Institut für Mathematik
Lehrstuhl für Angewandte Mathematik

Dissertation

zur Erlangung des akademischen Grades *doctor rerum naturalium* (Dr. rer. nat.) in der Wissenschaftsdisziplin *Angewandte Mathematik* eingereicht an der Mathematisch-Naturwissenschaftlichen Fakultät der Universität Potsdam

---

Correlation Based Bayesian Modeling

with applications in
Travel Time Tomography,
Seismic Source Inversion and
Magnetic Field Modeling.

Stefan Mauerberger

Potsdam, 4. February 2022

Hauptbetreuer: Prof. Dr. Matthias Holschneider
Gutachter: Prof. Dr. Catherine Constable
Prof. Dr. Torsten Dahm

Published online on the
Publication Server of the University of Potsdam:
https://doi.org/10.25932/publishup-53782
https://nbn-resolving.org/urn:nbn:de:kobv:517-opus4-537827

This work would not have been possible without the support of the
Helmholtz Graduate Research School *GeoSim*
and the
German Research Foundation (DFG), grant 388291411.

Summary

The motivation for this work was the question of reliability and robustness of seismic tomography. The problem is that many earth models exist which can describe the underlying ground motion records equally well. Most algorithms for reconstructing earth models provide a solution, but rarely quantify their variability. If there is no way to verify the imaged structures, an interpretation is hardly reliable. The initial idea was to explore the space of equivalent earth models using Bayesian inference. However, it quickly became apparent that the rigorous quantification of tomographic uncertainties could not be accomplished within the scope of a dissertation.

In order to maintain the fundamental concept of statistical inference, less complex problems from the geosciences are treated instead. This dissertation aims to anchor Bayesian inference more deeply in the geosciences and to transfer knowledge from applied mathematics. The underlying idea is to use well-known methods and techniques from statistics to quantify the uncertainties of inverse problems in the geosciences. This work is divided into three parts:

Part I introduces the necessary mathematics and should be understood as a kind of *toolbox*. With a physical application in mind, this section provides a compact summary of all methods and techniques used. The introduction of Bayesian inference makes the beginning. Then, as a special case, the focus is on regression with Gaussian processes under linear transformations. The chapters on the derivation of covariance functions and the approximation of non-linearities are discussed in more detail.

Part II presents two proof of concept studies in the field of seismology. The aim is to present the conceptual application of the introduced methods and techniques with moderate complexity. The example about traveltime tomography applies the approximation of non-linear relationships. The derivation of a covariance function using the wave equation is shown in the example of a damped vibrating string. With these two synthetic applications, a consistent concept for the quantification of modeling uncertainties has been developed.

Part III presents the reconstruction of the Earth's archeomagnetic field. This application uses the whole *toolbox* presented in Part I and is correspondingly complex. The modeling of the past 1000 years is based on real data and reliably quantifies the spatial modeling uncertainties. The statistical model presented is widely used and is under active development.

The three applications mentioned are intentionally kept flexible to allow transferability to similar problems. The entire work focuses on the non-uniqueness of inverse problems in the geosciences. It is intended to be of relevance to those interested in the concepts of Bayesian inference.

Zusammenfassung

Die Motivation für diese Arbeit war die Frage nach Verlässlichkeit und Belastbarkeit der seismischen Tomographie. Das Problem besteht darin, dass sehr viele Erdmodelle existieren welche die zugrundeliegenden seismischen Aufzeichnungen gleich gut beschreiben können. Die meisten Algorithmen zur Rekonstruktion von Erdmodellen liefern zwar eine Lösung, quantifizierten jedoch kaum deren Variabilität. Wenn es keine Möglichkeit gibt die abgebildeten Strukturen zu verifizieren, so ist eine Interpretation kaum verlässlich. Der ursprüngliche Gedanke war den Raum äquivalenter Erdmodelle mithilfe Bayesianische Inferenz zu erkunden. Es stellte sich jedoch schnell heraus, dass die vollständige Quantifizierung tomographischer Unsicherheiten im Rahmen einer Promotion nicht zu bewältigen ist.

Um das wesentliche Konzept der statistischen Inferenz beizubehalten werden stattdessen weniger komplexe Problemstellungen aus den Geowissenschaften behandelt. Diese Dissertation hat das Ziel die Bayesianische Inferenz tiefer in den Geowissenschaften zu verankern und Wissen aus der angewandten Mathematik zu transferieren. Die zugrundeliegende Idee besteht darin auf bekannte Methoden und Techniken der Statistik zurückzugreifen um die Unsicherheiten inverser Probleme in den Geowissenschaften zu quantifizieren. Diese Arbeit gliedert sich in drei Teile:

Teil I führt die notwendige Mathematik ein und soll als eine Art *Werkzeugkasten* verstanden werden. In Hinblick auf eine physikalische Anwendung bietet dieser Abschnitt eine kompakte Zusammenfassung aller eingesetzter Methoden und Techniken. Den Anfang macht die Einführung der Bayesianische Inferenz. Danach steht als Spezialfall die Regression mit Gauß-Prozessen unter linearen Transformationen im Vordergrund. Die Kapitel zur Herleitung von Kovarianzfunktionen und die Approximation von Nichtlinearitäten gehen etwas weiter in die Tiefe.

Teil II präsentiert zwei Konzeptstudien aus dem Bereich der Seismologie. Ziel ist es bei moderater Komplexität die prinzipielle Anwendung der eingeführten Methoden und Techniken zu präsentieren. Das Beispiel zur Laufzeittomographie wendet die Näherungsmethoden für nichtlineare Zusammenhänge an. Die Herleitung einer Kovarianzfunktion mithilfe der Wellengleichung ist am Beispiel der gedämpften Saitenschwingung gezeigt. Mit diesen beiden synthetischen Anwendungen wurde ein konsistentes Konzept zur Quantifizierung von Modellierungsunsicherheiten erarbeitet.

Teil III präsentiert die Rekonstruktion des archeomagnetischen Feldes unserer Erde. Diese Anwendung nutzt den gesamten *Werkzeugkasten* aus Teil I und ist entsprechend umfangreich. Die Modellierung der vergangenen 1000 Jahre basiert auf echten Daten und quantifiziert zuverlässig die räumlichen Modellierungsunsicherheiten. Das präsentierte statistische Modell findet breite Anwendung und wird aktiv weiter entwickelt.

Die drei genannten Anwendungen sind bewusst flexibel gehalten um die Übertragbarkeit auf ähnliche Problemstellungen zu ermöglichen. Die gesamte Arbeit legt den Fokus auf die nicht-Eindeutigkeit inverser Probleme in den Geowissenschaften. Sie will für all Jene von Relevanz sein, die sich für die Konzepte der Bayesianischen Inferenz interessieren.

Contents

I. Theory 1

1. Bayesian Inference 3
1.1. Model, Data and Prediction . . . 4
1.2. Gaussian Process Regression . . . 7
1.2.1. Inference . . . 9
1.3. Gaussian Linear Model . . . 12

2. Reproducing Kernel Hilbert Spaces 19
2.1. Regularized Least Squares . . . 28
2.2. Duality GP and RKHS . . . 31
2.3. Kernel Parameters . . . 33

3. Non-Gaussian Likelihoods 37
3.1. Bayesian Posterior . . . 38
3.2. Gaussian Process Approximation . . . 41
3.3. Laplace Approximation . . . 43
3.3.1. Mixture Approach . . . 45
3.4. Local Likelihood Approximations . . . 47
3.4.1. Assumed Density Filtering . . . 47
3.4.2. Expectation Propagation Algorithm . . . 49

II. Seismology 53

4. Travel Time Inversion 55
4.1. Direct Waves . . . 56
4.2. Reflected Waves . . . 59
4.3. Successive Approximation . . . 61
4.4. Conclusion . . . 64
4.5. Outlook & Further Thoughts . . . 64

5. Seismic Source Inversion 67
5.1. White Noise . . . 71
5.2. Fourier Analysis . . . 73
5.3. Filtered White Noise . . . 77
5.4. Outlook & Further Thoughts . . . 81

III. Geomagnetism **85**

6. Correlation Based Snapshot Models of the Archeomagnetic Field **87**
6.1. Modelling Concept . . . 89
6.1.1. Magnetic Field Model . . . 89
6.1.2. Inference . . . 91
6.1.3. Observational functionals & Linearization . . . 93
6.1.4. Measurement Errors . . . 94
6.2. Bayesian Update System . . . 96
6.2.1. Complete Records . . . 96
6.2.2. Incomplete Records . . . 98
6.2.3. Synthetic Tests . . . 99
6.3. Model Parameters . . . 100
6.3.1. Uninformative Dipole . . . 100
6.3.2. Compound Distribution . . . 102
6.3.3. Marginal Likelihood . . . 103
6.4. Application . . . 104
6.4.1. Numeric Integration . . . 106
6.4.2. Vector field predictions . . . 108
6.4.3. Declination, Inclination and Intensity . . . 108
6.4.4. Predictions at the Core-Mantle Boundary . . . 110
6.4.5. Gauss Coefficients . . . 111
6.4.6. Spatial Power Spectrum . . . 114
6.4.7. Dipole Moment . . . 115
6.5. Conclusions and Perspectives . . . 118

Bibliography **121**

Part I.

Theory

The following three chapters may be understood as a *toolbox* providing a brief recap of methods and techniques used in later applications. For further reading an overview on references and literature are provided.

Chapter 1 introduces Bayesian inversion and Gaussian process regression. The discussion is completed by easy to digest examples and leads to what is known as the *linear Gaussian model*.

Chapter 2 gives a brief review on methods that are closely related to Gaussian process regression. It presents a functional analytic point of view how to systematically derive a covariance function from first principles.

Chapter 3 deals with the treatment of non-linear likelihood functions. The limits of Bayesian inference are identified and selected algorithms for approximate Bayesian inference are presented.

1. Bayesian Inference

This Chapter is a very concise recap of a systematic approach in solving inverse problems using statistical techniques. The reader is assumed to have a fair degree of familiarity with basic probability theory. A light version on the fundamentals may be found at Calvetti and Somersalo [2007, Sec. 1.2]. As for any Bayesian approach, the conditional probability is the corner stone on which the theory is built upon. Rather than arguing the foundation of statistics, I deliberately pass on the abstract theory of probability spaces and present selected concepts and results. For a rigorous discussion the reader shall refer any standard text-book on probability theory e.g. Schmidt [2011].

To introduce the notations I am going to adopt throughout, a summary of the basic mathematics of Bayesian inference is presented. Only continuous real-valued random variables (RV) with corresponding probability density function (PDF) are considered. From a physical point of view this may be understood as the restriction to classical mechanics where phenomena are modeled in the continuous domain ignoring the microscopic scale. References on applied Bayesian inference that I used are Bui-Thanh [2012], Gelman et al. [2013, Chp. 1], Calvetti and Somersalo [2007, Chp. 2] and Bolstad [2007].

Statistical inference is the process of deducing properties of an underlying distribution by analysis of data [Cook, 2008]. According to Calvetti and Somersalo [2007] the problem of statistical inference may be phrased as follows:

> Determine the underlying distribution which gives rise for realizations.

In other words, we are looking for a RV X compatible with realizations $Y = y$. To formalize statistical inference, let X and Y denote continuous random variables with joint PDF p_{XY} and according marginal PDFs p_X and p_Y. The conditional distribution of X given the realization $Y = y$ can be written as

$$p_{X|Y}(x \mid y) = \frac{p_{XY}(x, y)}{p_Y(y)} \,. \tag{1.1}$$

By symmetry considerations, the conditional PDF of X under $Y = y$ is given by

$$p_{Y|X}(y \mid x)p_X(x) = p_{XY}(x, y) = p_{X|Y}(x \mid y)p_Y(y) \tag{1.2}$$

which is known as the *general product rule*, permitting to calculate joint PDFs. Therefrom the well-known Bayes formula follows

Theorem 1 (Bayes' law)**.** *The PDF of X knowing the realization y of Y is*

$$p_{X|Y}(x \mid y) = \frac{p_{Y|X}(y \mid x)\, p_X(x)}{p_Y(y)}$$

and is referred to as the posterior distribution.

The PDF $p_{X|Y}(x|y)$ – in words, the distribution of X knowing that Y got the value y assigned – is the solution of the Bayesian inverse problem. That simple formula encapsulates the technical core of Bayesian inference. Because of their importance the PDFs on the right hand side of Theorem 1 got special names assigned:

$p_Y(y)$ is called the **prior predictive distribution**. *Prior* because it is not conditional on a previous realization and *predictive* because it is the distribution for a quantity that is observable. It may be understood as a normalizing constant as it does not depend on X.

$p_{Y|X}(y \mid x)$ is typically considered as a function in x and dubbed **likelihood function** $L(x|y)$. Regarded as a function in x it is no longer a density. The likelihood function describes how the *noisy* realizations deviate from the underlying idealized model. The realization y affects the posterior PDF only through the likelihood function.

$p_X(x)$ is called the **prior distribution**. It can be quite approximate and expresses the belief about X before making an inference. In other words, the prior PDF encodes our knowledge on X in beforehand, regardless of any realization y of Y. There is no universal rule constructing a prior PDF, however, the more realistic the prior the better. On the other hand, if we are data-rich the effect of the prior distribution will be small and data will *swap out* vague a priori assumptions.

The posterior distribution summarizes our prior believes after incorporating the evidence and may be interpreted as a compromise between data and prior knowledge. It combines the likelihood and the prior, and captures everything we know. However, it is subjective to the chosen model and the a priori assumption does not necessarily represent reality. The difficulty and primary task of any application of Bayes' law is to develop appropriate statistical models for X and Y, to determine the afore mentioned three PDFs and finally to perform computations.

1.1. Model, Data and Prediction

The introduction of Bayesian inference with RVs X and Y is rather abstract. In the interest of application, the following gives a more physical interpretation and translates Bayes' law into the notion of *observations* and *predictions* subject to an underlying *system*. Typically the aforementioned realization y is identified with measurements whereas X denotes the quantity of interest.

Consider some physical system that relates observations and predictions. An example – that is addressed in Chapter 5 – is to infer the action that let a string vibrate. Let $\tilde{m}$ denote reality, the true system we do not know. Measurements $d = (d_1, \dots, d_n)$ are recorded to learn about the system. An observable is described by a synthetic formula the so-called observational functional $f[\cdot]$ representing our physical understanding of measurements [Parker, 1994, Sec. 1.04 and p. 32]. In fact, observations are not deterministic but corrupted by measurement noise. Measurements are realizations of the noise model conditional to reality

$$d \sim D|\tilde{m} = f[\tilde{m}] + E \tag{1.3}$$

still $\tilde{m}$ represents the real but unknown model and E a to be know error distribution. Throughout, mutually independent and purely additive noise models are considered [1]. Notice that the actual values measured are denoted by d whereas $D = (D_1, \dots D_n)$ refers to the according RV.

Within the Bayesian paradigm, randomness is used to express the lack of information. Therefore, the system is considered a random quantity $\tilde{m} \to M$ describing our ignorance. The a priori distribution p_M encodes our belief about possible values of the system before taking a look at the data. In fact, construction of the a priori distribution is subject to each individual problem. Again, the observed data must not have any influence on the choice of the prior distribution [Bolstad, 2007, Chp. 8]. Incorporating our ignorance about the system, the stochastic data model reads

$$D_i = f_i[M] + E_i \tag{1.4}$$

with to be known error model E. The understanding of physics may be partial and/or uncertain and the noise-level will not exclusively account for the inaccuracy in observations but also for the incompleteness of theory. Therefore, E is also called *residual term* and is typically not identified with the measurement's bare error model. Since E also accounts for those parts of theory we do not know, E and M become dependent. To simplify matters, in the following any correlations amongst E and M are neglected.

As measurements may be performed at will, non-observable properties are typically the quantities in question. In other words, the intention is to predict quantities we did not and/or cannot measure. Let $G = \{G_1, \dots, G_k\}$ denote the to be predicted quantity. In analogy to observations, predictions are assumed as functionals in M

$$G_j = h_j[M]. \tag{1.5}$$

Although it might appear natural to directly model G, introducing the system M often allows a convenient formulation of the statistical model. Another reason for introducing predictive functionals is that it might be easier making a priori assumptions about M.

Having a consent on the physical setting Bayes' law shall be applied and we are confronted to determine the prior density p_G, the prior predictive density p_D and the likelihood function $L(g|d)$.

Likelihood Retrieving the likelihood is relatively easy since the error model is assumed to be known. As we assume additive noise and keep $G = g$ fixed we have

$$p_{D|G}(d \mid h[m]) = p_{f[M]+E|M}(d \mid m) = p_{f[m]+E}(d) = p_E(d - f[m]) \tag{1.6}$$

where the relation $p_{X+a}(x) = \partial_x \mathbb{P}[X + a < x] = \partial_x \mathbb{P}[X < x - a] = p_X(x - a)$ was used. Remember, $h[\cdot]$ is a function in M; if the RV $G = g$ is considered fixed this actually means to fix the underlying system $M = m$.

[1] Although it is common to assume independence of M and E, in general this is neither necessary nor easily justified. Kaipio and Somersalo [2005, Sec. 3.2.1 and 3.2.2] give a discussion on dependent noise and other explicit noise models.

Prior distribution To determine the prior distribution p_G – i.e. the transformation of M by h – is in general challenging. Even simple transformations of variables with simple distributions can lead to variables with complex distributions [Siegrist, 2015, Sec. 2.7]. The prior distribution of M is known through its PDF p_M and similarly we would like to find the PDF $p_{h[M]}$. If h is of good-nature there is a formula for p_G:

Theorem 2 (Change of Variables)**.** *Let M be a continuous distribution and $G = h[M]$ a locally injective transformation $h\colon U \subset \mathbb{R}^k \to \mathbb{R}^k$. Branches of the inverse $\{C_i\}_{i\in I}$ are disjoint intervals such that the restriction $h|_{C_i}$ is continuous and differentiable with ${h|_{C_i}}' \neq 0$ and $\mathrm{P}(\mathbb{R} \setminus \sum C_i) = 0$. Then G's density is given by*

$$p_G(g) = p_{h[M]}(g) = \sum_i p_M\big(h|_{C_i}^{-1}(g)\big) \left|\det \mathcal{J}_{h|_{C_i}^{-1}}(g)\right| \mathbf{1}_{\{h(C_i)\}}(g)$$

where $\mathcal{J}$ denotes the Jacobian matrix and $\mathbf{1}_{\{\cdot\}}$ the indicator function.

A profound discussion including proofs may be found in Schmidt [2011, Secs. 12 and 13]. Notice, the map in Theorem 2 demands equality in dimensions. Predictions $h[M]$ and the model M will certainly not share same dimensions. A way around that limitation is, to add dummy arguments to achieve same dimensionality and marginalize subsequently.

Prior predictive distribution Since we assume M and E independent, the density of their sum is given by a convolution [2] and we have

$$p_D(d) = p_{(f[M]+E)}(d) = (p_{f[M]} * p_E)(d) \tag{1.7}$$

see Grinstead and Snell [1998, Thm. 7.1]. The difficulty to find the PDF of $f[M]$ remains. There is a way around transforming M by f. As we already derived the distribution of G, marginalizing the joint PDF p_{DG} helps. Using the product rule (Eq. 1.2) yields

$$p_D(d) = \int p_{DG}(d, g)\,\mathrm{d}g = \int p_{D|G}(d \mid g) p_G(g)\,\mathrm{d}g = \int p_E(d - f[m]) p_G(h[m])\,\mathrm{d}m \tag{1.8}$$

requiring an integration which is not much of a relief. Depending on the prior chosen, there may not necessarily exist a closed form solution.

If those three PDFs are at hand, the density of G posterior to a set observations d is given by Bayes' law

$$p_{G|D}(g \mid d) = \frac{p_{D|G}(d \mid g)\, p_G(g)}{p_D(d)}\,. \tag{1.9}$$

If G is uni- or bivariate, visualising the posterior PDF is simple and the interpretation is intuitive. In case of high dimensional predictions a direct visualization of a posterior PDF with several variates is no longer possible. For interpretation, typically characteristics such as marginals and/or moments are requested. However – in most applications – calculating moments and/or marginals calls for multidimensional numeric integration.

To conclude, the outlined theory on Bayesian inverse problem is rather universal but the solution is quite demanding. In general the posterior density and its characteristics do not possess a solution of closed form. Various techniques and methods have been

[2] For a general discussion of sums of dependent random variables see e.g. Schmidt [2011, Sec. 13.5].

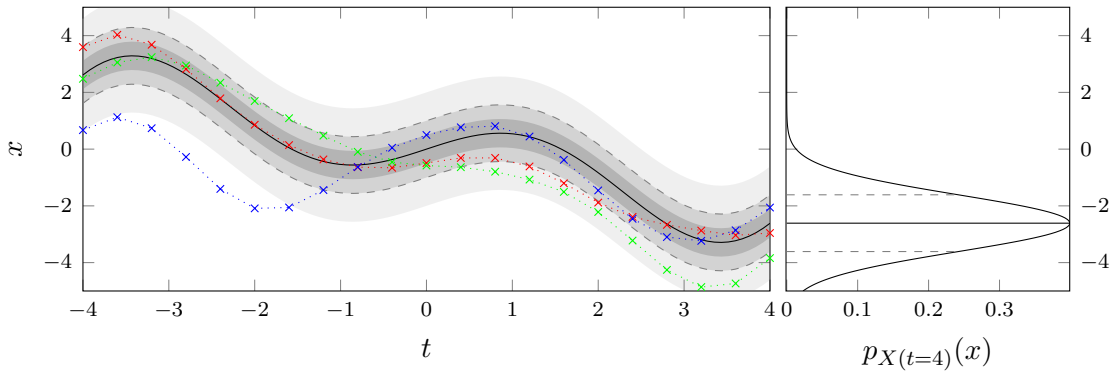

Figure 1.1.: A GP of mean $\mu_X(t) = t\cos(t)$ (——) and SE covariance (see Eq. 1.14 and Fig. 1.2). The gray shaded area illustrates half, once and twice the standard deviation. At equally space locations, realizations (·×·, ·×·, ·×·) are drawn from the corresponding MVN distribution.

developed determining posterior distributions, all suffering either from excessive sampling and/or numeric integration [Liu, 2008]. Nonetheless, with rigorous simplifications we can facilitate the Bayesian inverse problem. In order to find an algebraic solution, let surrender universality by making the following three assumptions: The setting is restricted to normal distributed errors, a normal distributed system and affine transformations.

1.2. Gaussian Process Regression

This section introduces distributions over functions. Although it might seem difficult to represent a distribution over a function, within a setting that preserves normality it turns out that the distribution needs to be known only at a finite, but arbitrary, set of points.

In many applications collections of RVs are of interest and inference shall take place directly in the space of functions – e.g. a system evolving over time. Thus, we are interested in a statistical distribution over functions

$$\{X(t) : t \in T\} \tag{1.10}$$

dependent on an index set T. The dependent variable t will typically denote a point in time, or space, or time and space. Based on their properties, stochastic processes can be divided into various categories. The class of Gaussian processes is one of the most widely used families of stochastic processes for modeling dependent data [Davis, 2014].

Roughly speaking, a Gaussian process (GP) may be understood as the infinite dimensional generalization of the Gaussian distribution. Each point within the potentially continuous input space is associated with a normal distributed random variable. This idea has already been used for a long time under the name *Kriging* in spatial interpolation, although it seems having ignored the generality of that method [Williams, 1997]. Historically, the dependent variable of a GP is time t whereas the extension of GPs to higher dimensions – e.g. $\mathbb{R}^3$ – is often referred to as a Gaussian random field (GRF). Nowadays, most authors are using the terms GP and GRF interchangeably. The following brief recap of GPs is based upon Gelman et al. [2013, Chp. 21], Rasmussen and Williams [2006, Chp. 2] and Murphy [2012, Chp. 15] from an applied point of view whereas Abrahamsen [1997] provides the background theory.

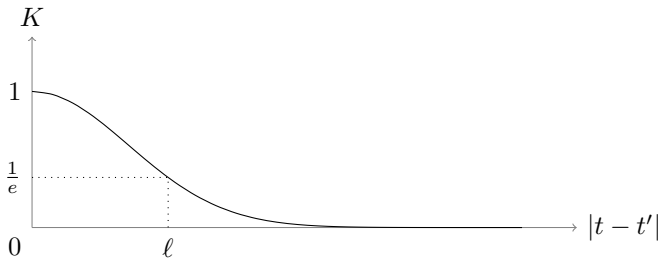

Figure 1.2.: The Gaussian kernel (Eq. 1.14) plotted as a function of distance.

Definition 1 (Gaussian Process)**.** *A time continuous stochastic process $X(t)$ is Gaussian if for every finite set of points $t_1, \ldots, t_n$ in the index set T*

$$\big(X(t_1), \ldots, X(t_n)\big)$$

is a multivariate normal distributed RV.

This is equivalent to saying any finite linear combination is normal distributed. The notation

$$X \sim \mathcal{GP}(\mu_X,\, K_X) \tag{1.11}$$

means that the random function $X(t)$ is distributed as a GP with mean function and covariance function

$$\mathbb{E}[X(t)] = \mu_X(t) \tag{1.12}$$
$$\mathrm{Cov}[X(t), X(t')] = K_X(t, t') \tag{1.13}$$

and we obviously require K_X to be a positive definite function (see Def. 3). The popularity of GPs stems primarily from the property that a GP is completely determined by its functions of first and second moment.

The key factor that controls the properties of a GP is the covariance function. It may be understood as a measure of how much two RVs change together and specifys the correlation between pairs of random variables. The smoothness of realizations of the GP and the shrinkage towards the mean is determined by the covariance function [Gelman et al., 2013, Sec. 21.1]. Selections of covariance functions, together with particular properties are discussed in Genton [2002], Murphy [2012, Chp. 14] or Rasmussen and Williams [2006, Chp. 4].

Example 1.1. A widely used covariance function is the squared exponential (SE) kernel

$$K(s,t) = \exp\left\{-\frac{|s-t|^2}{\ell^2}\right\} \tag{1.14}$$

of characteristic length-scale ℓ; also known as Gaussian kernel. Figure 1.2 illustrates K as a function of distance. An example GP is presented in Figure 1.1 . ◁

The definition of GPs implies that theorems on multivariate normal (MVN) distributions translate into the notion of GPs. References for MVN distributions are e.g. Gelman et al. [2013, p. 580] or Lindgren et al. [2013, A. 2] and in great depth Rao [1973, Ch. 8a], Muirhead [2005, Ch. 1] or Anderson [2003, Ch. 2].

Two properties that are reducing the dimensionality are of particular interest:

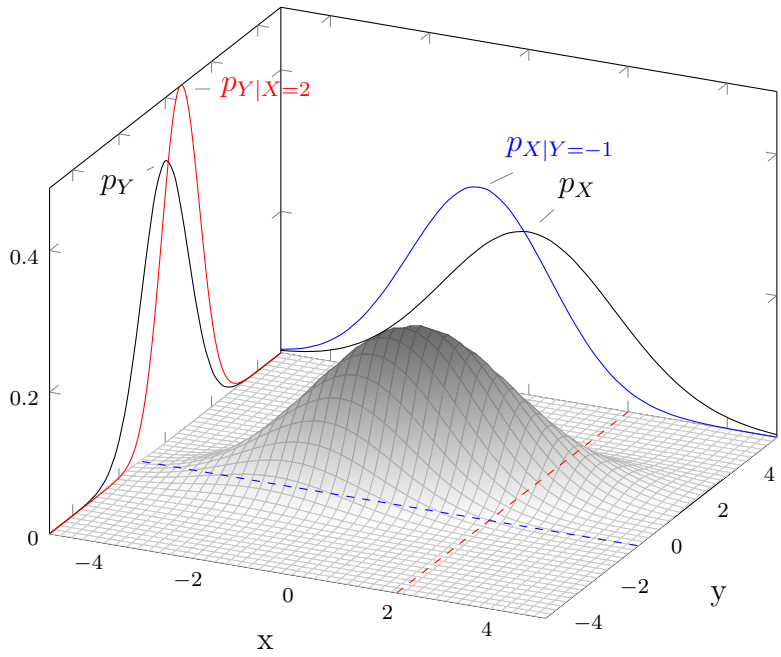

Figure 1.3.: Comparison of marginal and conditional densities for a centered bivariate normal distribution with variances $\sigma_X^2 = 3$, $\sigma_Y^2 = 1$ and covariance $\sigma_{XY} = 1$. Marginal densities are indicated by —— whereas conditional densities $X \mid Y = -1$ and $Y \mid X = 2$ are depicted by —— and ——. The surface plot is superelevated by factor 2.

Corollary 3 (Marginal)**.** *Let $X \sim \mathcal{GP}(\mu_X, K_X)$. If X is evaluated at a finite set of points $t_1, \ldots, t_n$, the resulting random vector is MVN distributed*

$$\boldsymbol{X} = \big(X(t_1), \ldots, X(t_n)\big) \sim \mathcal{N}_n(\boldsymbol{\mu}, \Sigma)$$

with mean $\boldsymbol{\mu} = \big(\mu_X(t_1), \ldots, \mu_X(t_n)\big)$ and covariance matrix $\Sigma = \{K_X(t_i, t_j)\}_{i,j=1,\ldots,n}$.

Marginalizing a GP is automatically implied by its definition. The conditional of a GP is the analog to the conditional of a MVN RV.

Corollary 4 (Conditional)**.** *Let $X \sim \mathcal{GP}(\mu_X, K_X)$. For a realization $X(s) = x_s$, the conditional distribution is a GP, as well. Mean and covariance read*

$$\begin{aligned} \mathbb{E}[X(r) \mid x_s] &= \mu_X(r) \quad + K_X(r,s)K_X^{-1}(s,s)\big(x_s - \mu_X(s)\big) \\ \operatorname{Cov}[X(r), X(t) \mid x_s] &= K_X(r,t) - K_X(r,s)K_X^{-1}(s,s)K_X(s,t)\,. \end{aligned}$$

A proof may be found in any of the aforementioned references. The conditional may be interpreted as a slice whereas the marginal is an average. The illustration of that interpretation is shown in Figure 1.3 comparing conditional and marginal densities. Speaking in terms of functions, the conditional distribution sorts out all those *incompatible* with the realization. Conditioning a GP is going to form the cornerstone of the modeling approach introduced in the following section that gives a Bayesian interpretation to Corollary 4.

1.2.1. Inference

In the following we discuss a method to perform Bayesian inference over functions themselves. The intention is to give a Bayesian interpretation to Corollary 4 i.e. the posterior

distribution of a GP. In the Bayesian formalism a prior distribution is specified, expressing our beliefs. Therefore, a GP is used to describe a distribution over functions

$$X \sim \mathcal{GP}(\mu_X,\, K_X) \tag{1.15}$$

with a priori mean function $\mu_X(t)$ and assumed covariance function $K_X(t, t')$. As evidence consider a finite set of realizations

$$x_s = \big\{(x_1, s_1), \ldots, (x_n, s_n)\big\} \tag{1.16}$$

with values x_i at times s_i, $i = 1, \ldots, n$. The knowledge that realizations provide about the function shall be incorporated to predict on the process at some test point r. According to the prior, the joint distribution of realizations and test points reads

$$\begin{pmatrix} X(s) \\ X(r) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_X(s) \\ \mu_X(r) \end{pmatrix}, \begin{bmatrix} K_X(s,s) & K_X(s,r) \\ K_X(r,s) & K_X(r,r) \end{bmatrix} \right) \tag{1.17}$$

with $s = (s_1, \ldots, s_n)$. To get the posterior distribution we need to restrict the joint prior distribution to solely contain those functions which are in agreement with realizations x_s. The posterior distribution is given by the conditional

$$X(r) \mid X(s) = x_s \tag{1.18}$$

and according to Corollary 4 the posterior is a GP as well. Conditional mean and covariance are given by

$$\mathbb{E}[\mu_X(r) \mid x_s] = \mu_X(r) \quad + K_X(r,s) K_X(s,s)^{-1} \big(x_s - \mu_X(s)\big) \tag{1.19}$$

$$\mathrm{Cov}[\mu_X(r), \mu_X(r') \mid x_s] = K_X(r,r') - K_X(r,s) K_X(s,s)^{-1} K_X(s,r') \,. \tag{1.20}$$

The posterior GP represents all those functions that are in agreement with the evidence x_s. This means that the modeling approach is precise at training points and satisfies the general interpolation problem. The prior mean function and covariance function are forming the basis of rules for interpolating values at points for which there are no observations. Interestingly, the posterior covariance does only depend on locations s but not on drawn values. That fact gives a powerful tool for survey design at hand. Further details on GP-regression may be found in Rasmussen and Williams [2006, Sec. 2.2 and 2.7] and – in the light of conjugate distributions – in Gelman et al. [2013, Chp. 21].

Example 1.2. The GP presented in Figure 1.1 serves to demonstrate inference using a single measurement. At location $s = -1$ the value recorded is

$$x_s = s \cos(s) = -\cos(-1) \,. \tag{1.21}$$

The covariance of choice is the SE kernel (Eq. 1.14) with $\ell = \sqrt{2}$. The a priori guess for the mean function reads

$$\mu_X(t) = -t \,. \tag{1.22}$$

According to Equations 1.19 and 1.20, the posterior process is given through

$$\mathbb{E}[X(r) \mid x_s] = -r \quad + \exp\big\{-\tfrac{1}{2}(r+1)^2\big\} \big(-\cos(-1) - 1\big) \tag{1.23}$$

$$\mathrm{Cov}[X(r), X(r') \mid x_s] = \exp\big\{-\tfrac{1}{2}(r-r')^2\big\} - \exp\big\{-(1+r)^2\big\} \,. \tag{1.24}$$

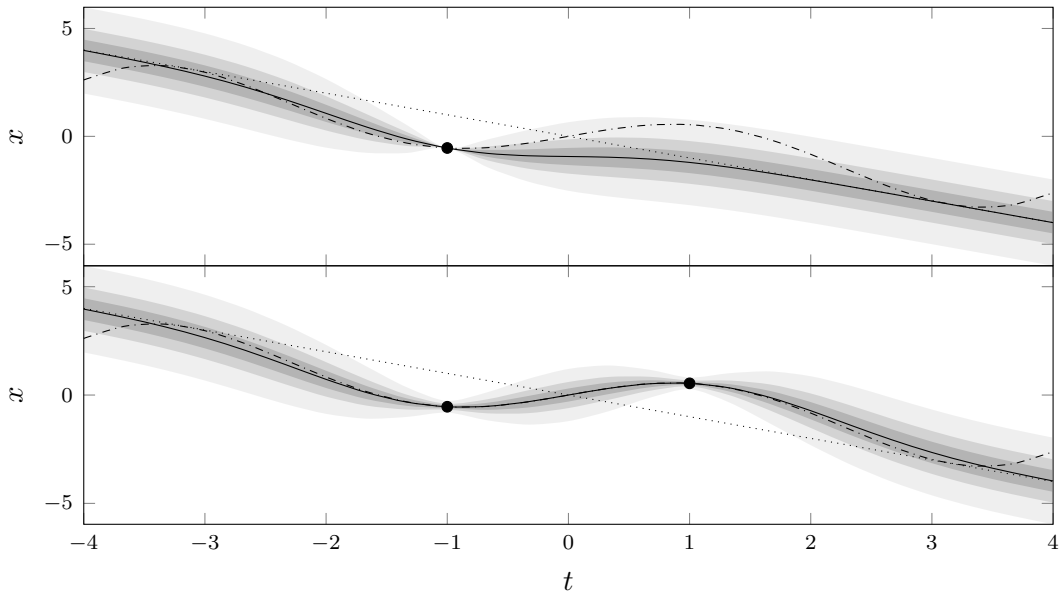

Figure 1.4.: The actual function $t\cos t$ (- - - -) is estimated by a posterior GP based upon precise evidence (•). The top panel considers single evidence $s_1 = -1$ whereas $s_2 = 1$ is added in the bottom panel. A priori parameters are mean $\mu_X(t) = -t$ (·····) and the SE kernel as covariance. The black solid line (——) indicates the posterior mean and the shaded gray area represent the 38%, 68% and 95% pointwise posterior credibility intervals.

