha})$:

$$\begin{aligned}
[\mathbf{S}_{\mathbf{b}}(\boldsymbol{\alpha})]_i =& \frac{\partial l(\boldsymbol{\alpha})}{\partial \mathbf{b}_i} \\
=& \frac{1}{2}\left\{ \left[\frac{\partial \mathbf{f}(\mathbf{X}, \mathbf{b})}{\partial \mathbf{b}_i}\right]^\top \mathbf{C}^{-1}(\boldsymbol{\theta})[\mathbf{Y} - \mathbf{f}(\mathbf{X}, \mathbf{b})] \right. \\
& \left. + [\mathbf{Y} - \mathbf{f}(\mathbf{X}, \mathbf{b})]^\top \mathbf{C}^{-1}(\boldsymbol{\theta}) \left[\frac{\partial \mathbf{f}(\mathbf{X}, \mathbf{b})}{\partial \mathbf{b}_i}\right] \right\} \\
=& \left[\frac{\partial \mathbf{f}(\mathbf{X}, \mathbf{b})}{\partial \mathbf{b}_i}\right]^\top \mathbf{C}^{-1}(\boldsymbol{\theta})[\mathbf{Y} - \mathbf{f}(\mathbf{X}, \mathbf{b})],
\end{aligned}$$

  where the last equality uses the fact that the transpose of a scalar is the

same scalar;

- the i-th element of $\mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\alpha})$:

$$\begin{aligned}
[\mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\alpha})]_i =& \frac{\partial l(\boldsymbol{\alpha})}{\partial \boldsymbol{\theta}_i} \\
=& -\frac{1}{2}\frac{1}{\left|\mathbf{C}(\boldsymbol{\theta})\right|}\frac{\partial\left|\mathbf{C}(\boldsymbol{\theta})\right|}{\partial \boldsymbol{\theta}_i} - \frac{1}{2}[\mathbf{Y}-\mathbf{f}(\mathbf{X},\,\mathbf{b})]^\top \frac{\partial \mathbf{C}^{-1}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i}[\mathbf{Y}-\mathbf{f}(\mathbf{X},\,\mathbf{b})] \\
=& -\frac{1}{2}\mathrm{tr}\left\{\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i}\right\} - \frac{1}{2}[\mathbf{Y}-\mathbf{f}(\mathbf{X},\,\mathbf{b})]^\top \frac{\partial \mathbf{C}^{-1}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i}[\mathbf{Y}-\mathbf{f}(\mathbf{X},\,\mathbf{b})] \\
=& -\frac{1}{2}\mathrm{tr}\left\{\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i}\right\} \\
&+ \frac{1}{2}[\mathbf{Y}-\mathbf{f}(\mathbf{X},\,\mathbf{b})]^\top \mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i}\mathbf{C}^{-1}(\boldsymbol{\theta})[\mathbf{Y}-\mathbf{f}(\mathbf{X},\,\mathbf{b})],
\end{aligned}$$

where the third and last steps use the following two matrix derivative identities:

$$\frac{\partial\left|\mathbf{C}(\boldsymbol{\theta})\right|}{\partial \boldsymbol{\theta}_i} = \left|\mathbf{C}(\boldsymbol{\theta})\right|\mathrm{tr}\left\{\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i}\right\} \tag{A.4}$$

and

$$\frac{\partial \mathbf{C}^{-1}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i} = -\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i}\mathbf{C}^{-1}(\boldsymbol{\theta}) \tag{A.5}$$

respectively from Petersen and Pedersen (2012);

- the ij-th element of $\mathbf{I}_{\mathbf{bb}}(\boldsymbol{\alpha})$:

$$\begin{aligned}
&[\mathbf{I}_{\mathbf{bb}}(\boldsymbol{\alpha})]_{ij} \\
=& \mathbb{E}\left[\frac{\partial l(\boldsymbol{\alpha})}{\partial \mathbf{b}_i}\frac{\partial l(\boldsymbol{\alpha})}{\partial \mathbf{b}_j}\right] \\
=& \mathbb{E}\left\{\left[\frac{\partial \mathbf{f}(\mathbf{X},\,\mathbf{b})}{\partial \mathbf{b}_i}\right]^\top \mathbf{C}^{-1}(\boldsymbol{\theta})[\mathbf{Y}-\mathbf{f}(\mathbf{X},\,\mathbf{b})]\left[\frac{\partial \mathbf{f}(\mathbf{X},\,\mathbf{b})}{\partial \mathbf{b}_j}\right]^\top \mathbf{C}^{-1}(\boldsymbol{\theta})[\mathbf{Y}-\mathbf{f}(\mathbf{X},\,\mathbf{b})]\right\} \\
=& \mathbb{E}\left\{\left[\frac{\partial \mathbf{f}(\mathbf{X},\,\mathbf{b})}{\partial \mathbf{b}_i}\right]^\top \mathbf{C}^{-1}(\boldsymbol{\theta})[\mathbf{Y}-\mathbf{f}(\mathbf{X},\,\mathbf{b})][\mathbf{Y}-\mathbf{f}(\mathbf{X},\,\mathbf{b})]^\top \mathbf{C}^{-1}(\boldsymbol{\theta})\left[\frac{\partial \mathbf{f}(\mathbf{X},\,\mathbf{b})}{\partial \mathbf{b}_j}\right]\right\} \\
=& \left[\frac{\partial \mathbf{f}(\mathbf{X},\,\mathbf{b})}{\partial \mathbf{b}_i}\right]^\top \mathbf{C}^{-1}(\boldsymbol{\theta})\,\mathbb{E}\left\{[\mathbf{Y}-\mathbf{f}(\mathbf{X},\,\mathbf{b})][\mathbf{Y}-\mathbf{f}(\mathbf{X},\,\mathbf{b})]^\top\right\}\mathbf{C}^{-1}(\boldsymbol{\theta})\left[\frac{\partial \mathbf{f}(\mathbf{X},\,\mathbf{b})}{\partial \mathbf{b}_j}\right] \\
=& \left[\frac{\partial \mathbf{f}(\mathbf{X},\,\mathbf{b})}{\partial \mathbf{b}_i}\right]^\top \mathbf{C}^{-1}(\boldsymbol{\theta})\,\mathbf{C}(\boldsymbol{\theta})\,\mathbf{C}^{-1}(\boldsymbol{\theta})\left[\frac{\partial \mathbf{f}(\mathbf{X},\,\mathbf{b})}{\partial \mathbf{b}_j}\right] \\
=& \left[\frac{\partial \mathbf{f}(\mathbf{X},\,\mathbf{b})}{\partial \mathbf{b}_i}\right]^\top \mathbf{C}^{-1}(\boldsymbol{\theta})\,\frac{\partial \mathbf{f}(\mathbf{X},\,\mathbf{b})}{\partial \mathbf{b}_j};
\end{aligned}$$

- the ij-th element of $\mathbf{I}_{\boldsymbol{\theta\theta}}(\boldsymbol{\alpha})$:

$$
\begin{aligned}
&[\mathbf{I}_{\boldsymbol{\theta\theta}}(\boldsymbol{\alpha})]_{ij} \\
=&\mathbb{E}\left[\frac{\partial l(\boldsymbol{\alpha})}{\partial \boldsymbol{\theta}_i}\frac{\partial l(\boldsymbol{\alpha})}{\partial \boldsymbol{\theta}_j}\right] \\
=&-\mathbb{E}\left[\frac{\partial^2 l(\boldsymbol{\alpha})}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j}\right] \\
=&\frac{1}{2}\mathbb{E}\left[\mathrm{tr}\left\{\frac{\partial \mathbf{C}^{-1}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j}\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i}+\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial^2 \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j}\right\}\right] \\
&-\frac{1}{2}\mathbb{E}\left[[\mathbf{Y}-\mathbf{f}(\mathbf{X},\,\mathbf{b})]^\top \frac{\partial}{\partial \boldsymbol{\theta}_j}\left(\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i}\mathbf{C}^{-1}(\boldsymbol{\theta})\right)[\mathbf{Y}-\mathbf{f}(\mathbf{X},\,\mathbf{b})]\right] \\
=&\frac{1}{2}\mathrm{tr}\left\{-\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j}\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i}+\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial^2 \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j}\right\} \\
&-\frac{1}{2}\mathbb{E}\left[\mathrm{tr}\left\{[\mathbf{Y}-\mathbf{f}(\mathbf{X},\,\mathbf{b})]^\top \frac{\partial}{\partial \boldsymbol{\theta}_j}\left(\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i}\mathbf{C}^{-1}(\boldsymbol{\theta})\right)[\mathbf{Y}-\mathbf{f}(\mathbf{X},\,\mathbf{b})]\right\}\right] \\
=&\frac{1}{2}\mathrm{tr}\left\{-\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j}\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i}+\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial^2 \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j}\right\} \\
&-\frac{1}{2}\mathbb{E}\left[\mathrm{tr}\left\{\frac{\partial}{\partial \boldsymbol{\theta}_j}\left(\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i}\mathbf{C}^{-1}(\boldsymbol{\theta})\right)[\mathbf{Y}-\mathbf{f}(\mathbf{X},\,\mathbf{b})][\mathbf{Y}-\mathbf{f}(\mathbf{X},\,\mathbf{b})]^\top\right\}\right] \\
=&\frac{1}{2}\mathrm{tr}\left\{-\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j}\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i}+\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial^2 \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j}\right\} \\
&-\frac{1}{2}\mathrm{tr}\left\{\frac{\partial}{\partial \boldsymbol{\theta}_j}\left(\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i}\mathbf{C}^{-1}(\boldsymbol{\theta})\right)\mathbb{E}\left\{[\mathbf{Y}-\mathbf{f}(\mathbf{X},\,\mathbf{b})][\mathbf{Y}-\mathbf{f}(\mathbf{X},\,\mathbf{b})]^\top\right\}\right\} \\
=&\frac{1}{2}\mathrm{tr}\left\{-\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j}\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i}+\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial^2 \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j}\right\} \\
&-\frac{1}{2}\mathrm{tr}\left\{\frac{\partial}{\partial \boldsymbol{\theta}_j}\left(\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i}\mathbf{C}^{-1}(\boldsymbol{\theta})\right)\mathbf{C}(\boldsymbol{\theta})\right\} \\
=&\frac{1}{2}\mathrm{tr}\left\{-\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j}\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i}+\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial^2 \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j}\right\} \\
&-\frac{1}{2}\mathrm{tr}\Big\{-\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j}\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i}+\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial^2 \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j} \\
&-\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i}\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j}\Big\} \\
=&\frac{1}{2}\mathrm{tr}\left\{\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i}\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j}\right\};
\end{aligned}
$$

- the ij-th element of $\mathbf{I}_{\mathbf{b}\boldsymbol{\theta}}(\boldsymbol{\alpha})$:

$$\begin{aligned}
[\mathbf{I}_{\mathbf{b}\boldsymbol{\theta}}(\boldsymbol{\alpha})]_{ij} =& \mathbb{E}\left[\frac{\partial l(\boldsymbol{\alpha})}{\partial \mathbf{b}_i}\frac{\partial l(\boldsymbol{\alpha})}{\partial \boldsymbol{\theta}_j}\right] \\
=& -\mathbb{E}\left[\frac{\partial^2 l(\boldsymbol{\alpha})}{\partial \mathbf{b}_i \partial \boldsymbol{\theta}_j}\right] \\
=& -\mathbb{E}\left\{\left[\frac{\partial \mathbf{f}(\mathbf{X},\,\mathbf{b})}{\partial \mathbf{b}_i}\right]^\top \frac{\partial \mathbf{C}^{-1}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j}[\mathbf{Y}-\mathbf{f}(\mathbf{X},\,\mathbf{b})]\right\} \\
=& -\left[\frac{\partial \mathbf{f}(\mathbf{X},\,\mathbf{b})}{\partial \mathbf{b}_i}\right]^\top \frac{\partial \mathbf{C}^{-1}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j}\mathbb{E}\left[\mathbf{Y}-\mathbf{f}(\mathbf{X},\,\mathbf{b})\right] \\
=& 0;
\end{aligned}$$

- Since $\mathbf{I}_{\boldsymbol{\theta}\mathbf{b}}(\boldsymbol{\alpha})$ and $\mathbf{I}_{\mathbf{b}\boldsymbol{\theta}}(\boldsymbol{\alpha})$ are symmetric, we have

$$\mathbf{I}_{\boldsymbol{\theta}\mathbf{b}}(\boldsymbol{\alpha}) = \mathbf{I}_{\mathbf{b}\boldsymbol{\theta}}^\top(\boldsymbol{\alpha}) = \mathbf{0}.$$

Replacing $\mathbf{I}_{\boldsymbol{\theta}\mathbf{b}}(\widehat{\boldsymbol{\alpha}}^{(k)})$ and $\mathbf{I}_{\mathbf{b}\boldsymbol{\theta}}(\widehat{\boldsymbol{\alpha}}^{(k)})$ by $\mathbf{0}$ in equation (2.12) proves the theorem.

A.4 Proof of Theorem 2.2

Denote by the following vector the model parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$, according to

$$\mathbf{a} = \begin{bmatrix}\boldsymbol{\gamma} \\ \boldsymbol{\theta}\end{bmatrix}.$$

Then, we have $l(\boldsymbol{\alpha}) = l(\mathbf{a},\,\boldsymbol{\beta})$, and given fixed $\mathbf{a}$ the log-likelihood function $l(\mathbf{a},\,\boldsymbol{\beta})$ is maximised when

$$\boldsymbol{\beta} = \left[\mathbf{g}^\top(\mathbf{X},\,\boldsymbol{\gamma})\mathbf{C}^{-1}(\boldsymbol{\theta})\mathbf{g}(\mathbf{X},\,\boldsymbol{\gamma})\right]^{-1}\left[\mathbf{g}^\top(\mathbf{X},\,\boldsymbol{\gamma})\mathbf{C}^{-1}(\boldsymbol{\theta})\mathbf{Y}\right]. \tag{A.6}$$

This allows the profile log-likelihood function

$$M(\mathbf{a}) = l(\mathbf{a},\,h(\mathbf{a})), \tag{A.7}$$

where

$$h(\mathbf{a}) \stackrel{\text{def}}{=} \left[\mathbf{g}^\top(\mathbf{X},\,\boldsymbol{\gamma})\mathbf{C}^{-1}(\boldsymbol{\theta})\mathbf{g}(\mathbf{X},\,\boldsymbol{\gamma})\right]^{-1}\left[\mathbf{g}^\top(\mathbf{X},\,\boldsymbol{\gamma})\mathbf{C}^{-1}(\boldsymbol{\theta})\mathbf{Y}\right].$$

Taking first order derivative with respect to $\mathbf{a}$ on both sides of (A.7), we have

$$\begin{aligned}\frac{\partial M(\mathbf{a})}{\partial \mathbf{a}} =& \frac{\partial l(\mathbf{a}, \boldsymbol{\beta})}{\partial \mathbf{a}}\bigg|_{\boldsymbol{\beta}=h(\mathbf{a})} + \left(\frac{\partial h(\mathbf{a})}{\partial \mathbf{a}^\top}\right)^\top \left[\frac{\partial l(\mathbf{a}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right]_{\boldsymbol{\beta}=h(\mathbf{a})} \\ =& \frac{\partial l(\mathbf{a}, \boldsymbol{\beta})}{\partial \mathbf{a}}\bigg|_{\boldsymbol{\beta}=h(\mathbf{a})}, \end{aligned} \tag{A.8}$$

where the last equality uses the fact that $\boldsymbol{\beta} = h(\mathbf{a})$ is the solution of

$$\frac{\partial l(\mathbf{a}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{0}.$$

Now if we evaluate $\mathbf{a}$ at its estimate $\widehat{\mathbf{a}}^{(k)}$ at iteration k, we have

$$\widehat{\boldsymbol{\beta}}^{(k)} = h\left(\widehat{\mathbf{a}}^{(k)}\right).$$

Then from equality (A.8), we obtain that

$$\frac{\partial M(\mathbf{a})}{\partial \mathbf{a}}\bigg|_{\mathbf{a}=\widehat{\mathbf{a}}^{(k)}} = \frac{\partial l(\mathbf{a}, \boldsymbol{\beta})}{\partial \mathbf{a}}\bigg|_{\mathbf{a}=\widehat{\mathbf{a}}^{(k)}, \boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}^{(k)}}. \tag{A.9}$$

Denote the Score function of $M(\mathbf{a})$ by

$$\mathbf{S}_M(\mathbf{a}) = \frac{\partial M(\mathbf{a})}{\partial \mathbf{a}}.$$

Then, we have from equality (A.9) that

$$\mathbf{S}_M(\widehat{\mathbf{a}}^{(k)}) = \mathbf{S}_{\mathbf{a}}(\widehat{\boldsymbol{\alpha}}^{(k)}).$$

Since

$$\frac{\partial l(\mathbf{a}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\bigg|_{\boldsymbol{\beta}=h(\mathbf{a})} = \mathbf{0},$$

taking derivative with respect to $\mathbf{a}$ gives

$$\frac{\partial^2 l(\mathbf{a}, \boldsymbol{\beta})}{\partial \mathbf{a} \partial \boldsymbol{\beta}^\top}\bigg|_{\boldsymbol{\beta}=h(\mathbf{a})} + \left(\frac{\partial h(\mathbf{a})}{\partial \mathbf{a}^\top}\right)^\top \left[\frac{\partial^2 l(\mathbf{a}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\beta}^\top}\right]_{\boldsymbol{\beta}=h(\mathbf{a})} = \mathbf{0}. \tag{A.10}$$

Taking expectation on both sides of (A.10), we have

$$\mathbb{E}\left[\frac{\partial^2 l(\mathbf{a}, \boldsymbol{\beta})}{\partial \mathbf{a} \partial \boldsymbol{\beta}^\top}\bigg|_{\boldsymbol{\beta}=h(\mathbf{a})}\right] + \left(\frac{\partial h(\mathbf{a})}{\partial \mathbf{a}^\top}\right)^\top \mathbb{E}\left[\frac{\partial^2 l(\mathbf{a}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\beta}^\top}\bigg|_{\boldsymbol{\beta}=h(\mathbf{a})}\right] = \mathbf{0}.$$

Evaluating $\mathbf{a}$ at $\widehat{\mathbf{a}}^{(k)}$ with

$$\widehat{\boldsymbol{\beta}}^{(k)} = h\left(\widehat{\mathbf{a}}^{(k)}\right),$$

we obtain

$$\mathbb{E}\left[\frac{\partial^2 l(\mathbf{a},\,\boldsymbol{\beta})}{\partial \mathbf{a}\partial \boldsymbol{\beta}^\top}\right]_{\mathbf{a}=\widehat{\mathbf{a}}^{(k)},\,\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}^{(k)}} + \left(\frac{\partial h(\mathbf{a})}{\partial \mathbf{a}^\top}\right)^\top\Bigg|_{\mathbf{a}=\widehat{\mathbf{a}}^{(k)}} \mathbb{E}\left[\frac{\partial^2 l(\mathbf{a},\,\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\,\partial \boldsymbol{\beta}^\top}\right]_{\mathbf{a}=\widehat{\mathbf{a}}^{(k)},\,\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}^{(k)}} = \mathbf{0}.$$

Thus,

$$\mathbf{I}_{\mathbf{a}\boldsymbol{\beta}}\left(\widehat{\boldsymbol{\alpha}}^{(k)}\right) + \left(\frac{\partial h(\mathbf{a})}{\partial \mathbf{a}^\top}\right)^\top\Bigg|_{\mathbf{a}=\widehat{\mathbf{a}}^{(k)}} \mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}\left(\widehat{\boldsymbol{\alpha}}^{(k)}\right) = \mathbf{0},$$

which gives

$$\left(\frac{\partial h(\mathbf{a})}{\partial \mathbf{a}^\top}\right)^\top\Bigg|_{\mathbf{a}=\widehat{\mathbf{a}}^{(k)}} = -\mathbf{I}_{\mathbf{a}\boldsymbol{\beta}}\left(\widehat{\boldsymbol{\alpha}}^{(k)}\right)\left[\mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}\left(\widehat{\boldsymbol{\alpha}}^{(k)}\right)\right]^{-1} \tag{A.11}$$

Taking derivative with respect to $\mathbf{a}$ on both sides of (A.8), we have

$$\frac{\partial^2 M(\mathbf{a})}{\partial \mathbf{a}\partial \mathbf{a}^\top} = \frac{\partial^2 l(\mathbf{a},\,\boldsymbol{\beta})}{\partial \mathbf{a}\partial \mathbf{a}^\top}\Bigg|_{\boldsymbol{\beta}=h(\mathbf{a})} + \left(\frac{\partial h(\mathbf{a})}{\partial \mathbf{a}^\top}\right)^\top\left[\frac{\partial^2 l(\mathbf{a},\,\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\,\partial \mathbf{a}^\top}\right]_{\boldsymbol{\beta}=h(\mathbf{a})} \tag{A.12}$$

Taking expectation on both sides of (A.12) and evaluating $\mathbf{a}$ at $\widehat{\mathbf{a}}^{(k)}$ with

$$\widehat{\boldsymbol{\beta}}^{(k)} = h\left(\widehat{\mathbf{a}}^{(k)}\right),$$

we obtain

$$\begin{aligned}\mathbb{E}\left[\frac{\partial^2 M(\mathbf{a})}{\partial \mathbf{a}\partial \mathbf{a}^\top}\right]_{\mathbf{a}=\widehat{\mathbf{a}}^{(k)}} =& \mathbb{E}\left[\frac{\partial^2 l(\mathbf{a},\,\boldsymbol{\beta})}{\partial \mathbf{a}\partial \mathbf{a}^\top}\right]_{\mathbf{a}=\widehat{\mathbf{a}}^{(k)},\,\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}^{(k)}} \\ &+ \left(\frac{\partial h(\mathbf{a})}{\partial \mathbf{a}^\top}\right)^\top\Bigg|_{\mathbf{a}=\widehat{\mathbf{a}}^{(k)}} \mathbb{E}\left[\frac{\partial^2 l(\mathbf{a},\,\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\,\partial \mathbf{a}^\top}\right]_{\mathbf{a}=\widehat{\mathbf{a}}^{(k)},\,\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}^{(k)}}.\end{aligned} \tag{A.13}$$

Denote the expected information matrix of $M(\mathbf{a})$ by

$$\mathbf{I}_M(\mathbf{a}) = -\mathbb{E}\left[\frac{\partial^2 M(\mathbf{a})}{\partial \mathbf{a}\partial \mathbf{a}^\top}\right].$$

Then, equation (A.13) gives

$$\mathbf{I}_M\left(\widehat{\mathbf{a}}^{(k)}\right) = \mathbf{I}_{\mathbf{a}\mathbf{a}}\left(\widehat{\boldsymbol{\alpha}}^{(k)}\right) + \left(\frac{\partial h(\mathbf{a})}{\partial \mathbf{a}^\top}\right)^\top\Bigg|_{\mathbf{a}=\widehat{\mathbf{a}}^{(k)}} \mathbf{I}_{\boldsymbol{\beta}\mathbf{a}}\left(\widehat{\boldsymbol{\alpha}}^{(k)}\right). \tag{A.14}$$

Plugging (A.11) into (A.14), we have

$$\mathbf{I}_M\left(\widehat{\mathbf{a}}^{(k)}\right) = \mathbf{I}_{\mathbf{a}\mathbf{a}}\left(\widehat{\boldsymbol{\alpha}}^{(k)}\right) - \mathbf{I}_{\mathbf{a}\boldsymbol{\beta}}\left(\widehat{\boldsymbol{\alpha}}^{(k)}\right)\mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1}\left(\widehat{\boldsymbol{\alpha}}^{(k)}\right)\mathbf{I}_{\boldsymbol{\beta}\mathbf{a}}\left(\widehat{\boldsymbol{\alpha}}^{(k)}\right).$$

The Scoring update scheme to find the estimate of $\mathbf{a}$ that maximises $M(\mathbf{a})$ is then given by

$$\begin{aligned}\widehat{\mathbf{a}}^{(k+1)} &= \widehat{\mathbf{a}}^{(k)} + \mathbf{I}_M^{-1}\left(\widehat{\mathbf{a}}^{(k)}\right)\mathbf{S}_M\left(\widehat{\mathbf{a}}^{(k)}\right) \\ &= \widehat{\mathbf{a}}^{(k)} + \left[\mathbf{I}_{\mathbf{aa}}\left(\widehat{\boldsymbol{\alpha}}^{(k)}\right) - \mathbf{I}_{\mathbf{a}\boldsymbol{\beta}}\left(\widehat{\boldsymbol{\alpha}}^{(k)}\right)\mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1}\left(\widehat{\boldsymbol{\alpha}}^{(k)}\right)\mathbf{I}_{\boldsymbol{\beta}\mathbf{a}}\left(\widehat{\boldsymbol{\alpha}}^{(k)}\right)\right]^{-1}\mathbf{S}_{\mathbf{a}}(\widehat{\boldsymbol{\alpha}}^{(k)}),\end{aligned} \tag{A.15}$$

where

$$\mathbf{I}_{\mathbf{aa}}\left(\widehat{\boldsymbol{\alpha}}^{(k)}\right) = \begin{bmatrix}\mathbf{I}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\left(\widehat{\boldsymbol{\alpha}}^{(k)}\right) & \mathbf{I}_{\boldsymbol{\gamma}\boldsymbol{\theta}}\left(\widehat{\boldsymbol{\alpha}}^{(k)}\right) \\ \mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\gamma}}\left(\widehat{\boldsymbol{\alpha}}^{(k)}\right) & \mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}\left(\widehat{\boldsymbol{\alpha}}^{(k)}\right)\end{bmatrix},$$

$$\mathbf{I}_{\mathbf{a}\boldsymbol{\beta}}\left(\widehat{\boldsymbol{\alpha}}^{(k)}\right) = \begin{bmatrix}\mathbf{I}_{\boldsymbol{\gamma}\boldsymbol{\beta}}\left(\widehat{\boldsymbol{\alpha}}^{(k)}\right) \\ \mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\beta}}\left(\widehat{\boldsymbol{\alpha}}^{(k)}\right)\end{bmatrix},$$

$$\mathbf{I}_{\boldsymbol{\beta}\mathbf{a}}\left(\widehat{\boldsymbol{\alpha}}^{(k)}\right) = \mathbf{I}_{\mathbf{a}\boldsymbol{\beta}}^{\top}\left(\widehat{\boldsymbol{\alpha}}^{(k)}\right)$$

and

$$\mathbf{S}_{\mathbf{a}}(\widehat{\boldsymbol{\alpha}}^{(k)}) = \begin{bmatrix}\mathbf{S}_{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\alpha}}^{(k)}) \\ \mathbf{S}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\alpha}}^{(k)})\end{bmatrix}.$$

By elementary calculations analogous to those used in Section A.3 of this appendix, we have that

- the i-th element of $\mathbf{S}_{\boldsymbol{\gamma}}(\boldsymbol{\alpha})$ is given by

$$[\mathbf{S}_{\boldsymbol{\gamma}}(\boldsymbol{\alpha})]_i = \left[\frac{\partial \mathbf{g}(\mathbf{X},\,\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}_i}\boldsymbol{\beta}\right]^{\top}\mathbf{C}^{-1}(\boldsymbol{\theta})[\mathbf{Y} - \mathbf{g}(\mathbf{X},\,\boldsymbol{\gamma})\boldsymbol{\beta}]\,;$$

- The i-th element of $\mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\alpha})$ is given by

$$\begin{aligned}[\mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\alpha})]_i = &-\frac{1}{2}\mathrm{tr}\left\{\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i}\right\} \\ &+ \frac{1}{2}[\mathbf{Y} - \mathbf{g}(\mathbf{X},\,\boldsymbol{\gamma})\boldsymbol{\beta}]^{\top}\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i}\mathbf{C}^{-1}(\boldsymbol{\theta})[\mathbf{Y} - \mathbf{g}(\mathbf{X},\,\boldsymbol{\gamma})\boldsymbol{\beta}];\end{aligned}$$

- $\mathbf{S}_{\boldsymbol{\beta}}(\boldsymbol{\alpha})$ is given by

$$\mathbf{S}_{\boldsymbol{\beta}}(\boldsymbol{\alpha}) = \mathbf{g}(\mathbf{X},\,\boldsymbol{\gamma})^{\top}\mathbf{C}^{-1}(\boldsymbol{\theta})[\mathbf{Y} - \mathbf{g}(\mathbf{X},\,\boldsymbol{\gamma})\boldsymbol{\beta}];$$

- the ij-th element of $\mathbf{I}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}(\boldsymbol{\alpha})$ is given by

$$[\mathbf{I}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}(\boldsymbol{\alpha})]_{ij} = \left[\frac{\partial \mathbf{g}(\mathbf{X}, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}_i}\boldsymbol{\beta}\right]^\top \mathbf{C}^{-1}(\boldsymbol{\theta})\, \frac{\partial \mathbf{g}(\mathbf{X}, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}_j}\boldsymbol{\beta};$$

- the ij-th element of $\mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\alpha})$ is given by

$$[\mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\alpha})]_{ij} = \frac{1}{2}\mathrm{tr}\left\{\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i}\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j}\right\};$$

- $\mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\alpha})$ is given by

$$\mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\alpha}) = \mathbf{g}(\mathbf{X}, \boldsymbol{\gamma})^\top \mathbf{C}^{-1}(\boldsymbol{\theta})\mathbf{g}(\mathbf{X}, \boldsymbol{\gamma});$$

- the i-th row of $\mathbf{I}_{\boldsymbol{\gamma}\boldsymbol{\beta}}(\boldsymbol{\alpha})$ $\big($or the i-th column of $\mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\gamma}}(\boldsymbol{\alpha})\big)$ is given by

$$[\mathbf{I}_{\boldsymbol{\gamma}\boldsymbol{\beta}}(\boldsymbol{\alpha})]_{i*} = [\mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\gamma}}(\boldsymbol{\alpha})]_{*i}^\top = \left[\frac{\partial \mathbf{g}(\mathbf{X}, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}_i}\boldsymbol{\beta}\right]^\top \mathbf{C}^{-1}(\boldsymbol{\theta})\mathbf{g}(\mathbf{X}, \boldsymbol{\gamma})\,;$$

- $\mathbf{I}_{\boldsymbol{\gamma}\boldsymbol{\theta}}(\boldsymbol{\alpha}) = \mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\gamma}}^\top(\boldsymbol{\alpha}) = \mathbf{0}$;

- $\mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\beta}}(\boldsymbol{\alpha}) = \mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\theta}}^\top(\boldsymbol{\alpha}) = \mathbf{0}$.

Replacing $\mathbf{I}_{\boldsymbol{\gamma}\boldsymbol{\theta}}(\boldsymbol{\alpha})$, $\mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\gamma}}(\boldsymbol{\alpha})$, $\mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\beta}}(\boldsymbol{\alpha})$ and $\mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\theta}}(\boldsymbol{\alpha})$ by $\mathbf{0}$ in (A.15), we obtain

$$\begin{bmatrix}\widehat{\boldsymbol{\gamma}}^{(k+1)} \\ \widehat{\boldsymbol{\theta}}^{(k+1)}\end{bmatrix} = \begin{bmatrix}\widehat{\boldsymbol{\gamma}}^{(k)} \\ \widehat{\boldsymbol{\theta}}^{(k)}\end{bmatrix} + \begin{bmatrix}\left[\mathbf{I}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\left(\widehat{\boldsymbol{\alpha}}^{(k)}\right) - \mathbf{I}_{\boldsymbol{\gamma}\boldsymbol{\beta}}\left(\widehat{\boldsymbol{\alpha}}^{(k)}\right)\mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1}\left(\widehat{\boldsymbol{\alpha}}^{(k)}\right)\mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\gamma}}\left(\widehat{\boldsymbol{\alpha}}^{(k)}\right)\right]^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}\left(\widehat{\boldsymbol{\alpha}}^{(k)}\right)\end{bmatrix}\begin{bmatrix}\mathbf{S}_{\boldsymbol{\gamma}}\left(\widehat{\boldsymbol{\alpha}}^{(k)}\right) \\ \mathbf{S}_{\boldsymbol{\theta}}\left(\widehat{\boldsymbol{\alpha}}^{(k)}\right)\end{bmatrix},$$

which yields

$$\widehat{\boldsymbol{\gamma}}^{(k+1)} = \widehat{\boldsymbol{\gamma}}^{(k)} + \left[\mathbf{I}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\left(\widehat{\boldsymbol{\alpha}}^{(k)}\right) - \mathbf{I}_{\boldsymbol{\gamma}\boldsymbol{\beta}}\left(\widehat{\boldsymbol{\alpha}}^{(k)}\right)\mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1}\left(\widehat{\boldsymbol{\alpha}}^{(k)}\right)\mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\gamma}}\left(\widehat{\boldsymbol{\alpha}}^{(k)}\right)\right]^{-1}\mathbf{S}_{\boldsymbol{\gamma}}\left(\widehat{\boldsymbol{\alpha}}^{(k)}\right) \tag{A.16}$$

$$\widehat{\boldsymbol{\theta}}^{(k+1)} = \widehat{\boldsymbol{\theta}}^{(k)} + \mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}\left(\widehat{\boldsymbol{\alpha}}^{(k)}\right)\mathbf{S}_{\boldsymbol{\theta}}\left(\widehat{\boldsymbol{\alpha}}^{(k)}\right). \tag{A.17}$$

Plugging estimates $\widehat{\boldsymbol{\gamma}}^{(k+1)}$ and $\widehat{\boldsymbol{\theta}}^{(k+1)}$ in (A.16) and (A.17) into (A.6), we obtain the updating equations for the estimate of $\boldsymbol{\beta}$, which concludes the proof.