The top panel of Figure 1.4 depicts the posterior process. At location $r = -1$, the prediction is precise. The standard deviation increases for input values that are distant from the record. The kernel's characteristic length scale determines the necking pattern. Apart from the data, the prior mean and standard deviation are reflected. ◁

Example 1.3. As a next step, let extend Example 1.2 by incorporating a second measurement at location $s_2 = 1$. The additional record to be worked in is

$$x_2 = s_2 \cos(s_2) = \cos(1) \; . \tag{1.25}$$

The a priori mean vector is determined by the prior mean function (Eq. 1.22)

$$\mu_X(s) = \begin{pmatrix} \mu_X(s_1) \\ \mu_X(s_2) \end{pmatrix} = \begin{pmatrix} -s_1 \\ -s_2 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \end{pmatrix} \; . \tag{1.26}$$

Evaluating the SE kernel (Eq. 1.14) at combinations of s_1 and s_2 results in

$$K_X(s,s) = \begin{bmatrix} K_X(s_1,s_1) & K_X(s_1,s_2) \\ K_X(s_2,s_1) & K_X(s_2,s_2) \end{bmatrix} = \begin{bmatrix} 1 & e^{-2} \\ e^{-2} & 1 \end{bmatrix} \tag{1.27}$$

and the correlations amongst design point r and records s are

$$K_X(r,s) = \begin{bmatrix} K_X(r,s_1) \\ K_X(r,s_2) \end{bmatrix} = \begin{bmatrix} \exp\{-\frac{1}{2}(r+1)^2\} \\ \exp\{-\frac{1}{2}(r-1)^2\} \end{bmatrix} \; . \tag{1.28}$$

According to Equations 1.19 and 1.20, the posterior mean and covariance are given by

$$\mathbb{E}[X(r)|x_s] = -r + \begin{bmatrix} K_X(r,-1) \\ K_X(r,1) \end{bmatrix} \begin{bmatrix} 1 & e^{-2} \\ e^{-2} & 1 \end{bmatrix}^{-1} \begin{pmatrix} -\cos(-1) - 1 \\ \cos(1) + 1 \end{pmatrix} \quad (1.29)$$

$$\mathrm{Cov}[X(r), X(r')|x_s] = K_X(r,r') - \begin{bmatrix} K_X(r,-1) \\ K_X(r,1) \end{bmatrix} \begin{bmatrix} 1 & e^{-2} \\ e^{-2} & 1 \end{bmatrix}^{-1} \begin{bmatrix} K_X(r,-1) \\ K_X(r,1) \end{bmatrix}^\top . \quad (1.30)$$

The bottom panel of Figure 1.4 depicts the posterior process. In analogy with Example 1.2 the pointwise posterior standard deviation shows a second constriction at s_2. ◁

The difficulty to perform regression with GPs is to determine a suitable covariance function. In many applications heuristics is addressed since the covariance may be considered as a rule for interpolation. As an application, Chapter 4 demonstrates how the SE kernel may be used for travel-time tomography. The theory for a systematic alternative that derives covariance functions from first principles is presented in Chapter 2. An example is the inference of the driving force of a vibrating string that is shown in Chapter 5 where the equation of motion serves to derive the covariance function. An apprlication is presented in Chapter 6 that addresses the modeling of the Earth's magnetic field with a covariance function derived from generalized geomagnetic energies.

1.3. Gaussian Linear Model

During the discussion of Bayesian inference several difficulties arose mainly due to the algebra and transformation of RVs. At the cost of loosing generality, these difficulties can be merged by normality assumptions and the restriction to affine transformations. The restriction to GPs and affine maps is an excessive limitation, however, comes with the advantage that computations required for inference are significantly lightened [Calvetti and Somersalo, 2007, Sec. 1.3].

Normal distributions and GPs are perhaps the most important distributions in probability and mathematical statistics, primarily because of the central limit theorem. Gaussian distributions are widely used to model physical measurements of all types that are subject to small, random errors [Siegrist, 2015, Sec. 2.7]. The central limit theorem justifies in many circumstances the use of Gaussian approximations which is why normally distributed RVs play a special role in statistics [Calvetti and Somersalo, 2007, Sec. 1.3].

The linear Gaussian model is the concept performing regression with GPs in combination with affine maps. The analog to affine transformations of normal RVs reads:

Corollary 5. *Let $X \sim \mathcal{GP}(\mu_X, K_X)$ and f an affine map. The transformation*

$$Y = f[X] \sim \mathcal{GP}(\mu_Y, K_Y)$$

is Gaussian with mean function and covariance function

$$\mu_Y = f\mu_X \,, \qquad K_Y = \tilde{f} K_X \tilde{f}^\dagger$$

where f's linear part[3] is denoted by $\tilde{f}$ and the $\dagger$-superscript indicates that the functional acts to the left i.e. with respect to the second entry of K_X.

[3] An affine map is composed of a linear and a translational part. When calculating the co-variance of an affine map the translational part drops out.

In other words, if a GP is passed through an affine transformation, the output is a GP as well. The prove may be found in Abrahamsen [1997]. Still, consider X a GP with mean μ_X and covariance K_X. Just to name three, examples for affine maps of GPs are:

Example 1.4 (Gradient)**.** Let $f[X] = \nabla X$. Then $\nabla X \sim \mathcal{GP}\big(\nabla\mu_X,\, \nabla K_X \nabla^\dagger\big)$ where $\nabla^\dagger$ acts on the second argument i.e. $\nabla K_X \nabla^\dagger(t,t') = \nabla_t \nabla_{t'} K_X(t,t')$. ◁

Example 1.5 (Integration)**.** Let $f[X] = \int_a^b X \,\mathrm{d}t$. Then $f[X]$ is normal distributed with mean $\mu_f = \int_a^b \mu_X(t)\,\mathrm{d}t$ and co-variance $K_f = \iint K_X(t,t')\,\mathrm{d}t\,\mathrm{d}t'$. A typical cases of application are inner products and the expansion w.r.t. some basis. ◁

Example 1.6 (Indicator)**.** Let $f[X] = \mathbf{1}_A[X]$ considered as an operator in X. Then, the resulting mean is $\mu_f = \mathbf{1}_A[\mu_X]$ and the covariance $K_f = \mathbf{1}_A\, K_X\, \mathbf{1}_A^\dagger$. An application is the restriction to a compact domain. ◁

Let us come back to the physical setting introduced in Section 1.1. The concept of Gaussian process regression and the idea making predictions with an underlying system subject to observations shall be brought together. To do so, we have to restricting ourselves to affine maps and normality assumptions. To calculate the posterior distribution, five quantities are to be determined. Given an a priori model, these are: The data's prior mean vector and covariance matrix, correlations amongst predictions and observations and the predictive prior mean and covariance.

As we do not know the underlying **model**, it is considered a random quantity. A priori, the underlying system M is assumed a GP

$$M \sim \mathcal{GP}(\mu_M,\, K_M) \tag{1.31}$$

with a priori mean function μ_M expressing the knowledge we already have without looking at the data and to be known covariance function K_M. Except for the error model, M is the actual random quantity we are looking at whereas measurements D and predictions G are borrowing their statistics through affine transformations.

Measurements are described through an affine map $f[\cdot]$, the observational functional. Recorded data are corrupted by measurement noise and the data model reads

$$D = f[M] + E\,. \tag{1.32}$$

Measurement errors E are assumed to be zero mean Gaussian with co-variance Σ_E. Due to Corollary 5, $f[M]$ is normal. Since linear combinations of independent Gaussian RVs are normal, the data distribution reads

$$D \sim \mathcal{N}\Big(f\mu_M,\, \tilde{f} K_M \tilde{f}^\dagger + \Sigma_E\Big) \tag{1.33}$$

where $\tilde{f}$ refers to the linear part of the affine map. Typically a set of n measurements is recorded and the data distribution is given by its mean vector and covariance matrix

$$\mathbb{E}[D] = \begin{pmatrix} f_1\mu_M \\ \vdots \\ f_n\mu_M \end{pmatrix}\,, \qquad \mathbb{V}[D] = \begin{bmatrix} \tilde{f}_1 K_M \tilde{f}_1^\dagger & \dots & \tilde{f}_1 K_M \tilde{f}_n^\dagger \\ \vdots & \ddots & \vdots \\ \tilde{f}_n K_M \tilde{f}_1^\dagger & \dots & \tilde{f}_n K_M \tilde{f}_n^\dagger \end{bmatrix} + \Sigma_E\,. \tag{1.34}$$

Measurement errors do not have to be independent and/or homoscedastic. Nonetheless, the independence of $f[M]$ and E is still assumed [4].

Predictions are treated similarly to measurements. The predictive functional $h[\cdot]$ is restricted to affine transformations. Due to Corollary 5, predictions $G = h[M]$ are normal distributed with mean and co-variance

$$\mathbb{E}[G] = h\mu_M \qquad \text{and} \qquad \mathbb{V}[G] = \tilde{h}K_M\tilde{h}^\dagger \tag{1.35}$$

where $\tilde{h}$ again refers to the linear part of h. In case of multiple predictions – say k – the mean vector and covariance matrix construct analogously to Equation 1.34 (except for the error term).

To adopt inference with GPs (see Sec. 1.2.1) **correlations** amongst predictions and observations need to be known. Predictions G and observations D are correlated through

$$\mathrm{Cov}[h[M], f[M]] = \tilde{h}K_M\tilde{f}^\dagger = \begin{bmatrix} \tilde{h}_1K_M\tilde{f}_1^\dagger & \dots & \tilde{h}_1K_M\tilde{f}_n^\dagger \\ \vdots & \ddots & \vdots \\ \tilde{h}_kK_M\tilde{f}_1^\dagger & \dots & \tilde{h}_kK_M\tilde{f}_n^\dagger \end{bmatrix} \tag{1.36}$$

a $k \times n$ cross-covariance matrix. Again, k refers to the number of predictions and n to the number of measurements. Any dependence amongst system and measurement noise is neglected although the residual term potentially accounts for incompleteness of theory.

The **posterior distribution** of G knowing d is normal. According to Equations 1.19 and 1.20 the conditional mean and covariance are given by

$$\mathbb{E}[G \mid d] = h\mu_M \quad + \tilde{h}K_M\tilde{f}^\dagger \left[\tilde{f}K_M\tilde{f}^\dagger + \Sigma_E\right]^{-1} \left(d - f\mu_M\right) \tag{1.37}$$

$$\mathbb{V}[G \mid d] = \tilde{h}K_M\tilde{h}^\dagger - \tilde{h}K_M\tilde{f}^\dagger \left[\tilde{f}K_M\tilde{f}^\dagger + \Sigma_E\right]^{-1} \tilde{f}K_M\tilde{h}^\dagger \,. \tag{1.38}$$

The posterior distribution in the linear Gaussian model possesses particularly nice properties. Due to measurement noise, the model is *not* required to perfectly interpolate the data. The posterior model will be smooth modulo error-level and kernel regularity. In contrast to Equation 1.9, not only the posterior PDF is right at hand but also first and second moment, given by algebraic equations. Marginalization is straightforward by considering sub-vectors and sub-matrices of conditional mean and covariance. If Equations 1.37 and 1.38 are implemented through Cholesky factorization the solution is backward stable at the machine's precision [Trefethen and Bau, 1997, Thm. 23.3]. The computational complexity of commonly used algorithms is $O(n^3)$ where n is the size of the data's covariance matrix [Rasmussen and Williams, 2006, Alg. 2.1]. This is inexpensive compared to high dimensional numeric integration showing exponential complexity in the widest sense (see Fig. 3.3).

Example 1.7. Let us revisit Example 1.2 . Still, the true, to be modeled function is

$$\tilde{m}(t) = t\cos t \,. \tag{1.39}$$

The a priori guess for mean function remains

$$\mu_M(t) = -t \tag{1.40}$$

[4] Although joint normality of M and E were sufficient.

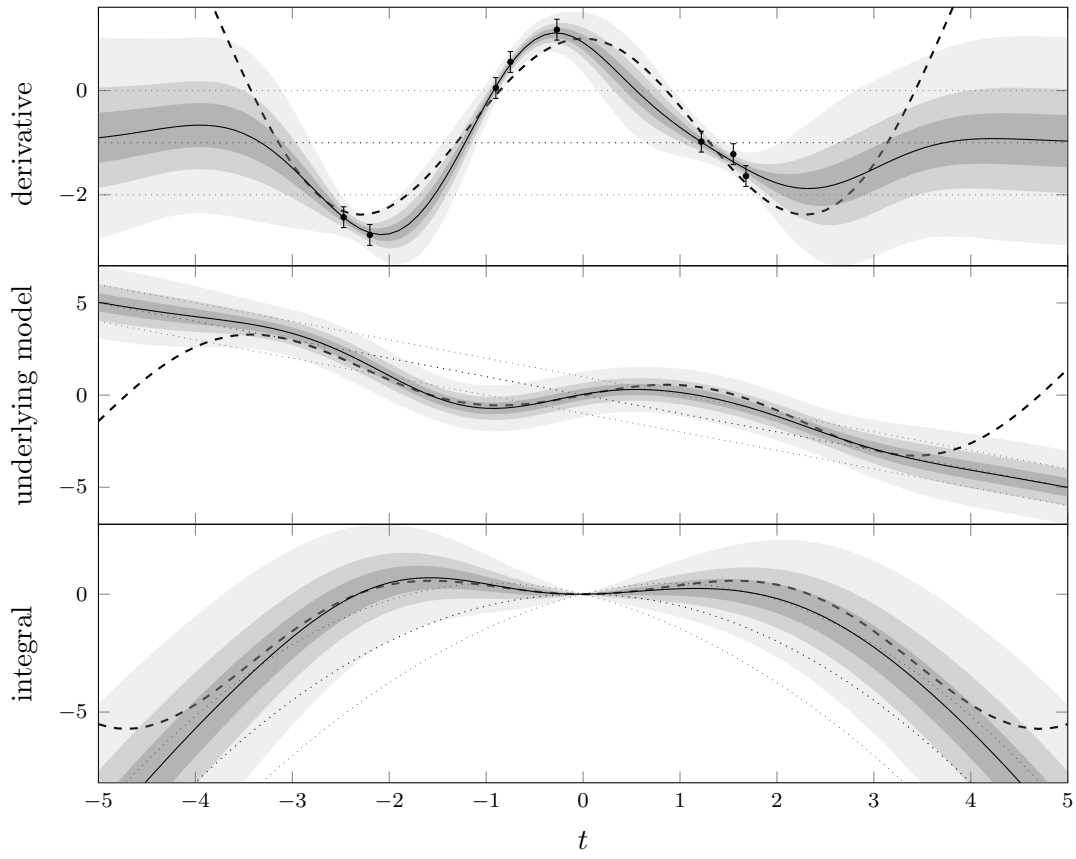

Figure 1.5.: GP regression with observational and predictive functionals. Observations are indicated by • in the top panel. A priori assumptions are shown by ····· (mean ± standard deviation) whereas actual curves are plotted by - - -. The posterior mean is depicted by —— and the gray shaded area represents the 38%, 68% and 95% pointwise confidence intervals.

and the SE kernel serves as covariance. In contrast, $\tilde{m}$ is not directly acquired but the pointwise derivative is recorded. The $n = 8$ records are corrupted by centered normal noise. With the observational functional $f[M] = \partial_t M$ the data model becomes

$$D_i = \partial_t M(t)\big|_{t_i} + E\,, \qquad i = 1, \ldots, n \tag{1.41}$$

with independent zero mean normal noise of standard deviation $\sigma_E = 0.20$. The data's a priori mean vector and covariance matrix are

$$f[\mu_M] = \left\{ \partial_t \mu_M(t_i) = -1 \;\middle|\; t \in T \right\} \tag{1.42}$$

$$fK_M f^\dagger = \left\{ \partial_s \partial_t K_M(s,t) = K_M(s,t)\left(1 - (s-t)^2\right) \;\middle|\; t, s \in T \right\} + \mathbb{I}\sigma_E^2\,. \tag{1.43}$$

The evidence considered is depicted in the top panel of Figure 1.5 . On purpose, locations T are arranged unevenly.

To demonstrate the flexibility of the method, we predict on the three different quantities. First of all, let us have a look at the **gradient** of the model. Thus, the predictive functional reads

$$h[M] = \partial_t M(t) \tag{1.44}$$

Correlations as well as the a priori covariance are given by Equation 1.43 . Conditional mean and covariance are given by Equations 1.37 and 1.38. The reconstruction of the

gradient is depicted in the top panel of Figure 1.5 . For comparison, the actual derivative

$$h[\tilde{m}](t) = \partial_t\big(t\cos t\big) = \cos t - t\sin t \tag{1.45}$$

is shown as well. The choice of a constant prior mean is pathetic as we a priori neglect any dynamics of the system. It is clear that the posterior distribution will be poor distant from evidence. Nonetheless, in a close neighbourhood to measurements reconstruction is reasonable, provided the assumed covariance is appropriate for interpolation.

As a next step we predict on the **underlying model** at its own. The predictive function is basically point evaluation

$$h[M] = M(s)\ . \tag{1.46}$$

Correlations amongst observations and predictions are given by

$$hK_M f^\dagger = \partial_t K_M(s,t) = (s-t)K_M(s,t) \tag{1.47}$$

and we explicitly know all quantities needed to calculate the posterior distribution of M (again Eqs. 1.37 and 1.38). The mid panel of Figure 1.5 shows the reconstruction of M. For comparison the true function (Eq. 1.39) is shown as well. The characteristic necking of the reconstruction is less prominent which is due to the $(s-t)$ term in Equation 1.47 . In other words, precisely at points of observation the kernel does not feature any correlations. Nonetheless, a clear reduction of the prior variance in the neighbourhood of observations is apparent. A noteworthy effect when observing derivatives is that any constant offset drops out. If observations were either clustered to the left or right of the origin the reconstruction will show an offset.

Finally, the predictive functional will be a **definite integral** and reads

$$h[M] = \int_0^s M(s')\,\mathrm{d}s' \tag{1.48}$$

considered as a function in s, the upper integration bound. The antiderivative of M's a priori mean is

$$\mu_{h[M]}(t) = \int_0^t -t'\,\mathrm{d}t' = -\tfrac{1}{2}t^2\ . \tag{1.49}$$

Correlations amongst predictions and observations calculate as follows

$$hK_M f^\dagger = \int_0^s \partial_t K_M(s',t)\,\mathrm{d}s' = -\int_0^s \partial_s K_M(s',t)\,\mathrm{d}s' = K_M(0,t) - K_M(s,t) \tag{1.50}$$

and the covariance for the predictive functional reads

$$\begin{aligned} hK_M h^\dagger &= \sqrt{\frac{\pi}{2}}\left[\int_0^s \operatorname{erf}\left(\frac{s'}{\sqrt{2}}\right)\,\mathrm{d}s' - \int_0^s \operatorname{erf}\left(\frac{s'-t}{\sqrt{2}}\right)\,\mathrm{d}s'\right] = \\ &= \sqrt{\tfrac{\pi}{2}}\left[s\operatorname{erf}\left(\tfrac{s}{\sqrt{2}}\right) + t\operatorname{erf}\left(\tfrac{t}{\sqrt{2}}\right) - (s-t)\operatorname{erf}\left(\tfrac{s-t}{\sqrt{2}}\right)\right] + K_M(s,0) + K_M(0,t) - 1 - K_M(s,t) \end{aligned} \tag{1.51}$$

where $\operatorname{erf}(\cdot)$ refers to the error-function. Ingredients for the calculus – the antiderivatives of SE kernel and error function – may be found in Bronstein et al. [2000, Eqs. 8.100a and 8.100d]. All quantities to calculate the posterior distribution are explicit and again

conditional mean and covariance are given through Equations 1.37 and 1.38. The bottom panel of Figure 1.5 presents the reconstruction of $h[M]$. For comparison the actual function

$$h[\tilde{m}] = \cos(t) + t\sin(t) - 1 \tag{1.52}$$

is shown as well. Evaluated at the origin ($s = 0$) the predictive functional is certainly zero. Even with a record at $t = 0$ there is no uncertainty as the predictive functional is precise at that location. The evidence we have is shifting the posterior mean towards $\tilde{m}$, however, off the origin no significant reduction of uncertainty occurs. Furthermore, it is clear that there is no necking close to observations any more. Any offset in M will be passed through the integral with consequences suffering from a quadratic effect and predictions will be totally off. ◁

Although, the restriction to affine maps is severe the Linear Gaussian model provides a powerful and flexible tool for uncertainty assessment. To obtain a reasonable posterior distribution there are two contradicting cases: Either the inversion is driven by strong data swapping vague prior assumptions. Or weak data are supported by particularly good a priori assumptions. A typically application will show something in between. Motivated by the central limit theorem, the Gaussian assumption is of moderate concern. Nonetheless, a restriction to affine maps is hardly satisfying. Section 3 is about approximations of non-linear functionals trying to increase generality while preserving the computational convenience.

2. Reproducing Kernel Hilbert Spaces

Until here inference with GPs was introduced from a purely statistical point of view. During the discussion, the covariance function remained vague and was only required to be symmetric and positive definite. However, the discussion left out the selection of covariance functions and attributed characteristics. The objective of the following is to establish a link between GP regression and the theory of reproducing kernel Hilbert spaces (RKHS). This connection will serve to either derive covariance functions from first principles or at least to analyze implied regularity. Changing the perspective from GP regression to RKHSs is beneficial as a *proper* Hilbert space defines structures relevant for solving inverse problems.

The following is a partial and brief review on RKHS theory summarizing key properties and consequences. To keep matters concise, technical details – such as existence, completeness and separability – are omitted. If provided, any proofs are just roughly sketched. A rigorous discussion may be found in e.g. Aronszajn [1950], Wahba [1990], Berlinet and Thomas-Agnan [2004] and Steinwart and Christmann [2008] in exhaustive depth. The reader is assumed to know basic facts about Hilbert space theory and functional analysis such as the Riesz lemma and the notion of bounded linear functionals. This background may be found in numerous books, e.g. Reed and Simon [1981].

Defined on a continuous index set T, consider a potentially infinite dimensional Hilbert space $\mathcal{H}$ of functions equipped with inner product $\langle \cdot, \cdot \rangle$ and introduced norm $\| \cdot \|$. For the following, all functions and RVs are assumed real valued. Hilbert spaces that satisfy certain *additional* properties are known as RKHSs.

Definition 2 (RKHS)**.** *$\mathcal{H}_K$ is called a RKHS if there exists a function $K : T \times T \to \mathbb{R}$ with the following properties:*

1. *For every t, $K_t = K(t, \cdot)$ belongs to $\mathcal{H}_K$.*
2. *K has the reproducing property $f(t) = \langle f, K_t \rangle_K$.*

K is called the reproducing kernel (RK) belonging to $\mathcal{H}_K$.

Since $K_s, K_t \in \mathcal{H}_K$ we have $K(s,t) = \langle K_s, K_t \rangle_K$ which is where the RK takes its name from. Having a single definition covering both RKHS and RK is pragmatic to shorten matters. Most authors, however, are using Theorem 7 as the defining property.

Theorem 6. *The RK that belongs to a RKHS is unique.*

If there were two RKs K and R, then $0 = f(t) - f(t) = \langle f, K_t \rangle - \langle f, R_t \rangle$, $\forall f$ and $\forall t$. Choosing $f = K_t - R_t$ yields $\|K_x - R_x\| = 0$ and thus $K = R$.

Evaluation functionals play a crucial role in RKHS theory.

Theorem 7. *$\mathcal{H}$ has a RK if and only if all evaluation functionals are linear and bounded.*

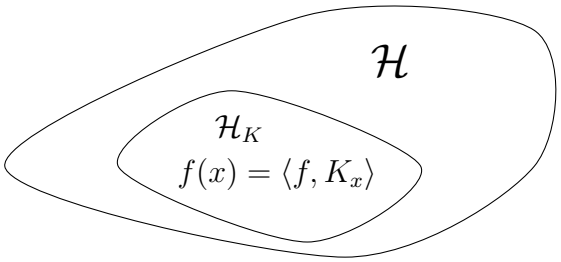

Figure 2.1.: Continuous evaluation can be interpreted as the restriction of a general Hilbert space with the additional condition of pointwise evaluation.

If $\mathcal{H}$ has a RK, the linearity of evaluation is governed by the linearity of the inner product and by the Cauchy-Schwarz inequality

$$|f(t)| = |\langle f, K_t \rangle| \leq \|f\| \|K_t\| \tag{2.1}$$

evaluation is bounded. Conversely, according to Riesz's lemma we have

$$f(t) = \langle f, L_t \rangle \tag{2.2}$$

since evaluation is a bounded and linear functional. Defining $K(s,t) = \delta_s[L_t] = \langle L_s, L_t \rangle_K$ is sufficient for the defining properties.

Continuous evaluation – as we will see in Example 2.1 – does not hold for Hilbert spaces in general. In the *smoother* RKHSs the kernel itself is the representer of evaluation and is an element of the RKHS.

Theorem 8. *Norm convergence in $\mathcal{H}_K$ implies pointwise convergence.*

If a series f_n converges to f in the RKHS norm, then

$$|f_n(t) - f(t)| = |\langle f_n - f, K_t \rangle| \leq |\|f_t - f\|_K \| K_t \| \|_K \to 0 \, . \tag{2.3}$$

In other words, when two functions are identical in the RHKS norm, they agree at every point. It is remarkable that norm convergence – as a *global* property – unveils local behavior.

Example 2.1. The Hilbert space L_2 of square integrable functions is not a RKHS. In L_2 the Dirac delta is the representer of evaluation. The Dirac delta is not an element of L_2 and evaluation is not continuous since

$$\lim_{n \to \infty} \delta_x[f_n] \neq \delta_x[\lim_{n \to \infty} f_n] \quad . \tag{2.4}$$

The elements of L_2 are equivalent classes. Loosely speaking, there are elements $f, \tilde{f}$ such that $f \neq \tilde{f}$ but $\|f\|_{L_2} = \|\tilde{f}\|_{L_2}$. The condition of square-integrability is not *strong* enough to make L_2 a RKHS. ◁

In RKHS theory kernels are the analogue to what the Dirac delta is for the L_2. A RKHS has properties which make it *well behaved* relative to a general Hilbert space. The characteristic property of a RKHS is that functions are pointwise defined and evaluation is represented by the kernel. In Definition 2, however, it remained vague what a RK is. The most important property is positive definiteness:

Definition 3 (Positive definite function)**.** *A symmetric function $K : T \times T \to \mathbb{R}$ is called a positive definite (PD) function if*

$$\sum c_i K(x_i, x_j) c_j \geq 0$$

holds for any $n \in \mathbb{N}$, $x_1, \cdots, x_n \in T$ and $c_1, \cdots, c_n \in \mathbb{R}$.

The concept of PD functions is closely related to positive semi-definite matrices.

Theorem 9. *Every RK belonging to a RKHS is a PD function.*

To check if the bi-linear form associated with Definition 3 is an inner product is an often used method to proof positive definiteness. Let K denote the RK, then

$$\sum c_i K(x_i, x_j) c_j = \langle \sum c_i K_{x_i}, \sum c_j K_{x_j} \rangle_K \geq 0 \,. \tag{2.5}$$

Example 2.2. In the same way it can be shown that

$$K(s,t) = \min(s,t) \tag{2.6}$$

defined on $s, t \in \mathbb{R}_+$ is PD. A plot of $K_s(t)$ is shown in the mid panel of Figure 2.2. For the derivative we have $\partial_t K(s,t) = \mathbf{1}_{0 \leq t \leq s}(t)$. The kernel may be expressed in terms of an inner product

$$K(s,t) = \int_{\mathbb{R}_+} \dot{K}_s(u) \dot{K}_t(u) \,\mathrm{d}u = \int_{\mathbb{R}_+} \mathbf{1}_{\{0 \leq u \leq s\}}(u)\, \mathbf{1}_{0 \leq u \leq t}(u) \,\mathrm{d}u \tag{2.7}$$

which is easily confirmed by standard calculus. This kernel is known to be the covariance of the Wiener process, a model for Brownian motion [Rios et al., 2012, Sec. 6.3]. ◁

The covariance being a PD function too, establishes a first link with GPs. For now, this connection remains just as a coincidence. Section 2.2, however, is going to introduce a duality between GPs and RKHSs.

Compared with Theorem 9 the following theorem goes in the other direction and constructs a RKHS from a given PD function:

Theorem 10 (Moore-Aronszajn)**.** *To every PD function P there corresponds a unique RKHS with RK P.*

The proof – details including uniqueness may be found in Berlinet and Thomas-Agnan [2004, Sec. 1.3] – is based on what is known under the term *kernel map construction*. Let P be a PD function and consider the space of all kernel expansions defined by

$$\mathcal{H}_P = \left\{ f(s) = \sum a_i P(s, t_i) \,\middle|\, t_i \in T,\, a_i \in \mathbb{R} \right\} \,. \tag{2.8}$$

Endowed with the inner product

$$\langle f, g \rangle_P = \sum_{i,j} a_i P(t_i, t_j) a_j \tag{2.9}$$

and by adjoining the limits of all Cauchy sequences, $\mathcal{H}_P$ forms a Hilbert space. Condition 1 of Definition 2 – $P(t_i, \cdot) \in \mathcal{H}_P$ – is clearly fulfilled under the kernel map construction. Condition 2 – the reproducing property – is checked by

$$\langle f, P_t \rangle_P = \langle \sum f_i P_i, P_t \rangle_P = \sum f_i \, P(s_i, t) = \sum f_i \, P(s_i, t) = f(t) \tag{2.10}$$

and boundedness directly follows from the Cauchy-Schwarz inequality. Thus, $\mathcal{H}_P$ is the RKHS and elements may be expressed in terms of a kernel expansion. In other words, there is only one PD function namely the RK and the notions of PD functions and RKs are in fact the same.

Since RKs are just PD functions it is intuitive that new RKs can be formed by combining existing PD functions. The simplest operation is scaling by positive scalars. Sums of PD functions are RKs, too. The difference of RKs is, however, not necessarily a RK since it may happen that $K_1(t,t) - K_2(t,t) < 0$. Products of PD functions are also RKs. Further details are given in the aforementioned textbooks.

Example 2.3 (Polynomial kernels)**.** Consider the space of polynomials of degree at most n with some basis $\big(b_0(x), \ldots, b_n(x)\big)$ not necessarily an ONB. Equipped with the inner product defined by

$$\langle g, h \rangle = \sum g_i h_i \; , \tag{2.11}$$

where g_i and h_i are the coefficient w.r.t. $b_i(x)$, this is a RKHS and the RK is given by

$$K(x, y) = \sum b_i(x) \, b_i(y) \; , \tag{2.12}$$

called *a* polynomial kernel. For any x fixed, it is clear that $K_x = K(x, \cdot)$ is a polynomial with coefficients $b_i(x)$ and belongs to the space. The reproducing property follows since

$$\langle f, K_x \rangle = \sum f_i b_i(x) = f(x) \tag{2.13}$$

where the f_i are again the coefficients w.r.t $b_i(x)$ and holds for any polynomial f.
It is noteworthy that the space of polynomials is spanned by various bases, e.g. monomials x^i or a Taylor basis about zero $\frac{x^i}{i!}$. However, the RKs that are constructed according to Equation 2.12 and according inner product are in general *not* identical. To see this, for a polynomial f let m_i be coefficients w.r.t. the monomials. Then $t_i = i! f_i$ are the coefficients w.r.t the Taylor type basis. The RKHS norms are not identical $\|f\|_M = \sum m_i^2$ whereas $\|f\|_T = \sum t_i^2 = \sum (i! m_i)^2$. ◁

From the above example (Ex. 2.3) we can deduce that every finite dimensional Hilbert space possesses a RK.

Example 2.4 (Exponential kernels)**.** The polynomial kernels presented in Example 2.3 may be understood as the sum of individual RKs. Interpreted as a power series with appropriate scaling gives raise to the exponential kernel

$$\sum \frac{s^k}{\sqrt{k!}} \frac{t^k}{\sqrt{k!}} \to \exp\{st\} \tag{2.14}$$

and also scaling of the arguments is straight forward. ◁

With the same arguments given in Examples 2.3 and 2.4, kernels may be formed from all functions which have a Taylor series of non-negative coefficients [Steinwart and Christmann, 2008, Lma. 4.8]. Although a sufficient condition, the following example shows that non-negative Taylor coefficients are not necessary.

Example 2.5 (SE Kernel)**.** The squared exponential kernel (Ex. 1.1) can be constructed with the above principles. Taking the product of exponential kernels (Ex. 2.4) with scaled arguments, the SE kernel is formed by

$$K_{\mathrm{SE}}(s,t) = \exp\{-\sigma s^2\}\exp\{-\sigma t^2\}\exp\{2\sigma st\} = \exp\{-\sigma(s-t)^2\} \tag{2.15}$$

and the Taylor series coefficients are not strictly positive but alternating. ◁

Since it is possible to form all sorts of different kernels, the question that raises is which properties do the according RKHS possess? An additional view of RKHSs may be obtained by characterizing the RK in terms of an integral operator and Mercer's theorem. Therefore, regard the kernel K as a bounded and linear integral operator

$$\mathrm{T}_K[f](\cdot) = \int K(\cdot, s) f(s)\,\mathrm{d}\mu(s) \tag{2.16}$$

and assume f real-valued potentially ranging through L_2. Since T_K is a linear and bounded operator, by the spectral calculus the notion of eigenvalues and eigenfunctions are at hand. A function ϕ_i that obeys the integral equation

$$\mathrm{T}_K[\phi_i](t) = \langle K_t, \phi_i\rangle_2 = \lambda_i \phi_i(t) \tag{2.17}$$

is called an eigenfunction of the kernel K with eigenvalue λ_i. The eigenfunctions are orthogonal with respect to the measure and are normalized such that $\langle \phi_i, \phi_j\rangle_2 = \delta_{ij}$. Mercer's theorem states that the spectral decomposition of the integral operator T_K yields a series representation of K in terms of T_K's eigenvalues and eigenfunctions.

Theorem 11 (Mercer)**.** *Let (T, μ) be a finite measure space and K be a* continuous *PD function. Then there exists a sequence of eigenvalues $\lambda_i > 0$ and normalized eigenfunctions ϕ_i of T_K such that*

$$K(s,t) = \sum \lambda_i \phi_i(x)\phi_i(y)$$

and the series convergence is absolute and uniform.

Mercer's theorem establishes a link between the RKHS inner product and the L_2 inner product. Consider a RK K with an eigenfunction expansion according to Theorem 11. Construct a Hilbert space comprised of linear combinations of the eigenfunctions

$$\left\{ f \in L_2 \;\middle|\; \sum \frac{\langle f, \phi_i\rangle_2}{\lambda_i} < \infty \right\} \tag{2.18}$$

with inner product

$$\langle f, g\rangle = \sum \frac{\langle f, \phi_i\rangle_2 \langle g, \phi_i\rangle_2}{\lambda_i}\,. \tag{2.19}$$

This Hilbert space is the RKHS corresponding to the RK. To see this, both conditions of Definition 2 must be met. Since $\lambda_i\phi_i(t)$ are the coefficients for K_t we have

$$\langle f, K_t\rangle = \sum \frac{\langle f, \phi_i\rangle_2 \langle K_t, \phi_i\rangle_2}{\lambda_i} = f(t) \tag{2.20}$$

$$\langle K_s, K_t\rangle = \sum \frac{\lambda_i\phi_i(t)\,\lambda_i\phi_i(t)}{\lambda_i} = K(s,t) \tag{2.21}$$

where Equation 2.17 was used. Due to uniqueness of the RK, this Hilbert space is in fact the RKHS. As already pointed out in Example 2.1, L_2 is not a RKHS in general, but for many kernels it contains the RKHS as subspace (see Figure 2.1).

Until here two particular bases of a RKHS were introduced. Any $f \in \mathcal{H}_K$ may be expressed through eigenfunctions of the RK as well as kernel expansion

$$f(t) = \sum f_i K(s_i, t) = \sum \tilde{f}_i \phi_i(t)\ . \tag{2.22}$$

In case of a kernel expansion the RKHS norm reads

$$\|f\|_K^2 = \sum f_i K(x_i, x_j) f_j = \ldots \tag{2.23}$$

whereas as an expansion in eigenfunctions the norm takes the form

$$\ldots = \sum \frac{\tilde{f}_i^2}{\lambda_i}\ . \tag{2.24}$$

For $\|f\|_K$ being finite, the sequence of coefficients $\tilde{f}_i = \langle f, \phi_i\rangle_2$ must decay at a sufficient rate. The *rougher* a function the slower their eigenvalue spectrum decays. In other words, the rate of decay of the eigenvalues imposes a smoothness condition on the space. It is important to distinguish clearly between the two inner products. While $\|\cdot\|_2$ refers to the distance with respect to some measure μ, the RKHS norm is a measure of the *roughness* of a function. For example, the eigenfunctions have $\|\phi_i\|_2 = 1$, but $\|\phi_i\|_K = \lambda_i^{-1/2}$ is becoming increasingly rough.

Compared with the n-dimensional quadratic form $\boldsymbol{f}^\top K_n^{-1}\boldsymbol{f}$ that appears in the MVN density, the squared norm $\|\cdot\|_K^2$ in the RKHS formalism can be thought of as a generalization to functions. Mercer's theorem may be understood as the infinite-dimensional extension of an eigenvalue decomposition. To see this, suppose that K_n is a symmetric and positive definite $n \times n$ matrix. Then there exists an orthonormal basis $\boldsymbol{e}_i$ with according eigenvalues λ_i that diagonalizes K_n. This means that $K_n = U\Lambda U^\top$ with unitary matrix $U = [\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n]$ and diagonal matrix $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. If this is multiplied out it simply reads $K_n = \sum \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^\top$, which is exactly like Mercer's theorem. If K_n is expressed in terms of the eigenvectors $\boldsymbol{e}_i$ then the quadratic form becomes $(Uf)^\top \Lambda^{-1}(Uf)$, which is in analogy to the inner product (see Equation 2.19). Compared with RKHSs, eigenvectors $\boldsymbol{e}_i$ become functions $\phi_i(t)$ as opposed to simple vectors.

Example 2.6. Let us revisit Example 2.2 – the Wiener process – and calculate the Mercer decomposition of

$$K(s,t) = \min(s,t) \tag{2.25}$$

defined on $s, t \in [0, T]$. The eigenvalue problem (Eq. 2.17) follows to read

$$\mathrm{T}_K[\phi](s) = \int_0^s t\phi(t)\,\mathrm{d}t + s\int_s^T \phi(t)\,\mathrm{d}t = \lambda\phi(s)\ . \tag{2.26}$$

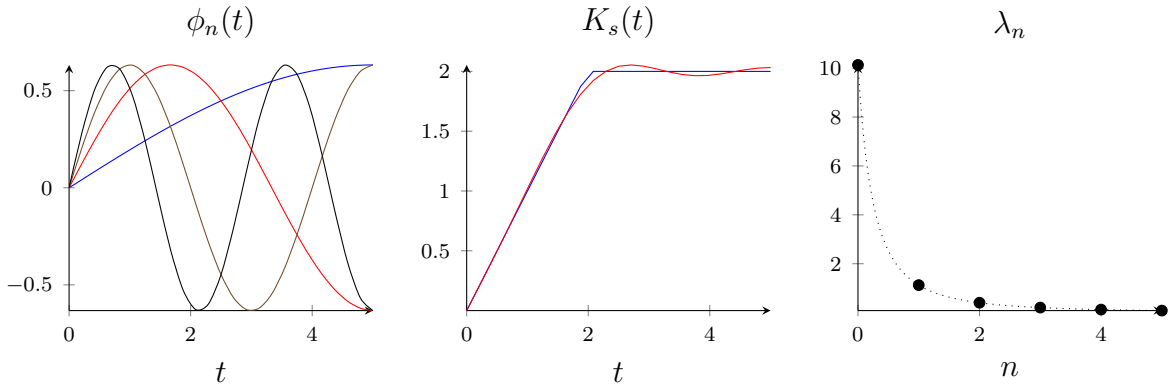

Figure 2.2.: Mercer decomposition of $\min(s,t)$ for $s = 2$ (see Ex. 2.6). The left panel illustrates the eigenfunctions for degrees $0, 1, 2, 3$. A comparison of $\min(s,t)$ with the expansion up to degree $n \leq 3$ is shown in the mid panel. The right panel depicts the spectrum of the eigenvalues.

For $s = 0$ the boundary condition $\phi_n(0) = 0$ is clear. Differentiating once yields

$$s\phi(s) + \int_s^T \phi(t)\,\mathrm{d}t - s\phi(s) = \lambda\dot{\phi}(s) \tag{2.27}$$

and for $s = T$ we find another boundary condition $\dot{\phi}_n(T) = 0$. A further differentiation brings up a well known ODE

$$\lambda\ddot{\phi}(s) = -\phi(s) \tag{2.28}$$

and the solution is of the form

$$\phi(s) = a\sin\left(\frac{s}{\sqrt{\lambda}}\right) + b\cos\left(\frac{s}{\sqrt{\lambda}}\right) \ . \tag{2.29}$$

Imposing both boundary conditions yields

$$b = 0 \qquad \text{and} \qquad \cos\left(\frac{T}{\sqrt{\lambda}}\right) = 0 \tag{2.30}$$

and the eigenvalues and normalized eigenfunctions are going to read

$$\lambda_n = \frac{T^2}{\pi^2\left(\frac{1}{2}+n\right)^2} \qquad \text{and} \qquad \phi_n(t) = \sqrt{\frac{2}{T}}\sin\left(\frac{t}{\sqrt{\lambda_n}}\right) \ . \tag{2.31}$$

Figure 2.2 illustrates the Mercer decomposition. ◁

Mercer's theorem may be generalized towards a continuous spectrum. This requires the restriction to stationary kernels that are RKs depending only on the lag $\tau = s - t$. The following theorem – quoted from Berlinet and Thomas-Agnan [2004, sec. 7.1] – establishes a link with the Fourier domain:

Theorem 12 (Bochner)**.** *The Fourier transform of a bounded positive measure on $\mathbb{R}^d$ is a continuous function of positive type. Conversely, any function of positive type is the Fourier transform of a bounded positive measure.*

Further details and a poof may also be found in Berlinet and Thomas-Agnan [2004, sec. 7.1]. If the positive measure has a spectral density $\lambda(\xi)$ then the stationary kernel and the spectral density are forming a Fourier pair. Considering the shift property of the Fourier transform yields

$$K(s,t) = K(s-t) = \int \lambda(\xi) \exp\{2i\pi\xi s\} \exp\{-2i\pi\xi t\} \, \mathrm{d}\xi \tag{2.32}$$

and the complex exponentials may be see as the eigenfunctions. That interpretation draws an analogy to Mercer's theorem where the sum is replaced by an integral. Expressed in terms of the lag, Equation 2.32 is also known as the Wiener-Khinchin theorem.

Example 2.6 already suggests that differentiability takes a significant role in constructing kernels that feature a certain regularity. The fundamental principle is to penalizes a function f in terms of variability. The following is a condensed from of what is detailed in Rasmussen and Williams [2006, sec. 6.2.1]. Define a norm that takes the derivatives up to order M into account

$$\|f\|_K = \sum_{m=0}^{M} a_m \int \left(\partial_x^m f(x)\right)^2 \, \mathrm{d}x \tag{2.33}$$

and all coefficients are assumed positive[1], i.e. $a_m > 0$. This norm imposes a regularity constraint similar to the principles already known for the kernel map construction. Equation 2.33 is transformed into the Fourier domain

$$\|f\|_K = \int \left(\sum a_m (2\pi\xi)^{2m}\right) \left|\hat{f}(\xi)\right|^2 \, \mathrm{d}\xi \tag{2.34}$$

which turns the derivatives into an algebraic equation. The key is to recognize that the complex exponentials $\exp\{2i\pi\xi\cdot\}$ are eigenfunctions of the derivative and the kernel's positive spectral density can directly be read off. The associated RK is determined with the help of Bochner's theorem. The kernel is given through an inverse Fourier transform

$$K(s,t) = K(\tau) = \int \frac{1}{\sum a_m (2\pi\xi)^{2m}} \exp\{2i\pi\xi\tau\} \, \mathrm{d}\xi \tag{2.35}$$

which follows from a comparison with Equation 2.32. As a consequence, the Fourier transformation is a powerful tool to analyse properties of a given kernel.

Example 2.7. It was confirmed in Example 2.5 that the SE covariance is a RK. However, neither an explicit form for the inner product nor the smothness of the RKHS are known. Since the SE kernel is stationary, the shift property of the Fourier transform yields

$$K(x,y) = \int \hat{K}(\xi) \exp\{2i\pi\xi x\} \exp\{-2i\pi\xi y\} \, \mathrm{d}\xi \; . \tag{2.36}$$

Using the Fourier pair $\exp\{-x^2/a\} \leftrightarrow \sqrt{\pi a} \exp\{-(\pi\xi)^2 a\}$, the spectral density is given by

$$\hat{K}(\xi) = \sqrt{2\pi}\ell \exp\left\{-2(\pi\xi\ell)^2\right\} \; . \tag{2.37}$$

[1] Vanishing coefficients are spanning a null space that needs separate treatment. An exhaustive discussion about Spline models may be found in Wahba [1990] and Rasmussen and Williams [2006, sec. 6.3]

The reproducing property becomes apparent through the insertion of $\frac{\hat{K}}{\hat{K}}$. Then, Equation 2.36 takes the form of an inner product

$$K(x,y) = \int \hat{K}^{-1}(\xi)\, \hat{K}_x(\xi)\hat{K}_y(\xi)\, \mathrm{d}\xi \ . \tag{2.38}$$

Consequently – within the Fourier domain – the inner product of the RKHS is given by

$$\langle g, h\rangle_K = \int \hat{g}(\xi)\hat{h}(\xi)\, \hat{K}^{-1}(\xi)\, \mathrm{d}\xi \ . \tag{2.39}$$

The spectral density $\hat{K}$ gives information about the smoothness of the RKHS. A function f belongs to a RKHS if the according norm is finite. Roughly speaking this means that $\hat{f}$ has to converge to zero at a rate faster than $\hat{K}^{1/2}$.
An expression in the original domain may be found by using the series expansion of the exponential function. Expanding by an imaginary unit and factoring out constants yields

$$K(x,y) = \frac{1}{\sqrt{2\pi}\ell} \sum_n \frac{\ell^{2n}}{n!2^n} \int (2i\pi\xi)^n \hat{K}_x(\xi)(-2i\pi\xi)^n \hat{K}_y(\xi)\, \mathrm{d}\xi \ . \tag{2.40}$$

Making use of the Plancherel theorem and identify the Fourier pair $\partial_x^n \leftrightarrow (2i\pi\xi)^n$ yields

$$K(x,y) = \frac{1}{\sqrt{2\pi}\ell} \sum \frac{\ell^{2n}}{n!2^n} \int \partial_z^n \hat{K}_x(z)\, \partial_z^n K_y(z)\, \mathrm{d}z \ . \tag{2.41}$$

Then – w.r.t the original domain – the inner product that corresponds with the SE kernel is given as a series expansion

$$\langle g, h\rangle = \frac{1}{\sqrt{2\pi}\ell} \sum \frac{\ell^{2n}}{n!2^n} \int \left(\partial_z^n g(z)\right) \left(\partial_z^n h(z)\right)\, \mathrm{d}z \ . \tag{2.42}$$

Consequently, elements of the according RKHS are very smooth. It is *not* enough to require all derivatives to exists. In addition, every derivative has to satisfy the condition of square integrability. And on top of that the series has to converge. Although the SE covariance is probably the most widely-used kernel function, in case of an application it should be ensured that such a strong smoothness assumption is appropriate. ◁

A slightly different approach to obtain the kernel implied by a differential operator is to address Green's function. The following example illustrates how to find the kernel corresponding to a given regularity assumption.

Example 2.8. For the domain $[0, T]$, determine the RK according to

$$\langle g, h\rangle_K = \int \dot{g}(z), \dot{h}(z)\, \mathrm{d}z \tag{2.43}$$

with zero initial conditions, i.e. $f(0) = 0$. The key is that a Green's function for the differential operator ∂_t is the Heaviside step function and the kernel results to be $K(x,y) = \min(x,y)$ i.e. the Wiener process. ◁

Chapter 5 details the approach via Green's function in greater depth using the example of a damped vibrating string.

2.1. Regularized Least Squares

A viewpoint closely related to RKHS theory is regularization. Without additional information, fitting a curve onto a set of points is ill posed because the associated optimization problem has infinitely many solutions. To uniquely determine such a solution it is necessary to introduce further constraints. The intuition behind finding a regularized solution is to specify a curve that is a trade-off (compromise) between a data-fit term and a norm of that function. For reference, the following discussion may be found in a number of sources, e.g. Rasmussen and Williams [2006, Sec. 6.2] or Steinwart and Christmann [2008].

Based on linear mixed models, the function f to be fitted is apportioned such that

$$f(t) = g(t) + h(t) \tag{2.44}$$

where g is adjustable and h is known i.e. a non-random effect. We wish to find a curve f that is a *good* estimate for a set of n training pairs

$$(y_1, x_1), \cdots, (y_n, x_n) \tag{2.45}$$

with pointwise observations y_i at locations x_i. The discrepancy between the prediction $f(x_i)$ and observations y_i is measured by the least squares loss, distinguishing between *good* and *bad*. Since f is an arbitrary function – in most applications the search domain will be an infinite-dimensional subspace of L_2 – an additional constraint penalizing the complexity of the function is needed. This is accomplished by considering functions within a RKHS and the additional cost is chosen to be the RKHS norm. Using the empiric risk, the objective functional is of the form

$$J[f] = \sum \big(y_i - f(x_i)\big)^2 + \lambda \|g\|_K \tag{2.46}$$

leading to the regularized minimization problem

$$\hat{f} \in \operatorname*{arg\,min}_{g \in \mathcal{H}_K} \big\{ J[f] \big\} \,. \tag{2.47}$$

The scaling parameter λ controls the trade-off between function smoothness (fidelity) and fitting error also known as the *residual sum of squares*. The smoothness assumptions on f are encoded by the RKHS and the data-fit term assesses the quality of the prediction. In general it is difficult to determine a value for λ. Section 2.3 mentions techniques how to choose λ.

Provided a value for λ is known, the representer theorem stats that the solution of minimization problem is given by a finite kernel expansion:

Theorem 13 (Representer)**.** *Given a RK K and a training set of n pairs $\big\{(y_i, x_i)\big\}_{i=1}^n$. Any minimizer*

$$\hat{g} \in \operatorname*{arg\,min}_{g \in \mathcal{H}_K} \Big\{ \sum \big(y_i - g(x_i)\big)^2 + \lambda \|g\|_K \Big\}$$

admits a representation of the form

$$\hat{g}(\cdot) = \sum_{i=1}^{n} \hat{a}_i K(\cdot, y_i)$$

with $\hat{a}_i \in \mathbb{R}$.

Remarkable about that theorem is that the representation of $\hat{g}$ reduces the infinite-dimensional optimization problem to minimize the coefficients $\hat{a}_1, \cdots, \hat{a}_n$, a finite dimensional optimization problem. Theorem 13 is a particular example of a family of results that are collectively referred to as *representer theorems*. Generalizations that may be found in the aforementioned references include: The representer theorem keeps its form considering an arbitrary but convex error function rather than the residual sum of squares. Instead of having a pointwise training set general observational functionals are possible. The regularizing parameter may be replaced by strictly monotonically increasing function of the RKHS norm.

Coming back to the residual sum of squares, the solution to the optimization problem (Eq. 2.47) is explicit. To keep equations concise vector/matrix notations are introduced:

$$\boldsymbol{y} = (y_1, \cdots, y_n)^\top \tag{2.48}$$
$$\boldsymbol{a} = (a_1, \cdots, a_n)^\top \tag{2.49}$$
$$\boldsymbol{h}_x = (h(x_1), \cdots, h(x_n))^\top \tag{2.50}$$
$$\boldsymbol{k}_{zx} = (K(z, x_1), \cdots, K(z, x_n)) \tag{2.51}$$
$$\mathrm{K}_{xx} = \{K(x_i, x_j)\}_{i,j=1,\cdots,n} \ . \tag{2.52}$$

Substituting f by a finite kernel expansion and using the reproducing property yields

$$J[\boldsymbol{a}] = (\boldsymbol{y} - \mathrm{K}_{xx}\boldsymbol{a})^2 + \lambda \boldsymbol{a}^\top K_{xx}\boldsymbol{a} \tag{2.53}$$

and the optimization problem reads

$$\hat{\boldsymbol{a}} \in \underset{\boldsymbol{a}\in\mathbb{R}^n}{\arg\min} \{J[\boldsymbol{a}]\} \ . \tag{2.54}$$

Differentiating and solving yields the minimizing coefficients

$$\hat{\boldsymbol{a}} = \left[(\mathrm{K}_{xx} + \lambda\mathbb{I})\,\mathrm{K}_{xx}\right]^{-1} \mathrm{K}_{xx}\,(\boldsymbol{y} - \boldsymbol{h}_x) \ , \tag{2.55}$$

a form that is well known from Tikhonov regularization and ordinary least squares. If K_{xx} is of full rank, the prediction that minimizes Equation 2.47 is given by

$$\hat{f}(z) = h(z) + \boldsymbol{k}_{zx}(\mathrm{K}_{xx} + \lambda\mathbb{I})^{-1}\,(\boldsymbol{y} - \boldsymbol{h}_x) \ . \tag{2.56}$$

Compared with GP regression the form of $\hat{f}$ is similar to the conditional mean (see Eq. 1.19). If pointwise observations corrupted by homoscedastic centered Gaussian noise of variance λ are assumed, the conditional mean and $\hat{f}$ are even identical. Section 2.2 demonstrates that this is no coincidence. An extension to bounded linear observational functionals is not far to seek and may be found in Wahba [1990].

Various data fitting methods may be unified as an optimization problem in a RKHS setting. Although most of them will not be mentioned here, the following three examples are intended to demonstrate the variety of methods regularization and RKHS theory are bridging across.

Example 2.9 (Polynomial Interpolation)**.** Let us proceed with the ongoing example where the signal is assumed to be

$$\tilde{f}(x) = x\cos(x) \tag{2.57}$$

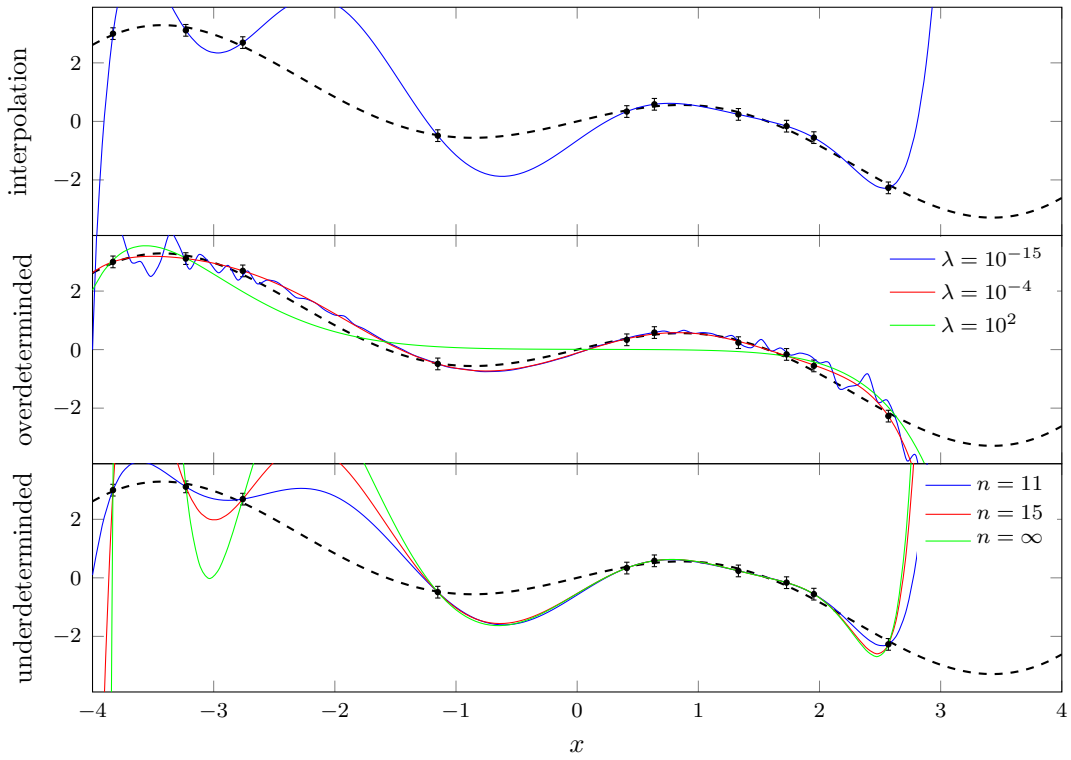

Figure 2.3.: The actual signal is shown by ---, the error bars are indicating the noise level and • refers to the training set. The top panel shows the interpolating polynomial (see Ex. 2.9). The smoothing effect of the regularizer on the least squares solution is depicted in the mid panel (see Ex. 2.10). The bottom panels illustrates the effect of the degrees of freedom on the reconstruction (see Ex. 2.11).

although this is not a polynomial. A training set of $n = 10$ records is acquired from $\tilde{f}$, corrupted by centered normal noise of standard deviation $\sigma = 0.2$. Finding the polynomial of degree at most n with coefficients $a_0, \ldots, a_{n-1}$ that passes through points

$$f(x_i) = \sum_{j=0}^{n-1} a_j {x_i}^j = y_i \tag{2.58}$$

is known as polynomial interpolation. Since the number of observations equals the number of coefficients this linear system of equations has a unique solution. Thus, there is no need to regularize – i.e. $\lambda = 0$. The interpolating polynomial is equivalent to the solution of the optimization problem given by Equation 2.56. The top panel of Figure 2.3 shows the interpolating polynomial, obtained using the kernel

$$K(x, y) = \sum_{j=0}^{n-1} \frac{(xy)^i}{i!} \ . \tag{2.59}$$

Any other kernel built off a polynomial basis of degree at most n (see Eq. 2.12) yields identical solutions, e.g. $K(x, y) = (xy + c)^n$ or $K(x, y) = \sum (xy)^i$. ◁

In statistical applications one is generally more interested in smoothing rather than interpolating data. The following example establishes a link to least squares techniques and illustrates the effect of the regularizing parameter.

Example 2.10 (Ordinary Least Squares & Ridge Regression)**.** Example 2.9 is modified such that there are more training points than degrees of freedom. The same kernel as in the previous example is used, with the difference that a polynomial of degree at most $n-1$ shall be found. In other words, the polynomial can not pass through all training points at the same time. That means the system of equations is overdetermined and the matrix K_{xx} becomes rank deficient. To compute the inverse, a non-zero λ is necessary. The smallest λ is determined through the condition number and the machine precision. The mid panel of Figure 2.3 illustrates varying λ through scales. It can clearly be seen that values close to the machine's precision are affected by numerical instabilities. On the other hand, the larger λ the smoother the reconstruction. ◁

Ordinary least squares regression, however, is not defined when the degrees of freedom exceed the number of observations. That limitation is solved by using a kernel based approach. Even passing to the limit of infinitely many degrees of freedom is possible.

Example 2.11. Still, we proceed with the setting introduced in Example 2.9. The aim of that example is to demonstrate the effect the kernel has on the reconstruction. This is best understood considering fewer data that degrees of freedom. In that case the matrix K_{xx} is of full rank and we set $\lambda = 0$. The kernel offers the opportunity passing to the infinite-dimensional case (see Ex 2.4). The bottom panel of Figure 2.3 illustrates the Kernel's effect by varying the degrees of freedom. The interpretation may be seen in the light of the usual Fourier transform: With growing polynomial degree there are more *high-frequency* contributions. Consequently, the fidelity of the reconstruction grows with the polynomial degree. ◁

Although the regularized solution gives only half the Gaussian process solution, the similarity between the conditional mean and the regularized solution motivates establishing a formal link between GPs and RKHSs.

2.2. Duality GP and RKHS

In this section the connection between GPs and RKHSs is described. Despite RKHSs being function spaces and sample paths being functions, a RKHS structure is not given to the space of all possible sample paths [Manton and Amblard, 2015, Sec. 7.2]. It is the covariance function of a GP that can be identified with the RK of a RKHS.

In order to establish this connection an *auxiliary* Hilbert Space is needed. Let $X(t)$ be a GP. The Hilbert space $\mathcal{H}_X$ generated by a process is the collection of RVs of the form

$$U = \sum u_i X(t_i) \tag{2.60}$$

equipped with the inner product

$$\langle U, V \rangle_X = \mathbb{E}[UV] \tag{2.61}$$

and with all the mean square limits adjoined. In other words, $\mathcal{H}_X$ represents all RVs attainable by linear operations, including limits.

The RKHS $\mathcal{H}_K$ is a representation of the GP due to the following theorem:

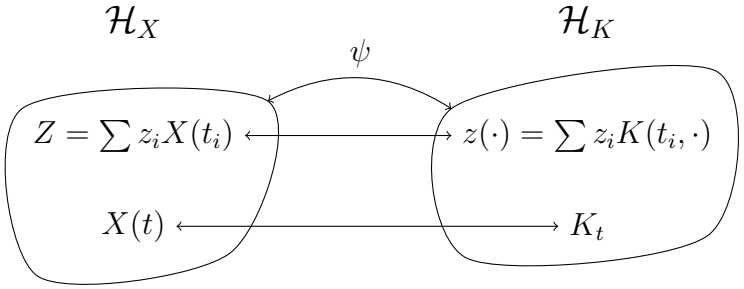

Figure 2.4.: The correspondence between GPs and RKHSs.

Theorem 14 (Loève)**.** *The Hilbert Space $\mathcal{H}_X$ generated by the GP $X(t)$ with covariance K is isometrically isomorphic to the RKHS $\mathcal{H}_K$ with RK K.*

In other words, each RV in $\mathcal{H}_X$ gets a function in $\mathcal{H}_K$ assigned. To see this correspondence, let ψ denote the map from $\mathcal{H}_X$ to $\mathcal{H}_K$ defined by

$$\psi\Big(\sum u_i X(t_i)\Big) = \sum u_i K(t_i, \cdot) \; . \tag{2.62}$$

This is an isometry since it preserves the linear structure of the inner products

$$\langle X(s), X(t)\rangle_X = \mathbb{E}[X(s)\, X(t)] = K(s,t) = \langle K_s, K_t\rangle_K \; . \tag{2.63}$$

The notion of $\mathcal{H}_X$ and $\mathcal{H}_K$ being isometrically isomorphic means that although the two Hilbert spaces consist of different elements (RVs vs. functions) they still share the same geometric structure. Figure 2.4 illustrates that one to one correspondence.

Once the isometric isomorphism is established, problems related to GPs can be translated into functional ones and vice versa. It is customary to use whichever of the spaces is more convenient for the problem. Inference is typically performed adopting a statistical point of view, whereas, functional analysis is addressed for deriving RKs.

For all $g, h \in \mathcal{H}_K$, consequences of Theorem 14 are

$$\langle \psi^{-1}(g), \psi^{-1}(h)\rangle_X = \langle g, h\rangle_K \tag{2.64}$$

$$\langle \psi^{-1}(g), X(t)\rangle_X = g(t) \tag{2.65}$$

$$\mathbb{E}\big[\psi^{-1}(g)\big] = \langle g, \mathbb{E}[X]\rangle_K \tag{2.66}$$

due to linearity and continuity. This can further be generalized. According to Riesz theorem, any bounded linear functional L has a representer $\eta \in \mathcal{H}_K$. By duality there is also a RV Z that corresponds to η and we have

$$\langle Z, X(t)\rangle_X = \mathbb{E}[ZX(t)] = \langle \eta, K_t\rangle_K = \eta(t) = LK_t \; . \tag{2.67}$$

Examples for such functionals are e.g. integrals, derivative or limiting sequences.

An application of the duality principle in combination with the Mercer decomposition leads to an important series representation of GPs:

Theorem 15 (Karhunen-Loève)**.** *Let $X(t)$ be a centered GP with continuous covariance K that possesses a Mercer decomposition (Thm. 11) of eigenfunctions ϕ_i and eigenvalues λ_i. Then, $X(t)$ admits a representation of the form*

$$X(t) = \sum \sqrt{\lambda_i} U_i \phi_i(t)$$

with U_i independent standard normal RVs and the convergence is uniform with respect to t and in quadratic mean.

To show this result both the duality and Mercer's theorem are used. Through the isometry ψ the GP is connected to the function

$$\phi\big(X(t)\big) = K(t,\cdot) = \sum \lambda_n \phi_n(t)\phi_n(\cdot) \; . \tag{2.68}$$

Conversely, the RV $\chi_n = \psi^{-1}(\sqrt{\lambda_n}\phi_n(t))$ is identified. Using Equations 2.66 and 2.19, first and second moments are given by

$$\mathbb{E}[\chi_n] = \langle \sqrt{\lambda_n}\phi_n, \mathbb{E}[X(t)]\rangle_K = 0 \tag{2.69}$$

$$\mathbb{E}[\chi_m \chi_n] = \sqrt{\lambda_m}\langle \phi_m, \phi_n\rangle_K \sqrt{\lambda_n} = 1 \tag{2.70}$$

and χ_n is identical to U_i. The general case of a non-zero mean process can be brought back to the case of a centered process by additively splitting off the mean.

The advantage of the Karhunen-Loève expansion is that it separates the dependence upon t and the randomness. The process is decorrelated in the sense that eigenfunctions ϕ_i are orthogonal and RVs U_i are independent. The division into functions and random contributions offers a way to analyse the RKHS norm of sample paths, yielding

$$\|X(t)\|_K^2 = \sum_{i,j} \sqrt{\lambda_i} U_i \langle \phi_i(t), \phi_j(t)\rangle_K U_j \sqrt{\lambda_j} = \sum U_i^2 \; . \tag{2.71}$$

That result indicates that the duality between GP and RKHS has its limitations. Following the heuristic argument given by Wahba [1990, p. 5] we have

$$\mathbb{E}\big[\|X(t)\|_K^2\big] = \sum \mathbb{E}\big[U_i^2\big] = \sum i \to \infty \; . \tag{2.72}$$

In other words, for more than a finite number of non-zero eigenvalues sample paths are not in $\mathcal{H}_K$ almost surely. Further details may be found in the afore mentioned references. Typically, the following example is given to demonstrate that the trajectories of a GP have a lower regularity than the associated RKHS.

Example 2.12. According to Example 2.8, the RKHS that corresponds to Brownian motion is the Sobolev space $H^1(0,T)$. Unlike elements of this Sobolev space, it is well known that sample paths of the Wiener process are almost surely not differentiable.

2.3. Kernel Parameters

For a selected kernel function, the predictive performance of GP regression depends exclusively on that covariance. However, in many applications, it is possible to only specify the structure of the covariance. The kernel still depends on unknown parameters whose values need to be determined. In order to bring GP regression into practice, techniques determining kernel parameters are essential. For the following discussion, references in use are Murphy [2012, Sec. 15.2.4] and Rasmussen and Williams [2006, Sec. 5.4].

Typically there is only vague information about kernel parameters. Therefore an empirical Bayesian approach is considered. Modulo a normalizing constant the parameter's posterior density is given by

$$p(\theta|d) \propto p(d|\theta)p(\theta) \tag{2.73}$$

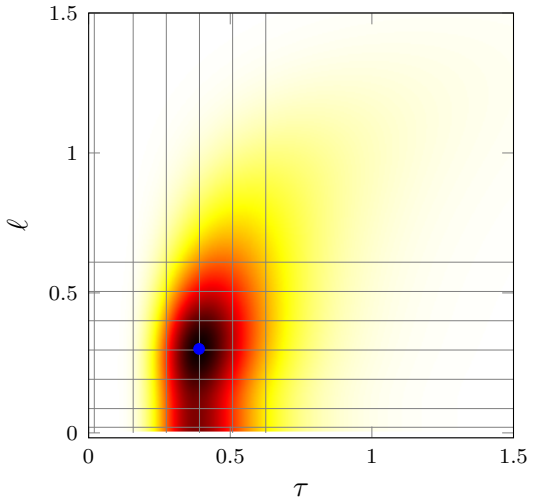

Figure 2.5.: Posterior density for parameters ℓ and τ of the SE kernel. A priori ℓ and τ are assumed uninformative. By • the MAP estimate is indicated. The grid refers to collation points used to integrate out kernel parameters.

where kernel parameters are denoted by θ and d refers to the data. Let adopt the Gaussian linear model i.e. linear observations with normal error model (see Sec. 1.3). Within that setting and for θ fix, the marginal likelihood (a.k. prior predictive distribution) is normal

$$p(d|\theta) = \mathcal{N}(\mu_\theta,\, K_\theta) \ . \tag{2.74}$$

with a priori mean μ_θ and covariance K_θ.

Using a point estimate is a simplistic approach to tune kernel parameters. In order to optimize θ across a parameter space Θ, maximize the Bayesian posterior density

$$\theta_{\mathrm{MAP}} \in \arg\max_{\Theta} \{p(d|\theta)p(\theta)\} \tag{2.75}$$

i.e. the maximum a posteriori (MAP) estimate. The derivative with respect to θ is tractable [Murphy, 2012, Eq. 15.24] and continuous optimization techniques may be used for estimation such as the differential evolution algorithm. The main argument against using a point estimate is that the actual inference does not take the parameters' uncertainties into account. Three more difficulties are worth mentioning: 2) In higher dimensions, optimization suffers from the curse of dimensionality. 3) Local minima may be a problem as the objective is not convex. 4) Computing K_θ^{-1} may be numerically expensive.

Example 2.13. The SE kernel (Eq. 1.14) depends upon two parameters, length scale ℓ and overall variance τ. The simplest approach to find a point estimate is to use an exhaustive search over a discrete grid. A priori, ℓ and τ are assumed positive but uninformative. The true signal in use and the measurement geometry with to be known error level are shown in Figure 2.6. Figure 2.5 shows the parameters' posterior density also indicating the MAP estimate

$$\ell_{\mathrm{MAP}} = 0.3 \qquad\qquad \tau_{\mathrm{MAP}} = 0.39 \ . \tag{2.76}$$

Having only 8 data points is not enough to decide about the kernel parameters with confidence. Nonetheless, the top panel of Figure 2.6 shows the according posterior GP. ◁

If the parameter's posterior density is diffuse, a point estimate is not truly satisfying. Although not of particular interest, the variabilities of kernel parameters should be taken into account. Performing GP regression based on θ_{MAP} certainly does not propagate parameter uncertainties. If the dimensionality of θ is small enough to perform numeric integration a hierarchical Bayesian approach is feasible. Therefore, kernel parameters are taking the role of a latent RV. The compound distribution is the result of marginalizing over the latent variable

$$p(f|d) = \int p(f|d,\theta)p(\theta|d)\,\mathrm{d}\theta\ . \tag{2.77}$$

integrating out any dependence on kernel parameters. The integrand is a product of two densities. For a certain choice of parameters, $p(f|d,\theta)$ is normal and obtained through GP regression. To keep track of $p(\theta|d)$ we add a hierarchical stage introduced above (Eq. 2.73). However, the mixing distribution $p(\theta|d)$ has no particular form and the integral must be evaluated by numerical methods. To compute the compound density requires two numeric integrations: First the actual marginalization and a second quadrature to suitably normalize. Although the compound distribution is not normal, calculating higher-order moments is one of its potentials. Mean and co-variance are given by

$$\mathbb{E}[f|d] = \int p(\theta|d)\ \mathbb{E}[f|d,\theta]\ \mathrm{d}\theta \tag{2.78}$$

$$\mathbb{V}[f|d] = \int p(\theta|d)\big(\mathbb{E}[f|d,\theta]^2 + \mathbb{V}[f|d,\theta]\,\big)\,\mathrm{d}\theta - \mathbb{E}[f|d]^2\ . \tag{2.79}$$

To calculate the Gaussian moment matching proxy three numeric integrations are necessary. In any case, the resulting variance is greater or equal than for any point estimate. If the parameter space is of moderate dimensionality, standard methods for numeric integration may be used. Nevertheless, the numerical costs may be enormous. One difficulty that arises is to explore and discretize the parameter space. The method called *Integrated Nested Laplace Approximation* (INLA) Rue et al. [2017] is known to be suited well to perform inference with latent Gaussian models.