Appendix B

Expressions for Proposition 3.3

B.1 Exponential Case

$$\begin{aligned}
\xi_{ik} =& \exp\left\{\frac{\sigma_k^2 + 2\gamma_k\left(w_{ik}^{\mathcal{T}} - \mu_k\right)}{2\gamma_k^2}\right\}\Phi\left(\frac{\mu_A - w_{ik}^{\mathcal{T}}}{\sigma_k}\right) \\
&+ \exp\left\{\frac{\sigma_k^2 - 2\gamma_k\left(w_{ik}^{\mathcal{T}} - \mu_k\right)}{2\gamma_k^2}\right\}\Phi\left(\frac{w_{ik}^{\mathcal{T}} - \mu_B}{\sigma_k}\right), \\
\zeta_{ijk} =& \begin{cases} h_\zeta\left(w_{ik}^{\mathcal{T}},\, w_{jk}^{\mathcal{T}}\right), & w_{jk}^{\mathcal{T}} \geq w_{ik}^{\mathcal{T}}, \\ h_\zeta\left(w_{jk}^{\mathcal{T}},\, w_{ik}^{\mathcal{T}}\right), & w_{jk}^{\mathcal{T}} < w_{ik}^{\mathcal{T}}, \end{cases} \\
\psi_{jk} =& \exp\left\{\frac{\sigma_k^2 + 2\gamma_k\left(w_{jk}^{\mathcal{T}} - \mu_k\right)}{2\gamma_k^2}\right\}\left[\mu_A\Phi\left(\frac{\mu_A - w_{jk}^{\mathcal{T}}}{\sigma_k}\right) + \frac{\sigma_k}{\sqrt{2\pi}}\exp\left\{-\frac{\left(w_{jk}^{\mathcal{T}} - \mu_A\right)^2}{2\sigma_k^2}\right\}\right] \\
&- \exp\left\{\frac{\sigma_k^2 - 2\gamma_k\left(w_{jk}^{\mathcal{T}} - \mu_k\right)}{2\gamma_k^2}\right\}\left[\mu_B\Phi\left(\frac{w_{jk}^{\mathcal{T}} - \mu_B}{\sigma_k}\right) - \frac{\sigma_k}{\sqrt{2\pi}}\exp\left\{-\frac{\left(w_{jk}^{\mathcal{T}} - \mu_B\right)^2}{2\sigma_k^2}\right\}\right],
\end{aligned}$$

where $\Phi(\cdot)$ denotes the cumulative density function of the standard normal;

$$\begin{aligned}
h_\zeta\left(x_1,\, x_2\right) =& \exp\left\{\frac{2\sigma_k^2 + \gamma_k\left(x_1 + x_2 - 2\mu_k\right)}{\gamma_k^2}\right\}\Phi\left(\frac{\mu_C - x_2}{\sigma_k}\right) \\
&+ \exp\left\{-\frac{x_2 - x_1}{\gamma_k}\right\}\left[\Phi\left(\frac{x_2 - \mu_k}{\sigma_k}\right) - \Phi\left(\frac{x_1 - \mu_k}{\sigma_k}\right)\right] \\
&+ \exp\left\{\frac{2\sigma_k^2 - \gamma_k\left(x_1 + x_2 - 2\mu_k\right)}{\gamma_k^2}\right\}\Phi\left(\frac{x_1 - \mu_D}{\sigma_k}\right);
\end{aligned}$$

and

$$\mu_A = \mu_k - \frac{\sigma_k^2}{\gamma_k}, \quad \mu_B = \mu_k + \frac{\sigma_k^2}{\gamma_k}, \quad \mu_C = \mu_k - \frac{2\sigma_k^2}{\gamma_k} \quad \text{and} \quad \mu_D = \mu_k + \frac{2\sigma_k^2}{\gamma_k}.$$

For notational convenience, in the above result we replace the index variable l in the subscript of ψ_{jl} by k, and $\mu_k(\mathbf{x}_k)$ and $\sigma_k(\mathbf{x}_k)$ by μ_k and σ_k. This change of notation is also applied in the remainder of the appendix.

B.2 Squared Exponential Case

$$\xi_{ik} = \frac{1}{\sqrt{1+2\sigma_k^2/\gamma_k^2}} \exp\left\{-\frac{\left(\mu_k - w_{ik}^{\mathcal{T}}\right)^2}{2\sigma_k^2+\gamma_k^2}\right\},$$

$$\zeta_{ijk} = \frac{1}{\sqrt{1+4\sigma_k^2/\gamma_k^2}} \exp\left\{-\frac{\left(\frac{w_{ik}^{\mathcal{T}}+w_{jk}^{\mathcal{T}}}{2} - \mu_k\right)^2}{\gamma_k^2/2+2\sigma_k^2} - \frac{\left(w_{ik}^{\mathcal{T}} - w_{jk}^{\mathcal{T}}\right)^2}{2\gamma_k^2}\right\},$$

$$\psi_{jk} = \frac{1}{\sqrt{1+2\sigma_k^2/\gamma_k^2}} \exp\left\{-\frac{\left(\mu_k - w_{jk}^{\mathcal{T}}\right)^2}{2\sigma_k^2+\gamma_k^2}\right\} \frac{2\sigma_k^2 w_{jk}^{\mathcal{T}} + \gamma_k^2\mu_k}{2\sigma_k^2+\gamma_k^2}.$$

B.3 Matérn-1.5 Case

$$\begin{aligned}
\xi_{ik} =& \exp\left\{\frac{3\sigma_k^2 + 2\sqrt{3}\gamma_k\left(w_{ik}^{\mathcal{T}} - \mu_k\right)}{2\gamma_k^2}\right\} \\
&\times \left[\mathbf{E}_1^\top \mathbf{\Lambda}_{11} \Phi\left(\frac{\mu_A - w_{ik}^{\mathcal{T}}}{\sigma_k}\right) + \mathbf{E}_1^\top \mathbf{\Lambda}_{12} \frac{\sigma_k}{\sqrt{2\pi}} \exp\left\{-\frac{(w_{ik}^{\mathcal{T}} - \mu_A)^2}{2\sigma_k^2}\right\}\right] \\
&+ \exp\left\{\frac{3\sigma_k^2 - 2\sqrt{3}\gamma_k\left(w_{ik}^{\mathcal{T}} - \mu_k\right)}{2\gamma_k^2}\right\} \\
&\times \left[\mathbf{E}_2^\top \mathbf{\Lambda}_{21} \Phi\left(\frac{w_{ik}^{\mathcal{T}} - \mu_B}{\sigma_k}\right) + \mathbf{E}_2^\top \mathbf{\Lambda}_{22} \frac{\sigma_k}{\sqrt{2\pi}} \exp\left\{-\frac{(w_{ik}^{\mathcal{T}} - \mu_B)^2}{2\sigma_k^2}\right\}\right], \\
\zeta_{ijk} =& \begin{cases} h_\zeta\left(w_{ik}^{\mathcal{T}}, w_{jk}^{\mathcal{T}}\right), & w_{jk}^{\mathcal{T}} \geq w_{ik}^{\mathcal{T}}, \\ h_\zeta\left(w_{jk}^{\mathcal{T}}, w_{ik}^{\mathcal{T}}\right), & w_{jk}^{\mathcal{T}} < w_{ik}^{\mathcal{T}}, \end{cases} \\
\psi_{jk} =& \exp\left\{\frac{3\sigma_k^2 + 2\sqrt{3}\gamma_k\left(w_{jk}^{\mathcal{T}} - \mu_k\right)}{2\gamma_k^2}\right\} \\
&\times \left[\mathbf{E}_1^\top \mathbf{\Lambda}_{61} \Phi\left(\frac{\mu_A - w_{jk}^{\mathcal{T}}}{\sigma_k}\right) + \mathbf{E}_1^\top \mathbf{\Lambda}_{62} \frac{\sigma_k}{\sqrt{2\pi}} \exp\left\{-\frac{(w_{jk}^{\mathcal{T}} - \mu_A)^2}{2\sigma_k^2}\right\}\right] \\
&- \exp\left\{\frac{3\sigma_k^2 - 2\sqrt{3}\gamma_k\left(w_{jk}^{\mathcal{T}} - \mu_k\right)}{2\gamma_k^2}\right\} \\
&\times \left[\mathbf{E}_2^\top \mathbf{\Lambda}_{71} \Phi\left(\frac{w_{jk}^{\mathcal{T}} - \mu_B}{\sigma_k}\right) + \mathbf{E}_2^\top \mathbf{\Lambda}_{72} \frac{\sigma_k}{\sqrt{2\pi}} \exp\left\{-\frac{(w_{jk}^{\mathcal{T}} - \mu_B)^2}{2\sigma_k^2}\right\}\right],
\end{aligned}$$

where

$$
\begin{aligned}
h_\zeta\left(x_1,\, x_2\right) = & \exp\left\{\frac{6\sigma_k^2+\sqrt{3}\gamma_k\left(x_1+x_2-2\mu_k\right)}{\gamma_k^2}\right\} \\
& \times\left[\mathbf{E}_3^\top\boldsymbol{\Lambda}_{31}\Phi\left(\frac{\mu_C-x_2}{\sigma_k}\right)+\mathbf{E}_3^\top\boldsymbol{\Lambda}_{32}\frac{\sigma_k}{\sqrt{2\pi}}\exp\left\{-\frac{(x_2-\mu_C)^2}{2\sigma_k^2}\right\}\right] \\
& +\exp\left\{-\frac{\sqrt{3}\left(x_2-x_1\right)}{\gamma_k}\right\}\left[\mathbf{E}_4^\top\boldsymbol{\Lambda}_{41}\left(\Phi\left(\frac{x_2-\mu_k}{\sigma_k}\right)-\Phi\left(\frac{x_1-\mu_k}{\sigma_k}\right)\right)\right. \\
& \left.+\mathbf{E}_4^\top\boldsymbol{\Lambda}_{42}\frac{\sigma_k}{\sqrt{2\pi}}\exp\left\{-\frac{(x_1-\mu_k)^2}{2\sigma_k^2}\right\}-\mathbf{E}_4^\top\boldsymbol{\Lambda}_{43}\frac{\sigma_k}{\sqrt{2\pi}}\exp\left\{-\frac{(x_2-\mu_k)^2}{2\sigma_k^2}\right\}\right] \\
& +\exp\left\{\frac{6\sigma_k^2-\sqrt{3}\gamma_k\left(x_1+x_2-2\mu_k\right)}{\gamma_k^2}\right\} \\
& \times\left[\mathbf{E}_5^\top\boldsymbol{\Lambda}_{51}\Phi\left(\frac{x_1-\mu_D}{\sigma_k}\right)+\mathbf{E}_5^\top\boldsymbol{\Lambda}_{52}\frac{\sigma_k}{\sqrt{2\pi}}\exp\left\{-\frac{(x_1-\mu_D)^2}{2\sigma_k^2}\right\}\right]
\end{aligned}
$$

and

- $\boldsymbol{\Lambda}_{11}=[1,\,\mu_A]^\top$, $\boldsymbol{\Lambda}_{12}=[0,\,1]^\top$, $\boldsymbol{\Lambda}_{21}=[1,\,-\mu_B]^\top$ and $\boldsymbol{\Lambda}_{22}=[0,\,1]^\top$;
- $\boldsymbol{\Lambda}_{31}=[1,\,\mu_C,\,\mu_C^2+\sigma_k^2]^\top$ and $\boldsymbol{\Lambda}_{32}=[0,\,1,\,\mu_C+x_2]^\top$;
- $\boldsymbol{\Lambda}_{41}=[1,\,\mu_k,\,\mu_k^2+\sigma_k^2]^\top$, $\boldsymbol{\Lambda}_{42}=[0,\,1,\,\mu_k+x_1]^\top$ and $\boldsymbol{\Lambda}_{43}=[0,\,1,\,\mu_k+x_2]^\top$;
- $\boldsymbol{\Lambda}_{51}=[1,\,-\mu_D,\,\mu_D^2+\sigma_k^2]^\top$ and $\boldsymbol{\Lambda}_{52}=[0,\,1,\,-\mu_D-x_1]^\top$;
- $\boldsymbol{\Lambda}_{61}=[\mu_A,\,\mu_A^2+\sigma_k^2]^\top$, $\boldsymbol{\Lambda}_{62}=[1,\,\mu_A+w_{jk}^{\mathcal{T}}]^\top$, $\boldsymbol{\Lambda}_{71}=[-\mu_B,\,\mu_B^2+\sigma_k^2]^\top$ and $\boldsymbol{\Lambda}_{72}=[1,\,-\mu_B-w_{jk}^{\mathcal{T}}]^\top$;
- $\mathbf{E}_1=\left[1-\frac{\sqrt{3}w_{ik}^{\mathcal{T}}}{\gamma_k},\,\frac{\sqrt{3}}{\gamma_k}\right]^\top$ and $\mathbf{E}_2=\left[1+\frac{\sqrt{3}w_{ik}^{\mathcal{T}}}{\gamma_k},\,\frac{\sqrt{3}}{\gamma_k}\right]^\top$;
- $\mathbf{E}_3=\left[1+\frac{3x_1x_2-\sqrt{3}\gamma_k\left(x_1+x_2\right)}{\gamma_k^2},\,\frac{2\sqrt{3}\gamma_k-3\left(x_1+x_2\right)}{\gamma_k^2},\,\frac{3}{\gamma_k^2}\right]^\top$;
- $\mathbf{E}_4=\left[1+\frac{\sqrt{3}\gamma_k\left(x_2-x_1\right)-3x_1x_2}{\gamma_k^2},\,\frac{3\left(x_1+x_2\right)}{\gamma_k^2},\,-\frac{3}{\gamma_k^2}\right]^\top$;
- $\mathbf{E}_5=\left[1+\frac{3x_1x_2+\sqrt{3}\gamma_k\left(x_1+x_2\right)}{\gamma_k^2},\,\frac{2\sqrt{3}\gamma_k+3\left(x_1+x_2\right)}{\gamma_k^2},\,\frac{3}{\gamma_k^2}\right]^\top$;
- $\mu_A=\mu_k-\frac{\sqrt{3}\sigma_k^2}{\gamma_k}$, $\mu_B=\mu_k+\frac{\sqrt{3}\sigma_k^2}{\gamma_k}$, $\mu_C=\mu_k-\frac{2\sqrt{3}\sigma_k^2}{\gamma_k}$, $\mu_D=\mu_k+\frac{2\sqrt{3}\sigma_k^2}{\gamma_k}$.

B.4 Matérn-2.5 Case

$$
\begin{aligned}
\xi_{ik} =& \exp\left\{\frac{5\sigma_k^2 + 2\sqrt{5}\gamma_k\left(w_{ik}^{\mathcal{T}} - \mu_k\right)}{2\gamma_k^2}\right\} \\
&\times \left[\mathbf{E}_1^\top \mathbf{\Lambda}_{11} \Phi\left(\frac{\mu_A - w_{ik}^{\mathcal{T}}}{\sigma_k}\right) + \mathbf{E}_1^\top \mathbf{\Lambda}_{12} \frac{\sigma_k}{\sqrt{2\pi}} \exp\left\{-\frac{(w_{ik}^{\mathcal{T}} - \mu_A)^2}{2\sigma_k^2}\right\}\right] \\
&+ \exp\left\{\frac{5\sigma_k^2 - 2\sqrt{5}\gamma_k\left(w_{ik}^{\mathcal{T}} - \mu_k\right)}{2\gamma_k^2}\right\} \\
&\times \left[\mathbf{E}_2^\top \mathbf{\Lambda}_{21} \Phi\left(\frac{w_{ik}^{\mathcal{T}} - \mu_B}{\sigma_k}\right) + \mathbf{E}_2^\top \mathbf{\Lambda}_{22} \frac{\sigma_k}{\sqrt{2\pi}} \exp\left\{-\frac{(w_{ik}^{\mathcal{T}} - \mu_B)^2}{2\sigma_k^2}\right\}\right],
\end{aligned}
$$

$$
\zeta_{ijk} = \begin{cases} h_\zeta\left(w_{ik}^{\mathcal{T}},\, w_{jk}^{\mathcal{T}}\right), & w_{jk}^{\mathcal{T}} \geq w_{ik}^{\mathcal{T}}, \\ h_\zeta\left(w_{jk}^{\mathcal{T}},\, w_{ik}^{\mathcal{T}}\right), & w_{jk}^{\mathcal{T}} < w_{ik}^{\mathcal{T}}, \end{cases}
$$

$$
\begin{aligned}
\psi_{jk} =& \exp\left\{\frac{5\sigma_k^2 + 2\sqrt{5}\gamma_k\left(w_{jk}^{\mathcal{T}} - \mu_k\right)}{2\gamma_k^2}\right\} \\
&\times \left[\mathbf{E}_1^\top \mathbf{\Lambda}_{61} \Phi\left(\frac{\mu_A - w_{jk}^{\mathcal{T}}}{\sigma_k}\right) + \mathbf{E}_1^\top \mathbf{\Lambda}_{62} \frac{\sigma_k}{\sqrt{2\pi}} \exp\left\{-\frac{(w_{jk}^{\mathcal{T}} - \mu_A)^2}{2\sigma_k^2}\right\}\right] \\
&- \exp\left\{\frac{5\sigma_k^2 - 2\sqrt{5}\gamma_k\left(w_{jk}^{\mathcal{T}} - \mu_k\right)}{2\gamma_k^2}\right\} \\
&\times \left[\mathbf{E}_2^\top \mathbf{\Lambda}_{71} \Phi\left(\frac{w_{jk}^{\mathcal{T}} - \mu_B}{\sigma_k}\right) + \mathbf{E}_2^\top \mathbf{\Lambda}_{72} \frac{\sigma_k}{\sqrt{2\pi}} \exp\left\{-\frac{(w_{jk}^{\mathcal{T}} - \mu_B)^2}{2\sigma_k^2}\right\}\right],
\end{aligned}
$$

where

$$
\begin{aligned}
h_\zeta\left(x_1,\, x_2\right) =& \exp\left\{\frac{10\sigma_k^2 + \sqrt{5}\gamma_k\left(x_1 + x_2 - 2\mu_k\right)}{\gamma_k^2}\right\} \\
&\times \left[\mathbf{E}_3^\top \mathbf{\Lambda}_{31} \Phi\left(\frac{\mu_C - x_2}{\sigma_k}\right) + \mathbf{E}_3^\top \mathbf{\Lambda}_{32} \frac{\sigma_k}{\sqrt{2\pi}} \exp\left\{-\frac{(x_2 - \mu_C)^2}{2\sigma_k^2}\right\}\right] \\
&+ \exp\left\{-\frac{\sqrt{5}\left(x_2 - x_1\right)}{\gamma_k}\right\} \left[\mathbf{E}_4^\top \mathbf{\Lambda}_{41} \left(\Phi\left(\frac{x_2 - \mu_k}{\sigma_k}\right) - \Phi\left(\frac{x_1 - \mu_k}{\sigma_k}\right)\right)\right. \\
&\left. + \mathbf{E}_4^\top \mathbf{\Lambda}_{42} \frac{\sigma_k}{\sqrt{2\pi}} \exp\left\{-\frac{(x_1 - \mu_k)^2}{2\sigma_k^2}\right\} - \mathbf{E}_4^\top \mathbf{\Lambda}_{43} \frac{\sigma_k}{\sqrt{2\pi}} \exp\left\{-\frac{(x_2 - \mu_k)^2}{2\sigma_k^2}\right\}\right] \\
&+ \exp\left\{\frac{10\sigma_k^2 - \sqrt{5}\gamma_k\left(x_1 + x_2 - 2\mu_k\right)}{\gamma_k^2}\right\} \\
&\times \left[\mathbf{E}_5^\top \mathbf{\Lambda}_{51} \Phi\left(\frac{x_1 - \mu_D}{\sigma_k}\right) + \mathbf{E}_5^\top \mathbf{\Lambda}_{52} \frac{\sigma_k}{\sqrt{2\pi}} \exp\left\{-\frac{(x_1 - \mu_D)^2}{2\sigma_k^2}\right\}\right]
\end{aligned}
$$

and

- $\mathbf{\Lambda}_{11} = [1,\, \mu_A,\, \mu_A^2 + \sigma_k^2]^\top$ and $\mathbf{\Lambda}_{12} = [0,\, 1,\, \mu_A + w_{ik}^{\mathcal{T}}]^\top$;
- $\mathbf{\Lambda}_{21} = [1,\, -\mu_B,\, \mu_B^2 + \sigma_k^2]^\top$ and $\mathbf{\Lambda}_{22} = [0,\, 1,\, -\mu_B - w_{ik}^{\mathcal{T}}]^\top$;
- $\mathbf{\Lambda}_{31} = [1,\, \mu_C,\, \mu_C^2 + \sigma_k^2,\, \mu_C^3 + 3\sigma_k^2\mu_C,\, \mu_C^4 + 6\sigma_k^2\mu_C^2 + 3\sigma_k^4]^\top$;
- $\mathbf{\Lambda}_{32} = [0,\, 1,\, \mu_C + x_2,\, \mu_C^2 + 2\sigma_k^2 + x_2^2 + \mu_C x_2,\, \mu_C^3 + x_2^3 + x_2\mu_C^2 + \mu_C x_2^2 + 3\sigma_k^2 x_2 + 5\sigma_k^2\mu_C]^\top$;
- $\mathbf{\Lambda}_{41} = [1,\, \mu_k,\, \mu_k^2 + \sigma_k^2,\, \mu_k^3 + 3\sigma_k^2\mu_k,\, \mu_k^4 + 6\sigma_k^2\mu_k^2 + 3\sigma_k^4]^\top$;
- $\mathbf{\Lambda}_{42} = [0,\, 1,\, \mu_k + x_1,\, \mu_k^2 + 2\sigma_k^2 + x_1^2 + \mu_k x_1,\, \mu_k^3 + x_1^3 + x_1\mu_k^2 + \mu_k x_1^2 + 3\sigma_k^2 x_1 + 5\sigma_k^2\mu_k]^\top$;
- $\mathbf{\Lambda}_{43} = [0,\, 1,\, \mu_k + x_2,\, \mu_k^2 + 2\sigma_k^2 + x_2^2 + \mu_k x_2,\, \mu_k^3 + x_2^3 + x_2\mu_k^2 + \mu_k x_2^2 + 3\sigma_k^2 x_2 + 5\sigma_k^2\mu_k]^\top$;
- $\mathbf{\Lambda}_{51} = [1,\, -\mu_D,\, \mu_D^2 + \sigma_k^2,\, -\mu_D^3 - 3\sigma_k^2\mu_D,\, \mu_D^4 + 6\sigma_k^2\mu_D^2 + 3\sigma_k^4]^\top$;
- $\mathbf{\Lambda}_{52} = [0,\, 1,\, -\mu_D - x_1,\, \mu_D^2 + 2\sigma_k^2 + x_1^2 + \mu_D x_1,\, -\mu_D^3 - x_1^3 - x_1\mu_D^2 - \mu_D x_1^2 - 3\sigma_k^2 x_1 - 5\sigma_k^2\mu_D]^\top$;
- $\mathbf{\Lambda}_{61} = [\mu_A,\, \mu_A^2 + \sigma_k^2,\, \mu_A^3 + 3\sigma_k^2\mu_A]^\top$;
- $\mathbf{\Lambda}_{62} = [1,\, \mu_A + w_{jk}^{\mathcal{T}},\, \mu_A^2 + 2\sigma_k^2 + \left(w_{jk}^{\mathcal{T}}\right)^2 + \mu_A w_{jk}^{\mathcal{T}}]^\top$;
- $\mathbf{\Lambda}_{71} = [-\mu_B,\, \mu_B^2 + \sigma_k^2,\, -\mu_B^3 - 3\sigma_k^2\mu_B]^\top$;
- $\mathbf{\Lambda}_{72} = [1,\, -\mu_B - w_{jk}^{\mathcal{T}},\, \mu_B^2 + 2\sigma_k^2 + \left(w_{jk}^{\mathcal{T}}\right)^2 + \mu_B w_{jk}^{\mathcal{T}}]^\top$;
- $\mathbf{E}_1 = \left[1 - \dfrac{\sqrt{5}w_{ik}^{\mathcal{T}}}{\gamma_k} + \dfrac{5\left(w_{ik}^{\mathcal{T}}\right)^2}{3\gamma_k^2},\, \dfrac{\sqrt{5}}{\gamma_k} - \dfrac{10w_{ik}^{\mathcal{T}}}{3\gamma_k^2},\, \dfrac{5}{3\gamma_k^2}\right]^\top$;
- $\mathbf{E}_2 = \left[1 + \dfrac{\sqrt{5}w_{ik}^{\mathcal{T}}}{\gamma_k} + \dfrac{5\left(w_{ik}^{\mathcal{T}}\right)^2}{3\gamma_k^2},\, \dfrac{\sqrt{5}}{\gamma_k} + \dfrac{10w_{ik}^{\mathcal{T}}}{3\gamma_k^2},\, \dfrac{5}{3\gamma_k^2}\right]^\top$;
- $\mathbf{E}_3 = [E_{30},\, E_{31},\, E_{32},\, E_{33},\, E_{34}]^\top$;
- $\mathbf{E}_4 = [E_{40},\, E_{41},\, E_{42},\, E_{43},\, E_{44}]^\top$;
- $\mathbf{E}_5 = [E_{50},\, E_{51},\, E_{52},\, E_{53},\, E_{54}]^\top$;

- $E_{30} = 1 + \dfrac{25x_1^2x_2^2 - 3\sqrt{5}\left(3\gamma_k^3 + 5\gamma_k x_1x_2\right)\left(x_1+x_2\right) + 15\gamma_k^2\left(x_1^2+x_2^2+3x_1x_2\right)}{9\gamma_k^4}$

  $E_{31} = \dfrac{18\sqrt{5}\gamma_k^3 + 15\sqrt{5}\gamma_k\left(x_1^2+x_2^2\right) - \left(75\gamma_k^2 + 50x_1x_2\right)\left(x_1+x_2\right) + 60\sqrt{5}\gamma_k x_1x_2}{9\gamma_k^4}$

  $E_{32} = \dfrac{5\left[5x_1^2 + 5x_2^2 + 15\gamma_k^2 - 9\sqrt{5}\gamma_k\left(x_1+x_2\right) + 20x_1x_2\right]}{9\gamma_k^4}$

  $E_{33} = \dfrac{10\left(3\sqrt{5}\gamma_k - 5x_1 - 5x_2\right)}{9\gamma_k^4}$ and $E_{34} = \dfrac{25}{9\gamma_k^4}$;

- $E_{40} = 1 + \dfrac{25x_1^2x_2^2 + 3\sqrt{5}\left(3\gamma_k^3 - 5\gamma_k x_1x_2\right)\left(x_2-x_1\right) + 15\gamma_k^2\left(x_1^2+x_2^2-3x_1x_2\right)}{9\gamma_k^4}$

  $E_{41} = \dfrac{5\left[3\sqrt{5}\gamma_k\left(x_2^2-x_1^2\right) + 3\gamma_k^2\left(x_1+x_2\right) - 10x_1x_2\left(x_1+x_2\right)\right]}{9\gamma_k^4}$

  $E_{42} = \dfrac{5\left[5x_1^2 + 5x_2^2 - 3\gamma_k^2 - 3\sqrt{5}\gamma_k\left(x_2-x_1\right) + 20x_1x_2\right]}{9\gamma_k^4}$

  $E_{43} = -\dfrac{50\left(x_1+x_2\right)}{9\gamma_k^4}$ and $E_{44} = \dfrac{25}{9\gamma_k^4}$;

- $E_{50} = 1 + \dfrac{25x_1^2x_2^2 + 3\sqrt{5}\left(3\gamma_k^3 + 5\gamma_k x_1x_2\right)\left(x_1+x_2\right) + 15\gamma_k^2\left(x_1^2+x_2^2+3x_1x_2\right)}{9\gamma_k^4}$

  $E_{51} = \dfrac{18\sqrt{5}\gamma_k^3 + 15\sqrt{5}\gamma_k\left(x_1^2+x_2^2\right) + \left(75\gamma_k^2 + 50x_1x_2\right)\left(x_1+x_2\right) + 60\sqrt{5}\gamma_k x_1x_2}{9\gamma_k^4}$

  $E_{52} = \dfrac{5\left[5x_1^2 + 5x_2^2 + 15\gamma_k^2 + 9\sqrt{5}\gamma_k\left(x_1+x_2\right) + 20x_1x_2\right]}{9\gamma_k^4}$

  $E_{53} = \dfrac{10\left(3\sqrt{5}\gamma_k + 5x_1 + 5x_2\right)}{9\gamma_k^4}$ and $E_{54} = \dfrac{25}{9\gamma_k^4}$;

- $\mu_A = \mu_k - \dfrac{\sqrt{5}\sigma_k^2}{\gamma_k}$, $\mu_B = \mu_k + \dfrac{\sqrt{5}\sigma_k^2}{\gamma_k}$, $\mu_C = \mu_k - \dfrac{2\sqrt{5}\sigma_k^2}{\gamma_k}$, $\mu_D = \mu_k + \dfrac{2\sqrt{5}\sigma_k^2}{\gamma_k}$.