Example 2.14. The setting and the prior over the kernel parameter ℓ and τ remain as in Example 2.13. In contrast – rather than using the MAP estimate – the SE kernel parameters are integrated out. The compound density is approximated by numeric integration

$$p(f|d) \approx \sum p(f|d,\theta_i)p(\theta_i|d)\Delta\theta_i \tag{2.80}$$

where Simpson's rule specifies the wights $\Delta\theta_i$. The coarse grid shown in Figure 2.5 refers to the 63 collocation points specified. The same numeric integration is in use to compute conditional mean (Eq. 2.78) and standard deviation (Eq. 2.79). In Figure 2.6 the bottom panel shows the moment matching Gaussian proxy with model parameters marginalized. In comparison with the reconstruction that is based on the MAP estimate (top panel) it can clearly be seen that the moment matching proxy deviates and features significantly larger uncertainties. ◁

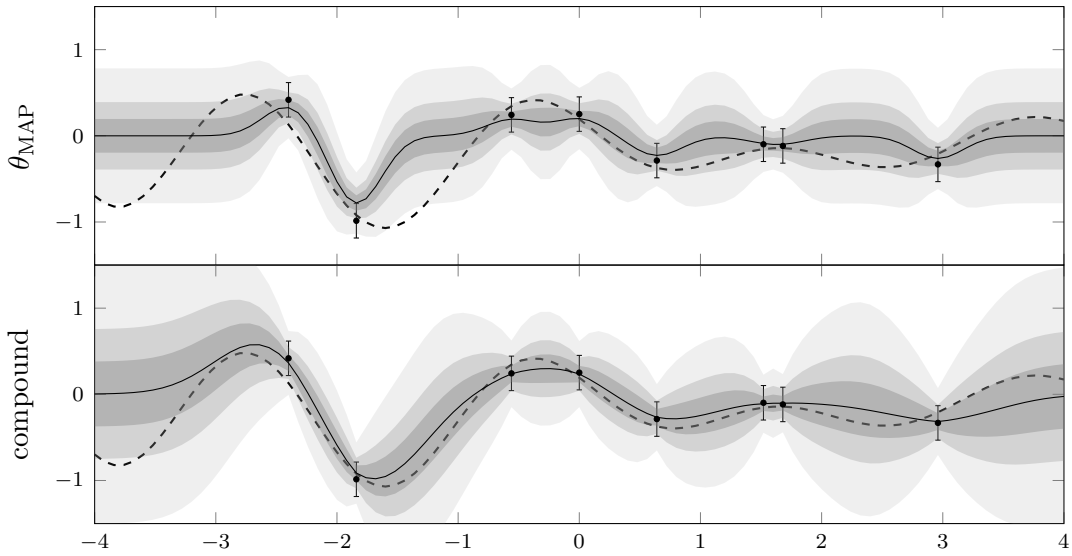

Figure 2.6.: The true signal is shown by - - - , error bars are referring to the noise level and • refers to the training set in use. The top panel shows the reconstruction using the MAP parameter estimate (Eq. 2.75). In the bottom panel kernel parameters are marginalized and the moment matching proxy is shown. The posterior mean is depicted by —— and the gray shaded area represents the 38%, 68% and 95% pointwise confidence intervals.

3. Non-Gaussian Likelihoods

The Gaussian linear model is attractive because of its algebraic convenience. On the contrary, the restriction to normal distributions and affine maps is not truly satisfying. Reality is unlikely to be captured well by such a simple statistical model. Many principles in physics show non-linear relations and measurement noise is not necessarily additive and normal distributed. However – at the time of writing – computing power is not sufficient to numerically solve the multidimensional integrals that are required to perform inference with non-Gaussian likelihoods. This motivates the consideration of approximate Bayesian inference that offers a more versatile modeling concept – e.g. accommodation of non-linear functionals – while preserving the algebraic convenience of the linear Gaussian model. Indeed, this is not possible in general but under certain circumstances an approximation may be valuable as guideline.

As motivation, assume X to be normally distributed. In general, non-linear maps will turn Gaussian RVs into non-Gaussian RVs. There are two well known strategies to approximate PDFs by normal densities. Let $Y = f(X)$, where f is a non-linear function. 1) Use a first-order approximation of f to preserve normality. 2) Project $f(X)$ onto the space of Gaussians by moment matching. The intuition behind both approaches shall be outlined by an example:

Example 3.1. Consider the quadratic function $f(x) = x^2$ and let $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$. According to Theorem 2, the density of $Y = X^2$ is given by

$$p_Y(y) = \frac{p_X(-\sqrt{y}) + p_X(\sqrt{y})}{2\sqrt{y}} \qquad \text{for} \qquad y \in (0, \infty) \tag{3.1}$$

since X^2 has values at the positive real line, only. First and second moment are given by

$$\mathbb{E}[Y] = \mu_X^2 + \sigma_X^2 \qquad \mathbb{V}[Y] = 2\sigma_X^4 + 4\mu_X^2\sigma_X^2 \ . \tag{3.2}$$

We aim to approximate p_Y by a normal density. The moment matching proxy is given by a normal density of mean $\mathbb{E}[Y]$ and variance $\mathbb{V}[Y]$. The 1st order approximation – linearised about $\tilde{x}$ – is given by $f(x) \approx f(\tilde{x}) + 2\tilde{x}(x - \tilde{x})$. Choosing μ_X as point of expansion, the approximate mean and variance read

$$\mathbb{E}[Y] \approx \mu_X^2 \qquad \mathbb{V}[Y] \approx 4\mu_X^2\sigma_X^2 \ . \tag{3.3}$$

Figure 3.1 illustrates the idea and compares the proxy densities. ◁

That introductory example already unveiled three major difficulties. (1) The moment matching strategy is easy to understand. However, depending on the map, calculating first and second moment may be exceedingly hard. (2) In contrast, computations to carry out a Taylor expansion are easy. Although mean and variance are right at hand, it is not apparent which point of expansion to favour. (3) There are arbitrary many normal

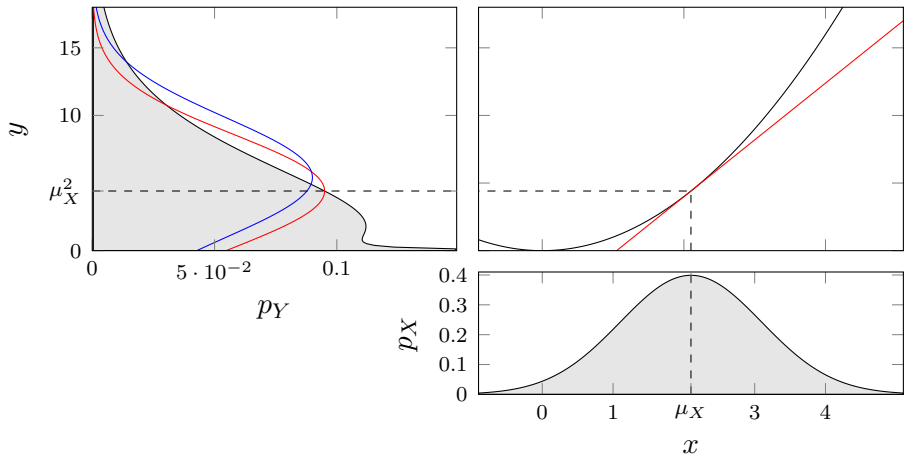

Figure 3.1.: Quadratic transformation of a Gaussian RV. At the bottom, the original PDF is shown. The top right panel indicates the non-linear map and its linearization. The moment matching (——) and linearised (——) proxy PDFs are shown in the top left panel together with the transformed PDF (——).

approximations. How to quantify the *goodness* of the approximation if the true density is not known. In the following, these difficulties and their consequences are repeatedly addressed and illustrated.

Besides Gaussian approximations, there exist a variety of methods dealing with non-linear inference. While Markov-Chain-Monte-Carlo techniques are maybe the most advanced and widely used class of approximate inference techniques it is not without problems such as convergence, the independence of samples and computational time [Seeger, 2004, Bishop, 2006, Rue et al., 2009]. Just to name few, complementary methods are particle filtering, unscented filters and the Expectation Maximization algorithm [Ghahramani, 2004, Sec. 5]. A whole class of alternatives are so-called variational techniques used for approximating intractable integrals arising in Bayesian inference [Murphy, 2012].

3.1. Bayesian Posterior

Before introducing approximations the exact Bayesian posterior is analysed to outline the setting and to point at difficulties arising. Still, the underlying a priori model $M(s)$ is assumed a Gaussian process (GP) with mean $\mu_M(s)$ and covariance $K_M(s,t)$. In contrast, the likelihood function is no longer assumed Gaussian. For a training set $\boldsymbol{d}$ with known error model, theoretically, the posterior process is given by

$$p_{M|D}(m|\boldsymbol{d}) = \frac{p_{D|M}(\boldsymbol{d}|m)p_M(m)}{p_D(\boldsymbol{d})} \tag{3.4}$$

where m refers to trajectories i.e. a function $m(s)$. Although appropriate as a concept, ranging through a space of functions is in most applications analytically not tractable. On top of that, computers of finite-memory and finite-precision do not admit to range through an infinite-dimensional function space. This means that further assumptions have to be made for a numerical implementation.

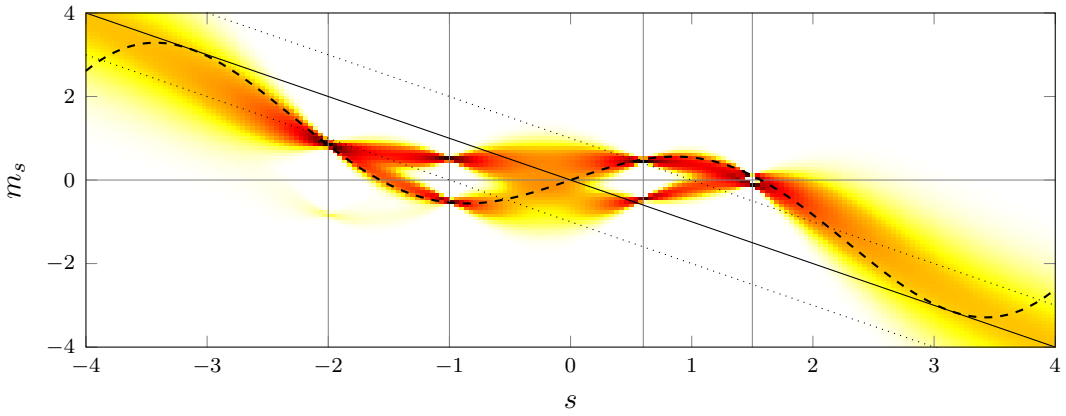

Figure 3.2.: The pseudo color plot refers to the pointwise posterior density according to Example 3.2. To improve the perception, a semi-logarithmic colormap is used. The range of the a priori density is linearly mapped from white to yellow. Values exceeding that range are shown at a logarithmic scale from yellow through red to black. The *true* signal is indicated by - - -. The a priori mean is shown by —— whereas ····· refers to ± standard deviation. Vertical lines are referring to locations where observations are acquired.

In contrast to previous sections measurements are restricted to pointwise observations. The training set of n records $\boldsymbol{d} = (d_1, \cdots, d_n)$ is acquired at points $\boldsymbol{t} = (t_i, \cdots, t_n)$. The likelihood depends on the process only via a finite set $M(t_1), \cdots, M(t_n)$ and reads

$$p_{D|M}(\boldsymbol{d}|m) = p_{D|M}(\boldsymbol{d}|m_{\boldsymbol{t}}) \; . \tag{3.5}$$

where $m_{\boldsymbol{t}}$ is a shortcut for $m(t_1), \cdots, m(t_n)$. The error model remains free of choice and the functions putting the a priori process and observations in relation may be non-linear but pointwise. The difficulty compared to GP regression is that the likelihood – regarded as a function in $m_{\boldsymbol{t}}$ – is *not* Gaussian shaped anymore.

Pointwise posterior predictions of the process are facilitated by assigning a latent role to the training points. Predictions at points of observation $M_{\boldsymbol{t}}|\boldsymbol{d}$ may be seen as *parameters* that carry dependencies along and are marginalized subsequently [Rios et al., 2012, Chp. 2]. The predictive distribution at a design point s (distinct from training points $\boldsymbol{t}$) is given by integrating out the latent variables

$$p_{M|D}\big(m(s)\big|\boldsymbol{d}\big) = \int p_{M|D}\big(m(s), m_{\boldsymbol{t}}\big|\boldsymbol{d}\big)\, \mathrm{d}m_{\boldsymbol{t}} \propto \int p_{D|M}(\boldsymbol{d}|m_{\boldsymbol{t}}) p_M\big(m(s), m_{\boldsymbol{t}}\big)\, \mathrm{d}m_{\boldsymbol{t}} \; . \tag{3.6}$$

except for normalization. Since the posterior is no longer a distribution over functions, a numerical implementation is conceivable. However, two extremely costly integrals have to be calculated: The marginalization and the normalization constant. Marginalizing the training set is one of the key problems. In the vast majority of cases the n-dimensional integrals are analytically not tractable.

The following toy example – numerically assessing the exact Bayesian posterior density – shall illustrate the enormous computational costs and will serve as reference for later comparison with approximations.

Example 3.2 (Exact Bayesian Inference)**.** As in the early examples the same *true* signal $\tilde{m}(t) = t \cos t$ is used. In contrast $n = 4$ pseudo records are generated off from its square.

Observations are assumed to be conditionally independent and the data model reads

$$D_i|\tilde{m} \sim \Gamma(k_i, \theta_i) \tag{3.7}$$

i.e. measurement errors are Gamma distributed. The error model is chosen such that by average the square of the signal is recorded at an error level of $\epsilon = 10\%$. In other words, the larger the value recorded the bigger the potential corruption. The parameters of the error model are given by

$$k_i = \epsilon^{-2} \qquad \text{and} \qquad \theta_i = \epsilon^2 m(t_i)^2 \tag{3.8}$$

which corresponds to a Gamma distribution that is scaled by the square of the signal. The according likelihood as a function in $m_{\boldsymbol{t}}$ reads

$$p(\boldsymbol{d}|m_{\boldsymbol{t}}) = \prod_i \frac{d_i^{k_i-1} e^{-\frac{d_i}{\theta_i}}}{\Gamma(k_i)\theta_i^{k_i}} \,. \tag{3.9}$$

The a priori GP remains as in the previous examples. The covariance is chosen to be the SE Kernel (Eq. 1.14) and the prior mean is kept a falling straight line (Eq. 1.22). Numerical integration is addressed to calculate the normalizing constant and to perform the marginalization [1]. SciPy's nested, general purpose adaptive quadrature is used to ease the implementation [Virtanen et al., 2020]. Figure 3.2 presents the resulting pointwise posterior density. Multimodality is clearly visible if records are acquired near the origin. The posterior distribution features two branches per record. This complexity is chosen on purpose to challenge the approximations discussed in the following. ◁

The computational complexity is dominated by the number of observations. Numeric integration suffers from the curse of dimensionality scaling computational costs at an untold rate [Bishop, 2006, sec. 1.4]. A training set that exceeds only a handful of records renders numerical integration utterly inconceivable. For multiple design points the effort further increases as the marginalization has to be performed for every individual point.

Example 3.3 (Computational Costs)**.** At the time of writing, it took an ordinary workstation featuring 8 parallel threads 3 d and 14 h to explore the posterior density shown in Figure 3.2 . To evaluate the density of the regular gird of size 200×100, every gridpoint requires to solve a 4-dimensional integral plus one for the normalizing constant. The adaptive quadrature scheme causes the computational costs to scale exponentially w.r.t. number of records [2]. Figure 3.3 illustrates how computational times blow up for calculating the normalization constant. By just Adding one more record, the computing time of Example 3.2 would have increased to about 258 d. ◁

The above example demonstrates that for large datasets the computational costs to directly solve Bayesian inference are prohibitive. Because the normalizing constant needs to be calculated *only* once, marginalizing the latent variables requires priority. Since GPs are the only precesses where marginalization is straight forward, a GP approximation appears to be a promising strategy.

[1] The conjugacy of Gaussian and Gamma distribution gives chance for further analytic steps. Because θ and m are sharing a quadratic relation further simplifications are not apparent. Howsoever, particular solutions are not the main interest.

[2] Although a fixed point quadrature scheme is less demanding, improved numerics can not evade the curse of dimensionality.

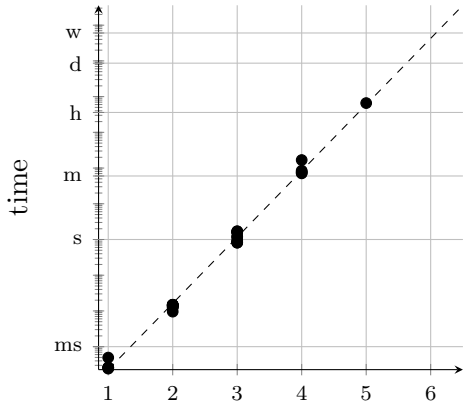

Figure 3.3.: Illustration of the computational costs depending on the number of observations. Computational costs of a nested general purpose adaptive quadrature scheme grow exponentially. The logarithmic scale ranges from a millisecond to one week.

3.2. Gaussian Process Approximation

The aim of this section is to approximate the Bayesian posterior distribution by a GP. A GP approximation is a reasonable approach since GPs are one of the few processes with which it is feasible to deal analytically. However, to find such an approximation a measure is required that quantifies the similarity of two distributions. One way to measure how a probability distribution is different from a reference distribution is defined as follows [Rasmussen and Williams, 2006, A. 5]:

Definition 4 (Kullback-Leibler Divergence)**.** *For distributions P and Q with densities $p(x)$ and $q(x)$,*

$$\mathrm{KL}(P\|Q) = \int p(x) \log\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x$$

is called the Kullback-Leibler (KL) divergence.

The KL divergence is non-negative but asymmetric and, thus, not a distance. A KL divergence of zero indicates that two distributions are identical. In the context of information theory, P represents the *true* distribution while Q refers to an approximation of P. An interpretation of the KL divergence is the amount of information lost when Q is used instead of P. The more the distributions differ the bigger the KL divergence.

The normal distribution Q – with mean $\tilde{\mu}$ and covariance $\tilde{\Sigma}$ – that is *closest* to some general distribution P may be found by optimizing the KL divergence [3]. The minimizing parameters are determined by differentiating $\mathrm{KL}(P\|Q)$ w.r.t. $\tilde{\mu}$ and $\tilde{\Sigma}$ and setting the expressions to zero. The resulting optimal parameters are given by

$$\partial_{\tilde{\mu}}\mathrm{KL}(P\|Q) := 0 \quad \rightsquigarrow \quad \tilde{\mu} = \mathbb{E}[P] \quad \text{and} \quad \partial_{\tilde{\Sigma}}\mathrm{KL}(P\|Q) := 0 \quad \rightsquigarrow \quad \tilde{\Sigma} = \mathbb{V}[P] \ , \quad (3.10)$$

the so-called moment matching Gaussian proxy [Rasmussen and Williams, 2006, A. 5].

To approximate the exact Bayesian posterior by its moment matching Gaussian requires to calculate posterior mean and covariance. Although, $\mathbb{E}[M|\boldsymbol{d}]$ and $\mathbb{V}[M|\boldsymbol{d}]$ are assumed

[3] An alternative is variational Bayesian inference where reference and proxy distributions of the KL divergence are reversed [Murphy, 2012, Chp. 21]. However, this work focuses mainly on distributional proxies rather than modal approximations.

intractable, taking advantage of the a priori process being Gaussian provides insight into the overall structure of the approximation: It is not necessary to inspect every design point individually. Assigning a latent role to the locations of observation is sufficient [4]. The a priori mean and covariance functions determine the behaviour in between.

To calculate mean and covariance it is beneficial to express the posterior density by

$$p(m_s|\boldsymbol{d}) = \int p(m_{\boldsymbol{t}}|\boldsymbol{d})\, p(m_s|m_{\boldsymbol{t}})\, \mathrm{d}m_{\boldsymbol{t}} \tag{3.11}$$

which is obtained by conditioning the a priori GP to $m_{\boldsymbol{t}}$. Let assume mean $\mathbb{E}[M_{\boldsymbol{t}}|\boldsymbol{d})]$ and covariance $\mathbb{V}[M_{\boldsymbol{t}}|\boldsymbol{d}]$ of the latent variables were numerically accessible. Then, for arbitrary design points s the posterior mean takes the familiar form

$$\mathbb{E}[M(s)|\boldsymbol{d}] = \mu_M(s) + K_M(s,\boldsymbol{t})K_M(\boldsymbol{t},\boldsymbol{t})^{-1}\left(\mathbb{E}[M_{\boldsymbol{t}}|\boldsymbol{d})] - \mu_M(\boldsymbol{t})\right) \tag{3.12}$$

since a priori the conditional mean $\mathbb{E}[M(s)|m_{\boldsymbol{t}}]$ is linear w.r.t. $m_{\boldsymbol{t}}$. The conditional covariance $\mathrm{Cov}[M(r), M(s)|m_{\boldsymbol{t}}]$ does only depend on the locations $\boldsymbol{t}$ but not on the actual values $m_{\boldsymbol{t}}$ and, thus, the posterior covariance follows to read

$$\begin{aligned}\mathrm{Cov}[M(r), M(s)|\boldsymbol{d}] = K_M(r,s) - \\ - K_M(r,\boldsymbol{t})K_M(\boldsymbol{t},\boldsymbol{t})^{-1}\left[K_M(\boldsymbol{t},\boldsymbol{t}) - \mathbb{V}[M_{\boldsymbol{t}}|\boldsymbol{d}]\right]K_M(\boldsymbol{t},\boldsymbol{t})^{-1}K_M(\boldsymbol{t},s)\ ,\end{aligned} \tag{3.13}$$

an algebraic form that is not entirely unknown. These formulae are similar but not identical to those known from GP regression [5].

The object of central importance are posterior mean and co-variance of the latent variables. Although the posterior mean and covariance of the training set are assumed intractable in most applications, it is noteworthy that knowing $\mathbb{E}[M_{\boldsymbol{t}}|\boldsymbol{d}]$ and $\mathbb{V}[M_{\boldsymbol{t}}|\boldsymbol{d}]$ arrives at a convenient algebraic form similar to the one known from GP regression. In other words, the benefit is that it is not necessary to use numeric integration for computing posterior mean and covariance at every individual design point. However, moment matching proxies are known to over-estimate the support if the posterior is multi modal or heavily tailed [Murphy, 2012, Sec. 21.2.2]. That effect is illustrated by revisiting the toy example:

Example 3.4 (Moment Matiching Proxy)**.** The setting remains as in the previous example (Ex. 3.2) with $n = 4$ records. In contrast, only mean and covariance of the latent variables are calculated. SciPy's general purpose adaptive quadrature scheme is again used for numerical integration. That requires a single n-dimensional quadrature to calculate the normalizing constant $p_D(\boldsymbol{d})$. Calculating the posterior mean $\mathbb{E}[M_{\boldsymbol{t}}|\boldsymbol{d}]$ requires to carry out n times a n-dimensional integral. The covariance matrix $\mathbb{V}[M_{\boldsymbol{t}}|\boldsymbol{d}]$ consist of $\frac{1}{2}n(n+1)$ independent entries, each requiring to evaluate a n-dimensional integral. In total, the computational complexity is even worse but, fortunately, independent of the number of design points. In comparison with Example 3.2 the time spent for numerical

[4] This raises the question whether it is necessary for latent variables and locations of observation to coincide? Conversely, asked: Is it possible to project the information held by the latent variables onto other locations e.g. a regular gird?

[5] Less effective, so-called sparse approximations lead to a form that one-to-one translates into the equations known from GP regression.

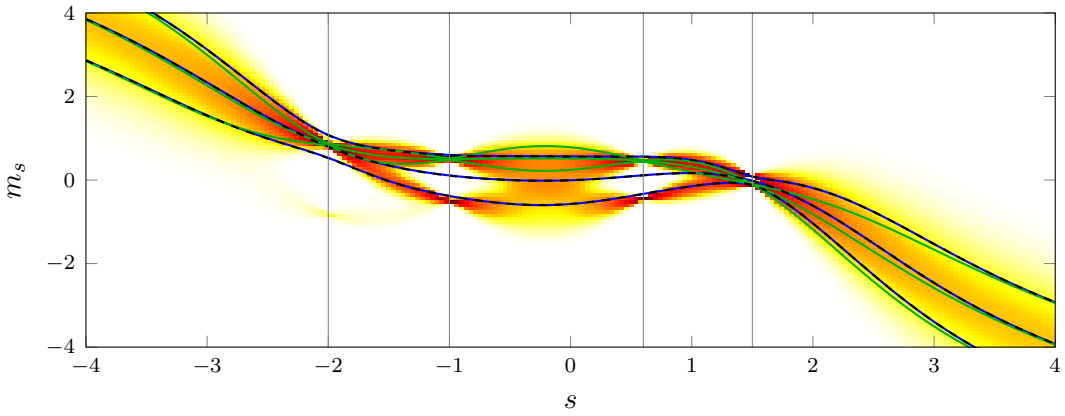

Figure 3.4.: Comparison of Gaussian approximations. The true posterior mean and pointwise standard deviation are indicated by ——. For the Laplace approximation, that is expanded around the MAP estimate, mean and standard deviation are shown by ——. It can be clearly seen that uncertainties are underestimated and the mean aligns with one of the branches. In contrast, with the mixture approach (- - -) no deviations are apparent by a visual inspection.

integration only adds up to a fraction because solely the latent variables have to be taken into account. Figure 3.4 presents the pointwise posterior mean and standard deviation. The posterior mean does not align with the most prominent mode of the Bayesian posterior. Instead, it lies in between of the modes, enclosed by a standard deviation that is large compared with the support of the individual branches. From Example 3.3, adressing the computational costs, it is clear that calculating mean and covariance becomes intractable when exceeding a hand full of records. ◁

The Gaussian moment matching proxy centers itself such that the coverage of the exact posterior is high. As presented in the above example, this is particularly evident in case of multimodality. This is because the forward KL divergence does not favour the proxy to feature high probability mass where the exact posterior does.

Although the moment matching GP proxy is optimal w.r.t. the KL divergence, this approach is ineffective since in most applications calculating first and second moment of the latent variables is not possible. Nonetheless, the general form of any kind of Gaussian approximation is given by Equations 3.12 and 3.13 [Seeger, 2004, Sec. 4]. These two equations are forming the foundation for less conserving but numerically tractable approximations. The subsequent sections resort to practical algorithms for approximative inference that scale well also for large scale applications.

3.3. Laplace Approximation

Laplace' method is an empirical but widely used approach, providing little control over optimality [e.g. Bishop, 2006, Murphy, 2012]. It replaces the high-dimensional integral by an optimization problem which may be easier to solve. This approximation can be used if the likelihood function is well peaked and concentrated to a *small* area.

The basic idea behind Laplace' method is to perform a second order Taylor series expansion of the log posterior PDF. A second order approximation fully captures the a priori GP. Therefore it is sufficient to expand the negative log likelihood function –

denoted by $\ell(\boldsymbol{d}|m_{\boldsymbol{t}})$ – in the latent variable $m_{\boldsymbol{t}}$ about some point $\hat{m}_{\boldsymbol{t}}$. The gradient and Hessian are abbreviated by

$$\boldsymbol{g} = -\nabla\ell(\boldsymbol{d}|m_{\boldsymbol{t}})\big|_{\hat{m}_{\boldsymbol{t}}} \qquad \text{and} \qquad \mathrm{H} = -\Delta\ell(\boldsymbol{d}|m_{\boldsymbol{t}})\big|_{\hat{m}_{\boldsymbol{t}}} \tag{3.14}$$

and the second order proxy of the log likelihood function is given by

$$\ell(\boldsymbol{d}|m_{\boldsymbol{t}}) \approx \ell(\boldsymbol{d}|\hat{m}_{\boldsymbol{t}}) - \boldsymbol{g}^\top(m_{\boldsymbol{t}} - \hat{m}_{\boldsymbol{t}}) - \frac{1}{2}(m_{\boldsymbol{t}} - \hat{m}_{\boldsymbol{t}})^\top \mathrm{H}(m_{\boldsymbol{t}} - \hat{m}_{\boldsymbol{t}}) \; . \tag{3.15}$$

The point of expansion remains as a free, yet unknown parameter vector. The definiteness of the Hessian depends on the choice of $\hat{m}_{\boldsymbol{t}}$. If $\hat{m}_{\boldsymbol{t}}$ is chose such that H is PD, then, the proxy likelihood as a function in $m_{\boldsymbol{t}}$ is Gaussian shaped with precision H and mean vector $\hat{m}_{\boldsymbol{t}} - \mathrm{H}^{-1}\boldsymbol{g}$. To see this, collect all constant terms and expand the square.

Gaussianity of the likelihood function is sufficient to arrive at a computationally convenient solution for some Gaussian proxy posterior. Since the second order expansion is Gaussian shaped w.r.t. the latent variables, the posterior is normal as well. This is the case because the product of two Gaussian functions is Gaussian shaped [Rasmussen and Williams, 2006, Eq. A.7]. Approximative mean and covariance of the latent variables $M_{\boldsymbol{t}}$ are given through the equations known from GP regression (Eqs. 1.19 and 1.20).

To predict at arbitrary design points r and s, the exact posterior mean and covariance (Eqs. 3.12 and 3.13) are combined with the latent variables' approximative mean and covariance. Under the Laplace approximation the conditional mean and covariance functions are given by

$$\mathbb{E}[M(s)|\boldsymbol{d}] \approx \mu_M(s) + K_M(s,\boldsymbol{t})\left(\mathrm{H}^{-1} + K_M(\boldsymbol{t},\boldsymbol{t})\right)^{-1}\left(\hat{m}_{\boldsymbol{t}} - \mathrm{H}^{-1}\boldsymbol{g} - \mu_M(\boldsymbol{t})\right) \tag{3.16}$$

$$\mathrm{Cov}[M(r), M(s)|\boldsymbol{d}] \approx K_M(r,s) - K_M(r,\boldsymbol{t})\left(\mathrm{H}^{-1} + K_M(\boldsymbol{t},\boldsymbol{t})\right)^{-1} K_M(\boldsymbol{t},s) \tag{3.17}$$

restoring the algebraic convenience of GP regression.

In order to bring Laplace' method into practice a point to perform the expansion about is missing. An obvious idea is to choose $\hat{m}_{\boldsymbol{t}}$ such that it minimizes the KL divergence. However, direct minimization turns out to be intractable because it involves averaging with respect to the exact distribution (see Sec. 3.2). In addition, a most likely non-linear system of equations would have to be solved. As a consequence, optimizing the KL divergence renders Laplace' method even more complicated than the moment matching Gaussian for a less accurate approximation. This is why a systematic approach determining the point of expansion is not rewarding.

The typically pursued strategy to choose $\hat{m}_{\boldsymbol{t}}$ is so-called mode seeking – i.e. find the MAP estimate, if unique. The objective is to find a posterior mode with high probability and wide support

$$\hat{m}_{\boldsymbol{t}} \in \arg\max\{p(m_{\boldsymbol{t}}|\boldsymbol{d})\} \; . \tag{3.18}$$

The problem of solving high dimensional integrals is replaced with computing the location of the mode which is an optimization problem. As a rule of thumb, mode seeking is faster to solve than numerical integration. Since $\hat{m}_{\boldsymbol{t}}$ is a mode, the gradient term of the Taylor expansion vanishes and the mean of the proxy likelihood function becomes $\hat{m}_{\boldsymbol{t}}$. However, finding the *global* maximum is not an easy task, especially in case of a high-dimensional parameter space featuring various saddle points and/or local maxima. The optimization problem may be rephrased as a non-linear root finding problem

$$0 := \nabla\ell(m_{\boldsymbol{t}}|\boldsymbol{d}) \qquad \leftrightarrow \qquad 0 := K_M(\boldsymbol{t},\boldsymbol{t})\boldsymbol{g}(m_{\boldsymbol{t}}) + m_{\boldsymbol{t}} - \mu_M(\boldsymbol{t}) \; . \tag{3.19}$$

In the majority of applications an analytic solution does not exists. Therefore numerical root finding is addressed which, however, is sensitive to the initial guesses. A concrete implementation using Newton's method and paying attention to stability considerations is presented in Rasmussen and Williams [2006, sec. 3.4].

Example 3.5 (Laplace Approximation)**.** The setting remains as in the previous examples (see Ex. 3.2). According to Equation 3.9, the gradient and Hessian of the log likelihood function follow to read

$$\boldsymbol{g} = \frac{2}{\epsilon^2}\left\{\frac{\hat{m}_i^2 - d_i}{\hat{m}_i^3}\right\}_{i=1,\dots,n}, \qquad \mathrm{H} = \frac{2}{\epsilon^2}\left\{\frac{3d_i - \hat{m}_i^2}{\hat{m}_i^4}\delta_{ij}\right\}_{i,j=1,\dots,n}. \tag{3.20}$$

The root finding strategy (Eq. 3.19) is addressed to determine $\hat{m}_{\boldsymbol{t}}$. To easy the implementation, SciPy's general purpose root finding procedure is used. To ensure that $\hat{m}_{\boldsymbol{t}}$ corresponds to the global maximum, every branch of the posterior must be taken into account. In total, the posterior density features 2^n branches because the square of the underlying signal is observed n-times. The permutations of $\pm\sqrt{d_i}$ are used as initial guesses. The global maximum is selected from the resulting 2^n roots. In comparison with the moment matching Gaussian, the computational costs for the root finding algorithm are negligible. But it is also true that the results differ greatly. It can clearly be seen in Figure 3.4 that the proxy mean aligns with one of the branches and the pointwise variance is considerably underestimated. ◁

A major weakness of Laplace' method is that it is essentially uncontrolled except that the proxy posterior mean aligns with one of the modes. Evaluating the Hessian at $\hat{m}_{\boldsymbol{t}}$ may lead to a poor approximation of the exact posterior density. The proxy covariance is likely to collapse to a very narrow mode since there is no mechanism to ensure spanning across all the modes. Another shortcoming of Laplace's method is exposed if the a priori covariance is large and/or a likelihood function that is highly non-Gaussian shaped. Then, a lot of probability mass is sent through regions that are far from where the approximation performs tolerably.

3.3.1. Mixture Approach

A Laplace approximation that is purely based on a single maximum of the exact posterior distribution can fail to capture important global properties. A Gaussian model is intrinsically unimodal and so unable to provide a good approximation to a multimodal PDF. Many of the distributions encountered in practice will be multimodal and so there will be different Laplace approximations according to which mode is being considered [Bishop, 2006, sec. 4.4]. In situations when the posterior distribution does not feature a clearly pronounced mode it is important to provide a more global approximation. A mixture approach can be pursued if there are multiple local maxima $\hat{m}_{\boldsymbol{t},j}$ present.

Let assume a set of local maxima is known that all in all covers the bulk of the exact posterior density . In the following, the approximated PDFs are indicated by $q(\cdot)$. The local maxima are marginalized by examining the individual Laplace approximations

$$p(m_s|\boldsymbol{d}) \approx \sum q(m_s|\boldsymbol{d}, \hat{m}_{\boldsymbol{t},j})\, p(\hat{m}_{\boldsymbol{t},j}|\boldsymbol{d}) \tag{3.21}$$

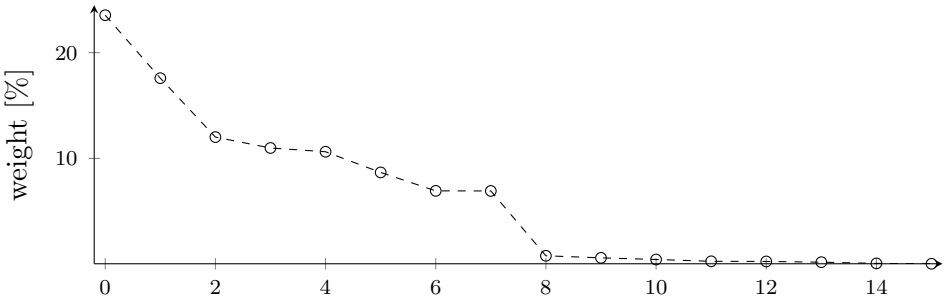

Figure 3.5.: Mixture weights for all branches in descending order.

where the general product rule is used and s refers to an arbitrary design point. Equation 3.21 has the form of a Gaussian mixture with weights $p(\hat{m}_{\boldsymbol{t},j}|\boldsymbol{d})$ because the individual Laplace approximations are normal distributed. To determine the weights an empirical Bayesian approach is addressed, adding a hierarchical stage. The weights are given by

$$p(\hat{m}_{\boldsymbol{t},j}|\boldsymbol{d}) \approx \frac{q(\boldsymbol{d}|\hat{m}_{\boldsymbol{t},j})p(\hat{m}_{\boldsymbol{t},j})}{Z} \,, \tag{3.22}$$

the so-called marginal posterior distribution. To simplify matters all points of expansion are assumed equally likely and mixing takes place only through the so-called marginal likelihood $q(\boldsymbol{d}|\hat{m}_{\boldsymbol{t},j})$. The overall normalizing constant is given by $Z = \sum q(\boldsymbol{d}|\hat{m}_{\boldsymbol{t},j})$ because in a mixture model the weights have to sum to one. Under the Laplace approximation the marginal likelihood may be see as the normalizing factors of the individual components (denominator of Bayes' law). According to Rasmussen and Williams [2006, Equation A.6], the marginal likelihood is given by

$$p(\boldsymbol{d}|\hat{m}_{\boldsymbol{t}}) = \frac{\exp\left\{-\frac{1}{2}(\hat{m}_{\boldsymbol{t}} - \mathrm{H}^{-1}\boldsymbol{g} - \mu_M(\boldsymbol{t}))\,(K_M(\boldsymbol{t},\boldsymbol{t})^{-1} + \mathrm{H})\,(\hat{m}_{\boldsymbol{t}} - \mathrm{H}^{-1}\boldsymbol{g} - \mu_M(\boldsymbol{t}))\right\}}{\sqrt{(2\pi)^n\,|K_M(\boldsymbol{t},\boldsymbol{t})^{-1} + \mathrm{H}|}} \tag{3.23}$$

depending on $\boldsymbol{d}$ through $\boldsymbol{g}$ and H. Because the mixing distributions are normal, calculating moments is simple. Mean and covariance are analogous to Equations 2.78 and 2.79.

Example 3.6 (Mixture Approach)**.** This is the continuation of Example 3.5. Rather than picking the MAP to linearize about – as shown in the previous example – all branches and their related maxima are taken into account. Figure 3.4 illustrates that by visual inspection, mean and standard deviation of the Gaussian mixture cannot be distinguished from the moment matching Gaussian proxy. It can clearly be seen in Figure 3.5 that the global maximum is not over-dominating the mixing weights. Without even knowing the exact posterior, this already raises doubts about choosing a single point of expansion. On the contrary it also shows that there is a certain number of components with little to negligible contribution, unnecessarily increasing computational efforts. The mixture approach meets the exact posterior much closer, but, computational costs increase by 2^n with respect to the number of observations. ◁

The main weakness of the mixture approach is that one usually does not know whether an optimization algorithm has found the decisive maxima. It requires exploring the entire space of latent variables which in itself is already a computational challenge, if feasible at

all. In most applications, the mixture approach is not practical because either the branches are not known and/or there are too many of them. In large scale applications the number of combinations can hardly be solved. There is no general strategy to determine the maxima of prime importance although the literature about root-finding and optimization problems is exhaustive. An alternative to Laplace' method are techniques that are based on local approximations that do not explicitly rely on the marginal likelihood function.

3.4. Local Likelihood Approximations

The following presents two techniques for approximate Bayesian inference, namely, assumed density filtering (ADF) and the expectation propagation (EP) algorithm. Although not a general purpose method, ADF is introduced first because of its intuitive nature. It may be understood as the groundwork necessary before introducing the more versatile EP algorithm. Compared to Laplace' method, these techniques require roughly similar computational costs, however, feature an improved accuracy. The following discussion settles on the special case of a GP model. With the Gaussian assumption, ADF and the EP algorithm may be seen as extensions of Kalman type filters. For a general discussion with respect to exponential families see e.g. Minka [2001], Seeger [2003], Bishop [2006], Murphy [2012] and/or Gelman et al. [2013].

3.4.1. Assumed Density Filtering

The key intuition behind ADF is twofold: 1) Rather than considering all the data at once, records are sequentially processed in terms of a Bayesian update system. 2) For every record an exact update step is perform but the intermediate posterior distributions are approximated by a tractable form [Murphy, 2012, sec. 18.5.3]. The following presents ADF for the special case assuming that the moment matching Gaussian is sufficient to approximate the intermediate steps.

The Bayesian update system is expressed as a recursive relation. The posterior distribution of the latent variables is gradually built, adding in new records one after the other. Following an arbitrary sequential ordering, the inclusion of a single record is given by

$$M_{\boldsymbol{t}}|\boldsymbol{d} \approx M_{\boldsymbol{t}}^{(j)} = M_{\boldsymbol{t}}^{(j-1)}|d_j \qquad \text{for} \qquad j = 1, \ldots, n \tag{3.24}$$

where the base case ($j = 0$) represents the a priori model. The superscript j indicates how many of the records are already worked in [6]. Performing exact update steps results in intermediate posterior distributions that are not tractable. To assume all the intermediate steps being Gaussian distributed

$$M_{\boldsymbol{t}}^{(j)} \sim \mathcal{N}\left(\mu_{\boldsymbol{t}}^{(j)}, \Sigma_{\boldsymbol{tt}}^{(j)}\right) \tag{3.25}$$

is the approximation upon which ADF is based. An interpretation of this recursion is that every intermediate proxy distribution serves as prior for the subsequent record.

The attempt to systematically derive $\mu_{\boldsymbol{t}}^{(j)}$ and $\Sigma_{\boldsymbol{tt}}^{(j)}$ results in similar problems as with the point of expansion in Laplace' method. A direct approach that tries to determine

[6] The superscript notation is chosen to keep equations concise, although, a vertical bar – to indicate a conditional – might appear more logical.

mean and co-variance such that the global KL divergence is minimized involves averaging with respect to the exact distribution and, thus, is assumed intractable. Even though not suited to characterize global optimality, the typically performed approximation is to minimize the local KL divergence for every individual record. This represents a much simpler problem to solve. At a local scale – i.e. only considering a single record – the KL divergence is minimized by the moment matching Gaussian proxy. The inclusion of a single record consists of two tasks: 1) Calculate exact posterior mean and variance for that very record, than, 2) perform a prediction on all the other latent variables.

Mean $\mu_j^{(j-1)}$ and standard deviation $\sigma_j^{(j-1)}$ from the previous step serve as prior for the inclusion of the next record. Because the tractable form is chosen to be the moment matching Gaussian, first and second moments are calculated from the exact posterior distribution

$$Z_j = \int p(d_j|m_j)\, p(m_j|\mu_j^{(j-1)}, \sigma_j^{(j-1)})\, \mathrm{d}m_j \tag{3.26}$$

$$\mathbb{E}\left[M_j^{(j-1)}\middle|d_j\right] = \frac{1}{Z_j}\int m_j p(d_j|m_j)\, p(m_j|\mu_j^{(j-1)}, \sigma_j^{(j-1)})\, \mathrm{d}m_j \tag{3.27}$$

$$\mathbb{V}\left[M_j^{(j-1)}\middle|d_j\right] = \frac{1}{Z_j}\int m_j^2 p(d_j|m_j)\, p(m_j|\mu_j^{(j-1)}, \sigma_j^{(j-1)})\, \mathrm{d}m_j - \mathbb{E}\left[M_j^{(j-1)}\middle|d_j\right]^2 \tag{3.28}$$

which, in general, do not possess explicit solutions. In many applications, one dimensional integrals are amenable to numerical approximations. Taking all n records into account, leads in total to $3n$ times one-dimensional integrals to be solved. Although numerical integration is required, the computational costs are cut down by carrying out one-dimensional integrals which are undemanding compared to a n-dimensional quadrature.

Once the moments of the j^{th} record are calculated, the so-called one-step ahead prediction follows. According to the moment matching Gaussian proxy (Eqs. 3.12 and 3.13), predictive mean and covariance are given by

$$\mu_{\boldsymbol{t}}^{(j)} = \mu_{\boldsymbol{t}}^{(j-1)} + \Sigma_{\boldsymbol{t}j}^{(j-1)} {\sigma_j^{(j-1)}}^{-2} \left(\mathbb{E}\left[M_j^{(j-1)}\middle|d_j\right] - \mu_j^{(j-1)}\right) \tag{3.29}$$

$$\Sigma_{\boldsymbol{tt}}^{(j)} = \Sigma_{\boldsymbol{tt}}^{(j-1)} + \Sigma_{\boldsymbol{t}j}^{(j-1)} {\sigma_j^{(j-1)}}^{-2} \left(\mathbb{V}\left[M_j^{(j-1)}\middle|d_j\right] - {\sigma_j^{(j-1)}}^2\right) {\sigma_j^{(j-1)}}^{-2} \Sigma_{j\boldsymbol{t}}^{(j-1)} \tag{3.30}$$

where $\Sigma_{\boldsymbol{t}j}^{(j-1)}$ refers to the cross-covariance of latent variables and j^{th} record i.e. the j^{th} row of the covariance matrix. In the consecutive step this proxy posterior serves as the prior to include the next datum. This procedure is repeated until all records are worked in.

When predicting at design points r and s, it is not necessary to carry those along while progressing. It is sufficient to model the latent variables and defer predictions as a post-processing step. The predictive mean is given by

$$\mathbb{E}[M(s)|\boldsymbol{d}] \approx \mu_M(s) + K_M(s, \boldsymbol{t}) K_M^{-1}(\boldsymbol{t}, \boldsymbol{t}) \left(\mu_{\boldsymbol{t}}^{(j)} - \mu_M(\boldsymbol{t})\right) \tag{3.31}$$

and the predictive covariance follows to read

$$\begin{aligned}\mathrm{Cov}[M(r), M(s)|\boldsymbol{d}] \approx\ & K_M(r, s) + \\ & + K_M(r, \boldsymbol{t}) K_M^{-1}(\boldsymbol{t}, \boldsymbol{t}) \left(\Sigma_{\boldsymbol{tt}}^{(j)} - K_M(\boldsymbol{t}, \boldsymbol{t})\right) K_M^{-1}(\boldsymbol{t}, \boldsymbol{t}) K_M(\boldsymbol{t}, s)\end{aligned} \tag{3.32}$$

where the a priori covariance function K_M may be seen as a means for interpolation. Because this property of ADF is not obvious at first sight it is briefly derived in the following. One way to show that Equations 3.31 and 3.32 are equivalent to carrying along design points is to start off from the hypothesis and use the recursive relation to replace $\mu_{\boldsymbol{t}}^{(j)}$ and $\Sigma_{\boldsymbol{tt}}^{(j)}$. Once all terms are factored out and the hypothesis for the $(j-1)^{\text{th}}$ step is identified, it remains to show that

$$K_M(r,\boldsymbol{t})K_M^{-1}(\boldsymbol{t},\boldsymbol{t})\Sigma_{\boldsymbol{t}j}^{(j-1)} = \Sigma_{rj}^{(j-1)} \tag{3.33}$$

holds. The key to recognize this relation is the cross-covariance with the latent variables

$$\Sigma_{x\boldsymbol{t}}^{(j)} = \Sigma_{x\boldsymbol{t}}^{(j-1)}{\Sigma_{\boldsymbol{tt}}^{(j-1)}}^{-1}\Sigma_{\boldsymbol{tt}}^{(j)} \tag{3.34}$$

which may be read off from Equation 3.13. This equation can be transformed into

$$\Sigma_{r\boldsymbol{t}}^{(j)}{\Sigma_{\boldsymbol{tt}}^{(j)}}^{-1} = \Sigma_{r\boldsymbol{t}}^{(j-1)}{\Sigma_{\boldsymbol{tt}}^{(j-1)}}^{-1} \equiv K_M(r,\boldsymbol{t})K_M^{-1}(\boldsymbol{t},\boldsymbol{t}) \tag{3.35}$$

and the quantity that mediates between design points and latent variables is invariant while progressing. This shows that modeling design points either while progressing or as a post-processing step is in fact equivalent. The information is solely stored in the latent variables. Performing predictions subsequently saves computational costs and memory demands, however, may also lead to a difficulty. It might be necessary to regularize the a priori covariance matrix because the inversion/factorization is hindered by numerical instabilities.

As a proof of concept, applying ADF to travel time tomography is presented in Section 4.3. The two step strategy that is presented in Section 6 about modeling the archeomagnetic field relies on a special case of ADF.

The major weakness of ADF is that the resulting approximation strongly depends on the sequential ordering [Seeger, 2003]. If the measurements are appreciably correlated a gradually set up posterior is sensitive to the order of succession. A record that was worked in early might turn out to be of greater value at the end of the chain and vice versa. Furthermore, because each datum is individually approximated, mutual interactions are weakened and the joint proxy posterior could well give a poor approximation [Bishop, 2006, sec. 10.7]. As with Laplace' method, ADF is an intuitive algorithm with little control about optimality. These shortcomings are addressed by the EP algorithm, which is introduced in the following.

3.4.2. Expectation Propagation Algorithm

The ADF chain is restricted to incorporate each record only once, not to contradict the fundamental concepts of stochastic inference. It was Minka in 2001 who observed that the basic ideas underlying ADF can be generalized. The concept behind the EP algorithm is to assign a record for re-inclusion while preserving contributions of the rest of the training set. EP may be seen as an iterative variant of ADF rendering the tractable posterior distribution independent of the sequential ordering constraint. By multiply passing over the data, EP adopts a more global point of view iteratively refining the tractable posterior distribution. Although EP in its most general form can be expressed for exponential families, the following assumes the tractable form to be Gaussian

$$M_{\boldsymbol{t}}|\boldsymbol{d} \approx M_{\boldsymbol{t}}^{(i)} \sim \mathcal{N}\left(\mu_{\boldsymbol{t}}^{(i)},\, \Sigma_{\boldsymbol{tt}}^{(i)}\right) \tag{3.36}$$

where i refers to an iteration until convergence. This differs from ADF where the index j points to a certain record. The increased accuracy stems from multiply passing through the data set which makes it necessary to compensate a previous inclusion while preserving what has been leaned from the other records.

Suppose we are at iteration i with proxy posterior mean $\mu_{\boldsymbol{t}}^{(i)}$ and covariance $\Sigma_{\boldsymbol{tt}}^{(i)}$, e.g. by passing through an ADF chain. The record d_j – which is assumed being worked in at the k-th step – is assigned for re-inclusion. This setting requires bookkeeping of three indices: i denotes to the current iteration, j designates a datum for re-inclusion and k refers to the previous step where d_j was already worked in. To consider d_j one more time, its contribution to the k-th step needs to be removed. The so-called *cavity* distribution is the i-th posterior divided by the k-th proxy likelihood

$$q(m_{\boldsymbol{t}}^{(i)}|\boldsymbol{d} \setminus d_j) = Z_j \frac{q(m_{\boldsymbol{t}}^{(i)}|\boldsymbol{d})}{q(d_j|m_j^{(k)})} \tag{3.37}$$

where densities are indicated by $q(\cdot)$ to emphasize that we are dealing with the Gaussian proxies and Z_j is the normalizing constant. This approach can be seen as kind of a reversal of Bayes' law but preserves correlations governed by the remaining records.

Because we have chosen the tractable form being normal distributed, all proxy likelihoods $q(d_j|m_j)$ must also be Gaussian shaped w.r.t m_j and thus Z_j is explicit. Mean and co-variance of the likelihood are obtained by comparing prior and proxy posterior at the k-th iteration. Using again the identity A.7 in Rasmussen and Williams [2006] we calculate mean and covariance

$$\Sigma_{lkh}^{(k)} = -\Sigma_{jj}^{(k)} - \Sigma_{jj}^{(k)} \left(\mathbb{V}\Big[M_j^{(k)} | d_j \Big] - \Sigma_{jj}^{(k)} \right)^{-1} \Sigma_{jj}^{(k)} \tag{3.38}$$

$$\mu_{lkh}^{(k)} = \mu_j^{(k)} - \Sigma_{jj}^{(k)} \left(\mathbb{V}\Big[M_j^{(k)} | d_j \Big] - \Sigma_{jj}^{(k)} \right)^{-1} \left(\mathbb{E}\Big[M_j^{(k)} | d_j \Big] - \mu_j^{(k)} \right) \tag{3.39}$$

where $\mathbb{E}\big[M_j^{(k)}|d_j\big]$ and $\mathbb{V}\big[M_j^{(k)}|d_j\big]$ are numerically calculated with respect to the *true* likelihood function (see Eqs. 3.26 to 3.28). In order to correct for the proxy likelihood we have to store μ_{lkh} and Σ_{lkh} for every individual record.

The quotient of Gaussian shaped functions is again Gaussian shaped since the denominator is less variant than the numerator in case of a GP regression. For the cavity distribution the Woodburry identity (a.k. matrix inversion lemma; Rasmussen and Williams [2006, A. 3]) is addressed and mean and covariance are given by

$$\Sigma_{tt}^{(i\setminus j)} = \left({\Sigma_{tt}^{(i)}}^{-1} - \mathbb{I}_{\boldsymbol{t}j} {\Sigma_{\text{lkh}}^{(k)}}^{-1} \mathbb{I}_{j\boldsymbol{t}} \right)^{-1} = \Sigma_{tt}^{(i)} - \Sigma_{tj}^{(i)} \left(\Sigma_{jj}^{(i)} - \Sigma_{\text{lkh}}^{(k)} \right)^{-1} \Sigma_{jt}^{(i)} \tag{3.40}$$

$$\mu_t^{(i\setminus j)} = \ldots \qquad = \mu_t^{(i)} + \Sigma_{tj}^{(i)} \left(\Sigma_{jj}^{(i)} - \Sigma_{\text{lkh}}^{(k)} \right)^{-1} \left(\mu_{\text{lkh}}^{(k)} - \mu_j^{(i)} \right) \tag{3.41}$$

where $\mathbb{I}_{\boldsymbol{t}j}$ refers to an identity that is padded with zeros. The calculation of the proxy likelihood and cavity distribution is the tricky part of the EP algorithm. The individual Gaussian moment matching proxies and the one-step ahead prediction are analogous to ADF and, thus, the heart of the EP algorithm is formed by the ADF equations.

The cavity distribution $M_{\boldsymbol{t}}^{(i\setminus j)}$ can safely serve as prior to again incorporate the j-th record because d_j's contribution is removed. The *new, better* moment matching Gaussian

proxy for the j-th record is numerically obtained through Equations 3.26 to 3.28, analogously to the ADF formalism. Likewise, projecting on the latent variables is the same as for ADF. The one step ahead prediction is given by

$$\mu_{\boldsymbol{t}}^{(i+1)} = \mu_{\boldsymbol{t}}^{(i\backslash j)} + \Sigma_{\boldsymbol{t}j}^{(i\backslash j)} \Sigma_{jj}^{(i\backslash j)^{-1}} \left(\mathbb{E}\Big[M_j^{(i\backslash j)} | d_j \Big] - \mu_j^{(i\backslash j)} \right) \tag{3.42}$$

$$\Sigma_{\boldsymbol{tt}}^{(i+1)} = \Sigma_{\boldsymbol{tt}}^{(i\backslash j)} + \Sigma_{\boldsymbol{t}j}^{(i\backslash j)} \Sigma_{jj}^{(i\backslash j)^{-1}} \left(\mathbb{V}\Big[M_j^{(i\backslash j)} | d_j \Big] - \Sigma_{jj}^{(i\backslash j)} \right) \Sigma_{jj}^{(i\backslash j)^{-1}} \Sigma_{j\boldsymbol{t}}^{(i\backslash j)} \tag{3.43}$$

and the base case $i = 0$ refers to the a priori assumption. Mean μ_{lkh} and (co)-variance Σ_{lkh} of the proxy likelihoods are carried along for every individual record and are iteratively optimized. If there is additional knowledge, both may be initialized with coarse starting values. Otherwise, starting with zero precision is analogous to an initial ADF chain. The iteration should be stopped when the parameters do not change any more. An often chosen termination criterion is the maximum of the absolute value for both mean and covariance.

Revisiting the ongoing example is intended to showcase increased accuracy together with the rapid convergence.

Example 3.7 (Expectation Propagation)**.** The setting remains just the same as in the previous examples. The first batch of the EP iteration is formed by an ordinary ADF chain. As in the previous examples SciPy's general purpose quadrature scheme is addressed for numerical integration. After just two more iterations the absolute difference in all entries of mean $\mu_{\boldsymbol{t}}^{(3)}$ and covariance $\Sigma_{\boldsymbol{tt}}^{(3)}$ is smaller than 10^{-4}. Taking the tolerance of 10^{-3} of the quadrature into account, this can be see as numerically identical. Because of the high accuracy an illustration is skipped. A comparison with the moment matching proxy would just yield the same view already provided in Figure 3.4.

The EP algorithm demonstrated empirical success in numerous applications e.g. the clutter problem [Minka, 2001], TrueSkill™ [Murphy, 2012], GP classification [Rasmussen and Williams, 2006] and many more. One disadvantage of EP is that there is no guarantee that the iterations will converge [Bishop, 2006]. If convergent the solution will be a fixed point obtained at tolerable computational costs (typically less than 5 iterations). However, precise convergence against the moment matching proxy is *not* a property of the algorithm. EP can offer local KL optimality depending on how the updates are performed. Although EP is not global KL optimal, the algorithm provides control that e.g. a linearization or Laplace's approximation are lacking.

Part II.

Seismology

Chapter 4 presents a new approach in travel time inversion which provides insight from a correlation-based Bayesian point of view. Therefore, the concept of Gaussian process regression is adopted to estimate a velocity model. The non-linear travel time integral is approximated and a heuristic covariance describes correlations amongst observations and a priori model. That approach assess a proxy of the Bayesian posterior distribution at ordinary computational costs.

Chapter 5 presents a systematic approach determining the spatial extent and duration of seismic sources. The concept of Gaussian processes regression is adopted to the classic inverse problem working backwards characterizing the external force that gave rise to recorded seismograms. The corner stone of the modeling approach is formed by the correlations amongst observations and heterogeneously distributed sources of continuous or transient nature.

4. Travel Time Inversion

A major application in seismology is the determination of seismic velocity models. Arrival times are the basic measurements of seismology reflecting the underlying velocity structures [1]. The fundamental data to deduce the velocity structure between source and receiver are the travel times of various seismic phases. In the following I am going to provide insight to travel time inversion from a statistical point of view. Therefore, the concept of Gaussian process regression is adopted using travel time observations to estimate the underlying velocity model.

For body waves, in a high frequency approximation, the transit times can be represented by ray theory. Continuous lines perpendicular to the propagation of wavefronts are called ray paths. For a wavefront generated from a point source at position $\boldsymbol{x}_s$ and a receiver located at $\boldsymbol{x}_r$, the travel time can be computed by an integral along the ray path

$$T = \int_{\boldsymbol{x}_s}^{\boldsymbol{x}_r} \frac{1}{c} \, \mathrm{d}s \tag{4.1}$$

with underlying velocity c. For heterogeneous velocity structures, rays are not straight because of refraction within the Earth. The solution to the ray path usually begins with the non-linear partial differential Eikonal equation

$$|\nabla T|^2 = \frac{1}{c^2} \tag{4.2}$$

which forms the basis for ray theoretical approaches. It is an approximate solution valid at high frequencies, neglecting gradients in the Lamé parameters and the wave amplitudes. Solving the two-point ray tracing problem is a non-trivial task. It is not subject of this work and we will negotiate that obstacle latter.

Based on observed travel times the aim is to invert for a velocity structure that explains the data. Travel times are in fact an average of the inverse of the velocity. Therefore, the velocity of the intervening structure can be inverted. Each travel time gives an integral constraint on the velocity between source and receiver. Hence, a set of travel times provides decisive information about the velocity structure.

A wide range of inverse methods for different structures and geometries have been established. There is a rich variety of textbooks discussing travel time inversion in greater depth. References used for that work are Shearer [2009] in the first place in combination with Gubbins [2004], Chapman [2004], Stein and Wysession [2003], Kennett [2002] and Aki and Richards [2002]. For a difference-based and discretized model, in particular the work by Malinverno and Parker [2006] is to be mentioned, which considers 1D travel time inversion from a statistical point of view.

In comparison to the literature mentioned, I am going to present an alternative approach. It is a correlation based Bayesian attempt estimating the velocity model's mean

[1] The useful information which may be deduced from amplitudes and waveforms are ignored. Interpretation of the full waveform of a seismogram is a relatively recent innovation.

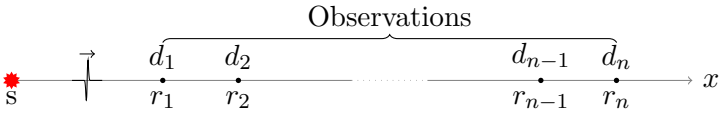

Figure 4.1.: Sketch of the source-receiver geometry.

and covariance, posterior to travel time observations. The following method may be interpreted as the continuous generalization of the Bayesian approach presented by Malinverno and Parker [2006] with a priori assumptions made over velocity rather than slowness. The techniques introduced in Chapter 1 – GP regression with non-linear functionals – are adopted for travel time inversion. The underlying correlation structure will be a heuristic choice with little physical motivation and the kernel may be understood as a means of interpolation. The aim of the subsequent examples is more about the concept rather than a real-world application.

4.1. Direct Waves

Consider only travel time measurements in a purely 1-dimensional setting [2]. To circumvent reflections a continuous velocity model is considered. Only direct waves exist and we do not have to solve for the ray path. The model is basically a straight line with actual propagation velocity $\tilde{c}(x) > 0$. At position s a point source is located and travel time measurements are gathered along the line at locations r_i, $i = 1, \dots, n$ with n the number of receivers. Figure 4.1 provides a sketch of the setting.

If the true velocity $\tilde{c}$ were known, travel times are calculated according to Equation 4.1 . The path to be integrated along is the model's coordinate axis and the observation functional is given by

$$T_{s,r}[\tilde{c}] = \int_s^r \frac{1}{\tilde{c}(x)} \, \mathrm{d}x \; . \tag{4.3}$$

Not to end up with negative values, flipping bounds of integration may be necessary if the source is on the right of the receiver. Picking and timing errors in the data are common. Actual values recorded are denoted by

$$d = \left\{ T_{s,r_i}[\tilde{c}] + e_i \mid i = 1, \dots, n \right\} \tag{4.4}$$

corrupted by an error term e_i. Although travel times may be observed at will, we do not know the true velocity model $\tilde{c}$. To proceed with a Bayesian approach, the velocity model is assumed a GRF determined by

$$\mathbb{E}[C(x)] = \mu_C(x) \qquad \text{and} \qquad \mathrm{Cov}[C(x), C(y)] = K_C(x, y) \; . \tag{4.5}$$

Modelling C as a GRF is questionable because because the velocity could at least theoretically become negative. To overcome concerns, let assume the average seismic velocity is significantly greater than zero and the variance is small compared to its mean. With respect to C, the observational functional is not linear. To stay with Gaussianity, a

[2] Not to be confused with a 3d medium only varying along one axis.

linearization of $T_{s,r}[\cdot]$ is performed (see Section 3). A Gaussian moment matching approximation is not possible since first and second moment of $1/x$, $X \sim \mathcal{N}(\mu, \sigma^2)$ do not exist [Robert, 1991]. To its first order, a Taylor expansion yields

$$\frac{1}{C} \approx \frac{1}{\mu_C} - \frac{1}{\mu_C^2}(C - \mu_C) \tag{4.6}$$

where μ_C – the a priori mean – serves as point of expansion. In its essence, this is nothing but the traditional approach in tomography [e.g. Gubbins, 2004]. It is adequate to just approximate $1/C$ as the integral is an affine transformation. The stochastic data model is approximated by

$$D = T_{s,r}[C] + E \approx \int_s^r \frac{2}{\mu_C(x)} - \frac{C(x)}{\mu_C(x)^2}\,\mathrm{d}x + E \tag{4.7}$$

with independent zero mean normal noise E of variance σ^2. To proceed with the approximated Bayesian posterior distribution we have to determine mean and covariance functions as well as correlations. The observation's approximated a priori mean reads

$$\mathbb{E}[D] = \mathbb{E}[T_{s,r}[C] + E] \approx \int_s^r \frac{2}{\mu_C(x)} - \frac{\mathbb{E}[C(x)]}{\mu_C(x)^2}\,\mathrm{d}x = \int_s^r \frac{1}{\mu_C(x)}\,\mathrm{d}x = T_{s,r}[\mu_C]\,. \tag{4.8}$$

The covariance for the observational functional is approximated by

$$\mathrm{Cov}[T_{s,r'}[C], T_{s,r}[C]] \approx \int_s^{r'} \int_s^r \frac{K_C(x,x')}{\mu_C(x)^2 \mu_C(x')^2}\,\mathrm{d}x'\,\mathrm{d}x\,. \tag{4.9}$$

Remember that the translational part drops out calculating the covariance. Considering measurement noise, the data covariance reads

$$\mathbb{V}[D] = \mathrm{Cov}[T_{s,r'}, T_{s,r}] + \mathrm{Cov}[E,E] \approx \int_s^{r'} \int_s^r \frac{K_C(x',x)}{\mu_C(x')^2 \mu_C(x)^2}\,\mathrm{d}x'\,\mathrm{d}x + \delta_{r'r}\sigma^2 \tag{4.10}$$

due to assumed independence of C and E and δ refers to the Kronecker delta. The predictive functional is the pointwise evaluation of C at locations y. Correlations of observations and model are

$$\mathrm{Cov}[C(y), T_{s,r}[C]] \approx \mathrm{Cov}\left[C(y), \int_s^r \frac{2}{\mu_C(x)} - \frac{C(x)}{\mu_C(x)^2}\,\mathrm{d}x\right] = -\int_s^r \frac{K_C(y,x)}{\mu_C(x)^2}\,\mathrm{d}x\,. \tag{4.11}$$

Notice that although C's correlation length may be short ranging, correlations of travel times are affecting the whole ray path (see Fig. 4.3 for an example).

We determined all quantities necessary to compute the approximated Bayesian posterior distribution of the velocity model. C's posterior mean and covariance are given by

$$\mathbb{E}[C(y)|d] = \mu_C(y) + \mathrm{Cov}[C(y), T_{s,r}]\,\mathbb{V}[D]^{-1}\big(T_{s,r}[\mu_C] - d\big) \tag{4.12}$$

$$\mathrm{Cov}[C(x), C(y)|d] = \mathrm{Cov}[C(x), C(y)] - \mathrm{Cov}[C(x), T_{s,r}]\,\mathbb{V}[D]^{-1}\mathrm{Cov}[T_{s,r}, C(y)] \tag{4.13}$$

according to Equations 1.37 and 1.38. It is worth emphasizing that we did not assimilate the non-linear observational functional (Eq. 4.3) but its approximation (Eq. 4.7).

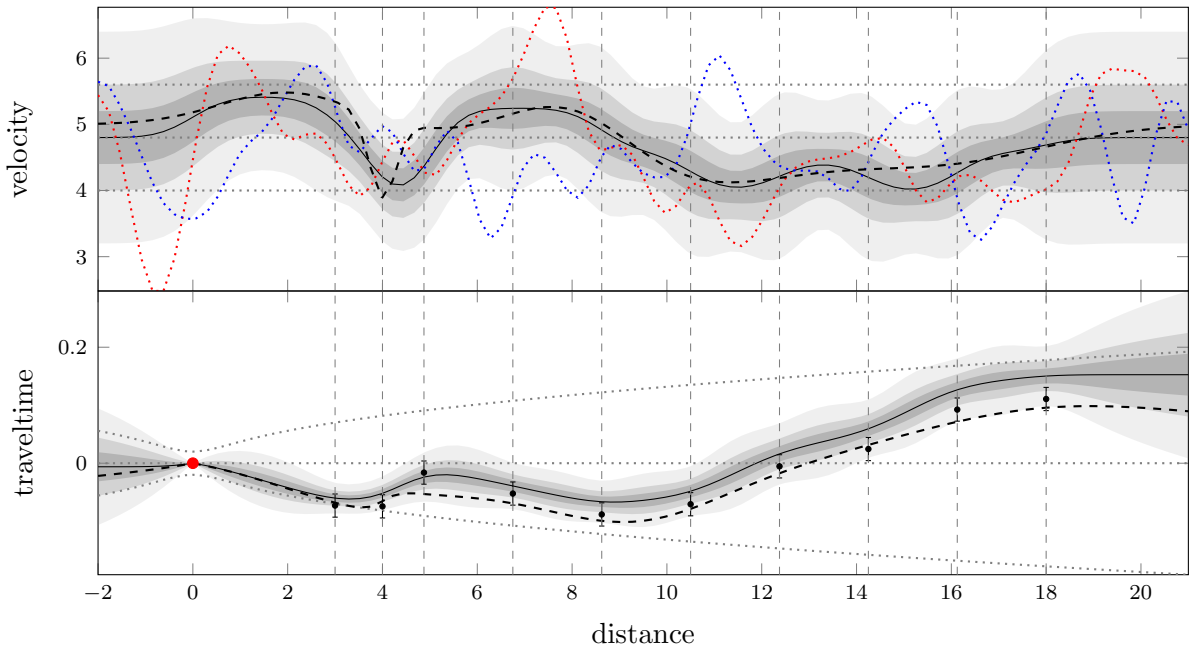

Figure 4.2.: Prediction of velocity structure (top) and travel time curve (bottom) with its linear trend removed. The posterior mean is shown by —— whereas half, once and twice the standard deviation is indicated by the gray shaded area. A priori assumptions are shown by ····· (mean ± standard deviation). For comparison the true velocity structure and travel time curve are plotted by - - -. Observations are depicted by • and • indicates the source location. In addition, realizations of the prior ····· and posterior ····· process are shown. Receiver locations are indicated by vertical lines.

To predict on travel time curves, again a linearization is necessary. Indeed, the predictive functional is $T_{s,y}[\cdot]$ and we again perform a 1^{st}-order Taylor expansion to make it linear. In contrast to observations, C's posterior mean (Eq. 4.12) serves as point of expansion. The approximated posterior mean and covariance are given by

$$\mathbb{E}[T_{s,y}|d] \approx \int_s^y \frac{1}{\mathbb{E}[C(x)|d]}\,\mathrm{d}x \tag{4.14}$$

$$\mathrm{Cov}[T_{s,y}, T_{s,y'}|d] \approx \int_s^y \int_s^{y'} \frac{\mathrm{Cov}[C(x), C(x')|d]}{\mathbb{E}[C(x)|d]^2\,\mathbb{E}[C(x')|d]^2}\,\mathrm{d}x\,\mathrm{d}x' \,. \tag{4.15}$$

Both estimates come with the deficit of approximating the non-linear functionals. If a priori assumptions are reasonably good, non-linearity is not a major concern. In Section 4.3 we discuss how to address non-linearity. Only considering direct waves is hardly satisfying. Lets have a look at an example before we go ahead with reflections at a discontinuity.

Example 4.1. For demonstration, an arbitrary but continuous velocity model $\tilde{c}$ is generated. It is composed of Gaussian blobs with a constant offset, depicted in Figure 4.2. To illustrate the limits of resolution the model shows a small scale wiggle at about 4. Synthetic travel times are calculated from that model and are corrupted by zero mean normal noise of variance $\sigma = 0.020$. The geometry of a single source accompanied by $n = 10$ receivers is outlined in Figure 4.2. For the receiver nearest to the source the

relative timing error is of about 5%. The velocity model C as well as travel times T will be predicted on a grid of 112 points.
Let assume there is an independent source of information suggesting a constant mean function of $\mu_C = 4.80$ with a deviation of ± 0.80 . Although it would be preferable to derive a covariance from principles of wave propagation, a heuristic approach is used. The assumption of a continuous velocity model motivates a smooth kernel function. The covariance of choice is the SE kernel (Eq. 1.14). The a priori knowledge already specifies the variance ($\tau = 0.80$) whereas the characteristic length scale is motivated by the sampling theorem. Half the mean free receiver spacing gives a characteristic length of $\ell = 0.83$ (being ℓ away from observations, the correlation drops to $1/e$). A priori assumptions are depicted in Figure 4.2. With the SE kernel and a constant a priori mean function an analytic solution for the integrals in Equations 4.9, 4.10 and 4.11 are at hand. Ingredients for the calculus – the antiderivatives of the SE kernel and the error function – may be found in Bronstein et al. [2000, Eqs. 8.100a and 8.100d].
Figure 4.2 shows the reconstruction of the velocity model. The posterior mean nicely follows the large scale features of $\tilde{c}$. Structures which are small compared to ℓ are not caught how ever densly receivers are spaced (notice the wiggle at about 4 with reduced receiver spacing). Although uncertainties shrank, a realization of the posterior process still shows significant variations in comparison to a realization of the prior process. If the kernel's correlation length were increased, the posterior process and its realizations become smoother. Interestingly, the posterior standard deviation shows minima in between the receivers. Velocity perturbations amongst source and receiver are affecting the travel time the most. This is due to the integral in the observational functional (Eq. 4.3) only putting a constraint on the average velocity.
The posterior travel time curve is also shown in Figure 4.2. For better illustration, the travel time curves are cleared of their linear trend. Therefore, the a priori mean is subtracted. Equations 4.14 and 4.15 do not possess a solution of explicit form. For numeric integration the cumulative trapezoidal rule is used. The overall reconstruction is not too good, showing late arrivals. Due to non-linearity, even at points of observation the deviation is large. Non-linearity has a substantial effect since a priori assumptions are pretty far from reality and the reconstruction is just of medium quality. ◁

4.2. Reflected Waves

The approach presented in the following shall consistently unify two distinct concepts: Discrete interfaces with an abrupt change in velocity and zones of continuous transition. Therefore, we still consider the oversimplified setting from Figure 4.1, however, with regard to a discontinuous velocity model. For simplicity let assume a velocity structure with single interface

$$\tilde{c}(x) = \begin{cases} \tilde{c}_l(x) & x < x_j \\ \tilde{c}_r(x) & x \geq x_j \end{cases} , \qquad c_l(x_j) \neq c_r(x_j) \tag{4.16}$$

where we are *certain* about the location x_j of the discontinuity [3]. In addition to direct waves the interface causes a reflected wave. Again, within a 1d model we are not concerned

[3] The location of an interface is typically another source of uncertainty.

about the ray path. The reflected wave basically travels from the interface back to the receiver and the travel time for a reflection reads

$$R_{s,r,j}[c] = T_{s,j}[c] + T_{j,r}[c] \ . \tag{4.17}$$

To consider travel times of reflections in the modeling approach we need to calculate the covariance for the reflected wave, the correlation amongst reflected and direct wave and the correlations amongst model and a reflected wave. That are

$$\mathrm{Cov}[R_{s,r,j}, R_{s,r',j}] = \mathbb{V}[T_{s,j}] + \mathrm{Cov}[T_{s,j}, T_{j,r'}] + \mathrm{Cov}[T_{j,r}, T_{s,j}] + \mathrm{Cov}[T_{j,r}, T_{j,r'}] \ , \tag{4.18}$$

$$\mathrm{Cov}[R_{s,r}, T_{s,r'}] = \mathrm{Cov}[T_{s,j}, T_{s,r'}] + \mathrm{Cov}[T_{j,r}, T_{s,r'}] \ , \tag{4.19}$$

$$\mathrm{Cov}[C(x), R_{s,r,j}] = \mathrm{Cov}[C(x), T_{s,j}] + \mathrm{Cov}[C(x), T_{j,r}] \tag{4.20}$$

and all covariances are determined by combinations of Equations 4.9 and 4.11 .