Appendix C

Proofs in Chapter 3

C.1 Proof of Theorem 3.1

In this section, we prove Theorem 3.1 by considering not only the multiplicative form of the kernel function but also the additive form given by

$$c(\mathbf{X}_i,\, \mathbf{X}_j) = \sum_{k=1}^{p} c_k(X_{ik},\, X_{jk}).$$

C.1.1 Derivation of μ_I

We first derive the expression for μ_I. Let $\mu_g(\mathbf{W}, \mathbf{z})$ and $\sigma_g^2(\mathbf{W}, \mathbf{z})$ be the mean and variance of the GP emulator $\widehat{g}$. Then, by the tower rule, we have

$$\mu_I = \mathbb{E}[\mu_g(\mathbf{W}, \mathbf{z})],$$

where the expectation is taken respect to $\mathbf{W}$. Replace $\mu_g(\mathbf{W}, \mathbf{z})$ by equation (3.4) with Assumption 1, we have

$$\begin{aligned}
\mu_I =& \mathbb{E}\left[\mathbf{W}^\top \widehat{\boldsymbol{\theta}} + \mathbf{h}(\mathbf{z})^\top \widehat{\boldsymbol{\beta}} + \mathbf{r}^\top(\mathbf{W},\, \mathbf{z})\mathbf{R}^{-1}\left(\mathbf{y}^{\mathcal{T}} - \mathbf{w}^{\mathcal{T}}\widehat{\boldsymbol{\theta}} - \mathbf{H}(\mathbf{z}^{\mathcal{T}})\widehat{\boldsymbol{\beta}}\right)\right] \\
=& \mathbb{E}\left[\mathbf{W}^\top\right]\widehat{\boldsymbol{\theta}} + \mathbf{h}(\mathbf{z})^\top \widehat{\boldsymbol{\beta}} + \mathbb{E}\left[\mathbf{r}^\top(\mathbf{W},\, \mathbf{z})\right]\mathbf{R}^{-1}\left(\mathbf{y}^{\mathcal{T}} - \mathbf{w}^{\mathcal{T}}\widehat{\boldsymbol{\theta}} - \mathbf{H}(\mathbf{z}^{\mathcal{T}})\widehat{\boldsymbol{\beta}}\right) \\
=& \boldsymbol{\mu}^\top \widehat{\boldsymbol{\theta}} + \mathbf{h}(\mathbf{z})^\top \widehat{\boldsymbol{\beta}} + \mathbf{I}^\top \mathbf{A}, \qquad \text{(C.1)}
\end{aligned}$$

where

- $\boldsymbol{\mu} = [\mu_1(\mathbf{x}_1), \ldots, \mu_d(\mathbf{x}_d)]^\top \in \mathbb{R}^{d\times 1}$;

- $\mathbf{A} = \mathbf{R}^{-1}\left(\mathbf{y}^{\mathcal{T}} - \mathbf{w}^{\mathcal{T}}\widehat{\boldsymbol{\theta}} - \mathbf{H}(\mathbf{z}^{\mathcal{T}})\widehat{\boldsymbol{\beta}}\right) \in \mathbb{R}^{m\times 1}$;
- $\left[\widehat{\boldsymbol{\theta}}^{\top}, \widehat{\boldsymbol{\beta}}^{\top}\right]^{\top} \stackrel{\text{def}}{=\!=} \left(\widetilde{\mathbf{H}}^{\top}\mathbf{R}^{-1}\widetilde{\mathbf{H}}\right)^{-1}\widetilde{\mathbf{H}}^{\top}\mathbf{R}^{-1}\mathbf{y}^{\mathcal{T}}$ with $\widetilde{\mathbf{H}} = \left[\mathbf{w}^{\mathcal{T}}, \mathbf{H}(\mathbf{z}^{\mathcal{T}})\right] \in \mathbb{R}^{m\times(d+q)}$;
- $\mathbf{I} = \mathbb{E}\left[\mathbf{r}(\mathbf{W}, \mathbf{z})\right] \in \mathbb{R}^{m\times 1}$ with its i-th element:

$$\begin{aligned} I_i =& \mathbb{E}\left[c(\mathbf{W}, \mathbf{w}_i^{\mathcal{T}})c(\mathbf{z}, \mathbf{z}_i^{\mathcal{T}})\right] \\ =& \mathbb{E}\left[c(\mathbf{W}, \mathbf{w}_i^{\mathcal{T}})\right] c(\mathbf{z}, \mathbf{z}_i^{\mathcal{T}}) \\ =& \prod_{k=1}^{d} \mathbb{E}\left[c_k(W_k, w_{ik}^{\mathcal{T}})\right] \prod_{k=1}^{p} c_k(z_k, z_{ik}^{\mathcal{T}}) \\ =& \prod_{k=1}^{d} \xi_{ik} \prod_{k=1}^{p} c_k(z_k, z_{ik}^{\mathcal{T}}) \end{aligned}$$

in case of multiplicative form, and

$$\begin{aligned} I_i =& \mathbb{E}\left[c(\mathbf{W}, \mathbf{w}_i^{\mathcal{T}}) + c(\mathbf{z}, \mathbf{z}_i^{\mathcal{T}})\right] \\ =& \mathbb{E}\left[c(\mathbf{W}, \mathbf{w}_i^{\mathcal{T}})\right] + c(\mathbf{z}, \mathbf{z}_i^{\mathcal{T}}) \\ =& \sum_{k=1}^{d} \mathbb{E}\left[c_k(W_k, w_{ik}^{\mathcal{T}})\right] + \sum_{k=1}^{p} c_k(z_k, z_{ik}^{\mathcal{T}}) \\ =& \sum_{k=1}^{d} \xi_{ik} + \sum_{k=1}^{p} c_k(z_k, z_{ik}^{\mathcal{T}}) \end{aligned}$$

in case of additive form, where

$$\xi_{ik} \stackrel{\text{def}}{=\!=} \mathbb{E}\left[c_k(W_k, w_{ik}^{\mathcal{T}})\right]$$

and in the derivation above we use the independence of $W_{i=1,\ldots,d}$.

C.1.2 Derivation of σ_I^2

We now derive the expression for the variance σ_I^2. Using the law of total variance, we have

$$\begin{aligned} \sigma_I^2 =& \mathbb{E}\left[\sigma_g^2(\mathbf{W}, \mathbf{z})\right] + \mathrm{Var}\left(\mu_g(\mathbf{W}, \mathbf{z})\right) \\ =& \mathbb{E}\left[\sigma_g^2(\mathbf{W}, \mathbf{z})\right] + \mathbb{E}\left[\mu_g^2(\mathbf{W}, \mathbf{z})\right] - \mathbb{E}\left[\mu_g(\mathbf{W}, \mathbf{z})\right]^2 \\ =& \mathbb{E}\left[\sigma_g^2(\mathbf{W}, \mathbf{z})\right] + \mathbb{E}\left[\mu_g^2(\mathbf{W}, \mathbf{z})\right] - \mu_I^2. \end{aligned} \tag{C.2}$$

1 Derivation of $\mathbb{E}\left[\mu_g^2(\mathbf{W}, \mathbf{z})\right]$

Replace $\mu_g(\mathbf{W}, \mathbf{z})$ by equation (3.4), we have

$$\begin{aligned}\mu_g(\mathbf{W}, \mathbf{z}) = &\left[\mathbf{W}^\top \widehat{\boldsymbol{\theta}} + \mathbf{h}(\mathbf{z})^\top \widehat{\boldsymbol{\beta}} + \mathbf{r}^\top(\mathbf{W}, \mathbf{z})\mathbf{R}^{-1}\left(\mathbf{y}^{\mathcal{T}} - \mathbf{w}^{\mathcal{T}}\widehat{\boldsymbol{\theta}} - \mathbf{H}(\mathbf{z}^{\mathcal{T}})\widehat{\boldsymbol{\beta}}\right)\right]^2 \\ =&\mathbf{W}^\top \widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^\top \mathbf{W} + \left(\mathbf{h}(\mathbf{z})^\top \widehat{\boldsymbol{\beta}}\right)^2 + 2\widehat{\boldsymbol{\theta}}^\top \mathbf{W}\mathbf{h}(\mathbf{z})^\top \widehat{\boldsymbol{\beta}} \\ &+ 2\widehat{\boldsymbol{\theta}}^\top \mathbf{W}\mathbf{r}^\top(\mathbf{W}, \mathbf{z})\mathbf{A} + 2\mathbf{h}(\mathbf{z})^\top \widehat{\boldsymbol{\beta}}\mathbf{r}^\top(\mathbf{W}, \mathbf{z})\mathbf{A} \\ &+ \mathbf{r}^\top(\mathbf{W}, \mathbf{z})\mathbf{A}\mathbf{A}^\top \mathbf{r}(\mathbf{W}, \mathbf{z}).\end{aligned}$$

Then, we have

$$\begin{aligned}\mathbb{E}\left[\mu_g(\mathbf{W}, \mathbf{z})^2\right] =&\mathbb{E}\left[\mathbf{W}^\top \widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^\top \mathbf{W}\right] + \left(\mathbf{h}(\mathbf{z})^\top \widehat{\boldsymbol{\beta}}\right)^2 + 2\widehat{\boldsymbol{\theta}}^\top \mathbb{E}\left[\mathbf{W}\right]\mathbf{h}(\mathbf{z})^\top \widehat{\boldsymbol{\beta}} \\ &+ 2\widehat{\boldsymbol{\theta}}^\top \mathbb{E}\left[\mathbf{W}\mathbf{r}^\top(\mathbf{W}, \mathbf{z})\right]\mathbf{A} + 2\mathbf{h}(\mathbf{z})^\top \widehat{\boldsymbol{\beta}}\mathbb{E}\left[\mathbf{r}^\top(\mathbf{W}, \mathbf{z})\right]\mathbf{A} \\ &+ \mathbb{E}\left[\mathbf{r}^\top(\mathbf{W}, \mathbf{z})\mathbf{A}\mathbf{A}^\top \mathbf{r}(\mathbf{W}, \mathbf{z})\right] \\ =&\mathbb{E}\left[\mathbf{W}^\top \widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^\top \mathbf{W}\right] + \left(\mathbf{h}(\mathbf{z})^\top \widehat{\boldsymbol{\beta}}\right)^2 + 2\widehat{\boldsymbol{\theta}}^\top \boldsymbol{\mu}\mathbf{h}(\mathbf{z})^\top \widehat{\boldsymbol{\beta}} \\ &+ 2\widehat{\boldsymbol{\theta}}^\top \mathbf{B}\mathbf{A} + 2\mathbf{h}(\mathbf{z})^\top \widehat{\boldsymbol{\beta}}\mathbf{I}^\top \mathbf{A} + \mathbb{E}\left[\mathbf{r}^\top(\mathbf{W}, \mathbf{z})\mathbf{A}\mathbf{A}^\top \mathbf{r}(\mathbf{W}, \mathbf{z})\right]\end{aligned}$$

The first expectation in the above equation can be solved as follow:

$$\begin{aligned}&\mathbb{E}\left[\mathbf{W}^\top \widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^\top \mathbf{W}\right] \\ =&\mathrm{tr}\left\{\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^\top \mathrm{var}(\mathbf{W})\right\} + \mathbb{E}_{\mathbf{W}}\left[\mathbf{W}\right]^\top \widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^\top \mathbb{E}_{\mathbf{W}}\left[\mathbf{W}\right] \\ =&\mathrm{tr}\left\{\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^\top \boldsymbol{\Omega}\right\} + \boldsymbol{\mu}^\top \widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^\top \boldsymbol{\mu} \\ =&\mathrm{tr}\left\{\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^\top \boldsymbol{\Omega}\right\} + \mathrm{tr}\left\{\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^\top \boldsymbol{\mu}\boldsymbol{\mu}^\top\right\} \\ =&\mathrm{tr}\left\{\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^\top \left(\boldsymbol{\mu}\boldsymbol{\mu}^\top + \boldsymbol{\Omega}\right)\right\}. \qquad \text{(C.3)}\end{aligned}$$

The second expectation can be solved in a similar manner:

$$\begin{aligned}&\mathbb{E}\left[\mathbf{r}^\top(\mathbf{W}, \mathbf{z})\mathbf{A}\mathbf{A}^\top \mathbf{r}(\mathbf{W}, \mathbf{z})\right] \\ =&\mathrm{tr}\left\{\mathbb{E}\left[\mathbf{r}^\top(\mathbf{W}, \mathbf{z})\mathbf{A}\mathbf{A}^\top \mathbf{r}(\mathbf{W}, \mathbf{z})\right]\right\} \\ =&\mathbb{E}\left[\mathrm{tr}\left\{\mathbf{r}^\top(\mathbf{W}, \mathbf{z})\mathbf{A}\mathbf{A}^\top \mathbf{r}(\mathbf{W}, \mathbf{z})\right\}\right] \\ =&\mathrm{tr}\left\{\mathbf{A}\mathbf{A}^\top \mathbb{E}\left[\mathbf{r}(\mathbf{W}, \mathbf{z})\mathbf{r}^\top(\mathbf{W}, \mathbf{z})\right]\right\} \\ =&\mathrm{tr}\left\{\mathbf{A}\mathbf{A}^\top \mathbf{J}\right\}. \qquad \text{(C.4)}\end{aligned}$$

Thus, we obtain that

$$\begin{aligned}\mathbb{E}\left[\mu_g(\mathbf{W},\mathbf{z})^2\right] =& \operatorname{tr}\left\{\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^\top \operatorname{var}(\mathbf{W})\right\} + \mathbb{E}\left[\mathbf{W}\right]^\top \widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^\top \mathbb{E}\left[\mathbf{W}\right] + \left(\mathbf{h}(\mathbf{z})^\top \widehat{\boldsymbol{\beta}}\right)^2 + 2\widehat{\boldsymbol{\theta}}^\top \boldsymbol{\mu}\mathbf{h}(\mathbf{z})^\top \widehat{\boldsymbol{\beta}} \\ &+ 2\widehat{\boldsymbol{\theta}}^\top \mathbf{B}\mathbf{A} + 2\mathbf{h}(\mathbf{z})^\top \widehat{\boldsymbol{\beta}}\mathbf{I}^\top \mathbf{A} + \operatorname{tr}\left\{\mathbf{A}\mathbf{A}^\top \mathbb{E}\left[\mathbf{r}(\mathbf{W},\,\mathbf{z})\mathbf{r}^\top(\mathbf{W},\,\mathbf{z})\right]\right\} \\ =& \operatorname{tr}\left\{\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^\top \left(\boldsymbol{\mu}\boldsymbol{\mu}^\top + \boldsymbol{\Omega}\right)\right\} + \left(\mathbf{h}(\mathbf{z})^\top \widehat{\boldsymbol{\beta}}\right)^2 + 2\widehat{\boldsymbol{\theta}}^\top \boldsymbol{\mu}\mathbf{h}(\mathbf{z})^\top \widehat{\boldsymbol{\beta}} \\ &+ 2\left[\widehat{\boldsymbol{\theta}}^\top \mathbf{B} + \mathbf{h}(\mathbf{z})^\top \widehat{\boldsymbol{\beta}}\mathbf{I}^\top\right]\mathbf{A} + \operatorname{tr}\left\{\mathbf{A}\mathbf{A}^\top \mathbf{J}\right\},\end{aligned}$$

where

- $\boldsymbol{\Omega} = \operatorname{diag}(\sigma_1^2(\mathbf{x}_1), \ldots, \sigma_d^2(\mathbf{x}_d)) \in \mathbb{R}^{d\times d}$;

- $\mathbf{B} = \mathbb{E}\left[\mathbf{W}\mathbf{r}^\top(\mathbf{W},\,\mathbf{z})\right] \in \mathbb{R}^{d\times m}$ with its lj-th element:

$$\begin{aligned}B_{lj} =& \mathbb{E}\left[W_l c(\mathbf{W},\,\mathbf{w}_j^{\mathcal{T}}) c(\mathbf{z},\,\mathbf{z}_j^{\mathcal{T}})\right] \\ =& \mathbb{E}\left[W_l c(\mathbf{W},\,\mathbf{w}_j^{\mathcal{T}})\right] c(\mathbf{z},\,\mathbf{z}_j^{\mathcal{T}}) \\ =& \mathbb{E}\left[W_l \prod_{k=1}^{d} c_k(W_k,\,w_{jk}^{\mathcal{T}})\right] \prod_{k=1}^{p} c_k(z_k,\,z_{jk}^{\mathcal{T}}) \\ =& \mathbb{E}\left[W_l c_l(W_l,\,w_{jl}^{\mathcal{T}})\right] \prod_{\substack{k=1\\k\neq l}}^{d} \mathbb{E}\left[c_k(W_k,\,w_{jk}^{\mathcal{T}})\right] \prod_{k=1}^{p} c_k(z_k,\,z_{jk}^{\mathcal{T}}) \\ =& \psi_{jl} \prod_{\substack{k=1\\k\neq l}}^{d} \xi_{jk} \prod_{k=1}^{p} c_k(z_k,\,z_{jk}^{\mathcal{T}})\end{aligned}$$

  in case of multiplicative form, and

$$\begin{aligned}B_{lj} =& \mathbb{E}\left[W_l c(\mathbf{W},\,\mathbf{w}_j^{\mathcal{T}}) + c(\mathbf{z},\,\mathbf{z}_j^{\mathcal{T}})\right] \\ =& \mathbb{E}\left[W_l c(\mathbf{W},\,\mathbf{w}_j^{\mathcal{T}})\right] + \mathbb{E}\left[W_l\right] c(\mathbf{z},\,\mathbf{z}_j^{\mathcal{T}}) \\ =& \mathbb{E}\left[W_l \sum_{k=1}^{d} c_k(W_k,\,w_{jk}^{\mathcal{T}})\right] + \mu_l \sum_{k=1}^{p} c_k(z_k,\,z_{jk}^{\mathcal{T}}) \\ =& \mathbb{E}\left[W_l c_l(W_l,\,w_{jl}^{\mathcal{T}})\right] + \mu_l \sum_{\substack{k=1\\k\neq l}}^{d} \mathbb{E}\left[c_k(W_k,\,w_{jk}^{\mathcal{T}})\right] + \mu_l \sum_{k=1}^{p} c_k(z_k,\,z_{jk}^{\mathcal{T}}) \\ =& \psi_{jl} + \mu_l \sum_{\substack{k=1\\k\neq l}}^{d} \xi_{jk} + \mu_l \sum_{k=1}^{p} c_k(z_k,\,z_{jk}^{\mathcal{T}})\end{aligned}$$

in case of additive form, in which

$$\psi_{jl} \stackrel{\text{def}}{=\!=} \mathbb{E}\left[W_l c_l(W_l,\, w_{jl}^{\mathcal{T}})\right];$$

- $\mathbf{J} = \mathbb{E}\left[\mathbf{r}(\mathbf{W},\, \mathbf{z})\mathbf{r}^{\top}(\mathbf{W},\, \mathbf{z})\right] \in \mathbb{R}^{m\times m}$ with its ij-th element:

$$\begin{aligned}
J_{ij} =&\mathbb{E}\Big[c(\mathbf{W},\, \mathbf{w}_i^{\mathcal{T}})c(\mathbf{z},\, \mathbf{z}_i^{\mathcal{T}})c(\mathbf{W},\, \mathbf{w}_j^{\mathcal{T}})c(\mathbf{z},\, \mathbf{z}_j^{\mathcal{T}})\Big] \\
=&\mathbb{E}\left[c(\mathbf{W},\, \mathbf{w}_i^{\mathcal{T}})c(\mathbf{W},\, \mathbf{w}_j^{\mathcal{T}})\right] c(\mathbf{z},\, \mathbf{z}_i^{\mathcal{T}})c(\mathbf{z},\, \mathbf{z}_j^{\mathcal{T}}) \\
=&\prod_{k=1}^{d} \mathbb{E}\left[c_k(W_k,\, w_{ik}^{\mathcal{T}})c_k(W_k,\, w_{jk}^{\mathcal{T}})\right] \prod_{k=1}^{p} c_k(z_k,\, z_{ik}^{\mathcal{T}})c_k(z_k,\, z_{jk}^{\mathcal{T}}) \\
=&\prod_{k=1}^{d} \zeta_{ijk} \prod_{k=1}^{p} c_k(z_k,\, z_{ik}^{\mathcal{T}})c_k(z_k,\, z_{jk}^{\mathcal{T}})
\end{aligned}$$

in case of multiplicative form, and

$$\begin{aligned}
J_{ij} =&\mathbb{E}\Big[\left(c(\mathbf{W},\, \mathbf{w}_i^{\mathcal{T}}) + c(\mathbf{z},\, \mathbf{z}_i^{\mathcal{T}})\right)\left(c(\mathbf{W},\, \mathbf{w}_j^{\mathcal{T}}) + c(\mathbf{z},\, \mathbf{z}_j^{\mathcal{T}})\right)\Big] \\
=&\mathbb{E}\left[c(\mathbf{W},\, \mathbf{w}_i^{\mathcal{T}})c(\mathbf{W},\, \mathbf{w}_j^{\mathcal{T}})\right] + \mathbb{E}\left[c(\mathbf{W},\, \mathbf{w}_i^{\mathcal{T}})\right] c(\mathbf{z},\, \mathbf{z}_j^{\mathcal{T}}) \\
&+ \mathbb{E}\left[c(\mathbf{W},\, \mathbf{w}_j^{\mathcal{T}})\right] c(\mathbf{z},\, \mathbf{z}_i^{\mathcal{T}}) + c(\mathbf{z},\, \mathbf{z}_i^{\mathcal{T}})c(\mathbf{z},\, \mathbf{z}_j^{\mathcal{T}}) \\
=&\sum_{\substack{k,l=1 \\ k\neq l}}^{d} \mathbb{E}\left[c_k(W_k,\, w_{ik}^{\mathcal{T}})\right] \mathbb{E}\left[c_l(W_l,\, w_{jl}^{\mathcal{T}})\right] \\
&+ \sum_{k=1}^{d} \mathbb{E}\left[c_k(W_k,\, w_{ik}^{\mathcal{T}})c_k(W_k,\, w_{jk}^{\mathcal{T}})\right] \\
&+ \sum_{k=1}^{d} \xi_{ik} \sum_{k=1}^{p} c_k(z_k,\, z_{jk}^{\mathcal{T}}) + \sum_{k=1}^{d} \xi_{jk} \sum_{k=1}^{p} c_k(z_k,\, z_{ik}^{\mathcal{T}}) \\
&+ \sum_{k=1}^{p} c_k(z_k,\, z_{ik}^{\mathcal{T}}) \sum_{k=1}^{p} c_k(z_k,\, z_{jk}^{\mathcal{T}}) \\
=&\sum_{\substack{k,l=1 \\ k\neq l}}^{d} \xi_{ik}\xi_{jl} + \sum_{k=1}^{d} \zeta_{ijk} + \sum_{k=1}^{d} \xi_{ik} \sum_{k=1}^{p} c_k(z_k,\, z_{jk}^{\mathcal{T}}) \\
&+ \sum_{k=1}^{d} \xi_{jk} \sum_{k=1}^{p} c_k(z_k,\, z_{ik}^{\mathcal{T}}) + \sum_{k=1}^{p} c_k(z_k,\, z_{ik}^{\mathcal{T}}) \sum_{k=1}^{p} c_k(z_k,\, z_{jk}^{\mathcal{T}})
\end{aligned}$$

in case of additive form, in which

$$\zeta_{ijk} \stackrel{\text{def}}{=\!=} \mathbb{E}\left[c_k(W_k,\, w_{ik}^{\mathcal{T}})c_k(W_k,\, w_{jk}^{\mathcal{T}})\right].$$

2 Derivation of $\mathbb{E}\left[\sigma_g^2(\mathbf{W}, \mathbf{z})\right]$

Replacing $\sigma_g^2(\mathbf{W}, \mathbf{z})$ by equation (3.5):

$$
\begin{aligned}
\mathbb{E}\left[\sigma_g^2(\cdot,\cdot)\right] =& \sigma^2\,\mathbb{E}\Big[1+\eta-\mathbf{r}^\top(\mathbf{W},\,\mathbf{z})\mathbf{R}^{-1}\mathbf{r}(\mathbf{W},\,\mathbf{z})+\Big(\mathbf{h}(\mathbf{W},\,\mathbf{z})-\widetilde{\mathbf{H}}^\top\mathbf{R}^{-1}\mathbf{r}(\mathbf{W},\,\mathbf{z})\Big)^\top \\
& \times\Big(\widetilde{\mathbf{H}}^\top\mathbf{R}^{-1}\widetilde{\mathbf{H}}\Big)^{-1}\Big(\mathbf{h}(\mathbf{W},\,\mathbf{z})-\widetilde{\mathbf{H}}^\top\mathbf{R}^{-1}\mathbf{r}(\mathbf{W},\,\mathbf{z})\Big)\Big] \\
=& \sigma^2\,(1+\eta)+\sigma^2\,\mathbb{E}\Bigg[\mathbf{h}^\top(\mathbf{W},\,\mathbf{z})\Big(\widetilde{\mathbf{H}}^\top\mathbf{R}^{-1}\widetilde{\mathbf{H}}\Big)^{-1}\mathbf{h}(\mathbf{W},\,\mathbf{z}) \\
& +\mathbf{r}^\top(\mathbf{W},\,\mathbf{z})\Big\{\mathbf{R}^{-1}\widetilde{\mathbf{H}}\Big(\widetilde{\mathbf{H}}^\top\mathbf{R}^{-1}\widetilde{\mathbf{H}}\Big)^{-1}\widetilde{\mathbf{H}}^\top\mathbf{R}^{-1}-\mathbf{R}^{-1}\Big\}\mathbf{r}(\mathbf{W},\,\mathbf{z}) \\
& -2\mathrm{tr}\Big\{\mathbf{h}^\top(\mathbf{W},\,\mathbf{z})\Big(\widetilde{\mathbf{H}}^\top\mathbf{R}^{-1}\widetilde{\mathbf{H}}\Big)^{-1}\widetilde{\mathbf{H}}^\top\mathbf{R}^{-1}\mathbf{r}(\mathbf{W},\,\mathbf{z})\Big\}\Bigg] \\
=& \sigma^2\,(1+\eta)+\sigma^2\,\mathbb{E}\left[\mathbf{h}^\top(\mathbf{W},\,\mathbf{z})\Big(\widetilde{\mathbf{H}}^\top\mathbf{R}^{-1}\widetilde{\mathbf{H}}\Big)^{-1}\mathbf{h}(\mathbf{W},\,\mathbf{z})\right] \\
& +\sigma^2\,\mathbb{E}\left[\mathbf{r}^\top(\mathbf{W},\,\mathbf{z})\Big\{\mathbf{R}^{-1}\widetilde{\mathbf{H}}\Big(\widetilde{\mathbf{H}}^\top\mathbf{R}^{-1}\widetilde{\mathbf{H}}\Big)^{-1}\widetilde{\mathbf{H}}^\top\mathbf{R}^{-1}-\mathbf{R}^{-1}\Big\}\mathbf{r}(\mathbf{W},\,\mathbf{z})\right] \\
& -2\sigma^2\,\mathbb{E}\left[\mathrm{tr}\Big\{\mathbf{h}^\top(\mathbf{W},\,\mathbf{z})\Big(\widetilde{\mathbf{H}}^\top\mathbf{R}^{-1}\widetilde{\mathbf{H}}\Big)^{-1}\widetilde{\mathbf{H}}^\top\mathbf{R}^{-1}\mathbf{r}(\mathbf{W},\,\mathbf{z})\Big\}\right] \\
=& \sigma^2\,\left[1+\eta+\mathrm{tr}\left\{\mathbf{CP}\right\}+\mathbf{G}^\top\mathbf{CG}+\mathrm{tr}\left\{\mathbf{QJ}\right\}-2\mathrm{tr}\Big\{\mathbf{C}\widetilde{\mathbf{H}}^\top\mathbf{R}^{-1}\mathbf{K}\Big\}\right],
\end{aligned}
$$

where

- $\mathbf{C}=\Big(\widetilde{\mathbf{H}}^\top\mathbf{R}^{-1}\widetilde{\mathbf{H}}\Big)^{-1}\in\mathbb{R}^{(d+q)\times(d+q)}$ with $\widetilde{\mathbf{H}}=\left[\mathbf{w}^{\mathcal{T}},\mathbf{H}(\mathbf{z}^{\mathcal{T}})\right]\in\mathbb{R}^{m\times(d+q)}$;
- $\mathbf{P}=\mathrm{Var}\left[\mathbf{h}(\mathbf{W},\,\mathbf{z})\right]=\mathrm{Var}\left[\left(\mathbf{W}^\top,\,\mathbf{h}(\mathbf{z})^\top\right)^\top\right]=\mathrm{blkdiag}(\boldsymbol{\Omega},\,\mathbf{0})\in\mathbb{R}^{(d+q)\times(d+q)}$;
- $\mathbf{G}=\mathbb{E}\left[\mathbf{h}(\mathbf{W},\,\mathbf{z})\right]=\mathbb{E}\left[\left(\mathbf{W}^\top,\,\mathbf{h}(\mathbf{z})^\top\right)^\top\right]=[\boldsymbol{\mu}^\top,\,\mathbf{h}(\mathbf{z})^\top]^\top\in\mathbb{R}^{(d+q)\times 1}$;
- $\mathbf{Q}=\mathbf{R}^{-1}\widetilde{\mathbf{H}}\Big(\widetilde{\mathbf{H}}^\top\mathbf{R}^{-1}\widetilde{\mathbf{H}}\Big)^{-1}\widetilde{\mathbf{H}}^\top\mathbf{R}^{-1}-\mathbf{R}^{-1}\in\mathbb{R}^{m\times m}$;
- $\mathbf{K}=\mathbb{E}\left[\mathbf{h}(\mathbf{W},\,\mathbf{z})\mathbf{r}^\top(\mathbf{W},\,\mathbf{z})\right]^\top=\left[\mathbf{B}^\top,\,\mathbf{I}\mathbf{h}(\mathbf{z})^\top\right]\in\mathbb{R}^{m\times(d+q)}$.

3 Derivation of μ_I^2

Using equation (3.4), we have

$$\begin{aligned}
\mu_I^2 =& \left(\boldsymbol{\mu}^\top\widehat{\boldsymbol{\theta}} + \mathbf{h}(\mathbf{z})^\top\widehat{\boldsymbol{\beta}} + \mathbf{I}^\top\mathbf{A}\right)\left(\boldsymbol{\mu}^\top\widehat{\boldsymbol{\theta}} + \mathbf{h}(\mathbf{z})^\top\widehat{\boldsymbol{\beta}} + \mathbf{I}^\top\mathbf{A}\right)^\top \\
=& \left(\boldsymbol{\mu}^\top\widehat{\boldsymbol{\theta}} + \mathbf{h}(\mathbf{z})^\top\widehat{\boldsymbol{\beta}} + \mathbf{I}^\top\mathbf{A}\right)\left(\widehat{\boldsymbol{\theta}}^\top\boldsymbol{\mu} + \widehat{\boldsymbol{\beta}}^\top\mathbf{h}(\mathbf{z}) + \mathbf{A}^\top\mathbf{I}\right) \\
=& \boldsymbol{\mu}^\top\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^\top\boldsymbol{\mu} + \left(\mathbf{h}(\mathbf{z})^\top\widehat{\boldsymbol{\beta}}\right)^2 + \mathbf{I}^\top\mathbf{A}\mathbf{A}^\top\mathbf{I} \\
&+ 2\widehat{\boldsymbol{\theta}}^\top\boldsymbol{\mu}\mathbf{h}(\mathbf{z})^\top\widehat{\boldsymbol{\beta}} + 2\widehat{\boldsymbol{\theta}}^\top\boldsymbol{\mu}\mathbf{I}^\top\mathbf{A} + 2\mathbf{h}(\mathbf{z})^\top\widehat{\boldsymbol{\beta}}\mathbf{I}^\top\mathbf{A} \\
=& \operatorname{tr}\left\{\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^\top\boldsymbol{\mu}\boldsymbol{\mu}^\top\right\} + \left(\mathbf{h}(\mathbf{z})^\top\widehat{\boldsymbol{\beta}}\right)^2 + \operatorname{tr}\left\{\mathbf{A}\mathbf{A}^\top\mathbf{I}\mathbf{I}^\top\right\} \\
&+ 2\widehat{\boldsymbol{\theta}}^\top\boldsymbol{\mu}\mathbf{h}(\mathbf{z})^\top\widehat{\boldsymbol{\beta}} + 2\left[\widehat{\boldsymbol{\theta}}^\top\boldsymbol{\mu} + \mathbf{h}(\mathbf{z})^\top\widehat{\boldsymbol{\beta}}\right]\mathbf{I}^\top\mathbf{A}
\end{aligned}$$

Finally, we obtain the expression for (C.2), which is given by

$$\begin{aligned}
\sigma_I^2 =& \operatorname{tr}\left\{\mathbf{A}\mathbf{A}^\top\mathbf{J}\right\} - \operatorname{tr}\left\{\mathbf{A}\mathbf{A}^\top\mathbf{I}\mathbf{I}^\top\right\} + 2\widehat{\boldsymbol{\theta}}^\top\mathbf{B}\mathbf{A} - 2\widehat{\boldsymbol{\theta}}^\top\boldsymbol{\mu}\mathbf{I}^\top\mathbf{A} + \operatorname{tr}\left\{\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^\top\boldsymbol{\Omega}\right\} \\
&+ \sigma^2\left(1 + \eta + \operatorname{tr}\{\mathbf{C}\mathbf{P}\} + \mathbf{G}^\top\mathbf{C}\mathbf{G} + \operatorname{tr}\{\mathbf{Q}\mathbf{J}\} - 2\operatorname{tr}\left\{\mathbf{C}\widetilde{\mathbf{H}}^\top\mathbf{R}^{-1}\mathbf{K}\right\}\right) \\
=& \mathbf{A}^\top\left(\mathbf{J} - \mathbf{I}\mathbf{I}^\top\right)\mathbf{A} + 2\widehat{\boldsymbol{\theta}}^\top\left(\mathbf{B} - \boldsymbol{\mu}\mathbf{I}^\top\right)\mathbf{A} + \operatorname{tr}\left\{\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^\top\boldsymbol{\Omega}\right\} \\
&+ \sigma^2\left(1 + \eta + \operatorname{tr}\{\mathbf{Q}\mathbf{J}\} + \mathbf{G}^\top\mathbf{C}\mathbf{G} + \operatorname{tr}\left\{\mathbf{C}\mathbf{P} - 2\mathbf{C}\widetilde{\mathbf{H}}^\top\mathbf{R}^{-1}\mathbf{K}\right\}\right). \quad \text{(C.5)}
\end{aligned}$$

This together with equation (C.1) completes the proof. In case that the trend is assumed constant, the expressions for μ_I and σ_I^2 can be simplified to the following:

$$\begin{aligned}
\mu_I =& \left(\mathbf{1}_m^\top\mathbf{R}^{-1}\mathbf{1}_m\right)^{-1}\mathbf{1}_m^\top\mathbf{R}^{-1}\mathbf{y}^{\mathcal{T}} + \mathbf{I}^\top\mathbf{A}, \\
\sigma_I^2 =& \mathbf{A}^\top\left(\mathbf{J} - \mathbf{I}\mathbf{I}^\top\right)\mathbf{A} + \sigma^2\left(1 + \eta + \operatorname{tr}\{\mathbf{Q}\mathbf{J}\} + \mathbf{C} - \operatorname{tr}\left\{2\mathbf{C}\mathbf{1}_m^\top\mathbf{R}^{-1}\mathbf{I}\right\}\right),
\end{aligned}$$

where

- $\mathbf{A} = \mathbf{R}^{-1}\left(\mathbf{y}^{\mathcal{T}} - \mathbf{1}_m\left(\mathbf{1}_m^\top\mathbf{R}^{-1}\mathbf{1}_m\right)^{-1}\mathbf{1}_m^\top\mathbf{R}^{-1}\mathbf{y}^{\mathcal{T}}\right)$;
- $\mathbf{Q} = \mathbf{R}^{-1}\mathbf{1}_m\mathbf{C}\mathbf{1}_m^\top\mathbf{R}^{-1} - \mathbf{R}^{-1}$;
- $\mathbf{C} = \left(\mathbf{1}_m^\top\mathbf{R}^{-1}\mathbf{1}_m\right)^{-1}$.