An intuitive approach is to encode the knowledge about the interface in terms of a piecewise a priori mean function. However, if we proceed with a covariance continuous across the interface, the two regimes are coupled through the kernel. Then, the reconstruction of C will also be continuous and the smoothness is determined by the kernel. With the knowledge of an interface, the big concern is to setup an appropriate covariance comprising the discontinuity. A simplistic approach constructing such a kernel is to a priori model the two velocity regimes independent of each other. Therefore, let the prior GRF of the overall model partition into distinct processes

$$C = \begin{cases} C_l & x < x_j \\ C_r & x \geq x_j \end{cases} \tag{4.21}$$

left and right of x_j. To accomplish independence the indicator function is used to truncate at x_j. Considered as an operator, the indicator function – e.g. $\mathbf{1}_{x<x_j}[\cdot]$ – is linear (additivity and homogeneity). The left hand a priori GRF reads

$$C_l \sim \mathcal{GP}(\mathbf{1}_{x<j}\, \mu_l,\ \mathbf{1}_{x<j}\, K_l\, \mathbf{1}_{y<j}) \tag{4.22}$$

with to be truncated mean function μ_l and kernel K_l. The right hand side GRF constructs accordingly. Remember that linear combinations of Gaussian are still normal and we have

$$C = C_l + C_r \ . \tag{4.23}$$

Notice that the assumption of independent regimes does not mean travel time observations being partly uncorrelated. Due to integration along the ray path correlations for travel times are ranging across the interface. To obtain the approximated Bayesian posterior we proceed as before (see Eqs. 4.12 and 4.13). Although a priori kernels left and right of x_j are independent, the posterior covariance will show correlations across the interface.

Being certain about the location of an interface is typically not the case. It is rather an additional source of uncertainty and shall be subject of further research. Nonetheless, some thoughts may be found in Section 4.5. Before we go ahead to better accommodate non-linearity, lets have a look at an example.

Example 4.2. Let us adopt the setting of Example 4.1. An interface at $x_j = 10$ is introduced to the previous model. The actual velocity structure is shown in Figure 4.4.

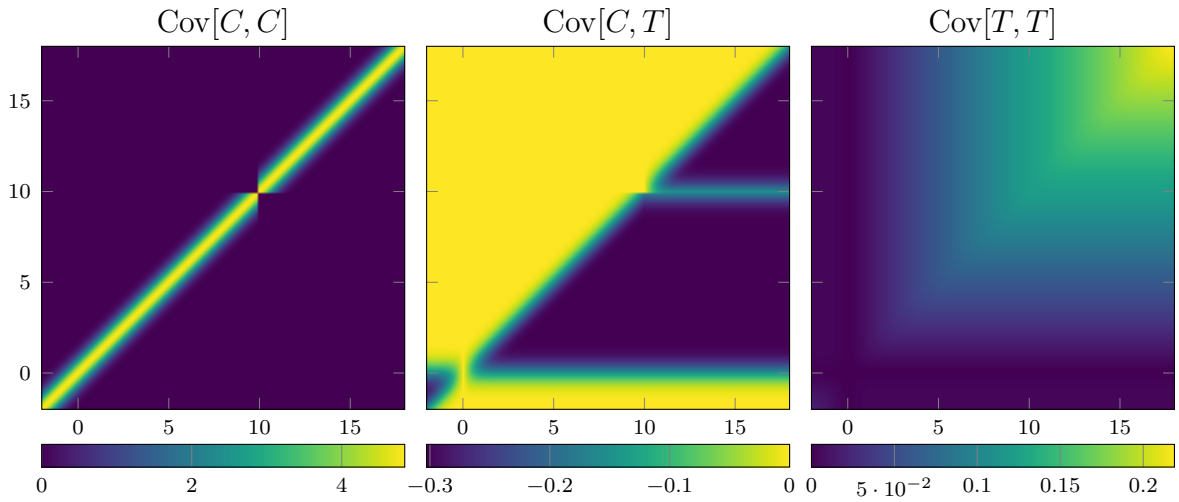

Figure 4.3.: Depiction of a priori correlation kernels. Although, the characteristic length of the truncated SE kernel (left panel) is short, correlations amongst model and travel times (mid panel) and the travel time covariance (right panel) are showing interdependencies along the whole ray path. The jump at $x_j = 10$ encoded in the kernels is clearly visible.

Travel times are generated from the discontinuous model and the receivers on the left of x_j are in addition recording the reflection. The kernel of choice is almost be the same as in Example 4.1, except for correlations across the interface. With a constant prior mean $\mu_0 = 4.86$ we have

$$\mu_C(x) = \mathbf{1}_{x<j}\,\mu_0 + \mathbf{1}_{x\geq j}\,\mu_0 = \mu_0 \tag{4.24}$$

$$K_C(x,y) = \mathbf{1}_{x<j}\,K_0(x,y)\,\mathbf{1}_{y<j} + \mathbf{1}_{x\geq j}\,K_0(x,y)\,\mathbf{1}_{y\geq j} \tag{4.25}$$

where K_0 refers to the SE kernel with characteristic length $\ell = 0.83$ and variance $\tau = 2.20$. In comparison to Example 4.1 the variance was increased to account for larger velocity variations. In Figure 4.3 depicts the three correlation kernels and the interface encoded is clearly visible.
Figure 4.4 shows the reconstruction of the discontinuity. The point to emphasize is that the posterior velocity model recovers the jump although not encoded in the a priori mean. The wiggle at about 4 is still over pronounced because of C's large correlation length. Interestingly, the posterior covariance shows very little correlations amongst left and right. Due to non-linearity the reconstruction is *not* truly satisfying. The a priori assumptions are even further off than in Example 4.1 and the posterior model shows late arrivals, i.e. average velocities are below the actual values. Although the inversion did not cover the interface's location, the height of the jump is nicely captured. ◁

4.3. Successive Approximation

To better accommodate non-linearity the successive approach introduce in Section 3.4.1 is adopted. The fundamental idea is the following: The closer the a priori mean to reality the better a Taylor expansion approximates. Therefore, a Bayesian updated system is established progressively improving the function used for the Taylor expansion. Only

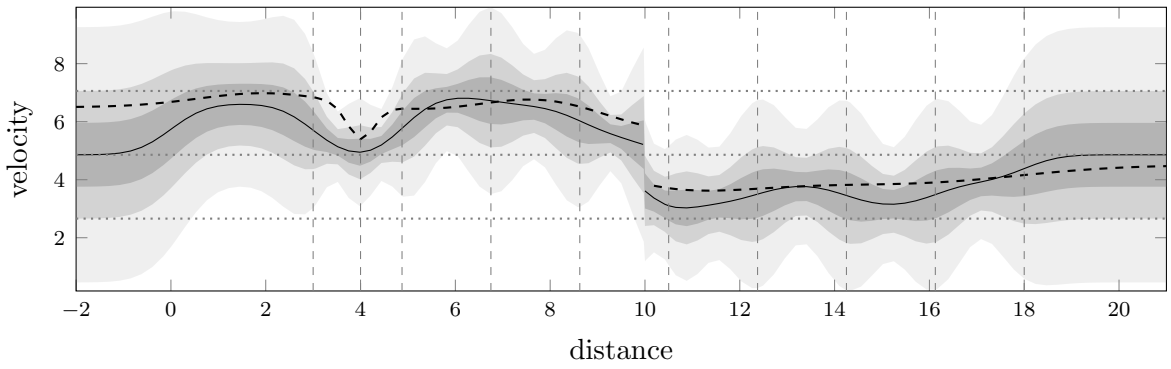

Figure 4.4.: Prediction of a velocity structure possessing a discontinuity at $x_j = 10$ using independent kernels. The posterior mean is shown by —— whereas half, once and twice the standard deviation is indicated by the gray shaded area. The true velocity structure $\tilde{c}$ is plotted by - - -. A priori assumptions are indicated by ····· (mean ± standard deviation). Receiver locations are shown by vertical lines.

incorporating a single measurement at a time gradually builds the posterior mean which in turn serves as point of expansion for the subsequent observation. Again, the posterior mean is successively setup and the Taylor expansion will perform the better the more observations are factored in. That approach is expected to perform well since there are far reaching correlations along the entire ray path.

Let extend our notations by a subscript indicating the observation incorporated to encode the idea outlined in the recursive relation. As an example C_6 refers to the velocity model posterior to $d_1, \ldots, d_6$. The a priori GRF is indicated by C_0. Then, the posterior mean and covariance of $C_{i+1} = C_i | d_{i+1}$ read

$$\mathbb{E}[C_{i+1}] = \mathbb{E}[C_i] + \operatorname{Cov}\big[C_i, T_{s,r_{i+1}}[C_i]\big]\, \mathbb{V}[D_{i+1}[C_i]]^{-1} \left(d_{i+1} - \mathbb{E}\big[T_{s,r_{i+1}}[C_i]\big]\right) \tag{4.26}$$

$$\mathbb{V}[C_{i+1}] = \mathbb{V}[C_i] - \operatorname{Cov}\big[C_i, T_{s,r_{i+1}}[C_i]\big]\, \mathbb{V}[D_{i+1}[C_i]]^{-1} \operatorname{Cov}\big[T_{s,r_{i+1}}[C_i], C_i\big] \tag{4.27}$$

where $i = 1, \ldots, n$ is referring to the total amount of data incorporated. Having considered all n measurements, the final model is denoted by C_n. Let's have a detailed look at what the terms in the recursive relation are referring to. The quantities $\mathbb{E}[C_i]$ and $\mathbb{V}[C_i]$ refer to the predecessor's mean and covariance. According to Equations 4.8, 4.10 and 4.11 the remaining terms become

$$\mathbb{E}\big[T_{s,r_{i+1}}[C_i]\big] = \mathbb{E}\left[\int_s^{r_{i+1}} \frac{2}{\mu_{C_i}(x)} - \frac{C_i(x)}{\mu_{C_i}(x)^2}\,\mathrm{d}x\right] = \int_s^{r_{i+1}} \frac{1}{\mu_{C_i}(x)}\,\mathrm{d}x \tag{4.28}$$

$$\operatorname{Cov}\big[C_i(x), T_{s,r_{i+1}}[C_i]\big] = -\int_s^{r_{i+1}} \frac{\operatorname{Cov}[C_i(x), C_i(y)]}{\mu_{C_i}(y)^2}\,\mathrm{d}y \tag{4.29}$$

$$\mathbb{V}[D_{i+1}[C_i]] = \mathbb{V}\big[T_{s,r_{i+1}}[C_i] + E\big] = \int_s^{r_{i+1}} \frac{1}{\mu_{C_i}(x)^2} \int_s^{r_{i+1}} \frac{\operatorname{Cov}[C_i(x), C_i(y)]}{\mu_{C_i}(y)^2}\,\mathrm{d}y\,\mathrm{d}x + \sigma^2 \tag{4.30}$$

where μ_{C_i} refers to $\mathbb{E}[C_i]$. To account for reflections, Equations 4.18, 4.19 and 4.20 need to be translated accordingly.

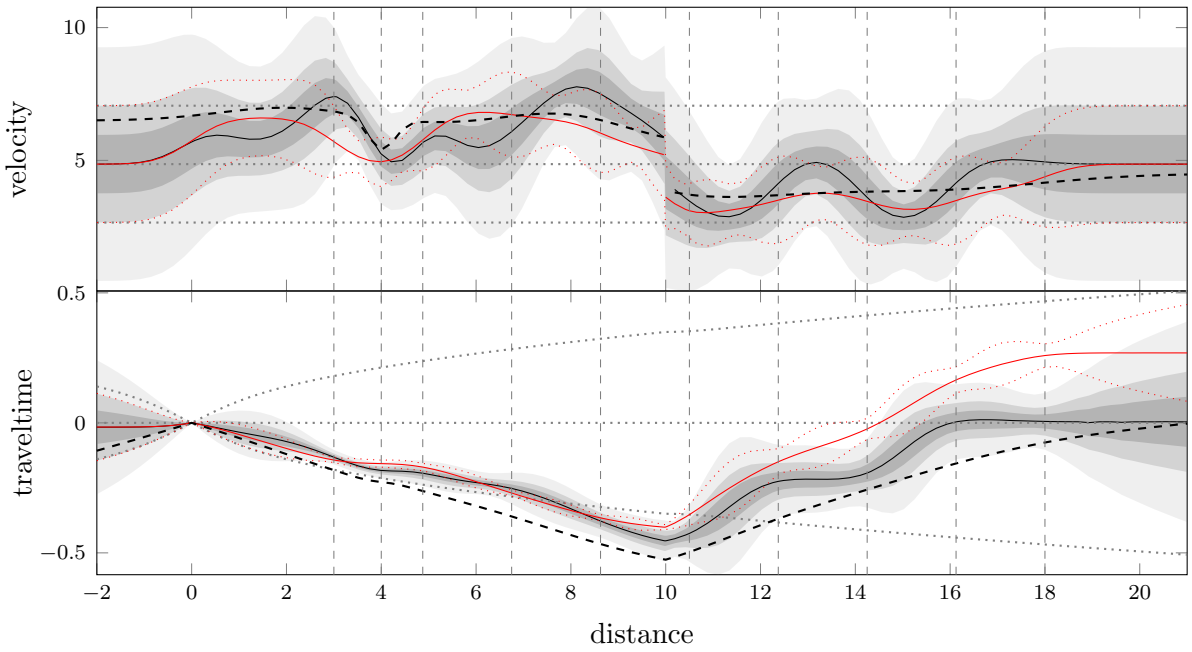

Figure 4.5.: Prediction of the velocity model (top) and travel time curve (bottom) with successively incorporated measurements. The posterior mean is shown by —— whereas half, once and twice the standard deviation is indicated by the gray shaded area. Travel time curves are cleared by their linear trends. A priori assumptions are indicated by ····· (mean ± standard deviation). Results from Example 4.2 (——; mean ± standard deviation) and the true travel time curve (- - -) are shown for comparison. Receiver locations are indicated by vertical lines.

The successive approach comes at the price of increased computational costs since mean functions, correlations and variances will not possess a solution of closed from. By using a fixed samples numeric integration scheme, this shall not be too much of a concern.

Example 4.3. We carry on with Example 4.2, however, incorporating measurements successively. A priori assumptions and observations remain the same. The succession order is the following: Direct waves first followed by reflections, both, from left to right. Computing correlations and variances calls for numeric integration. Therefore, Simpson's rule is used considering a grid of 115 points across the interval $[-2 : 21]$.
The top panel of Figure 4.5 shows the successive reconstruction of the velocity model. In contrast to Example 4.2, the posterior mean shows more complexity, nonetheless, uncertainties are at a similar level. Both attempts are catching the jump, however, at a different height. Meeting the falling edge of the wiggle at 4 does not mean much. By looking at those curves it is hard to judge which reconstruction outperforms the other.
The posterior travel time curve, however, helps to decide which attempt better assimilates the data. In the bottom panel of Figure 4.5 it can clearly be seen that the successive posterior travel time curve is closer to the true one. The successive approach replicates prior assumptions past the last observations and does not impose an offset. Although not explicitly demonstrated, the order of succession has a remarkable effect. If data are considered in reverse oder, reconstruction of the right hand side is in dire straits whereas the left hand side becomes very accurate. Nonetheless, the example shows that a successive approach is a way to better accommodate non-linearity.

4.4. Conclusion

The outlined approach provides a consistent tool for Bayesian travel time inversion. The presumed understanding of physics – the high frequency approximation – is determining the correlations amongst model and observations. In particular, the presence of posterior uncertainties can be used to judge the quality of the reconstruction. On top of uncertainties, discontinuities are modeled in a consistent manner. The key factor controlling the inversion is a correlation kernel appropriately accounting for interfaces. However, it is questionable if the SE kernel – used in the examples above – is well suited to describe correlations of velocity models. Non-linearity has a substantial effect. The presented successive approach accommodates non-linearity, however, with the need for further studies. As a non-parametric approach is addressed – in the sense considering a distribution over functions – the spatial resolution is not limited by a certain basis. Computational costs are at an ordinary level (compared with Markov chain Monte Carlo methods utterly cheap).
While accommodating non-linearity and assessing posterior uncertainties the presented modeling scheme is beyond ordinary approaches offering a promising vision. Still, it is a long way to go to bring the outlined concept into a real world application.

4.5. Outlook & Further Thoughts

In case of multiply triggering an artificial source, summing the records to increase the signal-to-noise ratio is called **stacking**, leading to reduced noise levels [Shearer, 2009]. For the presented modeling approach stacking raises a question to be investigated: At each receiver location, shall we either favour the accurate stack (sample mean and variance) or incorporate individual measurements in an order of succession that remains to be investigated? As the modeling approach presented partly qualifies for survey design, we can also deduce if it is preferable to use single but accurate measurements (pyrotechnics) over a stack of noisy records (sledgehammer).

It is straight forward to incorporate **further or multiple observations**. Think of the knowledge obtained about the subsurface from outcrops or exploration boreholes. Although not explicitly demonstrated, considering records of the velocity structure are no difficulty. Another example are data sets obtained from multiple sources.

In case of seismological applications typically the **source location is unknown** and its variability shall be jointly explored. Therefore, let us adopt and extend the approach presented by Stein [1999, Sec. 7.2], putting an independent normal prior over $s \to S \sim \mathcal{N}(\mu_S, \sigma_S^2)$. To invert for $C, S|d$, correlations amongst source location and observations need to be known. In general we can not say much about about $\mathrm{Cov}[S, D]$. To obtain a proxy a Taylor expansion helps and to its first order the expansion around μ_S reads

$$\mathrm{Cov}[S, D] \approx \mathrm{Cov}\Big[S, T_{\mu_S,r} + \partial_S T_{S,r}\big|_{\mu_S}(S - \mu_S)\Big] = \partial_S T_{S,r}\big|_{\mu_S}\sigma_S^2 \tag{4.31}$$

since $S \perp\!\!\!\perp C$ is assumed.

Accounting for **multiple phases** is straightforward. With identified reflections, a partition P of the velocity model needs to be composed such that blocks $A, B \in P$ are disjoint

and covering the whole domain of interest $\bigcup_{A\in P} A$. For each block A, setup an a priori GRF given by its mean μ_A and covariance K_A. The overall GRF is given by

$$C = \sum \mathbf{1}_A \, A \qquad\qquad A \sim \mathcal{GP}(\mu_A,\, K_A)\,, \quad A \in P \tag{4.32}$$

truncated at the boundaries to achieve independence amongst blocks. In complex heterogeneous media, determining the **ray path** is certainly neither a linear nor a simple task (Eq. 4.2) with the demand of computational resources. In a first attempt the model's mean shall serve to calculate ray paths. Since the mean progressively develops, in 2d and 3d applications an adaptive grid is needed. This however is just half the difficulty. An uncertain model renders the ray path uncertain as well. One of the open questions is how to deal with uncertain ray paths. A further source of uncertainty is **ignorance about discontinuities**. To jointly invert for interfaces, partition boundaries need to be expressed in terms of functions, e.g. a plane in a 3d application. Its uncertainty may again be described by a GRF and a Taylor expansion is used to linearize.

A problem not yet even mentioned is how to **estimate hyper parameters**. An ordinary maximum likelihood estimate appears adequate to estimate the a priori mean μ_C, the overall variance τ and the residual term σ. However, a clever strategy is needed to determine the characteristic length ℓ.

Last but certainly not least, the role of the **correlation kernel** shall be investigated. Although presumed physics is encoded in correlations amongst model and observations, a heuristic kernel is hardly satisfying. In the above examples the SE kernel was not well motivated and may be considered more as a rule for interpolation. Because the earth is almost a sphere, on a global the usage of the 3-dimensional Poisson kernel appears to be the natural choice and imposes the reconstruction to be harmonic.

5. Seismic Source Inversion

The problem of locating and describing earthquakes is one of the oldest challenges in seismology. The associated classic inverse problem is to work backwards characterizing the seismic event that gave raise for ground motion records [Stein, 1999, sec 7.2]. The preceding chapter was directed at finding a model of the Earth's subsurface but largely neglected the origin of the recorded wave field. In a seismological application not only the subsurface is uncertain but also duration and spatial extent of a seismic event. It actually were preferable to jointly invert for both the source mechanism and the medium through which the waves passed [Shearer, 2009, sec 5.7]. However, for the sake of convenience the subsurface is regarded to be known even though the Earth's elastic properties may inherit a substantial amount of uncertainty.

Although seismic source inversion has widespread applications – e.g. Heimann et al. [2018], Sadeghisorkhani et al. [2016], Donner et al. [2016], Fichtner et al. [2017], Ballesio et al. [2018] – credibility continues to be one of the pressing questions. Typically, a set of parameters is in use that aim to describe a seismic event. It is almost standard to report uncertainties of such parameters, nevertheless, some problems remain that need attention. Without being exhaustive, they are: The wave field is non-linearly related to some of the parameters – e.g. dip- and strike angle – rendering inference inherently difficult. A complex seismic event described by a large number of parameters makes sampling based estimation (e.g. MCMC) prohibitively expensive. The evolution and the spatial source geometry together with its uncertainties are not reflected in knowing variabilities of a partial parametrization only describing the far field.

The following work aims to present a conceptual and versatile framework for the characterization of earthquake sources. Therefore, a rigorous non-parametric Bayesian approach is addressed to consistently quantify modeling related uncertainties. It is not about determining parameters of point source approximations such as epicenter, moment tensors or finite fault rupture but the forcing function itself with spatial and temporal extent. The overall modelling concept is to combine the wave equation and Gaussian processes regression to infer the forcing function that gave raise for a wave field. Admittedly, the problem as a whole appears yet too complex to assess. To work towards the outlined goal, a fixed ends damped vibrating string subject to some external force serves as ongoing example. Although an oversimplification, it catches the essence of seismic source inversion at a comprehensive level.

Example 5.1 (Damped vibrating string)**.** A homogeneous fixed ends damped vibrating string subject to an external force f is considered. The equation of motion with respect to the displacement u reads

$$\partial_t^2 u(x,t) + 2\beta\partial_t u(x,t) - c^2\partial_x^2 u(x,t) = f(x,t) \tag{5.1}$$

with damping factor $\beta = 0.2$ and tension $c = 2$. Damping is caused by the reversing term proportional to the 1^{st} time derivative. Boundary and initial conditions are

$$u(x=0,t) = u(x=L,t) = 0 \qquad \text{and} \qquad u(x,t=0) = \partial_t u(x,t=0) = 0 \tag{5.2}$$

where $L = 8$ denotes the length of the string. In the following only weak damping is considered that is $\beta < \frac{\pi c}{L}$. Physical units are skipped in favour for staying concise. $\lhd$

The analogy of seismic source inversion and the damped vibrating string may be seen as follows. For a given Earth model the propagation of seismic waves is described in terms of the wave equation

$$\rho(x)\,\partial_t^2 u(x,t) - \nabla \cdot \sigma(x,t) = f(x,t) \tag{5.3}$$

that relates the displacement field u to the mass density ρ, the stress tensor σ and an external force f. At the Earth's surface the stress vanishes and at initial time displacement and velocity are zero. To obtain a complete set of equations, the stress tensor σ must be related to the displacement field u. Ignoring visco-elastic effects, the constitutive relation takes a simple form

$$\sigma(x,t) = C(x) : \nabla u(x,t) \tag{5.4}$$

with elastic parameters C, a 4^{th}-order tensor of 21 independent components [Fichtner, 2011]. By means of the double-dot product, a second order tensor and a fourth-order tensor are mapped to a *new* second-order tensor. Without anelasticity seismic waves from every earthquake ever occurred would still be reverberating until the accumulating vibrations shattered the Earth [Stein and Wysession, 2003, sec 3.7]. Though not a general problem at its own, dispersion due to damping seriously complicates the phenomenon under analysis [Orsingher, 1984]. Consult e.g. Aki and Richards [2002] for further details and derivations.

Solving Equations 5.3 and 5.4 is quite demanding for realistic media and source terms. It is beneficial to develop a notation that separates the source terms from all the other details of wave propagation. The so-called *Green function* is obtained by solving the wave equation for a point source that is $f(x,t) = \delta(x - x_0)\delta(t - t_0)$ where $\delta(\cdot)$ refers to the Dirac delta. A Green function is a tensor as a point source is considered individually for each direction. Provided a Green function, the wave field for a general source term is given by

$$u(x,t) = \iint G(x,t;y,s) \cdot f(y,s)\ \mathrm{d}y\,\mathrm{d}s \tag{5.5}$$

and the integral ranges over the support of f [Chapman, 2004, chp. 4]. The calculation of Green's function and the solution of the integral is a challenge in its own that in general requires numerical methods. Under the assumption that G can be computed, notice the power of this equation. Because it is linear, the ground motion resulting from a realistic forcing function is given as the superposition of individual force vectors. It already implies that knowledge of the wave field permits inference of the external force [Shearer, 2009, sec. 9.1].

Example 5.2 (Green's function)**.** For an infinite damped string the causal Green function may be written in terms of a Bessel function

$$G_\infty(x,t;x_0,t_0) = \frac{\exp\{-\beta(t-t_0)\}}{2c} \begin{cases} I_0\Big(\frac{\beta}{c}\sqrt{(\rho-\rho_0)(\eta-\eta_0)}\Big) & \rho < \rho_0 \text{ and } \eta_0 < \eta \\ 0 & \text{otherwise} \end{cases} \tag{5.6}$$

where $\eta = x + ct$ and $\rho = x - ct$ are denoting to so-called *light cone* coordinates and I_0 refers the modified Bessel function of the first kind of order 0 [Orsingher, 1984, Eq. 3].

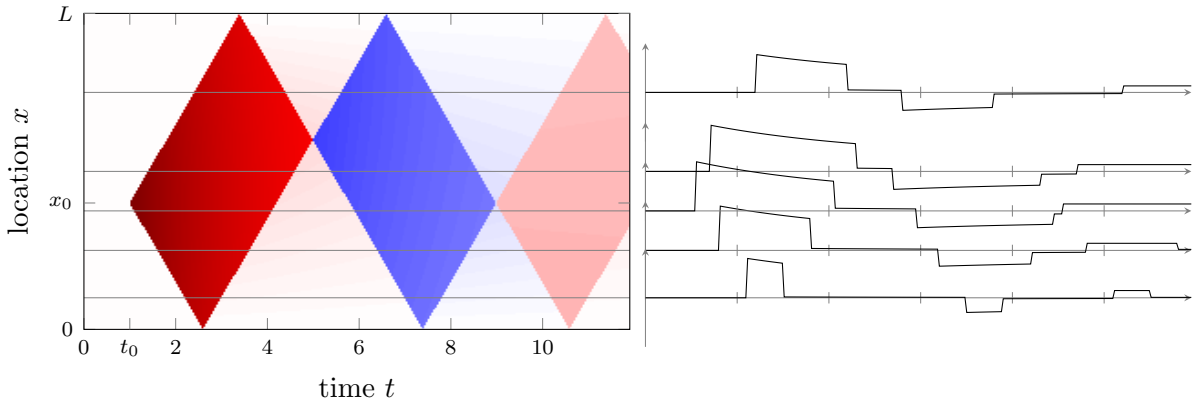

Figure 5.1.: The left panel shows a pseudo-color plot of the causal Green function. A diverging color-map is used, ranging from -1 (blue) to 1 (red). Selected traces are shown the right.

The causal Green function for the clamped string is established using the principle by d'Alembert. Therefore, consider a $2L$ periodic odd extension of the spatial Dirac pulse. The Green function follows as the consecutive and alternating superposition

$$G(x,t;x_0,t_0) = \sum_{n=-N}^{N} G_\infty(x,t;2nL+x_0,t_0) - G_\infty(x,t;2nL-x_0,t_0) \tag{5.7}$$

and the number of repetitions is truncatable with $N = \left\lceil t\frac{c}{L} \right\rceil$ by specifying the temporal domain of interest [1]. Figure 5.1 illustrates the causal Green function and the effect of dispersion due to damping is clearly visible. ◁

The following aims at inferring an unknown forcing function based on ground motion records. At seismic stations time series are recorded that are induced by a source distribution of unknown duration and unknown spatial extent. The data acquired at n stations are denoted by

$$d = \left\{ u(x_i,t) + e_i \;\middle|\; i = 1,\cdots,n \right\} \tag{5.8}$$

where measurement errors are indicated by e_i and x_i refers to the receiver locations. To keep equations concise the discrete nature of time series and measurement errors are omitted. For the rest of the section, d will refer to a general data set because evenly sampled time series is not a prerequisite.

Example 5.3 (Pseudo Data)**.** Synthetic data is generated to drive the inversions that are presented subsequently. The source pattern of choice is the truncated half period of a sine function, both, in space and time (see Fig. 5.3). The forcing function reads

$$f(x,t) = \mathbf{1}_{\{t_i \le t \le t_f\}} \sin\left(\frac{\pi(t-t_i)}{\Delta t}\right) \mathbf{1}_{\{x_l \le x \le x_r\}} \sin\left(\frac{\pi(x-x_l)}{\Delta x}\right) \tag{5.9}$$

where t_f and x_r are abbreviations for $t_i + \Delta t$ and $x_l + \Delta x$. The onset time is $t_i = 1$ with duration $\Delta t = 0.9$. The left edge is $x_l = 3.2$ with spatial extent $\Delta x = 0.52$. SciPy's [Virtanen et al., 2020] general purpose quadrature scheme is in use to solve the

[1] $\lceil x \rceil$ denotes the ceiling function that maps to the least integer greater than or equal to x.

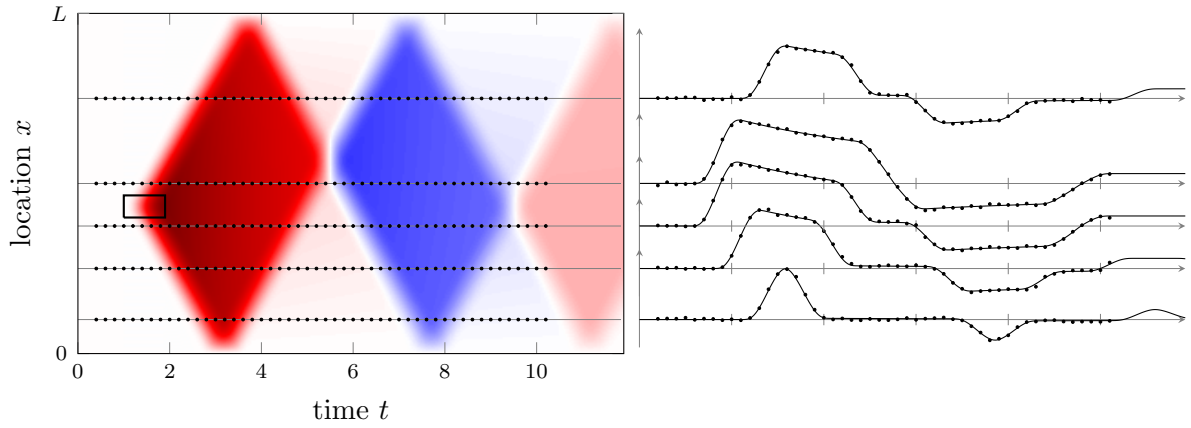

Figure 5.2.: The left panel shows a pseudo-color plot of the displacement. Black dots are indicating individual records. The square on top of the pseudo-color plot refers to the region of external action. Selected traces that are corrupted by normal noise are shown on the right.

integral in Equation 5.5. The data basis for the inversion is formed by $N_x = 5$ time series each of $N_t = 50$ samples. Values are corrupted by independent and centered normal noise such that the standard deviation is of about 1% of the signal's peak-to-peak value. The time increment of 0.2 is chosen such that it admits resolving the duration of the action. On the contrary the spatial extent of the force is small compared with the few locations specified. Figure 5.2 illustrates the displacement field together with traces recorded. ◁

Because a Bayesian approach is pursued, a priori assumptions are raised about the forcing function. To express our ignorance, the a priori force is assumed a GP

$$f \to F \sim \mathcal{GP}(\mu_F,\, K_F) \tag{5.10}$$

with a priori mean function $\mu_F(x,t)$ and assumed covariance structure $K_F(y,s;x,t)$. Since F is continuous with respect to space and time, earthquakes are *not* approximated by a point source but one with spatial and temporal extent. The a priori force distribution implies a wave field whose statistical properties are governed though the wave equation. For the modeling approach we exploit the fact that the integral with Green's function (Eq. 5.5) is a linear map in F. The resulting a priori wave field is a GP at its own right and fully determined by

$$\mathbb{E}[U(x,t)] = \iint G(x,t;\tilde{x},\tilde{t})\mu_F(\tilde{x},\tilde{t})\ \mathrm{d}\tilde{x}\,\mathrm{d}\tilde{t} \tag{5.11}$$

$$\mathrm{Cov}[U(y,s),U(x,t)] = \iint\iint G(y,s;\tilde{y},\tilde{s})K_F(\tilde{y},\tilde{s};\tilde{x},\tilde{t})\,G(x,t;\tilde{x},\tilde{t})\ \mathrm{d}\tilde{y}\,\mathrm{d}\tilde{s}\ \mathrm{d}\tilde{x}\,\mathrm{d}\tilde{t} \tag{5.12}$$

where we made use of the linearity of the expectation and the bi-linearity of the covariance. The decisive feature is that the space-time covariance function is deduced from physical principles i.e. Green's function. For the following it is taken for granted that at least a numeric proxy of a Green function is at hand and that the integrals can be computed. In fact, the overall modeling concept is preserved for any linear observational functional e.g. averages, velocities or recordings of strong motion sensors.

To stay within the Gaussian model, the recorded data set is assumed to be corrupted by centered, normal noise with error covariance matrix Σ_E. If we assume the a priori force

field and measurement noise statistically independent then the data model is normal as well. The data's a priori mean and covariance are given by

$$\mu_D = \{\mathbb{E}[U(x_i,t)] \,|\, i = 1,\cdots,n\} \tag{5.13}$$
$$\Sigma_D = \{\mathrm{Cov}[U(x_i,t), U(x_j,t)] \,|\, i,j = 1,\cdots,n\} + \Sigma_E \tag{5.14}$$

where discrete time steps are again not explicitly indicated. It is clear that the approach can handle gaps in the data and, again, does not require uniform sampling.

The corner stone of the modeling strategy is formed by the correlations amongst observations and to be predicted source field. For an arbitrary design point (y,s), the cross-covariance amongst forcing function and wave field reads

$$\mathrm{Cov}[F(y,s), U(x,t)] = \iint G(x,t;\tilde{x},\tilde{t}) K_F(y,s;\tilde{x},\tilde{t})\,\mathrm{d}\tilde{x}\,\mathrm{d}\tilde{t} \tag{5.15}$$

where we made again use of Equation 5.5 and the bi-linearity of the covariance. The according cross-covariance matrix is given by

$$\Sigma_{FD} = \big\{\mathrm{Cov}\big[F(\tilde{x},\tilde{t}), U(x_i,t)\big] \,\big|\, i = 1,\cdots,n\big\} \tag{5.16}$$

because measurement noise and a priori force are assumed independent. The advantage is that the understanding of physics serves to derive an appropriate correlation structure rather than using heuristics.