C.2 Proof of Proposition 3.2

Replace $\mu_g(\mathbf{W}, \mathbf{z})$ by equation (3.4) with Assumption 1, we have

$$\begin{aligned}\mathbb{E}_{W_{k\in\mathbb{S}^c}}\left[\mu_g(\mathbf{W}, \mathbf{z})\right] =& \mathbb{E}_{W_{k\in\mathbb{S}^c}}\left[\mathbf{W}^\top\right]\widehat{\boldsymbol{\theta}} + \mathbf{h}(\mathbf{z})^\top\widehat{\boldsymbol{\beta}} \\ &+ \mathbb{E}_{W_{k\in\mathbb{S}^c}}\left[\mathbf{r}^\top(\mathbf{W}, \mathbf{z})\right]\mathbf{R}^{-1}\left(\mathbf{y}^{\mathcal{T}} - \mathbf{w}^{\mathcal{T}}\widehat{\boldsymbol{\theta}} - \mathbf{H}(\mathbf{z}^{\mathcal{T}})\widehat{\boldsymbol{\beta}}\right) \\ =& \widetilde{\boldsymbol{\mu}}^\top\widehat{\boldsymbol{\theta}} + \mathbf{h}(\mathbf{z})^\top\widehat{\boldsymbol{\beta}} + \widetilde{\mathbf{I}}^\top\mathbf{A},\end{aligned}$$

where

- $\widetilde{\boldsymbol{\mu}} = \mathbb{E}_{W_{k\in\mathbb{S}^c}}\left[\mathbf{W}^\top\right] \in \mathbb{R}^{d\times 1}$ is a column vector with its k-th element:

$$\widetilde{\mu}_k = \begin{cases} W_k, & k \in \mathbb{S}, \\ \mu_k, & k \in \mathbb{S}^c; \end{cases}$$

- $\widetilde{\mathbf{I}} = \mathbb{E}_{W_{k\in\mathbb{S}^c}}\left[\mathbf{r}^\top(\mathbf{W}, \mathbf{z})\right] \in \mathbb{R}^{m\times 1}$ with its i-th element:

$$\begin{aligned}\widetilde{I}_i =& \mathbb{E}_{W_{k\in\mathbb{S}^c}}\left[c(\mathbf{W}, \mathbf{w}_i^{\mathcal{T}})c(\mathbf{z}, \mathbf{z}_i^{\mathcal{T}})\right] \\ =& \mathbb{E}_{W_{k\in\mathbb{S}^c}}\left[c(\mathbf{W}, \mathbf{w}_i^{\mathcal{T}})\right]c(\mathbf{z}, \mathbf{z}_i^{\mathcal{T}}) \\ =& \mathbb{E}_{W_{k\in\mathbb{S}^c}}\left[\prod_{k=1}^{d} c_k(W_k, w_{ik}^{\mathcal{T}})\right]\prod_{k=1}^{p} c_k(z_k, z_{ik}^{\mathcal{T}}) \\ =& \prod_{k\in\mathbb{S}} c_k(W_k, w_{ik}^{\mathcal{T}}) \prod_{k\in\mathbb{S}^c} \mathbb{E}_{W_k}\left[c_k(W_k, w_{ik}^{\mathcal{T}})\right] \prod_{k=1}^{p} c_k(z_k, z_{ik}^{\mathcal{T}}) \\ =& \prod_{k\in\mathbb{S}} c_k(W_k, w_{ik}^{\mathcal{T}}) \prod_{k\in\mathbb{S}^c} \xi_{ik} \prod_{k=1}^{p} c_k(z_k, z_{ik}^{\mathcal{T}}).\end{aligned}$$

Then, we have

$$\begin{aligned}V_1(\mathbb{S}) =& \mathrm{Var}_{W_{k\in\mathbb{S}}}\left(\widetilde{\boldsymbol{\mu}}^\top\widehat{\boldsymbol{\theta}} + \mathbf{h}(\mathbf{z})^\top\widehat{\boldsymbol{\beta}} + \widetilde{\mathbf{I}}^\top\mathbf{A}\right) \\ =& \mathrm{Var}_{W_{k\in\mathbb{S}}}\left(\widetilde{\boldsymbol{\mu}}^\top\widehat{\boldsymbol{\theta}} + \widetilde{\mathbf{I}}^\top\mathbf{A}\right) \\ =& \underbrace{\mathbb{E}_{W_{k\in\mathbb{S}}}\left[\left(\widetilde{\boldsymbol{\mu}}^\top\widehat{\boldsymbol{\theta}} + \widetilde{\mathbf{I}}^\top\mathbf{A}\right)^2\right]}_{\text{(C.6.1)}} - \underbrace{\left(\mathbb{E}_{W_{k\in\mathbb{S}}}\left[\widetilde{\boldsymbol{\mu}}^\top\widehat{\boldsymbol{\theta}} + \widetilde{\mathbf{I}}^\top\mathbf{A}\right]\right)^2}_{\text{(C.6.2)}}. \qquad \text{(C.6)}\end{aligned}$$

We first derive (C.6.1) as follow:

$$\begin{aligned}(\text{C.6.1}) =&\mathbb{E}_{W_{k\in\mathbb{S}}}\left[\widetilde{\boldsymbol{\mu}}^\top\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^\top\widetilde{\boldsymbol{\mu}}+\widetilde{\mathbf{I}}^\top\mathbf{A}\mathbf{A}^\top\widetilde{\mathbf{I}}+2\widehat{\boldsymbol{\theta}}^\top\widetilde{\boldsymbol{\mu}}\widetilde{\mathbf{I}}^\top\mathbf{A}\right]\\ =&\mathrm{tr}\left\{\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^\top\left(\boldsymbol{\mu}\boldsymbol{\mu}^\top+\widetilde{\boldsymbol{\Omega}}\right)\right\}+\mathrm{tr}\left\{\mathbf{A}\mathbf{A}^\top\mathbb{E}_{W_{k\in\mathbb{S}}}\left[\widetilde{\mathbf{I}}\widetilde{\mathbf{I}}^\top\right]\right\}+2\widehat{\boldsymbol{\theta}}^\top\mathbb{E}_{W_{k\in\mathbb{S}}}\left[\widetilde{\boldsymbol{\mu}}\widetilde{\mathbf{I}}^\top\right]\mathbf{A}\\ =&\mathrm{tr}\left\{\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^\top\left(\boldsymbol{\mu}\boldsymbol{\mu}^\top+\widetilde{\boldsymbol{\Omega}}\right)\right\}+\mathrm{tr}\left\{\mathbf{A}\mathbf{A}^\top\widetilde{\mathbf{J}}\right\}+2\widehat{\boldsymbol{\theta}}^\top\widetilde{\mathbf{B}}\mathbf{A},\end{aligned}\tag{C.7}$$

where the second step uses the derivations analogous to those used for equations (C.3) and (C.4); $\widetilde{\boldsymbol{\Omega}}=\mathrm{Var}_{W_{k\in\mathbb{S}}}(\widetilde{\boldsymbol{\mu}})\in\mathbb{R}^{d\times d}$ is a diagonal matrix with its k-th diagonal element given by $\widetilde{\boldsymbol{\Omega}}_k=\sigma_k^2(\mathbf{x}_k)\mathbb{1}_{\{k\in\mathbb{S}\}}$; and

- $\widetilde{\mathbf{B}}=\mathbb{E}_{W_{k\in\mathbb{S}}}\left[\widetilde{\boldsymbol{\mu}}\widetilde{\mathbf{I}}^\top\right]\in\mathbb{R}^{d\times m}$ with its lj-th element:

$$\begin{aligned}\widetilde{B}_{lj} =&\mathbb{E}_{W_{k\in\mathbb{S}}}\left[\widetilde{\mu}_l\prod_{k\in\mathbb{S}}c_k(W_k,\,w_{jk}^{\mathcal{T}})\prod_{k\in\mathbb{S}^{\mathsf{c}}}\xi_{jk}\prod_{k=1}^{p}c_k(z_k,\,z_{jk}^{\mathcal{T}})\right]\\ =&\mathbb{E}_{W_{k\in\mathbb{S}}}\left[\widetilde{\mu}_l\prod_{k\in\mathbb{S}}c_k(W_k,\,w_{jk}^{\mathcal{T}})\right]\prod_{k\in\mathbb{S}^{\mathsf{c}}}\xi_{jk}\prod_{k=1}^{p}c_k(z_k,\,z_{jk}^{\mathcal{T}})\\ =&\begin{cases}\mathbb{E}_{W_{k\in\mathbb{S}}}\left[W_lc_l(W_l,\,w_{jl}^{\mathcal{T}})\prod\limits_{\substack{k\in\mathbb{S}\\k\neq l}}c_k(W_k,\,w_{jk}^{\mathcal{T}})\right]\prod\limits_{k\in\mathbb{S}^{\mathsf{c}}}\xi_{jk}\prod\limits_{k=1}^{p}c_k(z_k,\,z_{jk}^{\mathcal{T}}), & l\in\mathbb{S}\\ \mathbb{E}_{W_{k\in\mathbb{S}}}\left[\mu_l\prod\limits_{k\in\mathbb{S}}c_k(W_k,\,w_{jk}^{\mathcal{T}})\right]\prod\limits_{k\in\mathbb{S}^{\mathsf{c}}}\xi_{jk}\prod\limits_{k=1}^{p}c_k(z_k,\,z_{jk}^{\mathcal{T}}), & l\in\mathbb{S}^{\mathsf{c}}\end{cases}\\ =&\begin{cases}\mathbb{E}_{W_l}\left[W_lc_l(W_l,\,w_{jl}^{\mathcal{T}})\right]\prod\limits_{\substack{k\in\mathbb{S}\\k\neq l}}\mathbb{E}_{W_k}\left[c_k(W_k,\,w_{jk}^{\mathcal{T}})\right]\prod\limits_{k\in\mathbb{S}^{\mathsf{c}}}\xi_{jk}\prod\limits_{k=1}^{p}c_k(z_k,\,z_{jk}^{\mathcal{T}}), & l\in\mathbb{S}\\ \mu_l\prod\limits_{k\in\mathbb{S}}\mathbb{E}_{W_k}\left[c_k(W_k,\,w_{jk}^{\mathcal{T}})\right]\prod\limits_{k\in\mathbb{S}^{\mathsf{c}}}\xi_{jk}\prod\limits_{k=1}^{p}c_k(z_k,\,z_{jk}^{\mathcal{T}}), & l\in\mathbb{S}^{\mathsf{c}}\end{cases}\\ =&\begin{cases}\psi_{jl}\prod\limits_{\substack{k\in\mathbb{S}\\k\neq l}}\xi_{jk}\prod\limits_{k\in\mathbb{S}^{\mathsf{c}}}\xi_{jk}\prod\limits_{k=1}^{p}c_k(z_k,\,z_{jk}^{\mathcal{T}}), & l\in\mathbb{S}\\ \mu_l\prod\limits_{k\in\mathbb{S}}\xi_{jk}\prod\limits_{k\in\mathbb{S}^{\mathsf{c}}}\xi_{jk}\prod\limits_{k=1}^{p}c_k(z_k,\,z_{jk}^{\mathcal{T}}), & l\in\mathbb{S}^{\mathsf{c}}\end{cases}\\ =&\begin{cases}\psi_{jl}\prod\limits_{\substack{k=1\\k\neq l}}^{d}\xi_{jk}\prod\limits_{k=1}^{p}c_k(z_k,\,z_{jk}^{\mathcal{T}}), & l\in\mathbb{S}\\ \mu_l\prod\limits_{k=1}^{d}\xi_{jk}\prod\limits_{k=1}^{p}c_k(z_k,\,z_{jk}^{\mathcal{T}}), & l\in\mathbb{S}^{\mathsf{c}};\end{cases}\end{aligned}$$

- $\widetilde{\mathbf{J}} = \mathbb{E}_{W_{k\in\mathbb{S}}}\left[\widetilde{\mathbf{I}}\widetilde{\mathbf{I}}^\top\right] \in \mathbb{R}^{m\times m}$ with its ij-th element:

$$\begin{aligned}
\widetilde{J}_{ij} =&\mathbb{E}_{W_{k\in\mathbb{S}}}\left[\prod_{k\in\mathbb{S}} c_k(W_k,\, w_{ik}^{\mathcal{T}}) \prod_{k\in\mathbb{S}^c} \xi_{ik} \prod_{k=1}^{p} c_k(z_k,\, z_{ik}^{\mathcal{T}}) \times \prod_{k\in\mathbb{S}} c_k(W_k,\, w_{jk}^{\mathcal{T}}) \prod_{k\in\mathbb{S}^c} \xi_{jk} \prod_{k=1}^{p} c_k(z_k,\, z_{jk}^{\mathcal{T}})\right] \\
=&\mathbb{E}_{W_{k\in\mathbb{S}}}\left[\prod_{k\in\mathbb{S}} c_k(W_k,\, w_{ik}^{\mathcal{T}}) c_k(W_k,\, w_{jk}^{\mathcal{T}}) \prod_{k\in\mathbb{S}^c} \xi_{ik}\xi_{jk} \prod_{k=1}^{p} c_k(z_k,\, z_{ik}^{\mathcal{T}}) c_k(z_k,\, z_{jk}^{\mathcal{T}})\right] \\
=&\mathbb{E}_{W_{k\in\mathbb{S}}}\left[\prod_{k\in\mathbb{S}} c_k(W_k,\, w_{ik}^{\mathcal{T}}) c_k(W_k,\, w_{jk}^{\mathcal{T}})\right] \prod_{k\in\mathbb{S}^c} \xi_{ik}\xi_{jk} \prod_{k=1}^{p} c_k(z_k,\, z_{ik}^{\mathcal{T}}) c_k(z_k,\, z_{jk}^{\mathcal{T}}) \\
=&\prod_{k\in\mathbb{S}} \mathbb{E}_{W_k}\left[c_k(W_k,\, w_{ik}^{\mathcal{T}}) c_k(W_k,\, w_{jk}^{\mathcal{T}})\right] \prod_{k\in\mathbb{S}^c} \xi_{ik}\xi_{jk} \prod_{k=1}^{p} c_k(z_k,\, z_{ik}^{\mathcal{T}}) c_k(z_k,\, z_{jk}^{\mathcal{T}}) \\
=&\prod_{k\in\mathbb{S}} \zeta_{ijk} \prod_{k\in\mathbb{S}^c} \xi_{ik}\xi_{jk} \prod_{k=1}^{p} c_k(z_k,\, z_{ik}^{\mathcal{T}}) c_k(z_k,\, z_{jk}^{\mathcal{T}}).
\end{aligned}$$

We now derive (C.6.2) as follow:

$$\begin{aligned}
(\text{C.6.2}) &= \left(\mathbb{E}_{W_{k\in\mathbb{S}}}\left[\widetilde{\boldsymbol{\mu}}^\top\right]\widehat{\boldsymbol{\theta}} + \mathbb{E}_{W_{k\in\mathbb{S}}}\left[\widetilde{\mathbf{I}}^\top\right]\mathbf{A}\right)^2 \\
&= \left(\boldsymbol{\mu}^\top\widehat{\boldsymbol{\theta}} + \mathbf{I}^\top\mathbf{A}\right)^2 \\
&= \boldsymbol{\mu}^\top\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^\top\boldsymbol{\mu} + \mathbf{A}^\top\mathbf{I}\mathbf{I}^\top\mathbf{A} + 2\widehat{\boldsymbol{\theta}}^\top\boldsymbol{\mu}\mathbf{I}^\top\mathbf{A}. \qquad (\text{C.8})
\end{aligned}$$

Plugging equations (C.7) and (C.8) back into equation (C.6), we obtain

$$\begin{aligned}
V_1(\mathbb{S}) =&\operatorname{tr}\left\{\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^\top\left(\boldsymbol{\mu}\boldsymbol{\mu}^\top + \widetilde{\boldsymbol{\Omega}}\right)\right\} + \operatorname{tr}\left\{\mathbf{A}\mathbf{A}^\top\widetilde{\mathbf{J}}\right\} \\
&+ 2\widehat{\boldsymbol{\theta}}^\top\widetilde{\mathbf{B}}\mathbf{A} - \left(\boldsymbol{\mu}^\top\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^\top\boldsymbol{\mu} + \mathbf{A}^\top\mathbf{I}\mathbf{I}^\top\mathbf{A} + 2\widehat{\boldsymbol{\theta}}^\top\boldsymbol{\mu}\mathbf{I}^\top\mathbf{A}\right) \\
=&\operatorname{tr}\left\{\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^\top\boldsymbol{\mu}\boldsymbol{\mu}\right\} + \operatorname{tr}\left\{\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^\top\widetilde{\boldsymbol{\Omega}}\right\} + \mathbf{A}^\top\widetilde{\mathbf{J}}\mathbf{A} \\
&+ 2\widehat{\boldsymbol{\theta}}^\top\widetilde{\mathbf{B}}\mathbf{A} - \boldsymbol{\mu}^\top\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^\top\boldsymbol{\mu} - \mathbf{A}^\top\mathbf{I}\mathbf{I}^\top\mathbf{A} - 2\widehat{\boldsymbol{\theta}}^\top\boldsymbol{\mu}\mathbf{I}^\top\mathbf{A} \\
=&\operatorname{tr}\left\{\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^\top\boldsymbol{\mu}\boldsymbol{\mu}\right\} + \operatorname{tr}\left\{\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^\top\widetilde{\boldsymbol{\Omega}}\right\} + \mathbf{A}^\top\widetilde{\mathbf{J}}\mathbf{A} \\
&+ 2\widehat{\boldsymbol{\theta}}^\top\widetilde{\mathbf{B}}\mathbf{A} - \operatorname{tr}\left\{\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^\top\boldsymbol{\mu}\boldsymbol{\mu}\right\} - \mathbf{A}^\top\mathbf{I}\mathbf{I}^\top\mathbf{A} - 2\widehat{\boldsymbol{\theta}}^\top\boldsymbol{\mu}\mathbf{I}^\top\mathbf{A} \\
=&\operatorname{tr}\left\{\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^\top\widetilde{\boldsymbol{\Omega}}\right\} + \mathbf{A}^\top\left(\widetilde{\mathbf{J}} - \mathbf{I}\mathbf{I}^\top\right)\mathbf{A} + 2\widehat{\boldsymbol{\theta}}^\top\left(\widetilde{\mathbf{B}} - \boldsymbol{\mu}\mathbf{I}^\top\right)\mathbf{A}.
\end{aligned}$$

In case that the trend is assumed constant, $V_1(\mathbb{S})$ can be simplified to the following expression:

$$V_1(\mathbb{S}) = \mathbf{A}^\top\left(\widetilde{\mathbf{J}} - \mathbf{I}\mathbf{I}^\top\right)\mathbf{A}.$$

C.3 Proof of Proposition 3.3

Lemma C.1 *Denote*

$$\Gamma[m] = \int_b^a \frac{x^m}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \mathrm{d}x$$

for $m \in \mathbb{N}_0$, *where* $a \in \mathbb{R}$, $b \in \mathbb{R}$, $\mu \in \mathbb{R}$ *and* $\sigma \in \mathbb{R}_{\geq 0}$. *Then, we have*

$$\begin{aligned}
\Gamma[0] =& \Phi\left(\frac{a-\mu}{\sigma}\right) - \Phi\left(\frac{b-\mu}{\sigma}\right), \\
\Gamma[1] =& \mu\left[\Phi\left(\frac{a-\mu}{\sigma}\right) - \Phi\left(\frac{b-\mu}{\sigma}\right)\right] + \frac{\sigma}{\sqrt{2\pi}}\left[\exp\left\{-\frac{(b-\mu)^2}{2\sigma^2}\right\} - \exp\left\{-\frac{(a-\mu)^2}{2\sigma^2}\right\}\right], \\
\Gamma[2] =& \left(\mu^2+\sigma^2\right)\left[\Phi\left(\frac{a-\mu}{\sigma}\right) - \Phi\left(\frac{b-\mu}{\sigma}\right)\right] \\
&+ \frac{(\mu+b)\sigma}{\sqrt{2\pi}} \exp\left\{-\frac{(b-\mu)^2}{2\sigma^2}\right\} - \frac{(\mu+a)\sigma}{\sqrt{2\pi}} \exp\left\{-\frac{(a-\mu)^2}{2\sigma^2}\right\}, \\
\Gamma[3] =& \left(\mu^3+3\mu\sigma^2\right)\left[\Phi\left(\frac{a-\mu}{\sigma}\right) - \Phi\left(\frac{b-\mu}{\sigma}\right)\right] \\
&+ \frac{(b^2+\mu b+\mu^2+2\sigma^2)\sigma}{\sqrt{2\pi}} \exp\left\{-\frac{(b-\mu)^2}{2\sigma^2}\right\} \\
&- \frac{(a^2+\mu a+\mu^2+2\sigma^2)\sigma}{\sqrt{2\pi}} \exp\left\{-\frac{(a-\mu)^2}{2\sigma^2}\right\}, \\
\Gamma[4] =& \left(\mu^4+3\sigma^4+6\mu^2\sigma^2\right)\left[\Phi\left(\frac{a-\mu}{\sigma}\right) - \Phi\left(\frac{b-\mu}{\sigma}\right)\right] \\
&+ \frac{(b^3+\mu^3+\mu^2 b+\mu b^2+3\sigma^2 b+5\sigma^2\mu)\sigma}{\sqrt{2\pi}} \exp\left\{-\frac{(b-\mu)^2}{2\sigma^2}\right\} \\
&- \frac{(a^3+\mu^3+\mu^2 a+\mu a^2+3\sigma^2 a+5\sigma^2\mu)\sigma}{\sqrt{2\pi}} \exp\left\{-\frac{(a-\mu)^2}{2\sigma^2}\right\},
\end{aligned}$$

where $\Phi(\cdot)$ *denotes the cumulative density function of the standard normal.*

Proof Denote

$$\kappa[m] = \int_t^s \frac{x^m}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} \mathrm{d}x$$

for $m \in \mathbb{N}_0$, where $s \in \mathbb{R}$ and $t \in \mathbb{R}$. Then via integration by parts, we have

$$\begin{aligned}
\kappa[m] =& \frac{1}{\sqrt{2\pi}}\left(-x^{m-1}e^{-\frac{x^2}{2}}\Big|_t^s + (m-1)\int_t^s x^{m-2}e^{-\frac{x^2}{2}}\mathrm{d}x\right) \\
=& \frac{1}{\sqrt{2\pi}}\left(t^{m-1}e^{-\frac{t^2}{2}} - s^{m-1}e^{-\frac{s^2}{2}}\right) + (m-1)\int_t^s x^{m-2}e^{-\frac{x^2}{2}}\mathrm{d}x \\
=& \frac{1}{\sqrt{2\pi}}\left(t^{m-1}e^{-\frac{t^2}{2}} - s^{m-1}e^{-\frac{s^2}{2}}\right) + (m-1)\kappa[m-2].
\end{aligned}$$

Thus, we have

$$\kappa[0] = \int_t^s \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} \mathrm{d}x = \Phi(s) - \Phi(t), \tag{C.9}$$

$$\begin{aligned}\kappa[1] =& \int_t^s \frac{x}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} \mathrm{d}x \\ =& -\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\Big|_t^s \\ =& \frac{1}{\sqrt{2\pi}}\left(e^{-\frac{t^2}{2}} - e^{-\frac{s^2}{2}}\right),\end{aligned} \tag{C.10}$$

$$\begin{aligned}\kappa[2] =& \frac{1}{\sqrt{2\pi}}\left(te^{-\frac{t^2}{2}} - se^{-\frac{s^2}{2}}\right) + \kappa[0] \\ =& \frac{1}{\sqrt{2\pi}}\left(te^{-\frac{t^2}{2}} - se^{-\frac{s^2}{2}}\right) + \Phi(s) - \Phi(t),\end{aligned} \tag{C.11}$$

and

$$\begin{aligned}\kappa[3] =& \frac{1}{\sqrt{2\pi}}\left(t^2e^{-\frac{t^2}{2}} - s^2e^{-\frac{s^2}{2}}\right) + 2\kappa[1] \\ =& \frac{1}{\sqrt{2\pi}}\left(t^2e^{-\frac{t^2}{2}} - s^2e^{-\frac{s^2}{2}}\right) + \frac{2}{\sqrt{2\pi}}\left(e^{-\frac{t^2}{2}} - e^{-\frac{s^2}{2}}\right),\end{aligned} \tag{C.12}$$

$$\begin{aligned}\kappa[4] =& \frac{1}{\sqrt{2\pi}}\left(t^3e^{-\frac{t^2}{2}} - s^3e^{-\frac{s^2}{2}}\right) + 3\kappa[2] \\ =& \frac{1}{\sqrt{2\pi}}\left(t^3e^{-\frac{t^2}{2}} - s^3e^{-\frac{s^2}{2}}\right) + \frac{3}{\sqrt{2\pi}}\left(te^{-\frac{t^2}{2}} - se^{-\frac{s^2}{2}}\right) + 3\left[\Phi(s) - \Phi(t)\right],\end{aligned} \tag{C.13}$$

where $\Phi(\cdot)$ denotes the cumulative density function of the standard normal.

Denote

$$\Gamma[m] = \int_b^a \frac{x^m}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \mathrm{d}x$$

for $m \in \mathbb{N}_0$, where $a \in \mathbb{R}$, $b \in \mathbb{R}$, $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}_{\geq 0}$. Let

$$s = \frac{x-\mu}{\sigma},$$

then we have

$$\Gamma[m] = \int_{\frac{b-\mu}{\sigma}}^{\frac{a-\mu}{\sigma}} \frac{(\sigma s + \mu)^m}{\sqrt{2\pi}} \exp\left\{-\frac{s^2}{2}\right\} \mathrm{d}s$$

for $m \in \mathbb{N}_0$. The lemma is subsequently proved by using equations (C.9), (C.10), (C.11), (C.12) and (C.13) for all $m \in \{0, \ldots, 4\}$. □

C.3.1 Derivation for exponential case

1 Derivation of ξ_{ik}

$$
\begin{aligned}
\xi_{ik} =& \mathbb{E}\left[c_k(W_k,\, w_{ik}^{\mathcal{T}})\right] \\
=& \int \exp\left\{-\frac{|w-w_{ik}^{\mathcal{T}}|}{\gamma_k}\right\}\frac{1}{\sigma_k\sqrt{2\pi}}\exp\left\{-\frac{(w-\mu_k)^2}{2\sigma_k^2}\right\}\mathrm{d}w \\
=& \int_{w_{ik}^{\mathcal{T}}}^{+\infty}\frac{1}{\sigma_k\sqrt{2\pi}}\exp\left\{-\frac{w-w_{ik}^{\mathcal{T}}}{\gamma_k}-\frac{(w-\mu_k)^2}{2\sigma_k^2}\right\}\mathrm{d}w \\
&+\int_{-\infty}^{w_{ik}^{\mathcal{T}}}\frac{1}{\sigma_k\sqrt{2\pi}}\exp\left\{\frac{w-w_{ik}^{\mathcal{T}}}{\gamma_k}-\frac{(w-\mu_k)^2}{2\sigma_k^2}\right\}\mathrm{d}w \\
=& \exp\left\{\frac{\sigma_k^2+2\gamma_k\left(w_{ik}^{\mathcal{T}}-\mu_k\right)}{2\gamma_k^2}\right\}\int_{w_{ik}^{\mathcal{T}}}^{+\infty}\frac{1}{\sigma_k\sqrt{2\pi}}\exp\left\{-\frac{(w-\mu_A)^2}{2\sigma_k^2}\right\}\mathrm{d}w \\
&+\exp\left\{\frac{\sigma_k^2-2\gamma_k\left(w_{ik}^{\mathcal{T}}-\mu_k\right)}{2\gamma_k^2}\right\}\int_{-\infty}^{w_{ik}^{\mathcal{T}}}\frac{1}{\sigma_k\sqrt{2\pi}}\exp\left\{-\frac{(w-\mu_B)^2}{2\sigma_k^2}\right\}\mathrm{d}w,
\end{aligned}
$$

where the last step is obtained by completing the square. Using Lemma C.1,

$$
\begin{aligned}
\xi_{ik} = \exp\left\{\frac{\sigma_k^2+2\gamma_k\left(w_{ik}^{\mathcal{T}}-\mu_k\right)}{2\gamma_k^2}\right\}\Phi\left(\frac{\mu_A-w_{ik}^{\mathcal{T}}}{\sigma_k}\right) \\
+\exp\left\{\frac{\sigma_k^2-2\gamma_k\left(w_{ik}^{\mathcal{T}}-\mu_k\right)}{2\gamma_k^2}\right\}\Phi\left(\frac{w_{ik}^{\mathcal{T}}-\mu_B}{\sigma_k}\right),
\end{aligned}
$$

where $\mu_A = \mu_k - \sigma_k^2/\gamma_k$ and $\mu_B = \mu_k + \sigma_k^2/\gamma_k$.

2 Derivation of ζ_{ijk}

$$
\begin{aligned}
\zeta_{ijk} =& \mathbb{E}\left[c_k(W_k,\, w_{ik}^{\mathcal{T}})c_k(W_k,\, w_{jk}^{\mathcal{T}})\right] \\
=& \int \frac{1}{\sigma_k\sqrt{2\pi}}\exp\left\{-\frac{|w-w_{ik}^{\mathcal{T}}|}{\gamma_k}-\frac{|w-w_{jk}^{\mathcal{T}}|}{\gamma_k}-\frac{(w-\mu_k)^2}{2\sigma_k^2}\right\}\mathrm{d}w \\
=& \int_{w_{jk}^{\mathcal{T}}}^{+\infty}\frac{1}{\sigma_k\sqrt{2\pi}}\exp\left\{-\frac{w-w_{ik}^{\mathcal{T}}}{\gamma_k}-\frac{w-w_{jk}^{\mathcal{T}}}{\gamma_k}-\frac{(w-\mu_k)^2}{2\sigma_k^2}\right\}\mathrm{d}w \qquad \text{(C.14)} \\
&+\int_{w_{ik}^{\mathcal{T}}}^{w_{jk}^{\mathcal{T}}}\frac{1}{\sigma_k\sqrt{2\pi}}\exp\left\{-\frac{w-w_{ik}^{\mathcal{T}}}{\gamma_k}-\frac{w_{jk}^{\mathcal{T}}-w}{\gamma_k}-\frac{(w-\mu_k)^2}{2\sigma_k^2}\right\}\mathrm{d}w \qquad \text{(C.15)} \\
&+\int_{-\infty}^{w_{ik}^{\mathcal{T}}}\frac{1}{\sigma_k\sqrt{2\pi}}\exp\left\{-\frac{w_{ik}^{\mathcal{T}}-w}{\gamma_k}-\frac{w_{jk}^{\mathcal{T}}-w}{\gamma_k}-\frac{(w-\mu_k)^2}{2\sigma_k^2}\right\}\mathrm{d}w, \qquad \text{(C.16)}
\end{aligned}
$$

where $w_{ik}^{\mathcal{T}} \leq w_{jk}^{\mathcal{T}}$ is assumed.

By completing the square, term (C.14) can be rewritten as follow:

$$(\text{C.14}) = \exp\left\{\frac{2\sigma_k^2 + \gamma_k\left(w_{ik}^{\mathcal{T}} + w_{jk}^{\mathcal{T}} - 2\mu_k\right)}{\gamma_k^2}\right\} \int_{w_{jk}^{\mathcal{T}}}^{+\infty} \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left\{-\frac{(w-\mu_C)^2}{2\sigma_k^2}\right\} \mathrm{d}w,$$

where

$$\mu_C = \mu_k - \frac{2\sigma_k^2}{\gamma_k}.$$

Then by Lemma C.1, we obtain

$$(\text{C.14}) = \exp\left\{\frac{2\sigma_k^2 + \gamma_k\left(w_{ik}^{\mathcal{T}} + w_{jk}^{\mathcal{T}} - 2\mu_k\right)}{\gamma_k^2}\right\} \Phi\left(\frac{\mu_C - w_{jk}^{\mathcal{T}}}{\sigma_k}\right).$$

Since term (C.16) can be rewritten as

$$\begin{aligned}(\text{C.16}) &= \int_{-\infty}^{w_{ik}^{\mathcal{T}}} \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left\{-\frac{w_{ik}^{\mathcal{T}} - w}{\gamma_k} - \frac{w_{jk}^{\mathcal{T}} - w}{\gamma_k} - \frac{(w-\mu_k)^2}{2\sigma_k^2}\right\} \mathrm{d}w \\ &= \int_{-w_{ik}^{\mathcal{T}}}^{+\infty} \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left\{-\frac{w + w_{ik}^{\mathcal{T}}}{\gamma_k} - \frac{w + w_{jk}^{\mathcal{T}}}{\gamma_k} - \frac{(w+\mu_k)^2}{2\sigma_k^2}\right\} \mathrm{d}w,\end{aligned}$$

the form of which allows us to obtain solution of term (C.16) by simply using that of term (C.14). Thus, we have

$$(\text{C.16}) = \exp\left\{\frac{2\sigma_k^2 - \gamma_k\left(w_{ik}^{\mathcal{T}} + w_{jk}^{\mathcal{T}} - 2\mu_k\right)}{\gamma_k^2}\right\} \Phi\left(\frac{w_{ik}^{\mathcal{T}} - \mu_D}{\sigma_k}\right),$$

where

$$\mu_D = \mu_k + \frac{2\sigma_k^2}{\gamma_k}.$$

Term (C.15) is obtained as follow:

$$\begin{aligned}(\text{C.15}) &= \int_{w_{ik}^{\mathcal{T}}}^{w_{jk}^{\mathcal{T}}} \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left\{-\frac{w_{jk}^{\mathcal{T}} - w_{ik}^{\mathcal{T}}}{\gamma_k} - \frac{(w-\mu_k)^2}{2\sigma_k^2}\right\} \mathrm{d}w \\ &= \exp\left\{-\frac{w_{jk}^{\mathcal{T}} - w_{ik}^{\mathcal{T}}}{\gamma_k}\right\} \int_{w_{ik}^{\mathcal{T}}}^{w_{jk}^{\mathcal{T}}} \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left\{-\frac{(w-\mu_k)^2}{2\sigma_k^2}\right\} \mathrm{d}w \\ &= \exp\left\{-\frac{w_{jk}^{\mathcal{T}} - w_{ik}^{\mathcal{T}}}{\gamma_k}\right\} \left[\Phi\left(\frac{w_{jk}^{\mathcal{T}} - \mu_k}{\sigma_k}\right) - \Phi\left(\frac{w_{ik}^{\mathcal{T}} - \mu_k}{\sigma_k}\right)\right],\end{aligned}$$

where the last step uses Lemma C.1. Therefore, we obtain that

$$\begin{aligned}\zeta_{ijk} =& \exp\left\{\frac{2\sigma_k^2+\gamma_k\left(w_{ik}^{\mathcal{T}}+w_{jk}^{\mathcal{T}}-2\mu_k\right)}{\gamma_k^2}\right\}\Phi\left(\frac{\mu_C-w_{jk}^{\mathcal{T}}}{\sigma_k}\right)\\ &+\exp\left\{-\frac{w_{jk}^{\mathcal{T}}-w_{ik}^{\mathcal{T}}}{\gamma_k}\right\}\left[\Phi\left(\frac{w_{jk}^{\mathcal{T}}-\mu_k}{\sigma_k}\right)-\Phi\left(\frac{w_{ik}^{\mathcal{T}}-\mu_k}{\sigma_k}\right)\right]\\ &+\exp\left\{\frac{2\sigma_k^2-\gamma_k\left(w_{ik}^{\mathcal{T}}+w_{jk}^{\mathcal{T}}-2\mu_k\right)}{\gamma_k^2}\right\}\Phi\left(\frac{w_{ik}^{\mathcal{T}}-\mu_D}{\sigma_k}\right) \qquad \text{(C.17)}\end{aligned}$$

for $w_{ik}^{\mathcal{T}} \le w_{jk}^{\mathcal{T}}$. Observe that

$$\mathbb{E}\left[c_k(W_k,\, w_{ik}^{\mathcal{T}})c_k(W_k,\, w_{jk}^{\mathcal{T}})\right] = \mathbb{E}\left[c_k(W_k,\, w_{jk}^{\mathcal{T}})c_k(W_k,\, w_{ik}^{\mathcal{T}})\right],$$

Thus, the expression for ζ_{ijk} when $w_{ik}^{\mathcal{T}} > w_{jk}^{\mathcal{T}}$ is obtained by simply interchanging the positions of $w_{ik}^{\mathcal{T}}$ and $w_{jk}^{\mathcal{T}}$ in formula (C.17).

3 Derivation of ψ_{jk}

$$\begin{aligned}\psi_{jk} =& \mathbb{E}\left[W_k c_k(W_k,\, w_{jk}^{\mathcal{T}})\right]\\ =& \int \exp\left\{-\frac{|w-w_{jk}^{\mathcal{T}}|}{\gamma_k}\right\}\frac{w}{\sigma_k\sqrt{2\pi}}\exp\left\{-\frac{(w-\mu_k)^2}{2\sigma_k^2}\right\}\mathrm{d}w\\ =& \int_{w_{jk}^{\mathcal{T}}}^{+\infty}\frac{w}{\sigma_k\sqrt{2\pi}}\exp\left\{-\frac{w-w_{jk}^{\mathcal{T}}}{\gamma_k}-\frac{(w-\mu_k)^2}{2\sigma_k^2}\right\}\mathrm{d}w\\ &+\int_{-\infty}^{w_{jk}^{\mathcal{T}}}\frac{w}{\sigma_k\sqrt{2\pi}}\exp\left\{\frac{w-w_{jk}^{\mathcal{T}}}{\gamma_k}-\frac{(w-\mu_k)^2}{2\sigma_k^2}\right\}\mathrm{d}w\\ =& \exp\left\{\frac{\sigma_k^2+2\gamma_k\left(w_{jk}^{\mathcal{T}}-\mu_k\right)}{2\gamma_k^2}\right\}\int_{w_{jk}^{\mathcal{T}}}^{+\infty}\frac{w}{\sigma_k\sqrt{2\pi}}\exp\left\{-\frac{(w-\mu_A)^2}{2\sigma_k^2}\right\}\mathrm{d}w\\ &+\exp\left\{\frac{\sigma_k^2-2\gamma_k\left(w_{jk}^{\mathcal{T}}-\mu_k\right)}{2\gamma_k^2}\right\}\int_{-\infty}^{w_{jk}^{\mathcal{T}}}\frac{w}{\sigma_k\sqrt{2\pi}}\exp\left\{-\frac{(w-\mu_B)^2}{2\sigma_k^2}\right\}\mathrm{d}w,\end{aligned}$$

where the last step is obtained by completing the square. Thus, by Lemma C.1 we have

$$\begin{aligned}\psi_{jk} =& \exp\left\{\frac{\sigma_k^2+2\gamma_k\left(w_{jk}^{\mathcal{T}}-\mu_k\right)}{2\gamma_k^2}\right\}\left[\mu_A\Phi\left(\frac{\mu_A-w_{jk}^{\mathcal{T}}}{\sigma_k}\right)+\frac{\sigma_k}{\sqrt{2\pi}}\exp\left\{-\frac{\left(w_{jk}^{\mathcal{T}}-\mu_A\right)^2}{2\sigma_k^2}\right\}\right]\\ &+\exp\left\{\frac{\sigma_k^2-2\gamma_k\left(w_{jk}^{\mathcal{T}}-\mu_k\right)}{2\gamma_k^2}\right\}\left[-\mu_B\Phi\left(\frac{w_{jk}^{\mathcal{T}}-\mu_B}{\sigma_k}\right)+\frac{\sigma_k}{\sqrt{2\pi}}\exp\left\{-\frac{\left(w_{jk}^{\mathcal{T}}-\mu_B\right)^2}{2\sigma_k^2}\right\}\right].\end{aligned}$$

C.3.2 Derivation for squared exponential case

1 Derivation of ξ_{ik}

$$\begin{aligned}
\xi_{ik} =& \mathbb{E}\left[c_k(W_k,\, w_{ik}^{\mathcal{T}})\right] \\
=& \int \exp\left\{-\left(\frac{w - w_{ik}^{\mathcal{T}}}{\gamma_k}\right)^2\right\} \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left\{-\frac{(w-\mu_k)^2}{2\sigma_k^2}\right\} \mathrm{d}w \\
=& \int \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left\{-\frac{\left(w - w_{ik}^{\mathcal{T}}\right)^2}{\gamma_k^2} - \frac{(w-\mu_k)^2}{2\sigma_k^2}\right\} \mathrm{d}w \\
=& \exp\left\{-\frac{\left(\mu_k - w_{ik}^{\mathcal{T}}\right)^2}{2\sigma_k^2 + \gamma_k^2}\right\} \\
& \times \int \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left\{-\frac{2\sigma_k^2 + \gamma_k^2}{2\sigma_k^2\gamma_k^2}\left[w - \frac{2\sigma_k^2 w_{ik}^{\mathcal{T}} + \gamma_k^2\mu_k}{2\sigma_k^2 + \gamma_k^2}\right]^2\right\} \mathrm{d}w,
\end{aligned}$$

where the last step is obtained by completing the square. Consequently,

$$\begin{aligned}
\xi_{ik} =& \frac{1}{\sqrt{1 + 2\sigma_k^2/\gamma_k^2}} \exp\left\{-\frac{\left(\mu_k - w_{ik}^{\mathcal{T}}\right)^2}{2\sigma_k^2 + \gamma_k^2}\right\} \\
& \times \int \frac{\sqrt{2\sigma_k^2 + \gamma_k^2}}{\sigma_k\gamma_k\sqrt{2\pi}} \exp\left\{-\frac{2\sigma_k^2 + \gamma_k^2}{2\sigma_k^2\gamma_k^2}\left[w - \frac{2\sigma_k^2 w_{ik}^{\mathcal{T}} + \gamma_k^2\mu_k}{2\sigma_k^2 + \gamma_k^2}\right]^2\right\} \mathrm{d}w \\
=& \frac{1}{\sqrt{1 + 2\sigma_k^2/\gamma_k^2}} \exp\left\{-\frac{\left(\mu_k - w_{ik}^{\mathcal{T}}\right)^2}{2\sigma_k^2 + \gamma_k^2}\right\},
\end{aligned}$$

where the last step uses the fact that the integral in the first step equals to one because it integrates the probability density function of a normal distribution with mean and variance equal to

$$\frac{2\sigma_k^2 w_{ik}^{\mathcal{T}} + \gamma_k^2\mu_k}{2\sigma_k^2 + \gamma_k^2} \quad \text{and} \quad \frac{\sigma_k^2\gamma_k^2}{2\sigma_k^2 + \gamma_k^2}$$

respectively.

2 Derivation of ζ_{ijk}

$$\begin{aligned}
\zeta_{ijk} =& \mathbb{E}\left[c_k(W_k,\, w_{ik}^{\mathcal{T}})c_k(W_k,\, w_{jk}^{\mathcal{T}})\right] \\
=& \int \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left\{-\frac{\left(w - w_{ik}^{\mathcal{T}}\right)^2}{\gamma_k^2} - \frac{\left(w - w_{jk}^{\mathcal{T}}\right)^2}{\gamma_k^2} - \frac{(w-\mu_k)^2}{2\sigma_k^2}\right\} \mathrm{d}w.
\end{aligned}$$

By applying the completing in square, we can obtain the following:

$$\zeta_{ijk} = \frac{1}{\sqrt{1+4\sigma_k^2/\gamma_k^2}} \exp\left\{ -\frac{\left(\frac{w_{ik}^{\mathcal{T}}+w_{jk}^{\mathcal{T}}}{2} - \mu_k\right)^2}{\gamma_k^2/2 + 2\sigma_k^2} - \frac{\left(w_{ik}^{\mathcal{T}} - w_{jk}^{\mathcal{T}}\right)^2}{2\gamma_k^2} \right\} \times \int \frac{1}{\sigma_* \sqrt{2\pi}} \exp\left\{ -\frac{(w-\mu_*)^2}{2\sigma_*^2} \right\} \mathrm{d}w,$$

where

$$\mu_* = \frac{2\sigma_k^2 \left(w_{ik}^{\mathcal{T}} + w_{jk}^{\mathcal{T}}\right) + \gamma_k^2 \mu_k}{4\sigma_k^2 + \gamma_k^2}$$

and

$$\sigma_*^2 = \frac{\sigma_k^2 \gamma_k^2}{4\sigma_k^2 + \gamma_k^2}.$$

Thus, we have

$$\zeta_{ijk} = \frac{1}{\sqrt{1+4\sigma_k^2/\gamma_k^2}} \exp\left\{ -\frac{\left(\frac{w_{ik}^{\mathcal{T}}+w_{jk}^{\mathcal{T}}}{2} - \mu_k\right)^2}{\gamma_k^2/2 + 2\sigma_k^2} - \frac{\left(w_{ik}^{\mathcal{T}} - w_{jk}^{\mathcal{T}}\right)^2}{2\gamma_k^2} \right\}.$$

3 Derivation of ψ_{jk}

$$\begin{aligned}
\psi_{jk} &= \mathbb{E}\left[W_k c_k(W_k,\, w_{jk}^{\mathcal{T}})\right] \\
&= \int \frac{w}{\sigma_k \sqrt{2\pi}} \exp\left\{ -\frac{\left(w - w_{jk}^{\mathcal{T}}\right)^2}{\gamma_k^2} - \frac{(w-\mu_k)^2}{2\sigma_k^2} \right\} \mathrm{d}w \\
&= \frac{1}{\sqrt{1+2\sigma_k^2/\gamma_k^2}} \exp\left\{ -\frac{\left(\mu_k - w_{jk}^{\mathcal{T}}\right)^2}{2\sigma_k^2 + \gamma_k^2} \right\} \int \frac{w}{\sigma_* \sqrt{2\pi}} \exp\left\{ -\frac{(w-\mu_*)^2}{2\sigma_*^2} \right\} \mathrm{d}w,
\end{aligned}$$

where the last step is obtained by completing in square; and

$$\mu_* = \frac{2\sigma_k^2 w_{jk}^{\mathcal{T}} + \gamma_k^2 \mu_k}{2\sigma_k^2 + \gamma_k^2} \quad \text{and} \quad \sigma_*^2 = \frac{\sigma_k^2 \gamma_k^2}{2\sigma_k^2 + \gamma_k^2}.$$

Realising that the integral

$$\int \frac{w}{\sigma_* \sqrt{2\pi}} \exp\left\{ -\frac{(w-\mu_*)^2}{2\sigma_*^2} \right\} \mathrm{d}w$$

is in fact the expectation of a normal random variable with mean μ_* and variance σ_*^2, we have

$$\psi_{jk} = \frac{1}{\sqrt{1+2\sigma_k^2/\gamma_k^2}} \exp\left\{ -\frac{\left(\mu_k - w_{jk}^{\mathcal{T}}\right)^2}{2\sigma_k^2 + \gamma_k^2} \right\} \frac{2\sigma_k^2 w_{jk}^{\mathcal{T}} + \gamma_k^2 \mu_k}{2\sigma_k^2 + \gamma_k^2}.$$

C.3.3 Derivation for Matérn-1.5 case

1 Derivation of ξ_{ik}

$$
\begin{aligned}
\xi_{ik} =& \mathbb{E}\left[c_k(W_k,\, w_{ik}^{\mathcal{T}})\right] \\
=& \int \left(1 + \frac{\sqrt{3}|w - w_{ik}^{\mathcal{T}}|}{\gamma_k}\right) \\
& \quad \times \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left\{-\frac{\sqrt{3}|w - w_{ik}^{\mathcal{T}}|}{\gamma_k} - \frac{(w-\mu_k)^2}{2\sigma_k^2}\right\} \mathrm{d}w \\
=& \int_{w_{ik}^{\mathcal{T}}}^{+\infty} \left(1 + \frac{\sqrt{3}\left(w - w_{ik}^{\mathcal{T}}\right)}{\gamma_k}\right) \\
& \quad \times \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left\{-\frac{\sqrt{3}\left(w - w_{ik}^{\mathcal{T}}\right)}{\gamma_k} - \frac{(w-\mu_k)^2}{2\sigma_k^2}\right\} \mathrm{d}w \qquad \text{(C.18)} \\
& + \int_{-\infty}^{w_{ik}^{\mathcal{T}}} \left(1 + \frac{\sqrt{3}\left(w_{ik}^{\mathcal{T}} - w\right)}{\gamma_k}\right) \\
& \quad \times \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left\{\frac{\sqrt{3}\left(w - w_{ik}^{\mathcal{T}}\right)}{\gamma_k} - \frac{(w-\mu_k)^2}{2\sigma_k^2}\right\} \mathrm{d}w. \qquad \text{(C.19)}
\end{aligned}
$$

We first calculate term (C.18) by completing in square:

$$
\begin{aligned}
\text{(C.18)} = \exp&\left\{\frac{3\sigma_k^2 + 2\sqrt{3}\gamma_k\left(w_{ik}^{\mathcal{T}} - \mu_k\right)}{2\gamma_k^2}\right\} \\
& \times \int_{w_{ik}^{\mathcal{T}}}^{+\infty} \left[E_{11}w + E_{10}\right] \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left\{-\frac{(w-\mu_A)^2}{2\sigma_k^2}\right\},
\end{aligned}
$$

where

$$
E_{10} = 1 - \frac{\sqrt{3}w_{ik}^{\mathcal{T}}}{\gamma_k}, \quad E_{11} = \frac{\sqrt{3}}{\gamma_k} \quad \text{and} \quad \mu_A = \mu_k - \frac{\sqrt{3}\sigma_k^2}{\gamma_k}.
$$

By Lemma C.1, we then obtain

$$
\begin{aligned}
\text{(C.18)} = \exp&\left\{\frac{3\sigma_k^2 + 2\sqrt{3}\gamma_k\left(w_{ik}^{\mathcal{T}} - \mu_k\right)}{2\gamma_k^2}\right\} \\
& \times \left[\mathbf{E}_1^\top \mathbf{\Lambda}_{11} \Phi\left(\frac{\mu_A - w_{ik}^{\mathcal{T}}}{\sigma_k}\right) + \mathbf{E}_1^\top \mathbf{\Lambda}_{12} \frac{\sigma_k}{\sqrt{2\pi}} \exp\left\{-\frac{(w_{ik}^{\mathcal{T}} - \mu_A)^2}{2\sigma_k^2}\right\}\right],
\end{aligned}
$$

where

$$
\mathbf{E}_1 = [E_{10},\, E_{11}]^\top, \quad \mathbf{\Lambda}_{11} = [1,\, \mu_A]^\top \quad \text{and} \quad \mathbf{\Lambda}_{12} = [0,\, 1]^\top.
$$

Term (C.19) can be rewritten as follow:

$$\begin{aligned}(\text{C.19}) &= \int_{-\infty}^{w_{ik}^{\mathcal{T}}} \left(1 + \frac{\sqrt{3}\left(w_{ik}^{\mathcal{T}} - w\right)}{\gamma_k}\right) \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left\{\frac{\sqrt{3}\left(w - w_{ik}^{\mathcal{T}}\right)}{\gamma_k} - \frac{(w-\mu_k)^2}{2\sigma_k^2}\right\} \mathrm{d}w \\ &= \int_{-w_{ik}^{\mathcal{T}}}^{+\infty} \left(1 + \frac{\sqrt{3}\left(w + w_{ik}^{\mathcal{T}}\right)}{\gamma_k}\right) \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left\{-\frac{\sqrt{3}\left(w + w_{ik}^{\mathcal{T}}\right)}{\gamma_k} - \frac{(w+\mu_k)^2}{2\sigma_k^2}\right\} \mathrm{d}w,\end{aligned}$$

the form of which allows us to obtain solution of term (C.19) by simply using that of term (C.18). Thus, we have

$$\begin{aligned}(\text{C.19}) = \exp&\left\{\frac{3\sigma_k^2 - 2\sqrt{3}\gamma_k\left(w_{ik}^{\mathcal{T}} - \mu_k\right)}{2\gamma_k^2}\right\} \\ &\left[\mathbf{E}_2^\top \mathbf{\Lambda}_{21} \Phi\left(\frac{w_{ik}^{\mathcal{T}} - \mu_B}{\sigma_k}\right) + \mathbf{E}_2^\top \mathbf{\Lambda}_{22} \frac{\sigma_k}{\sqrt{2\pi}} \exp\left\{-\frac{(w_{ik}^{\mathcal{T}} - \mu_B)^2}{2\sigma_k^2}\right\}\right],\end{aligned}$$

where

$$\mathbf{E}_2 = [E_{20},\, E_{21}]^\top, \quad \mathbf{\Lambda}_{21} = [1,\, -\mu_B]^\top \quad \text{and} \quad \mathbf{\Lambda}_{22} = [0,\, 1]^\top$$

with

$$E_{20} = 1 + \frac{\sqrt{3}w_{ik}^{\mathcal{T}}}{\gamma_k}, \quad E_{21} = \frac{\sqrt{3}}{\gamma_k} \quad \text{and} \quad \mu_B = \mu_k + \frac{\sqrt{3}\sigma_k^2}{\gamma_k}.$$

Finally, we have

$$\begin{aligned}\xi_{ik} =& \exp\left\{\frac{3\sigma_k^2 + 2\sqrt{3}\gamma_k\left(w_{ik}^{\mathcal{T}} - \mu_k\right)}{2\gamma_k^2}\right\} \\ &\times \left[\mathbf{E}_1^\top \mathbf{\Lambda}_{11} \Phi\left(\frac{\mu_A - w_{ik}^{\mathcal{T}}}{\sigma_k}\right) + \mathbf{E}_1^\top \mathbf{\Lambda}_{12} \frac{\sigma_k}{\sqrt{2\pi}} \exp\left\{-\frac{(w_{ik}^{\mathcal{T}} - \mu_A)^2}{2\sigma_k^2}\right\}\right] \\ &+ \exp\left\{\frac{3\sigma_k^2 - 2\sqrt{3}\gamma_k\left(w_{ik}^{\mathcal{T}} - \mu_k\right)}{2\gamma_k^2}\right\} \\ &\times \left[\mathbf{E}_2^\top \mathbf{\Lambda}_{21} \Phi\left(\frac{w_{ik}^{\mathcal{T}} - \mu_B}{\sigma_k}\right) + \mathbf{E}_2^\top \mathbf{\Lambda}_{22} \frac{\sigma_k}{\sqrt{2\pi}} \exp\left\{-\frac{(w_{ik}^{\mathcal{T}} - \mu_B)^2}{2\sigma_k^2}\right\}\right].\end{aligned}$$

2 Derivation of ζ_{ijk}

$$\begin{aligned}\zeta_{ijk} =& \mathbb{E}\left[c_k(W_k,\, w_{ik}^{\mathcal{T}}) c_k(W_k,\, w_{jk}^{\mathcal{T}})\right] \\ =& \int \left(1 + \frac{\sqrt{3}|w - w_{ik}^{\mathcal{T}}|}{\gamma_k}\right)\left(1 + \frac{\sqrt{3}|w - w_{jk}^{\mathcal{T}}|}{\gamma_k}\right) \\ &\times \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left\{-\frac{\sqrt{3}|w - w_{ik}^{\mathcal{T}}| + \sqrt{3}|w - w_{jk}^{\mathcal{T}}|}{\gamma_k} - \frac{(w-\mu_k)^2}{2\sigma_k^2}\right\} \mathrm{d}w.\end{aligned}$$

Assume that $w_{ik}^{\mathcal{T}} \leq w_{jk}^{\mathcal{T}}$, we have

$$\zeta_{ijk} = \int_{w_{jk}^{\mathcal{T}}}^{+\infty} \left(1 + \frac{\sqrt{3}(w - w_{ik}^{\mathcal{T}})}{\gamma_k}\right)\left(1 + \frac{\sqrt{3}(w - w_{jk}^{\mathcal{T}})}{\gamma_k}\right) \times \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left\{-\frac{\sqrt{3}(w - w_{ik}^{\mathcal{T}}) + \sqrt{3}(w - w_{jk}^{\mathcal{T}})}{\gamma_k} - \frac{(w-\mu_k)^2}{2\sigma_k^2}\right\} \mathrm{d}w \tag{C.20}$$

$$+ \int_{w_{ik}^{\mathcal{T}}}^{w_{jk}^{\mathcal{T}}} \left(1 + \frac{\sqrt{3}(w - w_{ik}^{\mathcal{T}})}{\gamma_k}\right)\left(1 + \frac{\sqrt{3}(w_{jk}^{\mathcal{T}} - w)}{\gamma_k}\right) \times \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left\{-\frac{\sqrt{3}(w - w_{ik}^{\mathcal{T}}) + \sqrt{3}(w_{jk}^{\mathcal{T}} - w)}{\gamma_k} - \frac{(w-\mu_k)^2}{2\sigma_k^2}\right\} \mathrm{d}w \tag{C.21}$$

$$+ \int_{-\infty}^{w_{ik}^{\mathcal{T}}} \left(1 + \frac{\sqrt{3}(w_{ik}^{\mathcal{T}} - w)}{\gamma_k}\right)\left(1 + \frac{\sqrt{3}(w_{jk}^{\mathcal{T}} - w)}{\gamma_k}\right) \times \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left\{-\frac{\sqrt{3}(w_{ik}^{\mathcal{T}} - w) + \sqrt{3}(w_{jk}^{\mathcal{T}} - w)}{\gamma_k} - \frac{(w-\mu_k)^2}{2\sigma_k^2}\right\} \mathrm{d}w. \tag{C.22}$$

We first calculate term (C.20) by expanding the product of two brackets after the integral sign:

$$\text{(C.20)} = \int_{w_{jk}^{\mathcal{T}}}^{+\infty} (E_{32}w^2 + E_{31}w + E_{30}) \times \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left\{-\frac{\sqrt{3}(w - w_{ik}^{\mathcal{T}}) + \sqrt{3}(w - w_{jk}^{\mathcal{T}})}{\gamma_k} - \frac{(w-\mu_k)^2}{2\sigma_k^2}\right\} \mathrm{d}w,$$

where

$$E_{30} = 1 + \frac{3w_{ik}^{\mathcal{T}}w_{jk}^{\mathcal{T}} - \sqrt{3}\gamma_k\left(w_{ik}^{\mathcal{T}} + w_{jk}^{\mathcal{T}}\right)}{\gamma_k^2}, \quad E_{31} = \frac{2\sqrt{3}\gamma_k - 3\left(w_{ik}^{\mathcal{T}} + w_{jk}^{\mathcal{T}}\right)}{\gamma_k^2}$$

and $E_{32} = 3/\gamma_k^2$. Then by completing in square, we have

$$\text{(C.20)} = \exp\left\{\frac{6\sigma_k^2 + \sqrt{3}\gamma_k\left(w_{ik}^{\mathcal{T}} + w_{jk}^{\mathcal{T}} - 2\mu_k\right)}{\gamma_k^2}\right\} \times \int_{w_{jk}^{\mathcal{T}}}^{+\infty} (E_{32}w^2 + E_{31}w + E_{30}) \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left\{-\frac{(w-\mu_C)^2}{2\sigma_k^2}\right\} \mathrm{d}w,$$

where

$$\mu_C = \mu_k - 2\sqrt{3}\frac{\sigma_k^2}{\gamma_k}.$$

Using Lemma C.1 and arranging terms, we obtain

$$\begin{aligned}(\text{C.20}) = &\exp\left\{\frac{6\sigma_k^2 + \sqrt{3}\gamma_k\left(w_{ik}^{\mathcal{T}} + w_{jk}^{\mathcal{T}} - 2\mu_k\right)}{\gamma_k^2}\right\} \\ &\times\left[\mathbf{E}_3^\top \mathbf{\Lambda}_{31}\Phi\left(\frac{\mu_C - w_{jk}^{\mathcal{T}}}{\sigma_k}\right) + \mathbf{E}_3^\top \mathbf{\Lambda}_{32}\frac{\sigma_k}{\sqrt{2\pi}}\exp\left\{-\frac{(w_{jk}^{\mathcal{T}} - \mu_C)^2}{2\sigma_k^2}\right\}\right],\end{aligned}$$

where

$$\mathbf{E}_3 = [E_{30},\, E_{31},\, E_{32}]^\top, \quad \mathbf{\Lambda}_{31} = [1,\, \mu_C,\, \mu_C^2{+}\sigma_k^2]^\top \quad \text{and} \quad \mathbf{\Lambda}_{32} = [0,\, 1,\, \mu_C{+}w_{jk}^{\mathcal{T}}]^\top.$$

The derivation of term (C.21) is analogue to that of term (C.20). By expanding the product of two brackets after the integral sign, we have

$$\begin{aligned}(\text{C.21}) = &\int_{w_{ik}^{\mathcal{T}}}^{w_{jk}^{\mathcal{T}}}(E_{42}w^2 + E_{41}w + E_{40}) \\ &\frac{1}{\sigma_k\sqrt{2\pi}}\exp\left\{-\frac{\sqrt{3}(w - w_{ik}^{\mathcal{T}}) + \sqrt{3}(w_{jk}^{\mathcal{T}} - w)}{\gamma_k} - \frac{(w-\mu_k)^2}{2\sigma_k^2}\right\}\mathrm{d}w,\end{aligned}$$

where

$$E_{40} = 1{+}\frac{\sqrt{3}\gamma_k\left(w_{jk}^{\mathcal{T}} - w_{ik}^{\mathcal{T}}\right) - 3w_{ik}^{\mathcal{T}}w_{jk}^{\mathcal{T}}}{\gamma_k^2}, \quad E_{41} = \frac{3\left(w_{ik}^{\mathcal{T}} + w_{jk}^{\mathcal{T}}\right)}{\gamma_k^2} \quad \text{and} \quad E_{42} = -\frac{3}{\gamma_k^2}.$$

Then by completing in square, we have

$$\begin{aligned}(\text{C.21}) = &\exp\left\{-\frac{\sqrt{3}\left(w_{jk}^{\mathcal{T}} - w_{ik}^{\mathcal{T}}\right)}{\gamma_k}\right\} \\ &\int_{w_{ik}^{\mathcal{T}}}^{w_{jk}^{\mathcal{T}}}(E_{42}w^2 + E_{41}w + E_{40})\frac{1}{\sigma_k\sqrt{2\pi}}\exp\left\{-\frac{(w-\mu_k)^2}{2\sigma_k^2}\right\}\mathrm{d}w.\end{aligned}$$