Referring to Section 1.3, the force's posterior distribution is directly accessible. The distribution of F posterior to the data set d is normal and given by conditional mean and conditional covariance

$$\mathbb{E}[F|d] = \mathbb{E}[F] + \Sigma_{FD}\Sigma_D^{-1}\big(d - \mu_D\big) \tag{5.17}$$
$$\mathbb{V}[F|d] = \mathbb{V}[F] - \Sigma_{FD}\Sigma_D^{-1}\Sigma_{DF}\ . \tag{5.18}$$

The most probable force field is given as the posterior mean whereas uncertainties are described by the posterior covariance. Inference takes place with respect to heterogeneously distributed sources of continuous or transient nature. The approach can deal with earthquake data as well as ambient noise records. The key challenge is to establish a suitable a priori force distribution. Localized seismicity calls for careful and elaborate prior selection. On the contrary, in case of ambient noise records a white noise prior force field appears appropriate at first sight.

5.1. White Noise

If there is no prior knowledge about the source characteristics, it is reasonable to impose a force of uninterrupted white noise impulses. Nonetheless, the white noise assumption shall only serve as a starting point and will receive refinements later on. Choosing an a priori force of Gaussian white noise

$$\mathbb{E}[F] = 0 \qquad \text{and} \qquad \mathrm{Cov}[F(x,t), F(y,s)] = \delta(t-s)\delta(x-y) \tag{5.19}$$

results in fairly simple computations as the Dirac-delta causes many of the integrals to collapse. The unbounded Dirac delta is used as a symbolic way of writing rather than

an appropriate covariance. From a theoretical point of view, the corresponding integrals must be understood in the sense of the stochastic Paley-Wiener integral.

To calculate the posterior distribution, the data's covariance and the cross-covariance with the forcing function need to be known. The a priori wave field borrows its statistical properties from uninterrupted white noise. Since F is centered the a priori displacement is also of zero mean. The assumed covariance for the displacement field reads

$$\operatorname{Cov}[U(y,s),U(x,t)] = \iint G(y,s;z,u)\,G(x,t;z,u)\;\mathrm{d}z\,\mathrm{d}u \tag{5.20}$$

and the cross-covariance amongst forcing function and wave field is simply given by

$$\operatorname{Cov}[F(y,s),U(x,t)] = G(x,t;y,s)\;. \tag{5.21}$$

Notice that the argument's order flipped. These equations allow the construction of the data covariance matrix and the cross-covariance matrix. Posterior mean and covariance are calculated as described above.

Example 5.4 (White Noise & Displacement)**.** Let us proceed with the ongoing example of a damped vibrating string, a priori submitted to uninterrupted white noise impulses. The according kernel for the displacement is time stationary and does not depend on knowing when the action starts. It features an upper bound

$$\operatorname{Cov}[U(x,t),U(y,s)] \leq \frac{L}{16\beta c^2} \tag{5.22}$$

which results as the equilibrium between damping an excitation. Observations driving the inversion are the displacement records introduced in Example 5.3. Figure 5.1 illustrates Green's function which serves a cross-covariance. SciPi's general purpose quadrature scheme is used to calculate the data's covariance. Figure 5.3 compares the actual action and the posterior mean. The reconstruction with edges and corners is not truly satisfying but the time window agrees quite well. Due to coarse spatial sampling the posterior mean is extensive and misses localization. The poor recovery is no surprise considering that the actual force cannot be a realization of the a priori assumption. ◁

Most seismometers, however, measure velocity rather than displacement [Shearer, 2009, p. 261]. In order to increase the overlap with seismology it is of interest to have the displacement velocity under observation. In analogy to Equation 5.5, the velocity field may be modeled in terms of the time derivative

$$v(x,t) = \iint \partial_t G(x,t;y,s) f(y,s)\,\mathrm{d}y\,\mathrm{d}s\;. \tag{5.23}$$

The derivative is a linear operation and thus also well suited as an observable. To directly calculate the time derivative of Green's function it is beneficial to address the so-called velocity–stress formulation of the wave equation [Fichtner, 2011, sec. 2.2.2]. The cross-covariance amongst a white noise force field and the velocity field is given by

$$\operatorname{Cov}[F(y,s),V(x,t)] = \partial_t G(x,t;y,s) \tag{5.24}$$

and the data covariance translates accordingly. Because of the abrupt onset of Green's function at the edges of the light cone, unbounded delta fences are introduced by the time derivative. The following example is intended to illustrate why the white noise assumption is inappropriate for inference using velocity records.

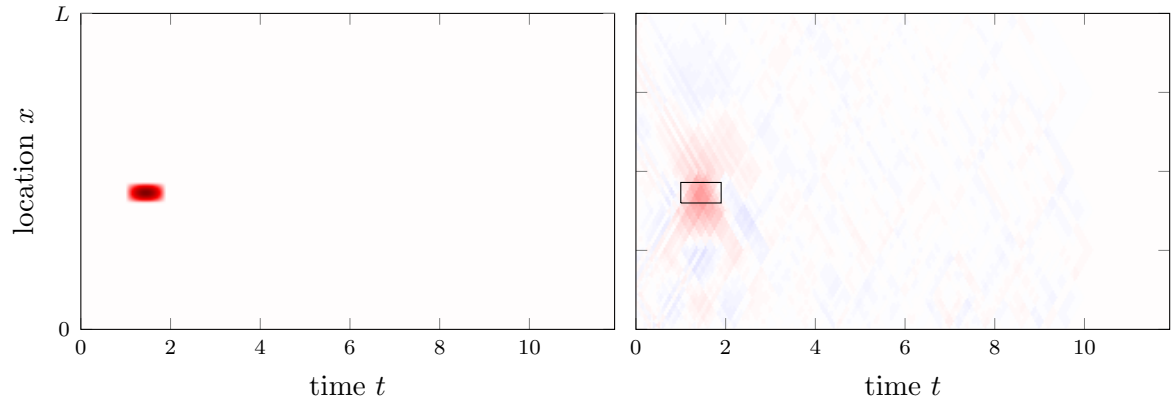

Figure 5.3.: Reconstruction of the external force using a white noise a priori force field. The actual forcing function is shown on the left. The right panel illustrates the posterior mean reconstructed using corrupted displacement records (see Ex. 5.3). A diverging color-map is in use that ranges from -1 (blue) to 1 (red). The black square indicates the area of external action.

Example 5.5 (White Noise & Velocity)**.** We again consider the damped vibrating string but with respect to displacement velocity. The calculation of the cross-covariance already shows that the white noise assumption is unsuitable for inference with velocities records. The time derivative of the Green function of the infinite string is given by

$$\partial_t G_\infty(x,t;\tilde{x},\tilde{t}) = c\big(\delta(\tilde{\rho}-\rho)+\delta(\eta-\tilde{\eta})-\beta\big)G_\infty(x,t;\tilde{x},\tilde{t})+ \\ +\frac{\beta\exp\big\{-\beta(t-\tilde{t})\big\}}{2c}I_1\left(\frac{\beta}{c}\sqrt{(\tilde{\rho}-\rho)(\eta-\tilde{\eta})}\right)\frac{c(t-\tilde{t})}{\sqrt{(\tilde{\rho}-\rho)(\eta-\tilde{\eta})}} \quad (5.25)$$

where I_1 refers to the Bessel function of order one. This result may be derived by imposing the secondary constraints $\rho < \rho_0$ and $\eta_0 < \eta$ in terms of two Heaviside step functions. At the onset of the light cone the time derivative introduces unbounded delta fences. The principle by d'Alembert is again used to establish a Green function for the clamped string. For a single observation the resulting cross-covariance is illustrated by Figure 5.4. As a consequence, the posterior mean consists of numerous of those delta fences which is the reason why the white noise assumption is too rough as a priori assumption. ◁

As a concept, a a prior force of white noise is appealing and instructive because of its simplicity. However there are fundamental shortcomings: Pointwise predictions of the posterior variance are unbound because the posterior covariance is dominated by the white noise prior. A spatial white noise prior distribution is too weak since in a seismological application the spatial coverage is typically coarse. A reasonable reconstruction is possible only if recorders are at least closely surrounding the source region. Nonetheless, white noise is a good starting point that requires refinements. The strategy for the following is to smoothen the white noise assumption in order to achieve compatibility with power spectra observed i.e. the use of a more informative a priori distribution.

5.2. Fourier Analysis

This is a brief and incomplete digression about the Fourier Transform and the Convolution Theorem. Details may be found in every textbook about time series analysis e.g. Gubbins

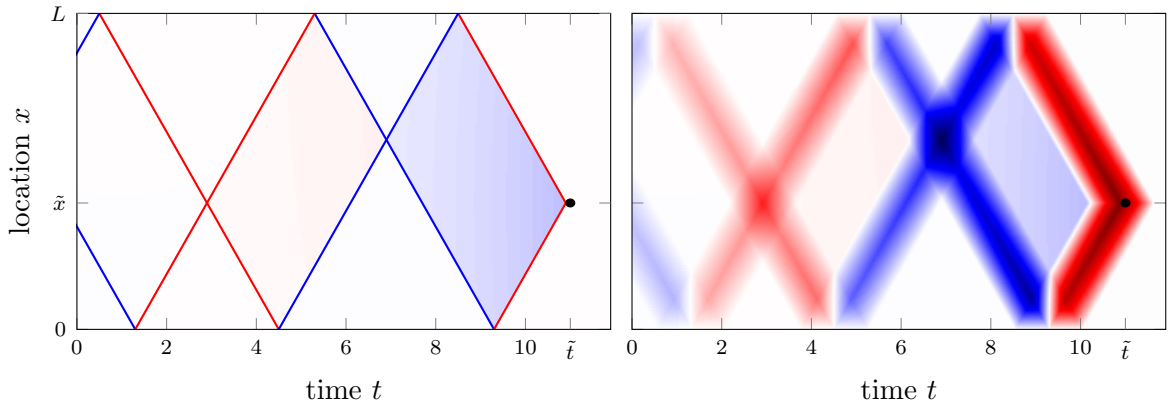

Figure 5.4.: The left panel illustrate the cross-covariance for a white noise force field with a single velocity record located at $x = 3.2$ and $t = 11$. Dirac fences at the onset of the light cones are indicated by red and blue lines. Due to causality, the forcing function is only correlated prior to the record. The right panel shows a filtered version of the white noise cross-covariance. Using $\mathrm{sinc}(\cdot)$ as transfer function turns the Dirac fences into triangular functions.

[2004], Chapman [2004, chp. 3] or Stein and Wysession [2003, chp. 6]. Throughout, the unitary definition of the Fourier transform is used. The Fourier analysis is useful for solving partial differential equations as the following example shows:

Example 5.6 (Helmholtz Equation)**.** A transformation into the frequency domain turns the temporal part of the wave equation into an algebraic equation

$$(\underbrace{4\pi^2\xi^2 - 4i\pi\xi\beta}_{-c^2k^2} - c^2\partial_x^2)u(x,\xi) = f(x,\xi) \tag{5.26}$$

where the Fourier pair $\partial_t \leftrightarrow 2i\pi\xi$ is used. Equation 5.26 is known as the Helmholtz equation. The Helmholtz equation is an eigenvalue problem of the Laplacian and quite elegant to solve. According to Fischer and Kaul [2014, §14], a Green function is given by

$$G_H(x,y) = \frac{1}{k\sin(kL)} \begin{cases} \sin(kx)\sin(k(y-L)) & x < y \\ \sin(ky)\sin(k(x-L)) & y \leq x \end{cases} \tag{5.27}$$

where k is a function in ξ and the boundary conditions are already satisfied. It is continuous but features an abrupt change at $x = y$. The transformation back into the time domain is achieved by a phase factor together with the shift property of the inverse Fourier transform. ◁

Performing computations in the frequency domain is beneficial because the subsurface is assumed static at the characteristic time scales of wave propagation. This is the reason why the temporal integral in Equation 5.5 can be written in terms of a convolution

$$u(x,t) = \iint G(x, t-s, y) f(y,s)\,\mathrm{d}s\,\mathrm{d}y \tag{5.28}$$

[see Chapman, 2004, sec. 4.5.2.3]. A convolution can also be described in the frequency domain because the Fourier transform of a convolution equals the product of Fourier

transforms. A synthetic seismogram resulting from a source is equivalently given by

$$u(x,t) = \int \mathcal{F}^{-1}\big[G(x,\xi,y)f(y,\xi)\big](t)\,\mathrm{d}y \tag{5.29}$$

where the letter ξ is used to indicate frequencies and $\mathcal{F}[\cdot]$ refers to the Fourier transform and its inverse. Within the Fourier domain it is easy to account for a time derivative. The velocity field is simply given by

$$v(x,t) = \int \mathcal{F}^{-1}[2i\pi\xi\, G(x,\xi,y)\, f(y,\xi)](t)\,\mathrm{d}y \tag{5.30}$$

since derivative and $2i\pi\xi$ are forming a Fourier pair. Using the fast Fourier transform (FFT) as an implementation – rather than a direct approach – admits rapid computations particularly for evenly spaced time series. Speaking in terms of computational costs, the FFT scales very well compared to the individual double integrals in Equation 5.5 [Stein and Wysession, 2003, chp. 6].

Example 5.7 (Velocity Data)**.** Pseudo records are generated analogously to Example 5.3 but the velocity field is derived using Fourier methods. The frequency representation of our external force of choice (Eq. 5.9) is given by

$$f(x,\xi) = \mathbf{1}_{\{x_l \le x \le x_r\}} \sin\left(\frac{\pi(x-x_l)}{\Delta x}\right) \cdot \\ \cdot\, \Delta t \exp\{-i\pi\xi(t_i+t_f)\}\, \frac{\operatorname{sinc}(\Delta t\xi - 1/2) + \operatorname{sinc}(\Delta t\xi + 1/2)}{2}\,. \tag{5.31}$$

The spectrum of the source time function is obtained by convolving the two Fourier pairs $\operatorname{rect}(\cdot) \leftrightarrow \operatorname{sinc}(\cdot)$ and $\sin(\cdot)$ as the sum of two Dirac-deltas. The velocity field is calculated using Equation 5.30 together with the frequency domain Green function, presented in Example 5.6. In this case the spatial integral can be solved analytically because the product of two $\sin(\cdot)$-functions possesses an elementary antiderivative. The solution is omitted because the case differentiation leads to very long equations not providing any further insight. The FFT is used to efficiently transform back into the time domain. To ensure almost trailing zeros, the spectrum is discretized such that the damping term causes a decay by a factor of 1000. Figure 5.5 shows the resulting velocity field along with the corrupted time series that are going to drive the subsequent inversions. The corruption is again drawn from a centered normal and the standard deviation is about 1% of the signals peak-to-peak value. ◁

Unfortunately, with a seismological application in mind, we can not carry over this approach to the spatial domain because the elastic modulus are in general inhomogeneous [Stein and Wysession, 2003, chp. 6.3, eq. 15]. As a consequence, a Fourier series expansion in the spatial domain depends on the specific locations and cannot – as in the time domain – be expressed as a function of distance [2]. Although we can not exploit the convolution theorem within the spatial domain, this does not prevent us from carrying

[2] A Fourier series expansion is chosen because the spatial domain is typically of compact support.

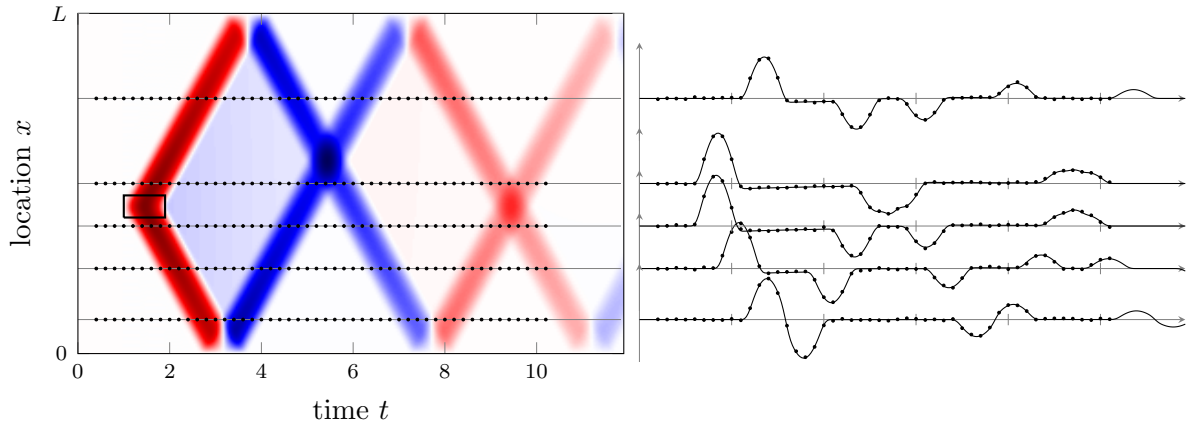

Figure 5.5.: The left panel shows a pseudo-color plot of the displacement velocity. Black dots are indicating individual records. The square on top of the pseudo-color plot refers to the region of external action. Selected traces that are corrupted by normal noise are shown on the right.

out a Fourier expansion with respect to y

$$u(x,t) = \sum_{m,n} \mathcal{F}^{-1}\big[G_n(x,\xi) \underbrace{\int \phi_n(y)\phi_m(y)\,\mathrm{d}y}_{\delta_{mn}} f_m(\xi)\big](t) = \\ = \mathcal{F}^{-1}\Big[\sum G_n(x,\xi) f_n(\xi)\Big](t) \quad (5.32)$$

where G_n and f_n are referring to the Fourier series coefficients. Because the Fourier basis is orthogonal, both, the integral and one of the two sums vanish. It is worth mentioning again, that every location x requires its *own* series expansion.

Example 5.8 (Series Expansion)**.** A Green function may also be derived using a series expansion as presented by Fischer and Kaul [2014, §6 sec. 3.7]. If we kick-off with the complex exponentials as fundamental solution and impose boundary conditions this leads to

$$\phi_n(x) = \sqrt{\frac{2}{L}} \sin\Big(\frac{n\pi}{L}x\Big) \quad (5.33)$$

which may be seen as the $\sin(\cdot)$-part of the $2L$ periodic orthonormal Fourier series. If we expand the Dirac-delta as Fourier series, then, the n-th mode of the causal Green function follows to read

$$G_n(x,\xi) = -\frac{\phi_n(x)}{4\pi^2\xi^2 - 4i\pi\xi\beta - c^2\frac{n^2\pi^2}{L^2}} \quad (5.34)$$

without contributing to the cos-type basis due to vanishing boundary conditions. In combination with the findings from Example 5.6, this result can be derived by differentiating the basis twice. ◁

To establish a link with the spherical Earth, the *solid spherical harmonics* are forming a complete basis. In analogy to a classical Fourier series, the spherical harmonics represent the fundamental modes of vibrations of a sphere. As a consequence, waves in a spherical Earth can be written as the sum of the Earth's normal modes [Stein and Wysession, 2003, sec. 2.9]. An expansion in spherical harmonics is analogous to a Fourier series on spheres and many results known from Fourier analysis translate accordingly [Gubbins, 2004, sec. 12.2].

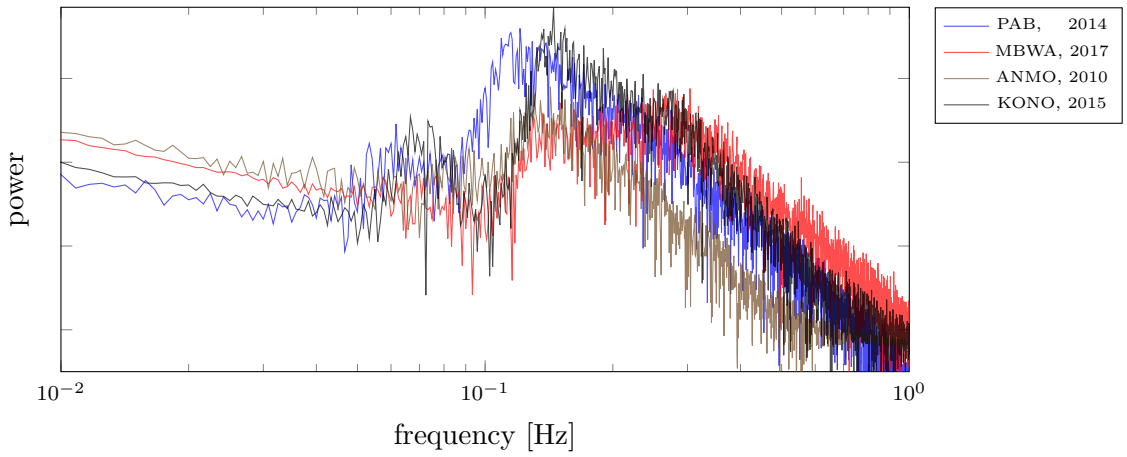

Figure 5.6.: Power spectra of microseismic noise derived from 20 minutes of data of the vertical component. The characteristic noise peaks associated with *microseism* at about 6 and 16 seconds are recognizable.

5.3. Filtered White Noise

As motivated by Examples 5.4 and 5.5, Gaussian white noise is unsuitable as an a priori assumption. The power spectrum of the ground motion that stems from white noise is not compatible with what is measured. If white noise were appropriate for modeling earthquakes, then the instrument corrected time series must show the sharp onset of Green's function once the light cone passes through rather than the earthquake rise. The observation of seismic noise is another indicator not to departure from white noise. It is well known that the power spectrum of seismic noise exhibits a characteristic known as *microseism* (see e.g. Shearer [2009, sec. 11.2] or Stein and Wysession [2003, sec. 6.6.3]). To provide an example, the power spectrum for selected stations is displayed in Figure 5.6 and the two predominant microseismic peaks are clearly visible. Power spectra everywhere on Earth feature characteristics which do not conform with the shape that stems from a forcing function of Gaussian white noise. If the shape of the power spectrum is known, this additional knowledge should be encoded in the a priori assumption. Thus, the previously made white noise assumption requires an adjusted such that the slope of the power spectra move closer to *reality*.

Many processes in the system Earth such as dispersion and/or attenuation have effects that can be approximated as a filter action on the wave field [Stein and Wysession, 2003, sec. 6.1]. To provide an example, an often addressed window function is:

Example 5.9 (Sinus cardinalis)**.** If we know that the external force took place within a finite time window, then $\operatorname{sinc}(\cdot)$ may be used as transfer function. This can be seen as follows: Let assume that t_i and $t_f = t_i + \Delta t$ are specifying the interval. Then, the Fourier transform of a boxcar function is given by

$$\mathcal{F}[\mathbf{1}_{\{t_i<t<t_f\}}](\xi) = \Delta t \operatorname{sinc}(\xi \Delta t) \exp\{-i\pi\xi(t_i + t_f)\} \tag{5.35}$$

and the center of the interval can be interpreted as a phase factor. A statement about the decay behavior of the amplitude spectrum can be made although the actual signal is not known. This principle translates one-to-one to a Fourier series expansion. ◁

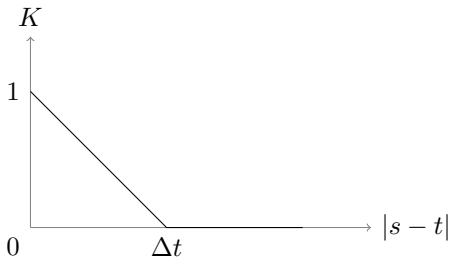

Figure 5.7.: Triangular covariance drawn as a function of the lag.

The concept of the following is to smooth the white noise assumption such that the resulting power spectra become compatible with additional knowledge such as microseism. Gaussianity – as the fundamental concept of that work – is preserved under stable, linear and time-invariant (LTI) filter, using the same arguments as in Section 1.3. As stated by the Wiener-Khinchin theorem for time stationary processes, the power spectrum and the auto-covariance function are forming a Fourier pair (see Eq. 2.32). As a consequence, a process can equivalently be characterized by its power spectrum or by its auto-covariance function. Loosely speaking, Gaussian white noise has a flat spectrum and puts equal power into each frequency. In turn, the power spectrum of filtered white noise is the squared amplitude spectrum of the response of the filter. Suppose $T_n(\cdot)$ denotes a suitable transfer function with respect to frequency domain and the index refers to a mode of a Fourier expansion. Then, the a priori cross power spectrum is given by

$$\mathrm{Cov}[F_m(\xi), F_n(\zeta)] = |T_n(\xi)|^2 \delta_{mn} \delta(\xi - \zeta) \tag{5.36}$$

where subscripts n and m are referring to individual modes and ξ and ζ frequencies. Because of the absolute value only the amplitude spectrum of the transfer function is of relevance and a potential phase factor vanishes. Transformed back into the space-time domain we have

$$\mathrm{Cov}[F(x,t), F(y,s)] = \mathcal{F}^{-1}\Big[\sum |T_n(\xi)|^2 \phi_n(y)\phi_n(x)\Big](t-s) \tag{5.37}$$

where again the shift property of the Fourier transform is used and $\phi_n(\cdot)$ refers to an orthonormal Fourier basis. One property worth mentioning is that filtering Gaussian white noise results in a process that features spatio-temporal correlations although uncorrelated with respect to the Fourier domain. Unlike Gaussian white noise, the decay behaviour of the transfer function must be chosen such that the resulting process is of finite power.

Example 5.10 (Triangular Covariance)**.** The sinc-function from the previous example transfers the white noise covariance into the triangular kernel [3]. This relation is established by the Fourier pair $\mathrm{sinc}^2(\cdot) \leftrightarrow \mathrm{tri}(\cdot)$. The triangular kernel can be interpreted as a weighted average and replaces the white noise impulses by triangular functions i.e. a Bartlett window [Gubbins, 2004, eq. 3.8]. The auto-covariance as a function of the lag is illustrated in Figure 5.7. Because the resulting power lacks normalization, the triangular kernel depends on two yet unknown parameters. The length of the interval Δt and the overall variance σ^2. ◁

[3] The phase factor in Example 5.9 is omitted because absolute value nullifies it.

To calculate the posterior mean and covariance, the cross-covariance and the data-covariance with respect to the smoothed a priori force are required. These are obtained analogously by applying the transfer function. The cross-covariance amongst forcing function and a velocity record follows to read

$$\mathrm{Cov}[F(y,s),V(x,t)] = \mathcal{F}^{-1}\Big[-2i\pi\xi \sum G_n(x,\xi)|T_n(\xi)|^2\phi_n(y)\Big](s-t) \tag{5.38}$$

where the minus sign results from the complex conjugation of the velocity. The cross-covariance matrix is composed as before. With respect to the time domain, the FFT in combination with the shift property admits an efficient implementation for any transfer function of choice. In contrast, the series expansion has to be carried out, individually for every pair of locations. The minimum receiver spacing together with a desired resolution admits to truncate the series. The expansion should be carried out until the Nyquist frequency is reached.

Example 5.11 (Smoothend cross-covariance)**.** Let us return to the damped vibrating string in combination with velocity records. In Example 5.5 it was shown that the white noise assumption is too rough to drive an inversion by velocity records. Using $\mathrm{sinc}(\cdot)$ as transfer function is reasonable if we know that the external force acted only for a finite period of time. For a single velocity record, the right panel in Figure 5.4 shows a smoothed cross-covariance structure. The Dirac fences at the edges of the light cones are replaced by overlapping triangular functions. This, however, brings up a problem because we do not know how strongly to smooth. ◁

As we can see from the above example, only the overall correlation structure is at hand but *not* appropriate parameters. Typically, transfer functions depend on a number of unknown parameters. Examples are correlation lengths and/or scaling factors to e.g. to adjust the overall variance. We can address this problem once we derive the data covariance. The covariance function for the velocity field calculates analogously

$$\mathrm{Cov}[V(y,s),V(x,t)] = \mathcal{F}^{-1}\Big[\sum (2\pi\xi)^2 G_n(y,\xi)\bar{G}_n(x,\xi)|T_n(\xi)|^2\Big](s-t) \tag{5.39}$$

and $\bar{(\cdot)}$ indicates the complex conjugate. When composing the data covariance matrix, the observation errors have to be taken into account (see Eq. 5.14). This brings up another problem because is not clear how to weight the smoothed covariance relative to the measurement errors. These parameters require estimation as part of an inference e.g. through addressing the so-called marginal likelihood (see Sec. 2.3).

Example 5.12 (Parameter search)**.** For both, the spatial and the temporal domain sinc is used as transfer function. That choice results in three model parameters that require estimation. They are the spatial extent Δx, temporal duration Δt and a scaling factor σ to adjust the overall kernel contribution. As described in Section 2.3, these parameters are searched by optimizing the logarithm of the so-called *marginal likelihood*. Therefore an ordinary grid-search is carried out and the resulting density is visualized in Figure 5.8. Because the bulk of the density is quite well concentrated, we settle for a point estimate rather than propagating the uncertainties. Estimated values are only to some extent comparable with the parameters of the external action of choice (see Ex. 5.3). The true duration of the action is $\Delta t = 0.9$ and the value recovered is $\widehat{\Delta t} = 0.585$. The estimate for

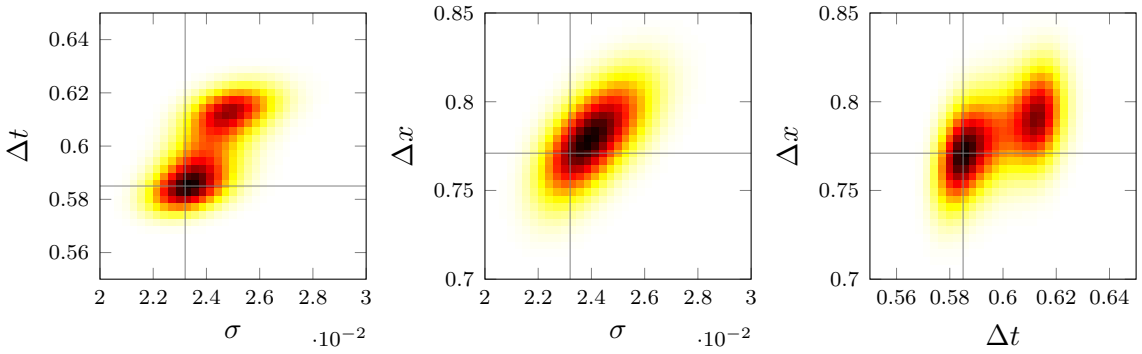

Figure 5.8.: Illustration of the marginal likelihood for the parameters of the transfer function in Example 5.12. In each panel one parameter out of three is integrated out. The marginalization is carried out by an ordinary Riemann sum.

the spatial extent is $\widehat{\Delta x} = 0.771$ whereas the *truth* is $\Delta x = 0.52$. Taking the slantening effect of the force's sinusoid into account makes the estimate appear reasonable. The scaling factor $\widehat{\sigma}$ has no direct interpretation and the associated uncertainty is discussed in the next example. ◁

Despite – or maybe because – of the rigorous Bayesian setting together with a consistent estimation of model parameters, we cannot expect to be able to reconstruct the external force almost exactly. The limiting factor are the discretized data and Shannon's sampling theorem. The precise reconstruction of a signal from discrete samples is only possible if the signal is of finite bandwidth [Gubbins, 2004, sec. 3.1]. In turn this means that a sampling rate of Δx does not allow recover of structures smaller than $2\Delta x$ i.e. the Nyquist rate. This principle serves as guideline and does not translate one-to-one because the inference utilizes Green's function to mediate between ground motion and external action. Nonetheless, with a coarse receiver spacing, spatial structures can only be resolved to a limited extent. The ongoing example of the damped vibrating string is chosen such that this effect is emphasized:

Example 5.13 (Posterior force)**.** Once the three model parameters are determined, we can estimate the external action that gave rise to the recorded string vibration. The force's posterior mean and (co)-variance are calculated according to Equations 5.17 and 5.18 and are illustrated in Figure 5.9. Compared to the white noise assumption the posterior mean has improved significantly. There are no more corners and edges and overall the forcing function shows improved localization. Considering the sampling theorem it becomes clear that the resolution of the reconstruction is limited. An ordinary low-pass Butterworth filter is applied to the actual force to improve the interpretation of the results. The corner-frequencies are roughly estimated as the arithmetic mean of the spatial and temporal sampling. As a result, the actual pulse becomes wider and flattens out. Figure 5.10 illustrates two cross sections through the center of the external action. Compared to the filtered force, the reconstruction is convincing and stays mostly within one standard deviation. The overshoots that can be seen in the time domain could be interpreted as a kind of a filter effect caused through the choice of the transfer function. It is of no concern that the low-pass filtered force does not always stay within error bounds because the uncertainties in the model parameters are not propagated. However, it is

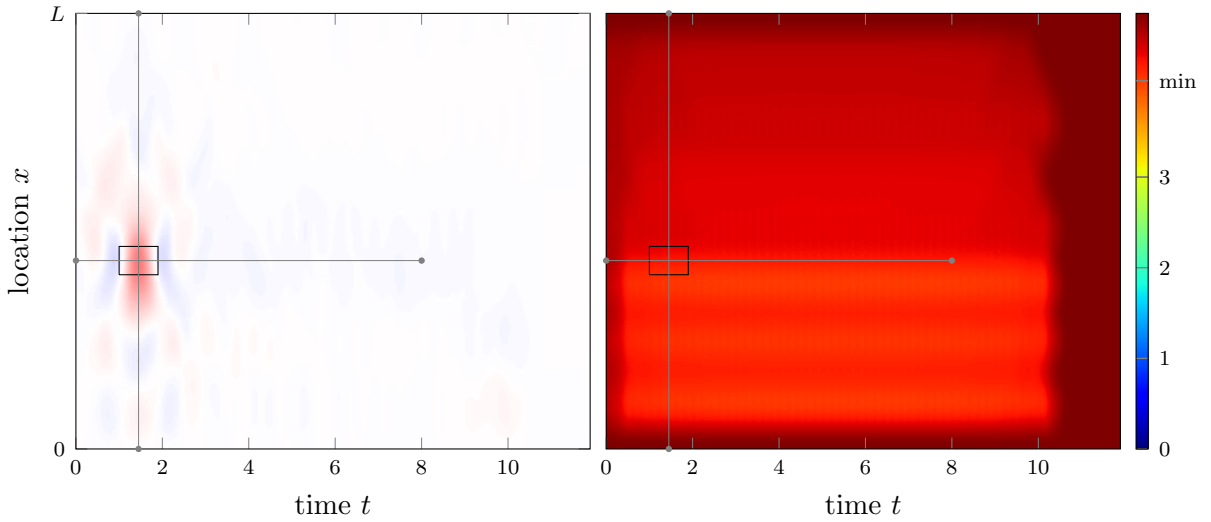

Figure 5.9.: Reconstruction of the external action using sinc as transfer function. The left panel illustrates the posterior mean with the same diverging colormap as in the previous figures. On the right, the posterior standard deviation is shown. The color scale is chosen to start at zero to illustrate the poor reduction in variance. At the locations of greatest variance reduction, the position of the observations can readily be guessed. The black square indicates the area of actual action and the gray lines are indicating cross sections that are detailed in Figure 5.10.

not clear why there is hardly any variance reduction achieved. Possible reasons are an underestimation of sigma, not propagating parameter uncertainties or a still inappropriate priori assumption. The classical Fourier series of the triangular kernel has both sin and cos contributions. In contrast, correlations that stem from a clamped vibrating string act only on the sin part. To use a covariance that has only sin contributions seems to be intuitive because one cannot pluck a camped string right at the ends. ◁

The prerequisite for filtering white noise is a parametric transfer function that can mimic the desired power spectrum. Coming back to the motivating example of microseism, a transfer function is needed whose square roughly approximates the characteristic noise peaks (see Fig. 5.6). An advantage using transfer functions is that the a priori process remains uncorrelated in the frequency domain. This is beneficial for calculating the covariances because an integral and a sum collapse. But it must also be said that using transfer functions alone is not truly satisfying in a seismological context. Although in the examples it was assumed that the force acts only within a finite interval, the a priori process spans the entire domain. It is often the case that the extent of a seismologically active area is known relatively accurately. The main microseism peak, for example, results from standing waves created in the open ocean [Shearer, 2009, sec. 11.2]. Because the receiver coverage is typically coarse, it would be desirable to encode the knowledge about the source region as part of the a priori distribution.