Using Lemma C.1 and arranging terms, we obtain

$$\begin{aligned}(\text{C.21}) = &\exp\left\{-\frac{\sqrt{3}\left(w_{jk}^{\mathcal{T}} - w_{ik}^{\mathcal{T}}\right)}{\gamma_k}\right\}\left[\mathbf{E}_4^\top \mathbf{\Lambda}_{41}\left(\Phi\left(\frac{w_{jk}^{\mathcal{T}} - \mu_k}{\sigma_k}\right) - \Phi\left(\frac{w_{ik}^{\mathcal{T}} - \mu_k}{\sigma_k}\right)\right)\right. \\ &\left. + \mathbf{E}_4^\top \mathbf{\Lambda}_{42}\frac{\sigma_k}{\sqrt{2\pi}}\exp\left\{-\frac{(w_{ik}^{\mathcal{T}} - \mu_k)^2}{2\sigma_k^2}\right\} - \mathbf{E}_4^\top \mathbf{\Lambda}_{43}\frac{\sigma_k}{\sqrt{2\pi}}\exp\left\{-\frac{(w_{jk}^{\mathcal{T}} - \mu_k)^2}{2\sigma_k^2}\right\}\right],\end{aligned}$$

where

$$\mathbf{E}_4 = [E_{40},\, E_{41},\, E_{42}]^\top, \quad \mathbf{\Lambda}_{41} = [1,\, \mu_k,\, \mu_k^2 + \sigma_k^2]^\top, \quad \mathbf{\Lambda}_{42} = [0,\, 1,\, \mu_k + w_{ik}^{\mathcal{T}}]^\top$$

and

$$\mathbf{\Lambda}_{43} = [0,\, 1,\, \mu_k + w_{jk}^{\mathcal{T}}]^\top.$$

Term (C.22) can then be computed in the following way:

$$\begin{aligned}
(\text{C.22}) &= \int_{-\infty}^{w_{ik}^{\mathcal{T}}} \left(1 + \frac{\sqrt{3}(w_{ik}^{\mathcal{T}} - w)}{\gamma_k}\right)\left(1 + \frac{\sqrt{3}(w_{jk}^{\mathcal{T}} - w)}{\gamma_k}\right) \\
&\qquad \times \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left\{-\frac{\sqrt{3}(w_{ik}^{\mathcal{T}} - w) + \sqrt{3}(w_{jk}^{\mathcal{T}} - w)}{\gamma_k} - \frac{(w-\mu_k)^2}{2\sigma_k^2}\right\} \mathrm{d}w \\
&= \int_{-w_{ik}^{\mathcal{T}}}^{+\infty} \left(1 + \frac{\sqrt{3}(w + w_{ik}^{\mathcal{T}})}{\gamma_k}\right)\left(1 + \frac{\sqrt{3}(w + w_{jk}^{\mathcal{T}})}{\gamma_k}\right) \\
&\qquad \times \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left\{-\frac{\sqrt{3}(w + w_{ik}^{\mathcal{T}}) + \sqrt{3}(w + w_{jk}^{\mathcal{T}})}{\gamma_k} - \frac{(w+\mu_k)^2}{2\sigma_k^2}\right\} \mathrm{d}w,
\end{aligned}$$

the form of which allows us to obtain solution of term (C.22) by simply using that of term (C.20). Thus, we have

$$\begin{aligned}
(\text{C.22}) &= \exp\left\{\frac{6\sigma_k^2 - \sqrt{3}\gamma_k\left(w_{ik}^{\mathcal{T}} + w_{jk}^{\mathcal{T}} - 2\mu_k\right)}{\gamma_k^2}\right\} \\
&\qquad \times \left[\mathbf{E}_5^\top \mathbf{\Lambda}_{51} \Phi\left(\frac{w_{ik}^{\mathcal{T}} - \mu_D}{\sigma_k}\right) + \mathbf{E}_5^\top \mathbf{\Lambda}_{52} \frac{\sigma_k}{\sqrt{2\pi}} \exp\left\{-\frac{(w_{ik}^{\mathcal{T}} - \mu_D)^2}{2\sigma_k^2}\right\}\right],
\end{aligned}$$

where

$$\mathbf{E}_5 = [E_{50},\, E_{51},\, E_{52}]^\top, \quad \mathbf{\Lambda}_{51} = [1,\, -\mu_D,\, \mu_D^2 + \sigma_k^2]^\top$$

and

$$\mathbf{\Lambda}_{52} = [0,\, 1,\, -\mu_D - w_{ik}^{\mathcal{T}}]^\top$$

with

- $E_{50} = 1 + \dfrac{3w_{ik}^{\mathcal{T}} w_{jk}^{\mathcal{T}} + \sqrt{3}\gamma_k\left(w_{ik}^{\mathcal{T}} + w_{jk}^{\mathcal{T}}\right)}{\gamma_k^2}$;
- $E_{51} = \dfrac{2\sqrt{3}\gamma_k + 3\left(w_{ik}^{\mathcal{T}} + w_{jk}^{\mathcal{T}}\right)}{\gamma_k^2}$;
- $E_{52} = \dfrac{3}{\gamma_k^2}$ and $\mu_D = \mu_k + 2\sqrt{3}\dfrac{\sigma_k^2}{\gamma_k}$.

Therefore, the expression for ζ_{ijk} when $w_{ik}^{\mathcal{T}} \leq w_{jk}^{\mathcal{T}}$ is given by

$$\begin{aligned}
\zeta_{ijk} =& \exp\left\{\frac{6\sigma_k^2 + \sqrt{3}\gamma_k\left(w_{ik}^{\mathcal{T}} + w_{jk}^{\mathcal{T}} - 2\mu_k\right)}{\gamma_k^2}\right\} \\
&\times \left[\mathbf{E}_3^\top \mathbf{\Lambda}_{31} \Phi\left(\frac{\mu_C - w_{jk}^{\mathcal{T}}}{\sigma_k}\right) + \mathbf{E}_3^\top \mathbf{\Lambda}_{32} \frac{\sigma_k}{\sqrt{2\pi}} \exp\left\{-\frac{(w_{jk}^{\mathcal{T}} - \mu_C)^2}{2\sigma_k^2}\right\}\right] \\
&+ \exp\left\{-\frac{\sqrt{3}\left(w_{jk}^{\mathcal{T}} - w_{ik}^{\mathcal{T}}\right)}{\gamma_k}\right\}\left[\mathbf{E}_4^\top \mathbf{\Lambda}_{41}\left(\Phi\left(\frac{w_{jk}^{\mathcal{T}} - \mu_k}{\sigma_k}\right) - \Phi\left(\frac{w_{ik}^{\mathcal{T}} - \mu_k}{\sigma_k}\right)\right)\right. \\
&\left. + \mathbf{E}_4^\top \mathbf{\Lambda}_{42} \frac{\sigma_k}{\sqrt{2\pi}} \exp\left\{-\frac{(w_{ik}^{\mathcal{T}} - \mu_k)^2}{2\sigma_k^2}\right\} - \mathbf{E}_4^\top \mathbf{\Lambda}_{43} \frac{\sigma_k}{\sqrt{2\pi}} \exp\left\{-\frac{(w_{jk}^{\mathcal{T}} - \mu_k)^2}{2\sigma_k^2}\right\}\right] \\
&+ \exp\left\{\frac{6\sigma_k^2 - \sqrt{3}\gamma_k\left(w_{ik}^{\mathcal{T}} + w_{jk}^{\mathcal{T}} - 2\mu_k\right)}{\gamma_k^2}\right\} \\
&\times \left[\mathbf{E}_5^\top \mathbf{\Lambda}_{51} \Phi\left(\frac{w_{ik}^{\mathcal{T}} - \mu_D}{\sigma_k}\right) + \mathbf{E}_5^\top \mathbf{\Lambda}_{52} \frac{\sigma_k}{\sqrt{2\pi}} \exp\left\{-\frac{(w_{ik}^{\mathcal{T}} - \mu_D)^2}{2\sigma_k^2}\right\}\right].
\end{aligned}$$

Observe that

$$\mathbb{E}\left[c_k(W_k,\, w_{ik}^{\mathcal{T}}) c_k(W_k,\, w_{jk}^{\mathcal{T}})\right] = \mathbb{E}\left[c_k(W_k,\, w_{jk}^{\mathcal{T}}) c_k(W_k,\, w_{ik}^{\mathcal{T}})\right].$$

Thus, the expression for ζ_{ijk} when $w_{ik}^{\mathcal{T}} > w_{jk}^{\mathcal{T}}$ is obtained by simply interchanging the positions of $w_{ik}^{\mathcal{T}}$ and $w_{jk}^{\mathcal{T}}$ in the above formula of ζ_{ijk} when $w_{ik}^{\mathcal{T}} \leq w_{jk}^{\mathcal{T}}$.

3 Derivation of ψ_{jk}

$$\begin{aligned}
\psi_{jk} =& \mathbb{E}\left[W_k c_k(W_k,\, w_{jk}^{\mathcal{T}})\right] \\
=& \int w\left(1 + \frac{\sqrt{3}|w - w_{jk}^{\mathcal{T}}|}{\gamma_k}\right) \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left\{-\frac{\sqrt{3}|w - w_{jk}^{\mathcal{T}}|}{\gamma_k} - \frac{(w - \mu_k)^2}{2\sigma_k^2}\right\} \mathrm{d}w \\
=& \int_{w_{jk}^{\mathcal{T}}}^{+\infty} \left(w + \frac{\sqrt{3}w\left(w - w_{jk}^{\mathcal{T}}\right)}{\gamma_k}\right) \\
&\times \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left\{-\frac{\sqrt{3}\left(w - w_{jk}^{\mathcal{T}}\right)}{\gamma_k} - \frac{(w - \mu_k)^2}{2\sigma_k^2}\right\} \mathrm{d}w \qquad \text{(C.23)} \\
&+ \int_{-\infty}^{w_{jk}^{\mathcal{T}}} \left(w + \frac{\sqrt{3}w\left(w_{jk}^{\mathcal{T}} - w\right)}{\gamma_k}\right) \\
&\times \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left\{\frac{\sqrt{3}\left(w - w_{jk}^{\mathcal{T}}\right)}{\gamma_k} - \frac{(w - \mu_k)^2}{2\sigma_k^2}\right\} \mathrm{d}w. \qquad \text{(C.24)}
\end{aligned}$$

We first calculate term (C.23) by arranging the terms in the bracket after the

integral sign and completing in square:

$$\text{(C.23)} = \exp\left\{\frac{3\sigma_k^2 + 2\sqrt{3}\gamma_k\left(w_{jk}^{\mathcal{T}} - \mu_k\right)}{2\gamma_k^2}\right\} \times \int_{w_{jk}^{\mathcal{T}}}^{+\infty} \left[E_{11}w^2 + E_{10}w\right] \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left\{-\frac{(w-\mu_A)^2}{2\sigma_k^2}\right\}.$$

By Lemma C.1, we then obtain

$$\text{(C.23)} = \exp\left\{\frac{3\sigma_k^2 + 2\sqrt{3}\gamma_k\left(w_{jk}^{\mathcal{T}} - \mu_k\right)}{2\gamma_k^2}\right\} \times \left[\mathbf{E}_1^\top \mathbf{\Lambda}_{61} \Phi\left(\frac{\mu_A - w_{jk}^{\mathcal{T}}}{\sigma_k}\right) + \mathbf{E}_1^\top \mathbf{\Lambda}_{62} \frac{\sigma_k}{\sqrt{2\pi}} \exp\left\{-\frac{(w_{jk}^{\mathcal{T}} - \mu_A)^2}{2\sigma_k^2}\right\}\right],$$

where

$$\mathbf{\Lambda}_{61} = [\mu_A,\, \mu_A^2 + \sigma_k^2]^\top \quad \text{and} \quad \mathbf{\Lambda}_{62} = [1,\, \mu_A + w_{jk}^{\mathcal{T}}]^\top.$$

Term (C.24) can be rewritten as follow:

$$\begin{aligned}\text{(C.24)} &= \int_{-\infty}^{w_{jk}^{\mathcal{T}}} \left(1 + \frac{\sqrt{3}\left(w_{jk}^{\mathcal{T}} - w\right)}{\gamma_k}\right) \times \frac{w}{\sigma_k\sqrt{2\pi}} \exp\left\{\frac{\sqrt{3}\left(w - w_{jk}^{\mathcal{T}}\right)}{\gamma_k} - \frac{(w-\mu_k)^2}{2\sigma_k^2}\right\} \mathrm{d}w \\ &= -\int_{-w_{jk}^{\mathcal{T}}}^{+\infty} \left(1 + \frac{\sqrt{3}\left(w + w_{jk}^{\mathcal{T}}\right)}{\gamma_k}\right) \times \frac{w}{\sigma_k\sqrt{2\pi}} \exp\left\{-\frac{\sqrt{3}\left(w + w_{jk}^{\mathcal{T}}\right)}{\gamma_k} - \frac{(w+\mu_k)^2}{2\sigma_k^2}\right\} \mathrm{d}w,\end{aligned}$$

the form of which allows us to obtain the solution of term (C.24) by simply using that of term (C.23). Thus, we have

$$\text{(C.24)} = -\exp\left\{\frac{3\sigma_k^2 - 2\sqrt{3}\gamma_k\left(w_{jk}^{\mathcal{T}} - \mu_k\right)}{2\gamma_k^2}\right\} \times \left[\mathbf{E}_2^\top \mathbf{\Lambda}_{71} \Phi\left(\frac{w_{jk}^{\mathcal{T}} - \mu_B}{\sigma_k}\right) + \mathbf{E}_2^\top \mathbf{\Lambda}_{72} \frac{\sigma_k}{\sqrt{2\pi}} \exp\left\{-\frac{(w_{jk}^{\mathcal{T}} - \mu_B)^2}{2\sigma_k^2}\right\}\right],$$

where

$$\mathbf{\Lambda}_{71} = [-\mu_B,\, \mu_B^2 + \sigma_k^2]^\top \quad \text{and} \quad \mathbf{\Lambda}_{72} = [1,\, -\mu_B - w_{jk}^{\mathcal{T}}]^\top.$$

Finally, we have

$$\begin{aligned}\psi_{jk} =& \exp\left\{\frac{3\sigma_k^2 + 2\sqrt{3}\gamma_k\left(w_{jk}^{\mathcal{T}} - \mu_k\right)}{2\gamma_k^2}\right\} \\ &\times \left[\mathbf{E}_1^\top \mathbf{\Lambda}_{61}\Phi\left(\frac{\mu_A - w_{jk}^{\mathcal{T}}}{\sigma_k}\right) + \mathbf{E}_1^\top \mathbf{\Lambda}_{62}\frac{\sigma_k}{\sqrt{2\pi}}\exp\left\{-\frac{(w_{jk}^{\mathcal{T}} - \mu_A)^2}{2\sigma_k^2}\right\}\right] \\ &- \exp\left\{\frac{3\sigma_k^2 - 2\sqrt{3}\gamma_k\left(w_{jk}^{\mathcal{T}} - \mu_k\right)}{2\gamma_k^2}\right\} \\ &\times \left[\mathbf{E}_2^\top \mathbf{\Lambda}_{71}\Phi\left(\frac{w_{jk}^{\mathcal{T}} - \mu_B}{\sigma_k}\right) + \mathbf{E}_2^\top \mathbf{\Lambda}_{72}\frac{\sigma_k}{\sqrt{2\pi}}\exp\left\{-\frac{(w_{jk}^{\mathcal{T}} - \mu_B)^2}{2\sigma_k^2}\right\}\right].\end{aligned}$$

C.3.4 Derivation for Matérn-2.5 case

1 Derivation of ξ_{ik}

$$\begin{aligned}\xi_{ik} =& \mathbb{E}\left[c_k(W_k,\, w_{ik}^{\mathcal{T}})\right] \\ =& \int\left(1 + \frac{\sqrt{5}|w - w_{ik}^{\mathcal{T}}|}{\gamma_k} + \frac{5(w - w_{ik}^{\mathcal{T}})^2}{3\gamma_k^2}\right) \\ &\times \frac{1}{\sigma_k\sqrt{2\pi}}\exp\left\{-\frac{\sqrt{5}|w - w_{ik}^{\mathcal{T}}|}{\gamma_k} - \frac{(w - \mu_k)^2}{2\sigma_k^2}\right\}\mathrm{d}w \\ =& \int_{w_{ik}^{\mathcal{T}}}^{+\infty}\left(1 + \frac{\sqrt{5}\left(w - w_{ik}^{\mathcal{T}}\right)}{\gamma_k} + \frac{5}{3}\left(\frac{w - w_{ik}^{\mathcal{T}}}{\gamma_k}\right)^2\right) \\ &\times \frac{1}{\sigma_k\sqrt{2\pi}}\exp\left\{-\frac{\sqrt{5}\left(w - w_{ik}^{\mathcal{T}}\right)}{\gamma_k} - \frac{(w - \mu_k)^2}{2\sigma_k^2}\right\}\mathrm{d}w \qquad \text{(C.25)} \\ &+ \int_{-\infty}^{w_{ik}^{\mathcal{T}}}\left(1 + \frac{\sqrt{5}\left(w_{ik}^{\mathcal{T}} - w\right)}{\gamma_k} + \frac{5}{3}\left(\frac{w - w_{ik}^{\mathcal{T}}}{\gamma_k}\right)^2\right) \\ &\times \frac{1}{\sigma_k\sqrt{2\pi}}\exp\left\{\frac{\sqrt{5}\left(w - w_{ik}^{\mathcal{T}}\right)}{\gamma_k} - \frac{(w - \mu_k)^2}{2\sigma_k^2}\right\}\mathrm{d}w. \qquad \text{(C.26)}\end{aligned}$$

We first calculate term (C.25) by arranging the terms in the bracket after the integral sign and completing the square:

$$\begin{aligned}\text{(C.25)} = \exp&\left\{\frac{5\sigma_k^2 + 2\sqrt{5}\gamma_k\left(w_{ik}^{\mathcal{T}} - \mu_k\right)}{2\gamma_k^2}\right\} \\ &\int_{w_{ik}^{\mathcal{T}}}^{+\infty}\left[E_{12}w^2 + E_{11}w + E_{10}\right]\frac{1}{\sigma_k\sqrt{2\pi}}\exp\left\{-\frac{(w - \mu_A)^2}{2\sigma_k^2}\right\},\end{aligned}$$

where

$$E_{10} = 1 - \frac{\sqrt{5}w_{ik}^{\mathcal{T}}}{\gamma_k} + \frac{5\left(w_{ik}^{\mathcal{T}}\right)^2}{3\gamma_k^2}, \quad E_{11} = \frac{\sqrt{5}}{\gamma_k} - \frac{10w_{ik}^{\mathcal{T}}}{3\gamma_k^2}, \quad E_{12} = \frac{5}{3\gamma_k^2}.$$

and

$$\mu_A = \mu_k - \frac{\sqrt{5}\sigma_k^2}{\gamma_k}.$$

By Lemma C.1, we then obtain

$$\begin{aligned}(\text{C.25}) = \exp&\left\{\frac{5\sigma_k^2 + 2\sqrt{5}\gamma_k\left(w_{ik}^{\mathcal{T}} - \mu_k\right)}{2\gamma_k^2}\right\} \\ &\times\left[\mathbf{E}_1^\top\mathbf{\Lambda}_{11}\Phi\left(\frac{\mu_A - w_{ik}^{\mathcal{T}}}{\sigma_k}\right) + \mathbf{E}_1^\top\mathbf{\Lambda}_{12}\frac{\sigma_k}{\sqrt{2\pi}}\exp\left\{-\frac{(w_{ik}^{\mathcal{T}} - \mu_A)^2}{2\sigma_k^2}\right\}\right],\end{aligned}$$

where

$$\mathbf{E}_1 = [E_{10},\, E_{11},\, E_{12}]^\top, \quad \mathbf{\Lambda}_{11} = [1,\, \mu_A,\, \mu_A^2 + \sigma_k^2]^\top, \quad \mathbf{\Lambda}_{12} = [0,\, 1,\, \mu_A + w_{ik}^{\mathcal{T}}]^\top.$$

Term (C.26) can be rewritten as follow:

$$\begin{aligned}(\text{C.26}) &= \int_{-\infty}^{w_{ik}^{\mathcal{T}}}\left(1 + \frac{\sqrt{5}\left(w_{ik}^{\mathcal{T}} - w\right)}{\gamma_k} + \frac{5}{3}\left(\frac{w - w_{ik}^{\mathcal{T}}}{\gamma_k}\right)^2\right) \\ &\qquad\times\frac{1}{\sigma_k\sqrt{2\pi}}\exp\left\{\frac{\sqrt{5}\left(w - w_{ik}^{\mathcal{T}}\right)}{\gamma_k} - \frac{(w - \mu_k)^2}{2\sigma_k^2}\right\}\mathrm{d}w \\ &= \int_{-w_{ik}^{\mathcal{T}}}^{+\infty}\left(1 + \frac{\sqrt{5}\left(w + w_{ik}^{\mathcal{T}}\right)}{\gamma_k} + \frac{5}{3}\left(\frac{w + w_{ik}^{\mathcal{T}}}{\gamma_k}\right)^2\right) \\ &\qquad\times\frac{1}{\sigma_k\sqrt{2\pi}}\exp\left\{-\frac{\sqrt{5}\left(w + w_{ik}^{\mathcal{T}}\right)}{\gamma_k} - \frac{(w + \mu_k)^2}{2\sigma_k^2}\right\}\mathrm{d}w,\end{aligned}$$

the form of which allows us to obtain solution of term (C.26) by simply using that of term (C.25). Thus, we have

$$\begin{aligned}(\text{C.26}) = \exp&\left\{\frac{5\sigma_k^2 - 2\sqrt{5}\gamma_k\left(w_{ik}^{\mathcal{T}} - \mu_k\right)}{2\gamma_k^2}\right\} \\ &\times\left[\mathbf{E}_2^\top\mathbf{\Lambda}_{21}\Phi\left(\frac{w_{ik}^{\mathcal{T}} - \mu_B}{\sigma_k}\right) + \mathbf{E}_2^\top\mathbf{\Lambda}_{22}\frac{\sigma_k}{\sqrt{2\pi}}\exp\left\{-\frac{(w_{ik}^{\mathcal{T}} - \mu_B)^2}{2\sigma_k^2}\right\}\right],\end{aligned}$$

where

$$\mathbf{E}_2 = [E_{20},\, E_{21},\, E_{22}]^\top, \quad \mathbf{\Lambda}_{21} = [1,\, -\mu_B,\, \mu_B^2 + \sigma_k^2]^\top \quad \text{and} \quad \mathbf{\Lambda}_{22} = [0,\, 1,\, -\mu_B - w_{ik}^{\mathcal{T}}]^\top$$

with

$$E_{20} = 1+\frac{\sqrt{5}w_{ik}^{\mathcal{T}}}{\gamma_k}+\frac{5\left(w_{ik}^{\mathcal{T}}\right)^2}{3\gamma_k^2}, \quad E_{21} = \frac{\sqrt{5}}{\gamma_k}+\frac{10w_{ik}^{\mathcal{T}}}{3\gamma_k^2}, \quad E_{22} = \frac{5}{3\gamma_k^2}, \quad \mu_B = \mu_k+\frac{\sqrt{5}\sigma_k^2}{\gamma_k}.$$

Thus, we have

$$\begin{aligned}\xi_{ik} =& \exp\left\{\frac{5\sigma_k^2+2\sqrt{5}\gamma_k\left(w_{ik}^{\mathcal{T}}-\mu_k\right)}{2\gamma_k^2}\right\}\\ &\times\left[\mathbf{E}_1^\top\mathbf{\Lambda}_{11}\Phi\left(\frac{\mu_A-w_{ik}^{\mathcal{T}}}{\sigma_k}\right)+\mathbf{E}_1^\top\mathbf{\Lambda}_{12}\frac{\sigma_k}{\sqrt{2\pi}}\exp\left\{-\frac{(w_{ik}^{\mathcal{T}}-\mu_A)^2}{2\sigma_k^2}\right\}\right]\\ &+\exp\left\{\frac{5\sigma_k^2-2\sqrt{5}\gamma_k\left(w_{ik}^{\mathcal{T}}-\mu_k\right)}{2\gamma_k^2}\right\}\\ &\times\left[\mathbf{E}_2^\top\mathbf{\Lambda}_{21}\Phi\left(\frac{w_{ik}^{\mathcal{T}}-\mu_B}{\sigma_k}\right)+\mathbf{E}_2^\top\mathbf{\Lambda}_{22}\frac{\sigma_k}{\sqrt{2\pi}}\exp\left\{-\frac{(w_{ik}^{\mathcal{T}}-\mu_B)^2}{2\sigma_k^2}\right\}\right].\end{aligned}$$

2 Derivation of ζ_{ijk}

Assume that $w_{ik}^{\mathcal{T}} \leq w_{jk}^{\mathcal{T}}$, we have

$$\begin{aligned}\zeta_{ijk} =& \int_{w_{jk}^{\mathcal{T}}}^{+\infty}\left(1+\frac{\sqrt{5}(w-w_{ik}^{\mathcal{T}})}{\gamma_k}+\frac{5}{3}\left(\frac{w-w_{ik}^{\mathcal{T}}}{\gamma_k}\right)^2\right)\left(1+\frac{\sqrt{5}(w-w_{jk}^{\mathcal{T}})}{\gamma_k}+\frac{5}{3}\left(\frac{w-w_{jk}^{\mathcal{T}}}{\gamma_k}\right)^2\right)\\ &\times\frac{1}{\sigma_k\sqrt{2\pi}}\exp\left\{-\frac{\sqrt{5}(w-w_{ik}^{\mathcal{T}})+\sqrt{5}(w-w_{jk}^{\mathcal{T}})}{\gamma_k}-\frac{(w-\mu_k)^2}{2\sigma_k^2}\right\}\mathrm{d}w \qquad \text{(C.27)}\\ &+\int_{w_{ik}^{\mathcal{T}}}^{w_{jk}^{\mathcal{T}}}\left(1+\frac{\sqrt{5}(w-w_{ik}^{\mathcal{T}})}{\gamma_k}+\frac{5}{3}\left(\frac{w-w_{ik}^{\mathcal{T}}}{\gamma_k}\right)^2\right)\left(1+\frac{\sqrt{5}(w_{jk}^{\mathcal{T}}-w)}{\gamma_k}+\frac{5}{3}\left(\frac{w-w_{jk}^{\mathcal{T}}}{\gamma_k}\right)^2\right)\\ &\times\frac{1}{\sigma_k\sqrt{2\pi}}\exp\left\{-\frac{\sqrt{5}(w-w_{ik}^{\mathcal{T}})+\sqrt{5}(w_{jk}^{\mathcal{T}}-w)}{\gamma_k}-\frac{(w-\mu_k)^2}{2\sigma_k^2}\right\}\mathrm{d}w \qquad \text{(C.28)}\\ &+\int_{-\infty}^{w_{ik}^{\mathcal{T}}}\left(1+\frac{\sqrt{5}(w_{ik}^{\mathcal{T}}-w)}{\gamma_k}+\frac{5}{3}\left(\frac{w-w_{ik}^{\mathcal{T}}}{\gamma_k}\right)^2\right)\left(1+\frac{\sqrt{5}(w_{jk}^{\mathcal{T}}-w)}{\gamma_k}+\frac{5}{3}\left(\frac{w-w_{jk}^{\mathcal{T}}}{\gamma_k}\right)^2\right)\\ &\times\frac{1}{\sigma_k\sqrt{2\pi}}\exp\left\{-\frac{\sqrt{5}(w_{ik}^{\mathcal{T}}-w)+\sqrt{5}(w_{jk}^{\mathcal{T}}-w)}{\gamma_k}-\frac{(w-\mu_k)^2}{2\sigma_k^2}\right\}\mathrm{d}w. \qquad \text{(C.29)}\end{aligned}$$

We first calculate term (C.27) by expanding the product of two brackets after the integral sign:

$$\begin{aligned}(\text{C.27}) = &\int_{w_{jk}^{\mathcal{T}}}^{+\infty} (E_{34}w^4 + E_{33}w^3 + E_{32}w^2 + E_{31}w + E_{30}) \\ &\times \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left\{-\frac{\sqrt{5}(w - w_{ik}^{\mathcal{T}}) + \sqrt{5}(w - w_{jk}^{\mathcal{T}})}{\gamma_k} - \frac{(w-\mu_k)^2}{2\sigma_k^2}\right\} \mathrm{d}w,\end{aligned}$$

where

$$\begin{aligned}E_{30} =& 1 + \Big[25\left(w_{ik}^{\mathcal{T}}\right)^2\left(w_{jk}^{\mathcal{T}}\right)^2 - 3\sqrt{5}\left(3\gamma_k^3 + 5\gamma_k w_{ik}^{\mathcal{T}} w_{jk}^{\mathcal{T}}\right)\left(w_{ik}^{\mathcal{T}} + w_{jk}^{\mathcal{T}}\right) \\ &+ 15\gamma_k^2\left(\left(w_{ik}^{\mathcal{T}}\right)^2 + \left(w_{jk}^{\mathcal{T}}\right)^2 + 3w_{ik}^{\mathcal{T}} w_{jk}^{\mathcal{T}}\right)\Big] \Big/ 9\gamma_k^4 \\ E_{31} =& \Big[18\sqrt{5}\gamma_k^3 + 15\sqrt{5}\gamma_k\left(\left(w_{ik}^{\mathcal{T}}\right)^2 + \left(w_{jk}^{\mathcal{T}}\right)^2\right) - 75\gamma_k^2\left(w_{ik}^{\mathcal{T}} + w_{jk}^{\mathcal{T}}\right) \\ &- 50 w_{ik}^{\mathcal{T}} w_{jk}^{\mathcal{T}}\left(w_{ik}^{\mathcal{T}} + w_{jk}^{\mathcal{T}}\right) + 60\sqrt{5}\gamma_k w_{ik}^{\mathcal{T}} w_{jk}^{\mathcal{T}}\Big] \Big/ 9\gamma_k^4 \\ E_{32} =& 5\Big[5\left(w_{ik}^{\mathcal{T}}\right)^2 + 5\left(w_{jk}^{\mathcal{T}}\right)^2 + 15\gamma_k^2 - 9\sqrt{5}\gamma_k\left(w_{ik}^{\mathcal{T}} + w_{jk}^{\mathcal{T}}\right) + 20 w_{ik}^{\mathcal{T}} w_{jk}^{\mathcal{T}}\Big] \Big/ 9\gamma_k^4 \\ E_{33} =& \frac{10\left(3\sqrt{5}\gamma_k - 5w_{ik}^{\mathcal{T}} - 5w_{jk}^{\mathcal{T}}\right)}{9\gamma_k^4} \quad \text{and} \quad E_{34} = \frac{25}{9\gamma_k^4}.\end{aligned}$$

Then by completing the square, we have

$$\begin{aligned}(\text{C.27}) = &\exp\left\{\frac{10\sigma_k^2 + \sqrt{5}\gamma_k\left(w_{ik}^{\mathcal{T}} + w_{jk}^{\mathcal{T}} - 2\mu_k\right)}{\gamma_k^2}\right\} \\ \times &\int_{w_{jk}^{\mathcal{T}}}^{+\infty} (E_{34}w^4 + E_{33}w^3 + E_{32}w^2 + E_{31}w + E_{30}) \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left\{-\frac{(w-\mu_C)^2}{2\sigma_k^2}\right\} \mathrm{d}w,\end{aligned}$$

where

$$\mu_C = \mu_k - 2\sqrt{5}\frac{\sigma_k^2}{\gamma_k}.$$

Using Lemma C.1 and arranging terms, we obtain

$$\begin{aligned}(\text{C.27}) = &\exp\left\{\frac{10\sigma_k^2 + \sqrt{5}\gamma_k\left(w_{ik}^{\mathcal{T}} + w_{jk}^{\mathcal{T}} - 2\mu_k\right)}{\gamma_k^2}\right\} \\ &\times \left[\mathbf{E}_3^\top \mathbf{\Lambda}_{31} \Phi\left(\frac{\mu_C - w_{jk}^{\mathcal{T}}}{\sigma_k}\right) + \mathbf{E}_3^\top \mathbf{\Lambda}_{32} \frac{\sigma_k}{\sqrt{2\pi}} \exp\left\{-\frac{(w_{jk}^{\mathcal{T}} - \mu_C)^2}{2\sigma_k^2}\right\}\right],\end{aligned}$$

where

- $\mathbf{E}_3 = [E_{30},\, E_{31},\, E_{32},\, E_{33},\, E_{34}]^\top$;
- $\mathbf{\Lambda}_{31} = [1,\, \mu_C,\, \mu_C^2 + \sigma_k^2,\, \mu_C^3 + 3\sigma_k^2\mu_C,\, \mu_C^4 + 6\sigma_k^2\mu_C^2 + 3\sigma_k^4]^\top$;
- $\mathbf{\Lambda}_{32} = [0,\, 1,\, \mu_C + w_{jk}^{\mathcal{T}},\, \mu_C^2 + 2\sigma_k^2 + \left(w_{jk}^{\mathcal{T}}\right)^2 + \mu_C w_{jk}^{\mathcal{T}},\, \mu_C^3 + \left(w_{jk}^{\mathcal{T}}\right)^3 + w_{jk}^{\mathcal{T}}\mu_C^2 + \mu_C\left(w_{jk}^{\mathcal{T}}\right)^2 + 3\sigma_k^2 w_{jk}^{\mathcal{T}} + 5\sigma_k^2\mu_C]^\top$.

The derivation of term (C.28) is analogue to that of term (C.27). By expanding the product of two brackets after the integral sign, we have

$$\begin{aligned}(\text{C.28}) = &\int_{w_{ik}^{\mathcal{T}}}^{w_{jk}^{\mathcal{T}}} (E_{44}w^4 + E_{43}w^3 + E_{42}w^2 + E_{41}w + E_{40}) \\ &\times \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left\{-\frac{\sqrt{5}(w - w_{ik}^{\mathcal{T}}) + \sqrt{5}(w_{jk}^{\mathcal{T}} - w)}{\gamma_k} - \frac{(w-\mu_k)^2}{2\sigma_k^2}\right\} \mathrm{d}w,\end{aligned}$$

where

$$\begin{aligned}E_{40} =& 1 + \Big[25\left(w_{ik}^{\mathcal{T}}\right)^2\left(w_{jk}^{\mathcal{T}}\right)^2 + 3\sqrt{5}\left(3\gamma_k^3 - 5\gamma_k w_{ik}^{\mathcal{T}} w_{jk}^{\mathcal{T}}\right)\left(w_{jk}^{\mathcal{T}} - w_{ik}^{\mathcal{T}}\right) \\ &+ 15\gamma_k^2\left(\left(w_{ik}^{\mathcal{T}}\right)^2 + \left(w_{jk}^{\mathcal{T}}\right)^2 - 3w_{ik}^{\mathcal{T}} w_{jk}^{\mathcal{T}}\right)\Big] \Big/ 9\gamma_k^4 \\ E_{41} =& 5\left[3\sqrt{5}\gamma_k\left(\left(w_{jk}^{\mathcal{T}}\right)^2 - \left(w_{ik}^{\mathcal{T}}\right)^2\right) + 3\gamma_k^2\left(w_{ik}^{\mathcal{T}} + w_{jk}^{\mathcal{T}}\right) - 10 w_{ik}^{\mathcal{T}} w_{jk}^{\mathcal{T}}\left(w_{ik}^{\mathcal{T}} + w_{jk}^{\mathcal{T}}\right)\right] \Big/ 9\gamma_k^4 \\ E_{42} =& 5\left[5\left(w_{ik}^{\mathcal{T}}\right)^2 + 5\left(w_{jk}^{\mathcal{T}}\right)^2 - 3\gamma_k^2 - 3\sqrt{5}\gamma_k\left(w_{jk}^{\mathcal{T}} - w_{ik}^{\mathcal{T}}\right) + 20 w_{ik}^{\mathcal{T}} w_{jk}^{\mathcal{T}}\right] \Big/ 9\gamma_k^4 \\ E_{43} =& -\frac{50\left(w_{ik}^{\mathcal{T}} + w_{jk}^{\mathcal{T}}\right)}{9\gamma_k^4} \quad \text{and} \quad E_{44} = \frac{25}{9\gamma_k^4}.\end{aligned}$$

Then by completing the square, we have

$$\begin{aligned}(\text{C.28}) =& \exp\left\{-\frac{\sqrt{5}\left(w_{jk}^{\mathcal{T}} - w_{ik}^{\mathcal{T}}\right)}{\gamma_k}\right\} \\ \times &\int_{w_{ik}^{\mathcal{T}}}^{w_{jk}^{\mathcal{T}}} (E_{44}w^4 + E_{43}w^3 + E_{42}w^2 + E_{41}w + E_{40}) \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left\{-\frac{(w-\mu_k)^2}{2\sigma_k^2}\right\} \mathrm{d}w.\end{aligned}$$

Using Lemma C.1 and arranging terms, we obtain

$$\begin{aligned}(\text{C.28}) =& \exp\left\{-\frac{\sqrt{5}\left(w_{jk}^{\mathcal{T}} - w_{ik}^{\mathcal{T}}\right)}{\gamma_k}\right\}\left[\mathbf{E}_4^\top \mathbf{\Lambda}_{41}\left[\Phi\left(\frac{w_{jk}^{\mathcal{T}} - \mu_k}{\sigma_k}\right) - \Phi\left(\frac{w_{ik}^{\mathcal{T}} - \mu_k}{\sigma_k}\right)\right]\right. \\ &\left. + \mathbf{E}_4^\top \mathbf{\Lambda}_{42}\frac{\sigma_k}{\sqrt{2\pi}}\exp\left\{-\frac{(w_{ik}^{\mathcal{T}} - \mu_k)^2}{2\sigma_k^2}\right\} - \mathbf{E}_4^\top \mathbf{\Lambda}_{43}\frac{\sigma_k}{\sqrt{2\pi}}\exp\left\{-\frac{(w_{jk}^{\mathcal{T}} - \mu_k)^2}{2\sigma_k^2}\right\}\right],\end{aligned}$$

where

- $\mathbf{E}_4 = [E_{40},\, E_{41},\, E_{42},\, E_{43},\, E_{44}]^\top$;
- $\mathbf{\Lambda}_{41} = [1,\, \mu_k,\, \mu_k^2 + \sigma_k^2,\, \mu_k^3 + 3\sigma_k^2\mu_k,\, \mu_k^4 + 6\sigma_k^2\mu_k^2 + 3\sigma_k^{4}]^\top$;
- $\mathbf{\Lambda}_{42} = [0,\, 1,\, \mu_k + w_{ik}^{\mathcal{T}},\, \mu_k^2 + 2\sigma_k^2 + \left(w_{ik}^{\mathcal{T}}\right)^2 + \mu_k w_{ik}^{\mathcal{T}},\, \mu_k^3 + \left(w_{ik}^{\mathcal{T}}\right)^3 + w_{ik}^{\mathcal{T}}\mu_k^2 + \mu_k \left(w_{ik}^{\mathcal{T}}\right)^2 + 3\sigma_k^2 w_{ik}^{\mathcal{T}} + 5\sigma_k^2\mu_k]^\top$;
- $\mathbf{\Lambda}_{43} = [0,\, 1,\, \mu_k + w_{jk}^{\mathcal{T}},\, \mu_k^2 + 2\sigma_k^2 + \left(w_{jk}^{\mathcal{T}}\right)^2 + \mu_k w_{jk}^{\mathcal{T}},\, \mu_k^3 + \left(w_{jk}^{\mathcal{T}}\right)^3 + w_{jk}^{\mathcal{T}}\mu_k^2 + \mu_k \left(w_{jk}^{\mathcal{T}}\right)^2 + 3\sigma_k^2 w_{jk}^{\mathcal{T}} + 5\sigma_k^2\mu_k]^\top$.

Term (C.29) can be computed in the following way:

$$
\begin{aligned}
(\text{C.29}) = \int_{-\infty}^{w_{ik}^{\mathcal{T}}} & \left(1 + \frac{\sqrt{5}(w_{ik}^{\mathcal{T}} - w)}{\gamma_k} + \frac{5}{3}\left(\frac{w - w_{ik}^{\mathcal{T}}}{\gamma_k}\right)^2\right) \\
& \times \left(1 + \frac{\sqrt{5}(w_{jk}^{\mathcal{T}} - w)}{\gamma_k} + \frac{5}{3}\left(\frac{w - w_{jk}^{\mathcal{T}}}{\gamma_k}\right)^2\right) \\
& \times \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left\{-\frac{\sqrt{5}(w_{ik}^{\mathcal{T}} - w) + \sqrt{5}(w_{jk}^{\mathcal{T}} - w)}{\gamma_k} - \frac{(w - \mu_k)^2}{2\sigma_k^2}\right\} \mathrm{d}w \\
= \int_{-w_{ik}^{\mathcal{T}}}^{+\infty} & \left(1 + \frac{\sqrt{5}(w + w_{ik}^{\mathcal{T}})}{\gamma_k} + \frac{5}{3}\left(\frac{w + w_{ik}^{\mathcal{T}}}{\gamma_k}\right)^2\right) \\
& \times \left(1 + \frac{\sqrt{5}(w + w_{jk}^{\mathcal{T}})}{\gamma_k} + \frac{5}{3}\left(\frac{w + w_{jk}^{\mathcal{T}}}{\gamma_k}\right)^2\right) \\
& \times \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left\{-\frac{\sqrt{5}(w + w_{ik}^{\mathcal{T}}) + \sqrt{5}(w + w_{jk}^{\mathcal{T}})}{\gamma_k} - \frac{(w + \mu_k)^2}{2\sigma_k^2}\right\} \mathrm{d}w,
\end{aligned}
$$

the form of which allows us to obtain solution of term (C.29) by simply using that of term (C.27). Thus, we have

$$
\begin{aligned}
(\text{C.29}) = & \exp\left\{\frac{10\sigma_k^2 - \sqrt{5}\gamma_k\left(w_{ik}^{\mathcal{T}} + w_{jk}^{\mathcal{T}} - 2\mu_k\right)}{\gamma_k^2}\right\} \\
& \times \left[\mathbf{E}_5^\top \mathbf{\Lambda}_{51} \Phi\left(\frac{w_{ik}^{\mathcal{T}} - \mu_D}{\sigma_k}\right) + \mathbf{E}_5^\top \mathbf{\Lambda}_{52} \frac{\sigma_k}{\sqrt{2\pi}} \exp\left\{-\frac{(w_{ik}^{\mathcal{T}} - \mu_D)^2}{2\sigma_k^2}\right\}\right],
\end{aligned}
$$

where

- $\mathbf{E}_5 = [E_{50},\, E_{51},\, E_{52},\, E_{53},\, E_{54}]^\top$;

- $\mathbf{\Lambda}_{51} = [1,\, -\mu_D,\, \mu_D^2 + \sigma_k^2,\, -\mu_D^3 - 3\sigma_k^2\mu_D,\, \mu_D^4 + 6\sigma_k^2\mu_D^2 + 3\sigma_k^4]^\top$;
- $\mathbf{\Lambda}_{52} = [0,\, 1,\, -\mu_D - w_{ik}^{\mathcal{T}},\, \mu_D^2 + 2\sigma_k^2 + \left(w_{ik}^{\mathcal{T}}\right)^2 + \mu_D w_{ik}^{\mathcal{T}},\, -\mu_D^3 - \left(w_{ik}^{\mathcal{T}}\right)^3 - w_{ik}^{\mathcal{T}}\mu_D^2 - \mu_D\left(w_{ik}^{\mathcal{T}}\right)^2 - 3\sigma_k^2 w_{ik}^{\mathcal{T}} - 5\sigma_k^2\mu_D]^\top$

with

$$
\begin{aligned}
E_{50} =& 1 + \Big[25\left(w_{ik}^{\mathcal{T}}\right)^2\left(w_{jk}^{\mathcal{T}}\right)^2 + 3\sqrt{5}\left(3\gamma_k^3 + 5\gamma_k w_{ik}^{\mathcal{T}} w_{jk}^{\mathcal{T}}\right)\left(w_{ik}^{\mathcal{T}} + w_{jk}^{\mathcal{T}}\right) \\
&+ 15\gamma_k^2\left(\left(w_{ik}^{\mathcal{T}}\right)^2 + \left(w_{jk}^{\mathcal{T}}\right)^2 + 3w_{ik}^{\mathcal{T}} w_{jk}^{\mathcal{T}}\right)\Big] \Big/ 9\gamma_k^4 \\
E_{51} =& \Big[18\sqrt{5}\gamma_k^3 + 15\sqrt{5}\gamma_k\left(\left(w_{ik}^{\mathcal{T}}\right)^2 + \left(w_{jk}^{\mathcal{T}}\right)^2\right) + 75\gamma_k^2\left(w_{ik}^{\mathcal{T}} + w_{jk}^{\mathcal{T}}\right) \\
&+ 50 w_{ik}^{\mathcal{T}} w_{jk}^{\mathcal{T}}\left(w_{ik}^{\mathcal{T}} + w_{jk}^{\mathcal{T}}\right) + 60\sqrt{5}\gamma_k w_{ik}^{\mathcal{T}} w_{jk}^{\mathcal{T}}\Big] \Big/ 9\gamma_k^4 \\
E_{52} =& 5\Big[5\left(w_{ik}^{\mathcal{T}}\right)^2 + 5\left(w_{jk}^{\mathcal{T}}\right)^2 + 15\gamma_k^2 + 9\sqrt{5}\gamma_k\left(w_{ik}^{\mathcal{T}} + w_{jk}^{\mathcal{T}}\right) + 20 w_{ik}^{\mathcal{T}} w_{jk}^{\mathcal{T}}\Big] \Big/ 9\gamma_k^4 \\
E_{53} =& \frac{10\left(3\sqrt{5}\gamma_k + 5w_{ik}^{\mathcal{T}} + 5w_{jk}^{\mathcal{T}}\right)}{9\gamma_k^4}, \quad E_{54} = \frac{25}{9\gamma_k^4} \quad \text{and} \quad \mu_D = \mu_k + 2\sqrt{5}\frac{\sigma_k^2}{\gamma_k}.
\end{aligned}
$$

Therefore, the expression for ζ_{ijk} when $w_{ik}^{\mathcal{T}} \leq w_{jk}^{\mathcal{T}}$ is given by

$$
\begin{aligned}
\zeta_{ijk} =& \exp\left\{\frac{10\sigma_k^2 + \sqrt{5}\gamma_k\left(w_{ik}^{\mathcal{T}} + w_{jk}^{\mathcal{T}} - 2\mu_k\right)}{\gamma_k^2}\right\} \\
&\times \left[\mathbf{E}_3^\top \mathbf{\Lambda}_{31}\Phi\left(\frac{\mu_C - w_{jk}^{\mathcal{T}}}{\sigma_k}\right) + \mathbf{E}_3^\top \mathbf{\Lambda}_{32}\frac{\sigma_k}{\sqrt{2\pi}}\exp\left\{-\frac{(w_{jk}^{\mathcal{T}} - \mu_C)^2}{2\sigma_k^2}\right\}\right] \\
&+ \exp\left\{-\frac{\sqrt{5}\left(w_{jk}^{\mathcal{T}} - w_{ik}^{\mathcal{T}}\right)}{\gamma_k}\right\} \\
&\times \left[\mathbf{E}_4^\top \mathbf{\Lambda}_{41}\left(\Phi\left(\frac{w_{jk}^{\mathcal{T}} - \mu_k}{\sigma_k}\right) - \Phi\left(\frac{w_{ik}^{\mathcal{T}} - \mu_k}{\sigma_k}\right)\right)\right. \\
&\left. + \mathbf{E}_4^\top \mathbf{\Lambda}_{42}\frac{\sigma_k}{\sqrt{2\pi}}\exp\left\{-\frac{(w_{ik}^{\mathcal{T}} - \mu_k)^2}{2\sigma_k^2}\right\} - \mathbf{E}_4^\top \mathbf{\Lambda}_{43}\frac{\sigma_k}{\sqrt{2\pi}}\exp\left\{-\frac{(w_{jk}^{\mathcal{T}} - \mu_k)^2}{2\sigma_k^2}\right\}\right] \\
&+ \exp\left\{\frac{10\sigma_k^2 - \sqrt{5}\gamma_k\left(w_{ik}^{\mathcal{T}} + w_{jk}^{\mathcal{T}} - 2\mu_k\right)}{\gamma_k^2}\right\} \\
&\times \left[\mathbf{E}_5^\top \mathbf{\Lambda}_{51}\Phi\left(\frac{w_{ik}^{\mathcal{T}} - \mu_D}{\sigma_k}\right) + \mathbf{E}_5^\top \mathbf{\Lambda}_{52}\frac{\sigma_k}{\sqrt{2\pi}}\exp\left\{-\frac{(w_{ik}^{\mathcal{T}} - \mu_D)^2}{2\sigma_k^2}\right\}\right],
\end{aligned}
$$

and interchanging positions of $w_{ik}^{\mathcal{T}}$ and $w_{jk}^{\mathcal{T}}$ gives the expression for ζ_{ijk} when $w_{ik}^{\mathcal{T}} > w_{jk}^{\mathcal{T}}$.

3 Derivation of ψ_{jk}

$$\begin{aligned}
\psi_{jk} &= \int w \left(1 + \frac{\sqrt{5}|w - w_{jk}^{\mathcal{T}}|}{\gamma_k} + \frac{5}{3}\left(\frac{w - w_{jk}^{\mathcal{T}}}{\gamma_k}\right)^2\right) \\
&\quad \times \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left\{-\frac{\sqrt{5}|w - w_{jk}^{\mathcal{T}}|}{\gamma_k} - \frac{(w-\mu_k)^2}{2\sigma_k^2}\right\} \mathrm{d}w \\
&= \int_{w_{jk}^{\mathcal{T}}}^{+\infty} \left(w + \frac{\sqrt{5}w\left(w - w_{jk}^{\mathcal{T}}\right)}{\gamma_k} + \frac{5w}{3}\left(\frac{w - w_{jk}^{\mathcal{T}}}{\gamma_k}\right)^2\right) \\
&\quad \times \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left\{-\frac{\sqrt{5}\left(w - w_{jk}^{\mathcal{T}}\right)}{\gamma_k} - \frac{(w-\mu_k)^2}{2\sigma_k^2}\right\} \mathrm{d}w \qquad \text{(C.30)} \\
&\quad + \int_{-\infty}^{w_{jk}^{\mathcal{T}}} \left(w + \frac{\sqrt{5}w\left(w_{jk}^{\mathcal{T}} - w\right)}{\gamma_k} + \frac{5w}{3}\left(\frac{w - w_{jk}^{\mathcal{T}}}{\gamma_k}\right)^2\right) \\
&\quad \times \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left\{\frac{\sqrt{5}\left(w - w_{jk}^{\mathcal{T}}\right)}{\gamma_k} - \frac{(w-\mu_k)^2}{2\sigma_k^2}\right\} \mathrm{d}w. \qquad \text{(C.31)}
\end{aligned}$$

We first calculate term (C.30) by arranging the terms in the bracket after the integral sign and completing the square:

$$\begin{aligned}
\text{(C.30)} &= \exp\left\{\frac{5\sigma_k^2 + 2\sqrt{5}\gamma_k\left(w_{jk}^{\mathcal{T}} - \mu_k\right)}{2\gamma_k^2}\right\} \\
&\quad \times \int_{w_{jk}^{\mathcal{T}}}^{+\infty} \left[E_{12}w^3 + E_{11}w^2 + E_{10}w\right] \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left\{-\frac{(w-\mu_A)^2}{2\sigma_k^2}\right\}.
\end{aligned}$$

By Lemma C.1, we then obtain

$$\begin{aligned}
\text{(C.30)} &= \exp\left\{\frac{5\sigma_k^2 + 2\sqrt{5}\gamma_k\left(w_{jk}^{\mathcal{T}} - \mu_k\right)}{2\gamma_k^2}\right\} \\
&\quad \times \left[\mathbf{E}_1^\top \mathbf{\Lambda}_{61} \Phi\left(\frac{\mu_A - w_{jk}^{\mathcal{T}}}{\sigma_k}\right) + \mathbf{E}_1^\top \mathbf{\Lambda}_{62} \frac{\sigma_k}{\sqrt{2\pi}} \exp\left\{-\frac{(w_{jk}^{\mathcal{T}} - \mu_A)^2}{2\sigma_k^2}\right\}\right],
\end{aligned}$$

where

- $\mathbf{\Lambda}_{61} = \left[\mu_A,\, \mu_A^2 + \sigma_k^2,\, \mu_A^3 + 3\sigma_k^2\mu_A\right]^\top$;
- $\mathbf{\Lambda}_{62} = \left[1,\, \mu_A + w_{jk}^{\mathcal{T}},\, \mu_A^2 + 2\sigma_k^2 + \left(w_{jk}^{\mathcal{T}}\right)^2 + \mu_A w_{jk}^{\mathcal{T}}\right]^\top$.

Term (C.31) can be rewritten as follow:

$$\begin{aligned}(\text{C.31}) = &\int_{-\infty}^{w_{jk}^{\mathcal{T}}} \left(1 + \frac{\sqrt{5}\left(w_{jk}^{\mathcal{T}} - w\right)}{\gamma_k} + \frac{5}{3}\left(\frac{w - w_{jk}^{\mathcal{T}}}{\gamma_k}\right)^2\right) \\ &\times \frac{w}{\sigma_k\sqrt{2\pi}} \exp\left\{\frac{\sqrt{5}\left(w - w_{jk}^{\mathcal{T}}\right)}{\gamma_k} - \frac{(w-\mu_k)^2}{2\sigma_k^2}\right\} \mathrm{d}w \\ = &-\int_{-w_{jk}^{\mathcal{T}}}^{+\infty} \left(1 + \frac{\sqrt{5}\left(w + w_{jk}^{\mathcal{T}}\right)}{\gamma_k} + \frac{5}{3}\left(\frac{w + w_{jk}^{\mathcal{T}}}{\gamma_k}\right)^2\right) \\ &\times \frac{w}{\sigma_k\sqrt{2\pi}} \exp\left\{-\frac{\sqrt{5}\left(w + w_{jk}^{\mathcal{T}}\right)}{\gamma_k} - \frac{(w+\mu_k)^2}{2\sigma_k^2}\right\} \mathrm{d}w,\end{aligned}$$

the form of which allows us to obtain solution of term (C.31) by using that of term (C.30). Thus, we have

$$\begin{aligned}(\text{C.31}) = &-\exp\left\{\frac{5\sigma_k^2 - 2\sqrt{5}\gamma_k\left(w_{jk}^{\mathcal{T}} - \mu_k\right)}{2\gamma_k^2}\right\} \\ &\times \left[\mathbf{E}_2^\top \mathbf{\Lambda}_{71} \Phi\left(\frac{w_{jk}^{\mathcal{T}} - \mu_B}{\sigma_k}\right) + \mathbf{E}_2^\top \mathbf{\Lambda}_{72} \frac{\sigma_k}{\sqrt{2\pi}} \exp\left\{-\frac{(w_{jk}^{\mathcal{T}} - \mu_B)^2}{2\sigma_k^2}\right\}\right],\end{aligned}$$

where

- $\mathbf{\Lambda}_{71} = [-\mu_B,\, \mu_B^2 + \sigma_k^2,\, -\mu_B^3 - 3\sigma_k^2\mu_B]^\top$;
- $\mathbf{\Lambda}_{72} = \left[1,\, -\mu_B - w_{jk}^{\mathcal{T}},\, \mu_B^2 + 2\sigma_k^2 + \left(w_{jk}^{\mathcal{T}}\right)^2 + \mu_B w_{jk}^{\mathcal{T}}\right]^\top$.

Thus, we have

$$\begin{aligned}\psi_{jk} = &\exp\left\{\frac{5\sigma_k^2 + 2\sqrt{5}\gamma_k\left(w_{jk}^{\mathcal{T}} - \mu_k\right)}{2\gamma_k^2}\right\} \\ &\times \left[\mathbf{E}_1^\top \mathbf{\Lambda}_{61} \Phi\left(\frac{\mu_A - w_{jk}^{\mathcal{T}}}{\sigma_k}\right) + \mathbf{E}_1^\top \mathbf{\Lambda}_{62} \frac{\sigma_k}{\sqrt{2\pi}} \exp\left\{-\frac{(w_{jk}^{\mathcal{T}} - \mu_A)^2}{2\sigma_k^2}\right\}\right] \\ &- \exp\left\{\frac{5\sigma_k^2 - 2\sqrt{5}\gamma_k\left(w_{jk}^{\mathcal{T}} - \mu_k\right)}{2\gamma_k^2}\right\} \\ &\times \left[\mathbf{E}_2^\top \mathbf{\Lambda}_{71} \Phi\left(\frac{w_{jk}^{\mathcal{T}} - \mu_B}{\sigma_k}\right) + \mathbf{E}_2^\top \mathbf{\Lambda}_{72} \frac{\sigma_k}{\sqrt{2\pi}} \exp\left\{-\frac{(w_{jk}^{\mathcal{T}} - \mu_B)^2}{2\sigma_k^2}\right\}\right].\end{aligned}$$

C.4 Proof of Proposition 3.4

C.4.1 Derivation of $\widetilde{\xi}_i$

$$\begin{aligned}
\widetilde{\xi}_i &= \mathbb{E}\left[c(\mathbf{W}, \mathbf{w}_i^{\mathcal{T}})\right] \\
&= \int \exp\left\{-\sum_{k=1}^{d} \frac{\left(w_k - w_{ik}^{\mathcal{T}}\right)^2}{\gamma_k^2}\right\} \\
&\quad \times \frac{1}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\mathbf{w}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{w}-\boldsymbol{\mu})\right\} \mathrm{d}\mathbf{w} \\
&= \int \exp\left\{-\frac{1}{2}(\mathbf{w}-\boldsymbol{\omega}_i^{\mathcal{T}})^\top \boldsymbol{\Lambda}^{-1}(\mathbf{w}-\boldsymbol{\omega}_i^{\mathcal{T}})\right\} \\
&\quad \times \frac{1}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\mathbf{w}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{w}-\boldsymbol{\mu})\right\} \mathrm{d}\mathbf{w},
\end{aligned}$$

where $\boldsymbol{\Lambda} = \mathrm{diag}(\frac{\gamma_1^2}{2}, \ldots, \frac{\gamma_d^2}{2}) \in \mathbb{R}^{d\times d}$ is a diagonal matrix.

By completing in squares, we then have

$$\begin{aligned}
\widetilde{\xi}_i = &\frac{1}{\sqrt{(2\pi)^d|\mathbf{M}^{-1}|}} \frac{1}{\sqrt{|\boldsymbol{\Sigma}\mathbf{M}|}} \\
&\times \int \exp\left\{-\frac{1}{2}(\mathbf{w}-\mathbf{M}^{-1}\mathbf{V})^\top \mathbf{M}(\mathbf{w}-\mathbf{M}^{-1}\mathbf{V}) + \frac{1}{2}(\mathbf{V}^\top\mathbf{M}^{-1}\mathbf{V} - R)\right\} \mathrm{d}\mathbf{w},
\end{aligned}$$

where $\mathbf{M} = \boldsymbol{\Sigma}^{-1}+\boldsymbol{\Lambda}^{-1}$, $\mathbf{V} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}+\boldsymbol{\Lambda}^{-1}\boldsymbol{\omega}_i^{\mathcal{T}}$ and $R = \boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}+(\boldsymbol{\omega}_i^{\mathcal{T}})^\top\boldsymbol{\Lambda}^{-1}\boldsymbol{\omega}_i^{\mathcal{T}}$.

By integrating out the probability density function of a multivariate normal distribution with mean $\mathbf{M}^{-1}\mathbf{V}$ and covariance matrix $\mathbf{M}^{-1}$, we have

$$\widetilde{\xi}_i = \frac{1}{\sqrt{|\boldsymbol{\Sigma}\mathbf{M}|}} \exp\left\{\frac{1}{2}(\mathbf{V}^\top\mathbf{M}^{-1}\mathbf{V} - R)\right\}$$

Using the Woodbury identity (Petersen and Pedersen, 2012), we have

$$\begin{aligned}
\mathbf{M}^{-1} &= \boldsymbol{\Sigma} - \boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\boldsymbol{\Lambda})^{-1}\boldsymbol{\Sigma} \\
\mathbf{M}^{-1} &= \boldsymbol{\Lambda} - \boldsymbol{\Lambda}(\boldsymbol{\Sigma}+\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}.
\end{aligned}$$

Thus, we obtain

$$\widetilde{\xi}_i = \frac{1}{\sqrt{|(\boldsymbol{\Lambda}+\boldsymbol{\Sigma})\boldsymbol{\Lambda}^{-1}|}} \exp\left\{-\frac{1}{2}(\boldsymbol{\omega}_i^{\mathcal{T}}-\boldsymbol{\mu})^\top(\boldsymbol{\Lambda}+\boldsymbol{\Sigma})^{-1}(\boldsymbol{\omega}_i^{\mathcal{T}}-\boldsymbol{\mu})\right\},$$

C.4.2 Derivation of $\widetilde{\zeta}_{ij}$

$$
\begin{aligned}
\widetilde{\zeta}_{ij} &= \mathbb{E}\left[c(\mathbf{W}, \mathbf{w}_i^{\mathcal{T}})c(\mathbf{W}, \mathbf{w}_j^{\mathcal{T}})\right] \\
&= \int \exp\left\{-\sum_{k=1}^{d} \frac{\left(w_k - w_{ik}^{\mathcal{T}}\right)^2}{\gamma_k^2} - \sum_{k=1}^{d} \frac{\left(w_k - w_{jk}^{\mathcal{T}}\right)^2}{\gamma_k^2}\right\} \\
&\quad \times \frac{1}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\mathbf{w} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{w} - \boldsymbol{\mu})\right\} \mathrm{d}\mathbf{w} \\
&= \int \exp\left\{-\sum_{k=1}^{d} \frac{2(w_k - w_{ik}^{\mathcal{T}})(w_k - w_{jk}^{\mathcal{T}})}{\gamma_k^2} - \sum_{k=1}^{d} \frac{\left(w_{ik}^{\mathcal{T}} - w_{jk}^{\mathcal{T}}\right)^2}{\gamma_k^2}\right\} \\
&\quad \times \frac{1}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\mathbf{w} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{w} - \boldsymbol{\mu})\right\} \mathrm{d}\mathbf{w} \\
&= \int \exp\left\{-\frac{1}{2}(\mathbf{w} - \boldsymbol{\omega}_i^{\mathcal{T}})^\top \boldsymbol{\Gamma}^{-1}(\mathbf{w} - \boldsymbol{\omega}_j^{\mathcal{T}}) - \frac{1}{4}(\boldsymbol{\omega}_i^{\mathcal{T}} - \boldsymbol{\omega}_j^{\mathcal{T}})^\top \boldsymbol{\Gamma}^{-1}(\boldsymbol{\omega}_i^{\mathcal{T}} - \boldsymbol{\omega}_j^{\mathcal{T}})\right\} \\
&\quad \times \frac{1}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\mathbf{w} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{w} - \boldsymbol{\mu})\right\} \mathrm{d}\mathbf{w} \\
&= \exp\left\{-\frac{1}{4}(\boldsymbol{\omega}_i^{\mathcal{T}} - \boldsymbol{\omega}_j^{\mathcal{T}})^\top \boldsymbol{\Gamma}^{-1}(\boldsymbol{\omega}_i^{\mathcal{T}} - \boldsymbol{\omega}_j^{\mathcal{T}})\right\} \frac{1}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}|}} \\
&\quad \times \int \exp\left\{-\frac{1}{2}\left[(\mathbf{w} - \boldsymbol{\omega}_i^{\mathcal{T}})^\top \boldsymbol{\Gamma}^{-1}(\mathbf{w} - \boldsymbol{\omega}_j^{\mathcal{T}}) + (\mathbf{w} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{w} - \boldsymbol{\mu})\right]\right\} \mathrm{d}\mathbf{w},
\end{aligned}
$$

where $\boldsymbol{\Gamma} = \mathrm{diag}(\frac{\gamma_1^2}{4}, \ldots, \frac{\gamma_d^2}{4}) \in \mathbb{R}^{d\times d}$ is a diagonal matrix. By completing in squares, we then have

$$
\begin{aligned}
\widetilde{\zeta}_{ij} &= \exp\left\{-\frac{1}{4}(\boldsymbol{\omega}_i^{\mathcal{T}} - \boldsymbol{\omega}_j^{\mathcal{T}})^\top \boldsymbol{\Gamma}^{-1}(\boldsymbol{\omega}_i^{\mathcal{T}} - \boldsymbol{\omega}_j^{\mathcal{T}})\right\} \frac{1}{\sqrt{(2\pi)^d|\mathbf{M}^{-1}|}} \frac{1}{\sqrt{|\boldsymbol{\Sigma}\mathbf{M}|}} \\
&\quad \times \int \exp\left\{-\frac{1}{2}(\mathbf{w} - \mathbf{M}^{-1}\mathbf{V})^\top \mathbf{M}(\mathbf{w} - \mathbf{M}^{-1}\mathbf{V}) + \frac{1}{2}(\mathbf{V}^\top \mathbf{M}^{-1}\mathbf{V} - R)\right\} \mathrm{d}\mathbf{w},
\end{aligned}
$$

where $\mathbf{M} = \boldsymbol{\Sigma}^{-1} + \boldsymbol{\Gamma}^{-1}$; $\mathbf{V} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\Gamma}^{-1}\boldsymbol{\omega}$ with $\boldsymbol{\omega} = \frac{1}{2}(\boldsymbol{\omega}_i^{\mathcal{T}} + \boldsymbol{\omega}_j^{\mathcal{T}})$; and $R = \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + (\boldsymbol{\omega}_i^{\mathcal{T}})^\top \boldsymbol{\Gamma}^{-1}\boldsymbol{\omega}_j^{\mathcal{T}}$.

By integrating out the probability density function of a multivariate normal distribution with mean $\mathbf{M}^{-1}\mathbf{V}$ and covariance matrix $\mathbf{M}^{-1}$, we have

$$
\begin{aligned}
\widetilde{\zeta}_{ij} = \exp\left\{-\frac{1}{4}(\boldsymbol{\omega}_i^{\mathcal{T}} - \boldsymbol{\omega}_j^{\mathcal{T}})^\top \boldsymbol{\Gamma}^{-1}(\boldsymbol{\omega}_i^{\mathcal{T}} - \boldsymbol{\omega}_j^{\mathcal{T}})\right\} \\
\times \frac{1}{\sqrt{|\boldsymbol{\Sigma}\mathbf{M}|}} \exp\left\{\frac{1}{2}(\mathbf{V}^\top \mathbf{M}^{-1}\mathbf{V} - R)\right\}.
\end{aligned}
$$

Using the Woodbury identity (Petersen and Pedersen, 2012), we have

$$\mathbf{M}^{-1} = \boldsymbol{\Sigma} - \boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \boldsymbol{\Gamma})^{-1}\boldsymbol{\Sigma}$$
$$\mathbf{M}^{-1} = \boldsymbol{\Gamma} - \boldsymbol{\Gamma}(\boldsymbol{\Sigma} + \boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}.$$

Thus, we obtain

$$\widetilde{\zeta}_{ij} = \exp\left\{-\frac{1}{8}(\boldsymbol{\omega}_i^{\mathcal{T}} - \boldsymbol{\omega}_j^{\mathcal{T}})^\top \boldsymbol{\Gamma}^{-1}(\boldsymbol{\omega}_i^{\mathcal{T}} - \boldsymbol{\omega}_j^{\mathcal{T}})\right\} \times \frac{1}{\sqrt{|(\boldsymbol{\Gamma} + \boldsymbol{\Sigma})\boldsymbol{\Gamma}^{-1}|}} \exp\left\{-\frac{1}{2}(\boldsymbol{\omega} - \boldsymbol{\mu})^\top (\boldsymbol{\Gamma} + \boldsymbol{\Sigma})^{-1}(\boldsymbol{\omega} - \boldsymbol{\mu})\right\}.$$

C.4.3 Derivation of $\widetilde{\psi}_{jl}$

$$\begin{aligned}
\widetilde{\psi}_{jl} &= \mathbb{E}\left[W_l c(\mathbf{W}, \mathbf{w}_j^{\mathcal{T}})\right] \\
&= \int w_l \exp\left\{-\sum_{k=1}^{d} \frac{\left(w_k - w_{jk}^{\mathcal{T}}\right)^2}{\gamma_k^2}\right\} \\
&\quad \times \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\mathbf{w} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{w} - \boldsymbol{\mu})\right\} \mathrm{d}\mathbf{w} \\
&= \int w_l \exp\left\{-\frac{1}{2}(\mathbf{w} - \boldsymbol{\omega}_i^{\mathcal{T}})^\top \boldsymbol{\Lambda}^{-1}(\mathbf{w} - \boldsymbol{\omega}_i^{\mathcal{T}})\right\} \\
&\quad \times \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\mathbf{w} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{w} - \boldsymbol{\mu})\right\} \mathrm{d}\mathbf{w},
\end{aligned}$$

where $\boldsymbol{\Lambda} = \mathrm{diag}(\frac{\gamma_1^2}{2}, \ldots, \frac{\gamma_d^2}{2}) \in \mathbb{R}^{d \times d}$ is a diagonal matrix.

By completing in squares, we then have

$$\widetilde{\psi}_{jl} = \frac{1}{\sqrt{(2\pi)^d |\mathbf{M}^{-1}|}} \frac{1}{\sqrt{|\boldsymbol{\Sigma}\mathbf{M}|}} \times \int w_l \exp\left\{-\frac{1}{2}(\mathbf{w} - \mathbf{M}^{-1}\mathbf{V})^\top \mathbf{M}(\mathbf{w} - \mathbf{M}^{-1}\mathbf{V}) + \frac{1}{2}(\mathbf{V}^\top \mathbf{M}^{-1}\mathbf{V} - R)\right\} \mathrm{d}\mathbf{w},$$

where $\mathbf{M} = \boldsymbol{\Sigma}^{-1} + \boldsymbol{\Lambda}^{-1}$, $\mathbf{V} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\Lambda}^{-1}\boldsymbol{\omega}_j^{\mathcal{T}}$ and $R = \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + (\boldsymbol{\omega}_j^{\mathcal{T}})^\top \boldsymbol{\Lambda}^{-1}\boldsymbol{\omega}_j^{\mathcal{T}}$.

By integrating out w_l with respect to the probability density function of a multivariate normal distribution with mean $\mathbf{M}^{-1}\mathbf{V}$ and covariance matrix $\mathbf{M}^{-1}$, we have

$$\widetilde{\psi}_{jl} = \frac{\mathbf{e}_l \mathbf{M}^{-1}\mathbf{V}}{\sqrt{|\boldsymbol{\Sigma}\mathbf{M}|}} \exp\left\{\frac{1}{2}(\mathbf{V}^\top \mathbf{M}^{-1}\mathbf{V} - R)\right\}.$$

Using the Woodbury identity (Petersen and Pedersen, 2012), we have

$$\mathbf{M}^{-1} = \boldsymbol{\Sigma} - \boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \boldsymbol{\Lambda})^{-1}\boldsymbol{\Sigma}$$
$$\mathbf{M}^{-1} = \boldsymbol{\Lambda} - \boldsymbol{\Lambda}(\boldsymbol{\Sigma} + \boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}.$$

Thus, we obtain

$$\widetilde{\psi}_{jl} = \frac{\mathbf{e}_l[\boldsymbol{\Lambda}(\boldsymbol{\Lambda} + \boldsymbol{\Sigma})^{-1}\boldsymbol{\mu} + \boldsymbol{\Sigma}(\boldsymbol{\Lambda} + \boldsymbol{\Sigma})^{-1}\boldsymbol{\omega}_j^{\mathcal{T}}]}{\sqrt{|(\boldsymbol{\Lambda} + \boldsymbol{\Sigma})\boldsymbol{\Lambda}^{-1}|}} \times \exp\left\{-\frac{1}{2}(\boldsymbol{\omega}_j^{\mathcal{T}} - \boldsymbol{\mu})^{\top}(\boldsymbol{\Lambda} + \boldsymbol{\Sigma})^{-1}(\boldsymbol{\omega}_j^{\mathcal{T}} - \boldsymbol{\mu})\right\},$$

which is

$$\widetilde{\psi}_{jl} = \mathbf{e}_l[\boldsymbol{\Lambda}(\boldsymbol{\Lambda} + \boldsymbol{\Sigma})^{-1}\boldsymbol{\mu} + \boldsymbol{\Sigma}(\boldsymbol{\Lambda} + \boldsymbol{\Sigma})^{-1}\boldsymbol{\omega}_j^{\mathcal{T}}]\,\widetilde{\xi}_j.$$

Bibliography

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y. and Zheng, X. (2015), 'TensorFlow: Large-scale machine learning on heterogeneous systems'. Software available from tensorflow.org.
URL: *https://www.tensorflow.org/*

Abrahamson, N. A., Silva, W. J. and Kamai, R. (2014), 'Summary of the ASK14 ground motion relation for active crustal regions', *Earthquake Spectra* **30**(3), 1025–1055.

Abrahamson, N. A. and Youngs, R. R. (1992), 'A stable algorithm for regression analyses using the random effects model', *Bulletin of the Seismological Society of America* **82**(1), 505–510.

Akkar, S. and Bommer, J. J. (2010), 'Empirical equations for the prediction of PGA, PGV, and spectral accelerations in Europe, the Mediterranean region, and the Middle East', *Seismological Research Letters* **81**(2), 195–206.

Andrianakis, I. and Challenor, P. G. (2012), 'The effect of the nugget on

Gaussian process emulators of computer models', *Computational Statistics & Data Analysis* **56**(12), 4215–4228.

Andrianakis, Y. and Challenor, P. G. (2009), Parameter Estimation and Prediction Using Gaussian Processes, Technical report, University of Southampton.

Arroyo, D. and Ordaz, M. (2010), 'Multivariate bayesian regression analysis applied to ground-motion prediction equations, part 1: theory and synthetic example', *Bulletin of the Seismological Society of America* **100**(4), 1551–1567.

Ba, S., Joseph, V. R. et al. (2012), 'Composite Gaussian process models for emulating expensive functions', *The Annals of Applied Statistics* **6**(4), 1838–1860.

Baptista, R., Marzouk, Y., Willcox, K. and Peherstorfer, B. (2018), 'Optimal approximations of coupling in multidisciplinary models', *AIAA Journal* **56**(6), 2412–2428.

Bauer, M., van der Wilk, M. and Rasmussen, C. E. (2016), Understanding probabilistic sparse Gaussian process approximations, *in* 'Advances in neural information processing systems', pp. 1533–1541.

Bilionis, I. and Zabaras, N. (2016), Bayesian uncertainty propagation using gaussian processes, *in* R. Ghanem, D. Higdon and H. Owhadi, eds, 'Handbook of Uncertainty Quantification', Springer International Publishing, pp. 1–45.

Bindi, D., Massa, M., Luzi, L., Ameri, G., Pacor, F., Puglia, R. and Augliera, P. (2014), 'Pan-european ground-motion prediction equations for the average horizontal component of pga, pgv, and 5%-damped psa at spectral periods up to 3.0 s using the resorce dataset', *Bulletin of Earthquake Engineering* **12**(1), 391–430.

Boore, D. M., Stewart, J., Seyhan, E. and Atkinson, G. M. (2014), 'NGA-West2 equations for predicting response spectral accelerations for shallow crustal earthquakes', *Earthquake Spectra* **30**(May), 106.

Brillinger, D. R. and Preisler, H. K. (1984), 'An exploratory analysis of the Joyner-Boore attenuation data', *Bulletin of the Seismological Society of America* **74**(4), 1441–1450.

Brillinger, D. R. and Preisler, H. K. (1985), 'Further analysis of the joyner-boore attenuation data', *Bulletin of the Seismological Society of America* **75**(2), 611–614.

Bui, T., Hernández-Lobato, D., Hernandez-Lobato, J., Li, Y. and Turner, R. (2016), Deep Gaussian processes for regression using approximate expectation propagation, *in* 'International conference on machine learning', pp. 1472–1481.

Campbell, K. W. and Bozorgnia, Y. (2014), 'NGA-West2 ground motion model for the average horizontal components of PGA, PGV, and 5% damped linear acceleration response spectra', *Earthquake Spectra* **30**(3), 1087–1114.

CEN (2005), 'Eurocode 8: Design of structures for earthquake resistance-part 1: general rules, seismic actions and rules for buildings', *Brussels: European Committee for Standardization* .

Chaudhuri, A., Lam, R. and Willcox, K. (2018), 'Multifidelity uncertainty propagation via adaptive surrogates in coupled multidisciplinary systems', *AIAA Journal* **56**(1), 235–249.

Chen, Y.-H. and Tsai, C.-C. P. (2002), 'A new method for estimation of the attenuation relationship with variance components', *Bulletin of the Seismological Society of America* **92**(5), 1984–1991.

Chiou, B. S. J. and Youngs, R. R. (2014), 'Update of the Chiou and Youngs NGA model for the average horizontal component of peak ground motion and response spectra', *Earthquake Spectra* **30**(3), 1117–1153.

Couvreur, C. (1997), The EM algorithm: A guided tour, *in* K. Warwick and

M. Kárný, eds, 'Computer intensive methods in control and signal processing', Birkhäuser, pp. 209–222.

Cressie, N. A. C. (1993), *Statistics for Spatial Data*, Wiley Interscience.

Cutajar, K., Pullin, M., Damianou, A., Lawrence, N. and González, J. (2019), 'Deep Gaussian processes for multi-fidelity modeling', *arXiv:1903.07320* .

Dalbey, K. R. (2013), Efficient and Robust Gradient Enhanced Kriging Emulators, Technical Report SAND2013–7022, Sandia National Laboratories: Albuquerque, NM, USA.

Damianou, A. and Lawrence, N. (2013), Deep Gaussian processes, *in* 'Artificial Intelligence and Statistics', pp. 207–215.

Das, R., Wason, H. and Sharma, M. (2011), 'Global regression relations for conversion of surface wave and body wave magnitudes to moment magnitude', *Natural Hazards* **59**(2), 801–810.

Das, R., Wason, H. and Sharma, M. (2012), 'Magnitude conversion to unified moment magnitude using orthogonal regression relation', *Journal of Asian Earth Sciences* **50**, 44–51.

Demidenko, E. (2013), *Mixed Models: Theory and Applications with R*, John Wiley & Sons.

Demmel, J. (1992), 'The componentwise distance to the nearest singular matrix', *SIAM Journal on Matrix Analysis and Applications* **13**(1), 10–19.

Di Giacomo, D., Bondár, I., Storchak, D. A., Engdahl, E. R., Bormann, P. and Harris, J. (2015), 'Isc-gem: Global instrumental earthquake catalogue (1900–2009), iii. re-computed m_s and m_b, proxy m_w, final magnitude composition and completeness assessment', *Physics of the Earth and Planetary Interiors* **239**, 33–47.

Douglas, J. and Edwards, B. (2016), 'Recent and future developments in earthquake ground motion estimation', *Earth-Science Reviews* **160**, 203–219.

Draper, N. R. and Smith, H. (2014), *Applied Regression Analysis*, John Wiley & Sons.

Dunlop, M. M., Girolami, M. A., Stuart, A. M. and Teckentrup, A. L. (2018), 'How deep are deep Gaussian processes?', *The Journal of Machine Learning Research* **19**(1), 2100–2145.

Duvenaud, D., Rippel, O., Adams, R. and Ghahramani, Z. (2014), Avoiding pathologies in very deep networks, *in* 'Artificial Intelligence and Statistics', pp. 202–210.

Esposito, S. and Iervolino, I. (2011), 'PGA and PGV spatial correlation models based on European multievent datasets', *Bulletin of the Seismological Society of America* **101**(5), 2532–2541.

Esposito, S. and Iervolino, I. (2012), 'Spatial correlation of spectral acceleration in European data', *Bulletin of the Seismological Society of America* **102**(6), 2781–2788.

Fazeley, H., Taei, H., Naseh, H. and Mirshams, M. (2016), 'A multi-objective, multidisciplinary design optimization methodology for the conceptual design of a spacecraft bi-propellant propulsion system', *Structural and Multidisciplinary Optimization* **53**(1), 145–160.

Fisher, R. A. (1925), Theory of statistical estimation, *in* 'Mathematical Proceedings of the Cambridge Philosophical Society', Vol. 22, Cambridge University Press, pp. 700–725.

Flury, B. (2013), *A first course in multivariate statistics*, Springer, New York.

Fricker, T. E., Oakley, J. E. and Urban, N. M. (2013), 'Multivariate Gaussian process emulators with nonseparable covariance structures', *Technometrics* **55**(1), 47–56.

Gao, X. and Wang, Y. (2013), 'Application of em algorithm in statistics natural language processing', *Research Journal of Applied Sciences, Engineering and Technology* **5**(10), 2969–2973.

Gill, P. E., Murray, W. and Wright, M. H. (1981), *Practical Optimization*, Emerald.

Goda, K. and Atkinson, G. M. (2009), 'Probabilistic characterization of spatially correlated response spectra for earthquakes in Japan', *Bulletin of the Seismological Society of America* **99**(5), 3003–3020.

Goda, K. and Atkinson, G. M. (2010), 'Intraevent spatial correlation of ground-motion parameters using SK-net data', *Bulletin of the Seismological Society of America* **100**(6), 3055–3067.

Goda, K. and Hong, H.-P. (2008), 'Spatial correlation of peak ground motions and response spectra', *Bulletin of the Seismological Society of America* **98**(1), 354–365.

Golub, G. H. and Van Loan, C. F. (2012), *Matrix Computations*, 4 edn, JHU Press.

Gramacy, R. B. and Lee, H. K. (2012), 'Cases for the nugget in modeling computer experiments', *Statistics and Computing* **22**(3), 713–722.

Gramacy, R. B. and Lee, H. K. H. (2008), 'Bayesian treed Gaussian process models with an application to computer modeling', *Journal of the American Statistical Association* **103**(483), 1119–1130.

Gu, M. and Berger, J. O. (2016), 'Parallel partial Gaussian process emulation for computer models with massive output', *The Annals of Applied Statistics* **10**(3), 1317–1347.

Gu, M., Wang, X. and Berger, J. O. (2018), 'Robust Gaussian stochastic process emulation', *The Annals of Statistics* **46**(6A), 3038–3066.

Hanks, T. C. and Kanamori, H. (1979), 'A moment magnitude scale', *Journal of Geophysical Research: Solid Earth* **84**(B5), 2348–2350.

Havasi, M., Hernández-Lobato, J. M. and Murillo-Fuentes, J. J. (2018), Inference in deep Gaussian processes using stochastic gradient hamiltonian monte carlo, *in* 'Advances in Neural Information Processing Systems', pp. 7506–7516.

Hawkins, E., Smith, R. S., Gregory, J. M. and Stainforth, D. A. (2016), 'Irreducible uncertainty in near-term climate projections', *Climate Dynamics* **46**(11-12), 3807–3819.

Hong, H. P., Zhang, Y. and Goda, K. (2009), 'Effect of spatial correlation on estimated ground-motion prediction equations', *Bulletin of the Seismological Society of America* **99**(2 A), 928–934.

Idriss, I. M. (2014), 'An NGA-West2 empirical model for estimating the horizontal spectral values generated by shallow crustal earthquakes', *Earthquake Spectra* **30**(3), 1155–1177.

Ipsen, I. C. and Lee, D. J. (2011), 'Determinant approximations', *arXiv:1105.0437* .

Jandarov, R., Haran, M., Bjørnstad, O. and Grenfell, B. (2014), 'Emulating a gravity model to infer the spatiotemporal dynamics of an infectious disease', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **63**(3), 423–444.

Jayaram, N. and Baker, J. W. (2009), 'Correlation model for spatially distributed ground-motion intensities', *Earthquake Engineering and Structural Dynamics* **38**(April), 1687–1708.

Jayaram, N. and Baker, J. W. (2010), 'Considering spatial correlation in mixed-effects regression and the impact on ground-motion models', *Bulletin of the Seismological Society of America* **100**(6), 3295–3303.

Johnstone, R. H., Chang, E. T., Bardenet, R., De Boer, T. P., Gavaghan, D. J., Pathmanathan, P., Clayton, R. H. and Mirams, G. R. (2016), 'Uncertainty and variability in models of the cardiac action potential: Can we build trustworthy models?', *Journal of Molecular and Cellular Cardiology* **96**, 49–62.

Joyner, W. B. and Boore, D. M. (1993), 'Methods for regression analysis of strong-motion data', *Bulletin of the Seismological Society of America* **83**(2), 469–487.

Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., Arblaster, J. M., Bates, S., Danabasoglu, G. and Edwards, J. (2015), 'The Community Earth System Model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability', *Bulletin of the American Meteorological Society* **96**(8), 1333–1349.

Kerby, B. (2016), Semivariogram Estimation: Asymptotic Theory and Applications, PhD thesis, The University of Utah.

Kingma, D. P. and Ba, J. (2015), Adam: A method for stochastic optimization, *in* Y. Bengio and Y. LeCun, eds, 'Proceedings of 3rd International Conference on Learning Representations, ICLR'.

Kodiyalam, S., Yang, R., Gu, L. and Tho, C.-H. (2004), 'Multidisciplinary design optimization of a vehicle system in a scalable, high performance computing environment', *Structural and Multidisciplinary Optimization* **26**(3-4), 256–263.

Komatitsch, D. and Tromp, J. (2002*a*), 'Spectral-element simulations of global seismic wave propagation-i. validation', *Geophysical Journal International* **149**(2), 390–412.

Komatitsch, D. and Tromp, J. (2002*b*), 'Spectral-element simulations of global seismic wave propagation–ii. three-dimensional models, oceans, rotation and self-gravitation', *Geophysical Journal International* **150**(1), 303–318.

Kyzyurova, K. N., Berger, J. O. and Wolpert, R. L. (2018), 'Coupling computer models through linking their statistical emulators', *SIAM/ASA Journal on Uncertainty Quantification* **6**(3), 1151–1171.

Lahiri, S. N., Lee, Y. and Cressie, N. (2002), 'On asymptotic distribution and asymptotic efficiency of least squares estimators of spatial variogram parameters', *Journal of Statistical Planning and Inference* **103**(1-2), 65–85.

Laird, N., Lange, N. and Stram, D. (1987), 'Maximum likelihood computations with repeated measures: application of the em algorithm', *Journal of the American Statistical Association* **82**(397), 97–105.

Laird, N. M. and Ware, J. H. (1982), 'Random-effects models for longitudinal data', *Biometrics* pp. 963–974.

Lindstrom, M. J. and Bates, D. M. (1990), 'Nonlinear mixed effects models for repeated measures data', *Biometrics* pp. 673–687.

Liu, H., Ong, Y.-S., Shen, X. and Cai, J. (2020), 'When Gaussian process meets big data: A review of scalable gps', *IEEE Transactions on Neural Networks and Learning Systems* .

Mardia, K. V. and Marshall, R. J. (1984), 'Maximum likelihood estimation of models for residual covariance in spatial regression', *Biometrika* **71**(1), 135–146.

Marque-Pucheu, S., Perrin, G. and Garnier, J. (2019), 'Efficient sequential experimental design for surrogate modeling of nested codes', *ESAIM: Probability and Statistics* **23**, 245–270.

Ming, D., Huang, C., Peters, G. W. and Galasso, C. (2019), 'An advanced estimation algorithm for ground-motion models with spatial correlation', *Bulletin of the Seismological Society of America* **109**(2), 541–566.

Minka, T. P. (2013), 'Expectation propagation for approximate Bayesian inference', *arXiv:1301.2294* .

Mohammadi, H., Riche, R. L., Durrande, N., Touboul, E. and Bay, X. (2016), 'An analytic comparison of regularization methods for Gaussian processes', *ArXiv Preprint* .

Montagna, S. and Tokdar, S. T. (2016), 'Computer emulation with nonstationary gaussian processes', *SIAM/ASA Journal on Uncertainty Quantification* **4**(1), 26–47.

Nocedal, J. and Wright, S. (2006), *Numerical optimization*, Springer.

Okada, Y. (1985), 'Surface deformation due to shear and tensile faults in a half-space', *Bulletin of the Seismological Society of America* **75**(4), 1135–1154.

Petersen, K. B. and Pedersen, M. S. (2012), *The Matrix Cookbook*, Technical University of Denmark, Lyngby, Denmark.

Pinheiro, J. and Bates, D. (2000), *Mixed-Effects Models in S and S-PLUS*, Springer, Madison, USA.

Pyzara, A., Bylina, B. and Bylina, J. (2011), The influence of a matrix condition number on iterative methods' convergence, *in* 'Computer Science and Information Systems (FedCSIS), 2011 Federated Conference on', IEEE, pp. 459–464.

Quiñonero-Candela, J. and Rasmussen, C. E. (2005), 'A unifying view of sparse approximate Gaussian process regression', *Journal of Machine Learning Research* **6**(Dec), 1939–1959.

Rainforth, T., Cornish, R., Yang, H., Warrington, A. and Wood, F. (2018), 'On nesting Monte Carlo estimators', *Proceedings of Machine Learning Research* **80**, 4267–4276.

Rasmussen, C. E. and Ghahramani, Z. (2002), Infinite mixtures of Gaussian process experts, *in* 'Advances in neural information processing systems', pp. 881–888.

Reguly, I. Z., Giles, D., Gopinathan, D., Quivy, L., Beck, J. H., Giles, M. B., Guillas, S. and Dias, F. (2018), 'The VOLNA-OP2 tsunami code (version 1.5)', *Geoscientific Model Development* **11**(11), 4621–4635.

Richter, C. F. (1935), 'An instrumental earthquake magnitude scale', *Bulletin of the Seismological Society of America* **25**(1), 1–32.

Rougier, J., Guillas, S., Maute, A. and Richmond, A. D. (2009), 'Expert knowledge and multivariate emulation: The thermosphere–ionosphere electrodynamics general circulation model (tie-gcm)', *Technometrics* **51**(4), 414–424.

Salimbeni, H. and Deisenroth, M. (2017), Doubly stochastic variational inference for deep Gaussian processes, *in* 'Advances in Neural Information Processing Systems', pp. 4588–4599.

Salmanidou, D., Guillas, S., Georgiopoulou, A. and Dias, F. (2017), 'Statistical emulation of landslide-induced tsunamis at the Rockall Bank, NE Atlantic', *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **473**(2200), 20170026.

Salter, J. M., Williamson, D. B., Scinocca, J. and Kharin, V. (2019), 'Uncertainty quantification for computer models with spatial output using calibration-optimal bases', *Journal of the American Statistical Association* pp. 1–24.

Sankararaman, S. and Mahadevan, S. (2012), 'Likelihood-based approach to multidisciplinary analysis under uncertainty', *Journal of Mechanical Design* **134**(3), 031008.

Sanson, F., Le Maitre, O. and Congedo, P. M. (2019), 'Systems of Gaussian process models for directed chains of solvers', *Computer Methods in Applied Mechanics and Engineering* **352**, 32–55.

Santiago, A., Aguado-Sierra, J., Zavala-Aké, M., Doste-Beltran, R., Gómez, S., Arís, R., Cajas, J. C., Casoni, E. and Vázquez, M. (2018), 'Fully coupled

fluid-electro-mechanical model of the human heart for supercomputers', *International Journal for Numerical Methods in Biomedical Engineering* **34**(12), e3140.

Santner, T. J., Williams, B. J., Notz, W. and Williams, B. J. (2003), *The Design and Analysis of Computer Experiments*, Springer, New York.

Seber, G. and Wild, C. (2003), *Nonlinear Regression*, Wiley Series in Probability and Statistics, Wiley.

Silverman, B. W. (1985), 'Some aspects of the spline smoothing approach to non-parametric regression curve fitting', *Journal of the Royal Statistical Society: Series B (Methodological)* **47**(1), 1–21.

Simpson, T. W., Mauery, T. M., Korte, J. J. and Mistree, F. (2001), 'Kriging models for global approximation in simulation-based multidisciplinary design optimization', *AIAA Journal* **39**(12), 2233–2241.

Snelson, E. and Ghahramani, Z. (2006), Sparse Gaussian processes using pseudo-inputs, *in* 'Advances in neural information processing systems', pp. 1257–1264.

Sokolov, V., Wenzel, F. and Kuo-Liang, W. (2010), 'Uncertainty and spatial correlation of earthquake ground motion in Taiwan', *TAO: Terrestrial, Atmospheric and Oceanic Sciences* **21**(6), 9.

Stein, M. L. (1999), *Interpolation of Spatial Data: Some Theory for Kriging*, Springer, New York.

Stucchi, M., Meletti, C., Montaldo, V., Crowley, H., Calvi, G. M. and Boschi, E. (2011), 'Seismic hazard assessment (2003-2009) for the Italian building code', *Bulletin of the Seismological Society of America* **101**(4), 1885–1911.

Tagade, P. M., Jeong, B.-M. and Choi, H.-L. (2013), 'A Gaussian process emulator approach for rapid contaminant characterization with an integrated multizone-CFD model', *Building and Environment* **70**, 232–244.

The USGS earthquake magnitude working group (2002), 'USGS earthquake magnitude policy (implemented on January 18, 2002)'. Available at: `https://web.archive.org/web/20160504144754/http://earthquake.usgs.gov/aboutus/docs/020204mag_policy.php`.

Thuiller, W., Guéguen, M., Renaud, J., Karger, D. N. and Zimmermann, N. E. (2019), 'Uncertainty in ensembles of global biodiversity scenarios', *Nature Communications* **10**(1), 1446.

Titsias, M. (2009), Variational learning of inducing variables in sparse Gaussian processes, *in* 'Artificial Intelligence and Statistics', pp. 567–574.

Trifunac, M. and Brady, A. (1976), 'Correlations of peak acceleration, velocity and displacement with earthquake magnitude, distance and site conditions', *Earthquake Engineering & Structural Dynamics* **4**(5), 455–471.

Trifunac, M. D. and Brady, A. G. (1975), 'A study on the duration of strong earthquake ground motion', *Bulletin of the Seismological Society of America* **65**(3), 581–626.

Ulrich, T., Vater, S., Madden, E. H., Behrens, J., van Dinther, Y., van Zelst, I., Fielding, E. J., Liang, C. and Gabriel, A.-A. (2019), 'Coupled, physics-based modeling reveals earthquake displacements are critical to the 2018 Palu, Sulawesi Tsunami', *Pure and Applied Geophysics* **176**(10), 4069–4109.

Vafa, K. (2016), Training deep Gaussian processes with sampling, *in* 'NIPS 2016 Workshop on Advances in Approximate Bayesian Inference'.

Vernon, I., Goldstein, M. and Bower, R. (2014), 'Galaxy formation: Bayesian history matching for the observable universe', *Statistical Science* pp. 81–90.

Volodina, V. and Williamson, D. (2020), 'Diagnostics-driven nonstationary emulators using kernel mixtures', *SIAM/ASA Journal on Uncertainty Quantification* **8**(1), 1–26.

Wald, D. J. and Allen, T. I. (2007), 'Topographic slope as a proxy for seismic site conditions and amplification', *Bulletin of the Seismological Society of America* **97**(5), 1379–1395.

Wang, M. and Takada, T. (2005), 'Macrospatial correlation model of seismic ground motions', *Earthquake spectra* **21**(4), 1137–1156.

Williams, C. K. and Rasmussen, C. E. (2006), *Gaussian processes for machine learning*, MIT press, Cambridge, MA.

Wolfe, P. (1969), 'Convergence conditions for ascent methods', *SIAM review* **11**(2), 226–235.

Wolfe, P. (1971), 'Convergence conditions for ascent methods. ii: Some corrections', *SIAM review* **13**(2), 185–188.

Wooldridge, J. M. (2010), *Econometric Analysis of Cross Section and Panel Data*, MIT Press.

Worden, C. B., Thompson, E. M., Baker, J. W., Bradley, B. A., Luco, N. and Wald, D. J. (2018), 'Spatial and spectral interpolation of ground-motion intensity measure observations', *Bulletin of the Seismological Society of America* **108**(2), 866–875.

Zhang, B., Konomi, B. A., Sang, H., Karagiannis, G. and Lin, G. (2015), 'Full scale multi-output Gaussian process emulator with nonseparable auto-covariance functions', *Journal of Computational Physics* **300**, 623–642.

Zhao, W., Wang, Y. and Wang, C. (2018), 'Multidisciplinary optimization of electric-wheel vehicle integrated chassis system based on steady endurance performance', *Journal of Cleaner Production* **186**, 640–651.

Zimmerman, D. L. and Stein, M. (2010), Classical geostatistical methods, *in* A. E. Gelfand, P. Diggle, P. Guttorp and M. Fuentes, eds, 'Handbook of Spatial Statistics', CRC Press, pp. 29–44.