5.4. Outlook & Further Thoughts

A desirable extension of the presented method would be to concentrate the a priori force field to a specific source region. As shown in Example 1.6 and demonstrated in Section 4.2,

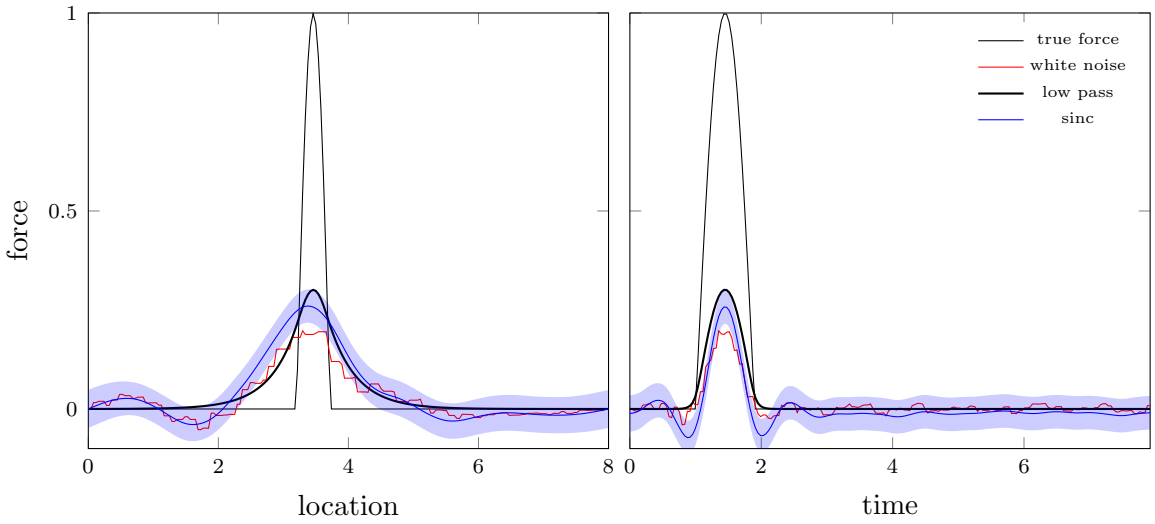

Figure 5.10.: Cross sections through the center of the force. Comparison of the actual force with the posterior mean ± one standard deviation. Every step in the white noise reconstruction (displacement based) corresponds to the onset of the back-projected light cones. Figure 5.9 illustrates the orientation of the cross sections.

an indicator function may be used to restrict a GP to a certain area. However, a sharp edged source region is hard to justify e.g. because the subsurface is not well known. A similar approach is the use of a taper function with a smooth transition to zero where it is known for sure that there are no sources. This will render meaningless solutions unlikely and deposit the energy at areas where activity is expected. If $g(x)$ denotes a suitable taper function, then, the a priori covariance for the external force becomes

$$\mathrm{Cov}[F(y,s),F(x,t)] = g(y)\,\mathcal{F}^{-1}\Big[\sum |T_n(\xi)|^2\phi_n(y)\phi_n(x)\Big](s-t)\,g(y)\;, \tag{5.40}$$

a non-stationary kernel (unless $g \equiv$ const.). To calculate the covariances for velocity records, an expansion of the product $G(x,\xi,z)g(z)$ may be addressed. The corresponding coefficients are given by a discrete convolution

$$Gg_n(x,\xi) = \big(G_{(\cdot)}(x,\xi) * g_{(\cdot)}\big)[n] = \sum G_m(x,\xi)g_{n-m} \tag{5.41}$$

where the g_n refer to the series coefficients of $g(z)$. In analogy to Equations 5.38 and 5.39, the cross- and data covariance follow to read

$$\mathrm{Cov}[F(y,s),V(x,t)] = \mathcal{F}^{-1}\Big[-2i\pi\xi\sum Gg_n(x,\xi)|T_n(\xi)|^2\phi_n(y)\Big](s-t) \tag{5.42}$$

and

$$\mathrm{Cov}[V(y,s),V(x,t)] = \mathcal{F}^{-1}\Big[\sum (2\pi\xi)^2 Gg_n(y,\xi)\bar{G}g_n(x,\xi)|T_n(\xi)|^2\Big](s-t)\;. \tag{5.43}$$

Because of the discrete convolution the covariance for velocity records ends up as a very unpleasant triple sum. It is within possible that this triple sum renders the approach prohibitive. To overcome that, there is chance that local stationary kernels offer a way to reduce the complexity. It is certainly worth a try to investigate whether it is beneficial to departure from local stationary white noise that is

$$g\left(\frac{x+y}{2}\right)\delta(x-y) \tag{5.44}$$

where, again, g is required to be a non-negative function of the centroid. The reason for this assumption is that local stationary kernels have a remarkable spectral representation

$$\mathcal{F}_{xy}\left[g\left(\frac{x+y}{2}\right)K(x-y)\right](\xi,\zeta) = \hat{K}\left(\frac{\xi+\zeta}{2}\right)\hat{g}(\xi-\zeta) \tag{5.45}$$

where K refers to a stationary kernel function and $\hat{(\cdot)}$ indicates the individual Fourier transforms [Genton, 2002, sec. 3]. The difficulty is to link this idea with the concept of using transfer functions. On top of all that, using a taper function will introduce additional model parameters that need to be determined. Because numerical optimization is in general difficult, the parameter search will become a bigger challenge as the number of model parameters grows.

Provided the calculation of the cross covariance and data covariance is feasible [4], then the numerical complexity is determined by a Cholesky factorization of the data covariance matrix. An implementation using the Cholesky factorization is only suitable for datasets of two to three thousand records because it scales with complexity $O(n^3)$. In addition, there is a massive memory requirement to store the covariance matrix that grows quadratically. To consider entire seismograms that easily sum up to millions of records is utterly out of the range. Nonetheless, there are very promising developments in the field of machine learning that circumvent both, the memory requirements and computational costs. Through bypassing the Cholesky decomposition, Gardner et al. [2018] succeeded in reducing the time complexity of the parameter search to almost $O(n^2)$, a tremendous acceleration. The key ingredient is the iterative conjugate gradients algorithm that does not need to keep the complete covariance matrix in main memory. By effectively using multi GPU parallelization, Wang et al. [2019] managed to estimate the model parameters for more than one million data points. This achievement demonstrates, that the presented approach – combined with data selection and reduction strategies – is within range. However, these promising approaches cannot hide the fact that a great deal of development work still needs to be done for an application.

[4] Calculating realistic Green's functions and solving the corresponding integrals is in itself a challenge and might demand a high performance computer.

Part III.

Geomagnetism

The following chapter was published as Mauerberger et al. [2020] in *Geophysical Journal International*. The conceptional and theoretical work was conducted by S. Mauerberger and M. Schanner, with significant assistance by M. Holschneider. M. Korte accounted for data selection as well as interpretation and embedding of the case study. S. Mauerberger prepared the manuscript with support of all authors. Software development and data processing were performed by M. Schanner with major contributions by S. Mauerberger. M. Korte and M. Holschneider supervised the findings of this work.

For the time stationary global geomagnetic field, a new modelling concept is presented. A Bayesian non-parametric approach provides realistic location dependent uncertainty estimates. Modelling related variabilities are dealt with systematically by making few subjective a priori assumptions. Rather than parametrizing the model by Gauss coefficients, a functional analytic approach is applied. The geomagnetic potential is assumed a Gaussian process to describe a distribution over functions. A priori correlations are given by an explicit kernel function with non-informative dipole contribution. A refined modelling strategy is proposed that accommodates non-linearities of archeomagnetic observables: First, a rough field estimate is obtained considering only sites that provide full field vector records. Subsequently, this estimate supports the linearization that incorporates the remaining incomplete records. The comparison of results for the archeomagnetic field over the past 1000 yr is in general agreement with previous models while improved model uncertainty estimates are provided.

6. Correlation Based Snapshot Models of the Archeomagnetic Field

Global geomagnetic field reconstructions of the past millennia are useful to investigate the geodynamo process or the complex interaction of the field with solar wind particles and cosmic rays, and they find application in archeomagnetic and palaeomagnetic dating. Reconstructions are typically built from volcanic and archeomagnetic samples collected at the Earth's surface providing records of the ancient Earth's magnetic field (EMF). Unfortunately, on a global scale records are clustered, unevenly distributed towards the Western Eurasian region and corrupted by various uncertainties. This considerably complicates the reconstruction of the ancient EMF.

Dating back to 1985, Gubbins and Bloxham were amongst the first to propose a Bayesian inference for modelling the EMF, already discussing non-linear observables and model uncertainties. Their parametrized implementation of a truncated spherical harmonic (SH) representation with norm optimization has become a widely used modelling scheme. The majority of historical and archeomagnetic field models published over the past years essentially rely upon the same inverse strategy. Early models such as Jackson et al. [2000], Constable et al. [2000], Korte and Constable [2003] provide estimates without quantifying uncertainties. More recent attempts – for example Korte et al. [2009], Licht et al. [2013], Hellio and Gillet [2018], Senftleben [2019] – describe variabilities by deriving ensembles of equivalent solutions. Roughly speaking, those models differ in two aspects: On the one hand, the error handling, data selection and outlier detection have been refined over the years [Licht et al., 2013]. On the other hand, different strategies are used to incorporate a priori knowledge. Early models are typically starting off from an axial dipole and are regularized by a physically motivated norm. Since regularized field models are known to underestimate uncertainties at small length scales [Gillet, 2019], more elaborate modelling concepts are under investigation. Recent attempts deduce a priori information including temporal dynamics from the statistics of satellite era models [e.g. Hellio and Gillet, 2018] or from geodynamo simulations [e.g. Sanchez et al., 2016]. Existing models, however, have in common that uncertainties related to modelling, in particular due to model parameters and the uneven data distribution, are not dealt with systematically.

This paper introduces an advanced concept to model snapshots of the EMF. This work should be considered as a first step towards a new inverse strategy in which the notion of modelling related uncertainties is well defined. Therefore, we adapt the correlation based inversion developed by Holschneider et al. [2016] that is known from modelling observatory and satellite data. Several modifications are required to adjust the concept to archeo- and palaeomagnetic data.

We pursue a fully Bayesian approach that determines the EMF's posterior distribution which simultaneously encodes the most probable field model and its uncertainties. To obtain the posterior distribution we use a functional analytic approach where inference takes place directly in the space of functions. Observables and quantities of interest are expressed in terms of functionals that act on the geomagnetic potential. Rather than

using a model that is parametrized by a finite SH basis, the geomagnetic potential is assumed a GP. The GP in use is non-parametric in the sense that it is a distribution over functions and is specified by a two-point covariance function.

From a parametric point of view, GPs have been used for a long time in modelling the EMF [e.g. Bouligand et al., 2005, Khokhlov et al., 2006], known under the term Giant Gaussian Process (GGP). That term was coined by Constable and Parker [1988] who proposed a GP based model focusing on the estimation of model parameters. Our approach may be seen as the functional analytic extension of the GGP model.

Our a priori distribution of the geomagnetic potential is characterized by its mean power spectral behaviour, which is represented by an explicit correlation function that takes all SH degrees into account. If a SH truncation was desired, transdimensional modelling [Livermore et al., 2018] may be applied to also infer the cut-off degree. Using an explicit kernel function – not truncated at a certain SH degree – circumvents that problem. This does not necessarily mean that our approach reaches a higher resolution at a global scale but, improves treatment of the uneven data coverage and allows the exploitation of the records to their fullest. In addition, a low SH degree truncation may lead to spurious oscillations and ringings if the data include pronounced local anomalies. We try to be the least subjective and specify the a priori field model using uninformative distributions, when possible. Our a priori model depends only on a single parameter that controls the a priori power spectral behaviour.

In the case of satellite and observatory data, EMF full vector components are observed directly and observables are linearly related to the geomagnetic potential. Thus the posterior distribution for the GP is explicit and may be computed using ordinary linear algebra. Archeomagnetic data, however, call for a refined modelling strategy that takes the non-linearity of declination, inclination and intensity into account. The majority of sites only have incomplete vector information so that a direct linearization of each record is not possible. Therefore, we propose a two step Bayesian update system: First, a rough field estimate is obtained considering only sites that provide complete field vector records. Subsequently, this estimate supports the linearization that incorporates the remaining incomplete records.

To demonstrate the potential of our modelling strategy we present a case study using archeomagnetic and volcanic data of the past 1000 yr. Joint maps of best prediction and point-wise uncertainty are presented, which allow an improved interpretation of the spatial field structure. Although our modelling is not based upon a SH basis, we predict Gauss coefficients and quantify their uncertainties. In addition, we calculate the posterior mean of the spatial power spectrum and estimate error bounds. Finally, we present the posterior probability density function (PDF) for the dipole strength and the location of the geomagnetic north pole. The latter results are obtained for records of the past millennium, coarsely sorted into 100 yr bins and arranged into a discrete time-series.

The structure of the document is as follows: Section 6.1 gives an overview of the modelling theory, which closely follows Holschneider et al. [2016]. First, we introduce the general construction of our dipole and non-dipole priors and correlations kernels (Section 6.1.1). We describe the link between observations and model, and how to obtain the posterior distribution from linear observations (Section 6.1.2), and then discuss the need for linearization of archeomagnetic observables (Section 6.1.3). The treatment of data uncertainties is laid out in Section 6.1.4. The following two Sections, 6.2 and 6.3, give details about the necessary adjustments to model archeo- and palaeomagnetic data.

Section 6.2 focuses on the two step strategy used to handle the non-linearities, formulated as a Bayesian update system and includes synthetic tests. Section 6.3 covers the translation of a priori uncertainties in the model parameters to the posterior. Finally, Section 6.4 provides the case study. We close the document by drawing conclusions and showing future perspectives in Section 6.5.

6.1. Modelling Concept

In this section we review the non-parametric and correlation-based modelling strategy [Holschneider et al., 2016] that underlies our Bayesian approach for modelling time-stationary snapshots of the EMF. We lay out our field model together with our a priori assumptions and establish the nomenclature that we adopt throughout this paper. We point out difficulties arising when working with non-linear observables such as archeo- and palaeomagnetic records and discuss the general treatment of data uncertainties.

6.1.1. Magnetic Field Model

When modelling the geomagnetic core field on archeo- to palaeomagnetic timescales, further contributions to the geomagnetic field – the crust, ionosphere and magnetosphere – are neglected since archeo- and palaeomagnetic measurement errors are assumed to be significantly larger than the respective field contributions [Constable and Korte, 2015].

Close to the surface – far from magnetic sources – the EMF, $\boldsymbol{B}$, can be approximated by the gradient of a scalar potential satisfying Laplace's equation

$$\boldsymbol{B} = -\nabla\Phi \quad , \qquad \nabla^2\Phi = 0 \ , \tag{6.1}$$

where Φ is referred to as the geomagnetic potential [Backus et al., 1996, chap. 4]. In contrast to the magnetic field, the potential is not directly observable. For an internal source and with respect to some reference sphere of radius R, the potential Φ at location $\boldsymbol{x}, |\boldsymbol{x}| > R$, can be expanded in SHs

$$\Phi(\boldsymbol{x}) = R\sum_{\ell}\left(\tfrac{R}{|\boldsymbol{x}|}\right)^{\ell+1}\sum_{-l\le m\le l} g_\ell^m\, Y_\ell^m(\hat{\boldsymbol{x}}) \tag{6.2}$$

where Y_ℓ^m refers to the real valued and Schmidt semi-normalized SH of degree l and order m with related Gauss coefficient g_ℓ^m. The dependence of g_ℓ^m on a reference radius R is not explicitly typed.

We use a spherical coordinate system with B_N pointing to geographic north, B_E to the east and B_Z vertically downward. The components of the magnetic field vector $\boldsymbol{B}$ at a location of radius r, colatitude θ and longitude ϕ, are

$$B_N = -\frac{1}{r}\frac{\partial\Phi}{\partial\theta} \ , \qquad B_E = \frac{1}{r\sin(\theta)}\frac{\partial\Phi}{\partial\phi} \ , \qquad B_Z = -\frac{\partial\Phi}{\partial r} \ . \tag{6.3}$$

The ellipticity of the Earth is neglected and geocentric coordinates are treated as if they were geodetic.

Motivated by magnetic field theory, we express our lack of knowledge by making a priori assumptions about all Gauss coefficients. Considering Holocene timescales, the dipole part

of the field does not have the same statistical distribution as the non-dipole part [Constable and Parker, 1988]. Thus, our model of choice is dipole dominated with additional random field contributions.

The dipole part is specified by the Gauss coefficients of degree $\ell = 1$. A priori the coefficients are assumed normal

$$\boldsymbol{g}_1 \sim \mathcal{N}(\bar{\boldsymbol{g}}_1, \Sigma_1) \ , \tag{6.4}$$

with mean vector $\bar{\boldsymbol{g}}_1$ and covariance matrix Σ_1, nine model parameters to be determined (three for the mean and six for the covariance). The subscript is hinting at the SH degree and is going to be generalized. The dipole potential is a GP with mean and covariance function

$$\mathbb{E}[\Phi_{\mathrm{DP}}] = \boldsymbol{Y}_1^\top \boldsymbol{g}_1 \ , \qquad \mathbb{V}[\Phi_{\mathrm{DP}}] = \boldsymbol{Y}_1^\top \Sigma_1 \boldsymbol{Y}_1 \tag{6.5}$$

where $\boldsymbol{Y}_1$ refers to the SH basis of degree $\ell = 1$, that is

$$\boldsymbol{Y}_1(\boldsymbol{x})^T = R\left(Y_1^0(\boldsymbol{x}), Y_1^1(\boldsymbol{x}), Y_1^{-1}(\boldsymbol{x})\right)\left(\tfrac{R}{|\boldsymbol{x}|}\right)^2 \ . \tag{6.6}$$

Section 6.3.1 deals with choosing a priori dipole parameters. The correlation pattern for independent and identically distributed (IID) Dipole coefficients is shown in the left-hand panel of Figure 6.1.

Our model of the non-dipole part is similar to the one proposed by Constable and Parker [1988] but within a Bayesian setting. The non-dipole potential is assumed a GP of zero mean. Rather than truncating the SH basis at a certain degree, a covariance function of closed form is used. Therefore, we adopt the kernel construction method from Holschneider et al. [2016, Sec. 4]. With respect to a reference sphere, the potential is characterized by its mean power spectral behaviour (see Sec. 6.4.6). The reference sphere may be seen as a virtual source region and the reference radius has no particular physical significance [Constable and Parker, 1988]. Except for the dipole, at the reference radius Gauss coefficients are assumed IID normal random variables. The corresponding Legendre type kernel is of the form

$$K_{\mathrm{L}}(\boldsymbol{x}, \boldsymbol{y}) = \lambda^2 R^2 \sum_{\ell=2}^{\infty} \left(\frac{R^2}{|\boldsymbol{x}||\boldsymbol{y}|}\right)^{\ell+1} \sum_{-l \le m \le l} Y_\ell^m(\hat{\boldsymbol{x}}) Y_\ell^m(\hat{\boldsymbol{y}}) \tag{6.7}$$

where we introduced λ, a scaling factor that controls the amount of the non-dipole contribution and its dimension. A priori, dipole and non-dipole parts are assumed statistically independent. According to Holschneider et al. [2016, Eq. 54], K_{L} can be expressed in closed form. We choose this kernel as it is computationally simple and depends only on two parameters, R and λ. The right-hand panel of Figure 6.1 depicts an a priori correlation pattern for the non-dipole potential.

In principle an SH decomposition is possible but computationally limited. To cover highly localized modelling errors, the covariance has to map the characteristic length scales present in the data. Small scale correlations among records are a valuable source of information that should enter the model. In case of an expansion, the highest SH degree must be chosen such that the smallest spatial wavelength coincides with distances between sites. If some sites were clustered with distances of about 100 km, an expansion

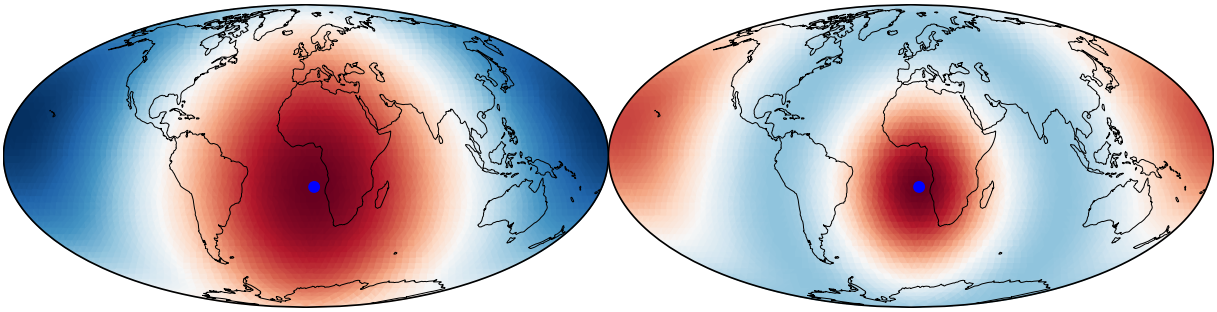

Figure 6.1.: Visualization of dipole (left-hand panel) and non-dipole (right-hand panel) correlation structures at the Earth's surface. The reference location is indicated by the blue dot. The dipole pattern corresponds to IID dipole coefficients. The apparent correlation of antipodal points stems from the fact that this kernel describes the fluctuations around the removed dipole, which is dominated by the quadrupole contributions. Overall cross correlations result from superimposing dipole and non-dipole parts. Correlations are normalized and a diverging colourmap is used, ranging from -1 (blue) to 1 (red).

up to degree $\ell \leq 175$ would be needed. Since high degree expansions are demanding, using an explicit kernel function is computationally beneficial and makes it feasible to adopt a global point of view while preserving the accuracy of local length scales.

The reference radius R controls the predominant slope of the a priori power spectrum. The smaller R, the smoother the a priori field at the surface. The scaling factor λ causes a shift along the axis of ordinates. Similar to Constable and Parker [1988], the power spectrum is used to tune the reference radius. R is chosen such that the prior mean power spectrum roughly conforms with the International Geomagnetic Reference Field (IGRF) [Thébault et al., 2015] time average from 1900 to 2020. Figure 6.2 depicts the alignment, carried out by visual inspection. Throughout we are going to use the fixed value $R = 2800\,\mathrm{km}$. In contrast, as λ is highly uncertain, it receives special treatment in Section 6.3.2. A reference radius below the core-mantle boundary (CMB) implies a non-erratic covariance structure at the CMB. A virtual source region within the outer core may seem unconventional but is becoming more popular. To give an example, Sanchez et al. [2016] are using complex correlation patterns at the CMB obtained from dynamo simulations as prior information.

The EMF is modelled as the negative gradient of the potential (Eq. 6.1). Differentiation is a linear operation and thus the field model is a GP as well. The a priori mean reads

$$\bar{\boldsymbol{B}}(\boldsymbol{x}) = -\nabla \boldsymbol{Y}_1^\top(\boldsymbol{x})\bar{\boldsymbol{g}}_1 = -\sum_{-1\leq m\leq 1} \nabla Y_1^m(\boldsymbol{x})\bar{g}_1^m \tag{6.8}$$

and the correlation kernel is composed of dipole and non-dipole covariance functions

$$K_{\boldsymbol{B}}(\boldsymbol{x},\boldsymbol{y}) = K_{\boldsymbol{B},\mathrm{DP}} + K_{\boldsymbol{B},\mathrm{ND}} = \nabla \boldsymbol{Y}_1^\top(\boldsymbol{x})\Sigma_1 \boldsymbol{Y}_1(\boldsymbol{y})\nabla^\dagger + \nabla K_\mathrm{L}(\boldsymbol{x},\boldsymbol{y})\nabla^\dagger \tag{6.9}$$

where the right-hand gradient acts on the left at the second argument.

6.1.2. Inference

In contrast to existing archeomagnetic field models we pursue a functional analytic approach and perform a regression directly in the space of functions. The corner stone of

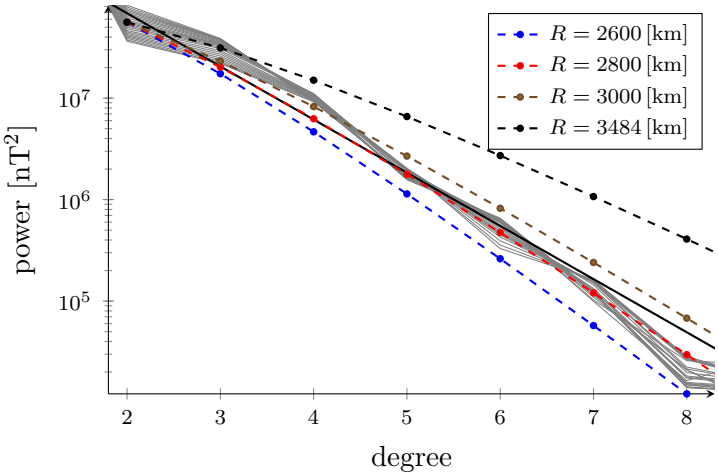

Figure 6.2.: The mean power spectral behaviour chosen a priori is indicated by red dots. As a reference, the power spectra of the IGRF from 1900 to 2020 are drawn in grey. For comparison, the solid black line refers to the mean spectrum suggested by Constable and Parker [1988].

our modelling strategy is formed by a GP regression. GP regression is also known in the field of geostatistics as Kriging [see Rasmussen and Williams, 2006, Sec. 2.2 and 2.8]. For linear $\boldsymbol{B}$-field observations this strategy has already been adopted by Holschneider et al. [2016]. We briefly recall the overall modelling concept before we introduce non-linear observations.

Suppose we observe the components B_N, B_E and B_Z of the EMF. Measurements are denoted by

$$o = \left\{ \boldsymbol{o}_i = \big(B_N(\boldsymbol{z}_i), B_E(\boldsymbol{z}_i), B_Z(\boldsymbol{z}_i)\big)^\top \right\}_{i=1,\dots,n} , \tag{6.10}$$

recorded at locations $\boldsymbol{z}_i$. Observations are corrupted by additive noise $\boldsymbol{E}_i$ and the data model reads

$$O = \{\boldsymbol{B}(\boldsymbol{z}_i) + \boldsymbol{E}_i\}_{i=1,\dots,n} . \tag{6.11}$$

The error model is assumed normal of zero mean and covariance Σ_E. Recorded values o are assumed a realization of O. The data's a priori mean vector and covariance matrix are

$$\bar{O} = \big\{\bar{\boldsymbol{B}}(\boldsymbol{z}_i)\big\}_{i=1,\dots,n} \qquad \text{and} \qquad \Sigma_O = \{K_{\boldsymbol{B}}(\boldsymbol{z}_i, \boldsymbol{z}_j)\}_{i,j=1,\dots,n} + \Sigma_E \tag{6.12}$$

with typically diagonal error-covariance matrix Σ_E.

To obtain information about the EMF, we need to compute $\boldsymbol{B}$'s posterior distribution. If we assume the $\boldsymbol{B}$-field and measurement errors are independent, the cross covariance matrix follows to read

$$\Sigma_{\boldsymbol{B}(\boldsymbol{x})O} = \mathrm{Cov}[\boldsymbol{B}(\boldsymbol{x}), O] = \{K_{\boldsymbol{B}}(\boldsymbol{x}, \boldsymbol{z}_i)\}_{i=1,\dots,n} \tag{6.13}$$

for any design point $\boldsymbol{x}$ outside the reference sphere. Since O and $\boldsymbol{B}$ are jointly Gaussian, the posterior distribution is normal as well. It is fully determined by the conditional mean and conditional covariance

$$\mathbb{E}[\boldsymbol{B}(\boldsymbol{x})|o] = \bar{\boldsymbol{B}}(\boldsymbol{x}) \quad + \Sigma_{\boldsymbol{B}O}\Sigma_O^{-1}\big(o - \bar{O}\big) \tag{6.14}$$

$$\mathrm{Cov}[\boldsymbol{B}(\boldsymbol{x}), \boldsymbol{B}(\boldsymbol{y})|o] = K_{\boldsymbol{B}}(\boldsymbol{x}, \boldsymbol{y}) - \Sigma_{\boldsymbol{B}O}\Sigma_O^{-1}\Sigma_{\boldsymbol{B}O}^\top . \tag{6.15}$$

Gauss coefficients are modelled analogously. Magnetic potential and Gauss coefficients are related through

$$g_\ell^m = \frac{2\ell+1}{4\pi R} \iint Y_\ell^m(\boldsymbol{x})\Phi(\boldsymbol{x})\,\mathrm{d}^2\boldsymbol{x}\ , \tag{6.16}$$

where integration is carried out over the sphere of radius R. The collection of Gauss coefficients up to SH degree ℓ is denoted by $\boldsymbol{g}_\ell$. Accordingly, Σ_ℓ refers to the prior covariance matrix of Gauss coefficients up to degree ℓ. Except for the dipole and according to Eq. 6.7, the a priori covariance Σ_ℓ is diagonal. At the reference radius the a priori variance is λ^2. The cross covariance matrix between $\boldsymbol{g}_\ell$ and the observations reads

$$\Sigma_{\ell O} = \mathrm{Cov}[\boldsymbol{g}_\ell, O] = \{-\Sigma_\ell \nabla \boldsymbol{Y}_\ell(\boldsymbol{z}_i)\}_{i=1,\dots,n} \tag{6.17}$$

and $\boldsymbol{Y}_\ell$ refers to the SH basis up to degree ℓ. Since Gauss coefficients and potential are linearly related, the posterior distribution is normal as well and, again, fully determined by the conditional mean and covariance

$$\mathbb{E}[\boldsymbol{g}_\ell|o] = \bar{g}_\ell\ + \Sigma_{\ell O}\Sigma_O^{-1}\big(o-\bar{O}\big) \tag{6.18}$$

$$\mathbb{V}[\boldsymbol{g}_\ell|o] = \Sigma_\ell - \Sigma_{\ell O}\Sigma_O^{-1}\Sigma_{\ell O}^\top\ . \tag{6.19}$$

It is worth mentioning that Bouligand et al. [2005] used geodynamo simulations to conclude that there are significant cross-correlations among the Gauss coefficients. Our statistical model places no restrictions on the posterior cross-correlations although the a priori assumptions are based on IID Gauss coefficients.

These are the formulae we are going to build our modelling strategy upon. However, they require an extension since archeomagnetic records and the magnetic potential are non-linearly related. Before elaborating our approach to this problem, let us first recall the observational functionals in question.

6.1.3. Observational functionals & Linearization

Archeomagnetic data is not provided in the form of Cartesian field vector components. The quantities that are determined in laboratory experiments are the two angles, declination (D) and inclination (I), and intensity (F), acquired at locations $\boldsymbol{z}$. These quantities and the vector components are non-linearly related. The corresponding functionals read

$$\mathbf{H} : \boldsymbol{B} \to \begin{pmatrix} D \\ I \\ F \end{pmatrix} = \begin{pmatrix} \arctan\left(\frac{B_E}{B_N}\right) \\ \arctan\left(\frac{B_Z}{F_H}\right) \\ \sqrt{B_N^2 + B_E^2 + B_Z^2} \end{pmatrix} \tag{6.20}$$

where the horizontal intensity

$$F_H = \sqrt{B_N^2 + B_E^2} \tag{6.21}$$

is introduced as an auxiliary variable [Backus et al., 1996, Eq. 1.2.1 – 1.2.4]. The components of $\mathbf{H}$ are referred to as observation functionals and are denoted by $\mathrm{H}_i[\boldsymbol{B}]$ for $i = D, I, F$. The inverse map to magnetic field vector components reads

$$\mathbf{H}^{-1} : \begin{pmatrix} D \\ I \\ F \end{pmatrix} \to \boldsymbol{B} = F \begin{pmatrix} \cos(I)\cos(D) \\ \cos(I)\sin(D) \\ \sin(I) \end{pmatrix}\ . \tag{6.22}$$

The inverse map is only unique if all three observables are at hand, that is three vector components relate uniquely to three observables and vice versa.

As already pointed out in the previous Section, the vital prerequisite for the modelling strategy is joint normality of observations and EMF. Certainly, the functionals D, I and F do not preserve $\boldsymbol{B}$'s normality, nor are the measurement errors Gaussian.

To adopt the modelling concept by Holschneider et al. [2016] we approximate D, I and F by a 1^{st} order Taylor expansion

$$\mathrm{H}_i[\boldsymbol{B}] \approx \mathrm{H}_i[\tilde{\boldsymbol{B}}] + \nabla \mathrm{H}_i[\tilde{\boldsymbol{B}}]^\top \big(\boldsymbol{B} - \tilde{\boldsymbol{B}}\big) \ , \tag{6.23}$$

where $\tilde{\boldsymbol{B}}$ refers to an arbitrary point of expansion (POE). This is the non-parametric counterpart compared with the approach presented by Gubbins and Bloxham [1985, Eq. 10]. The functionals approximating D, I and F arise to

$$D \approx \tilde{D} + \frac{1}{\tilde{F}_H^2} \begin{bmatrix} -\tilde{B}_E \\ \tilde{B}_N \\ 0 \end{bmatrix}^\top \boldsymbol{B} \ , \tag{6.24}$$

$$I \approx \tilde{I} + \frac{1}{\tilde{F}_H} \left(\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} - \frac{\tilde{B}_Z}{\tilde{F}} \frac{\tilde{\boldsymbol{B}}}{\tilde{F}} \right)^\top \boldsymbol{B} \ , \tag{6.25}$$

$$F \approx \frac{\tilde{\boldsymbol{B}}^\top}{\tilde{F}} \boldsymbol{B} \tag{6.26}$$

where $\tilde{D}$, $\tilde{I}$ and $\tilde{F}$ are referring to the 0^{th} order terms. From those equations it is obvious that performing an expansion about zero is not going to work. Existing models [e.g. Licht et al., 2013, Hellio and Gillet, 2018] typically use an axial dipole as initial POE. Since a Taylor expansion performs better the less $\boldsymbol{B}$ deviates from the POE, we propose not to use an axial dipole. Section 6.2 is dedicated to which point to linearize about.

Since $\boldsymbol{B}$ is assumed a GP and due to linearity, the approximating observational functionals are normally distributed. However, to actually achieve joint normality amongst observations and $\boldsymbol{B}$-field, measurement errors need a Gaussian proxy as well.

6.1.4. Measurement Errors

Although the observational functionals were linearised, the data model is still not normal since measurement errors are not necessarily Gaussian. It is common to characterize the uncertainty of archeomagnetic directions (D and I) using the von Mises-Fisher (vMF) distribution and the variability of intensity using the normal distribution [Love and Constable, 2003]. As long as intensity records report the standard error, linearizing F provides a normal proxy model which we are going to use for inference. However, we lack a Gaussian proxy for directional errors.

The commonly used approach is to calculate directional errors individually. The vMF distribution is parametrized by a concentration factor and a location parameter. The larger the value of the concentration factor, the more the distribution tends towards concentrating around the location parameter. Provided a large concentration factor and that the location parameters are not pointing towards high latitudes, the marginal probability densities for D and I are approximately Gaussian. Proxy errors for declination and

inclination are assumed independent, of zero mean and standard deviation

$$\sigma_I = \frac{57.3^\circ}{140}\alpha_{95} \qquad \text{and} \qquad \sigma_D = \frac{1}{\cos o_I}\sigma_I \tag{6.27}$$

[Suttie and Nilsson, 2019]. Typically directional records report the 95% confidence cone α_{95} of the vMF distribution. This is a pragmatic approach that does not necessarily reflect the Gaussian moment matching proxy, since correlations are dropped and – in general – the mean of a vMF distribution does not coincide with the location parameters. In case of isolated declinations this approach does not work and such records are rejected. Another drawback of a Gaussian proxy error model is the intolerance against outliers. In the context of optimization theory there exist more robust approaches e.g. Farquharson and Oldenburg [1998], Walker and Jackson [2008], Hellio et al. [2014]. For the vast majority of non-Gaussian likelihoods, however, no explicit solution to the Bayesian inverse problem is known.

Two philosophies have been used in previous work when modelling archeo- and palaeomagnetic data. Either very strict selection criteria are applied and these often contain tests that had not been applied in data published a few decades ago. As the absence of the test does not necessarily mean a result would not have passed the test, the other philosophy is to include as much data as possible without applying very strict criteria, aiming to increase the signal to noise ratio. However, it is likely that the reported measurement errors in several cases underestimate the true data uncertainties, which might contain systematic biases if corrections for, for example, cooling rate or anisotropy have not been performed.

For practical reasons, we adopt the second philosophy. To compensate for possibly non-conforming error estimates we introduce a scaling parameter ϵ. Then, the individual proxy uncertainties are given by

$$E_i \to \epsilon E_i \sim \mathcal{N}\big(0,\, \epsilon^2\sigma_i^2\big) \tag{6.28}$$

and ϵ is regarded as a model parameter. Although we do not know the specific value for ϵ, its order of magnitude is assumed to be one.

In addition, we introduce a residual term $\boldsymbol{P}$ that compensates for modelling related errors and accounts for observational biases. Amongst others, effects that our model does not include are temporal correlations, dating errors and crustal field anomalies. Therefore, we assume $\boldsymbol{P} \sim \mathcal{N}(0,\, \boldsymbol{I})$, that is uncorrelated standard normal at every pair of distinct sites. Our final data model becomes

$$o_i = \mathrm{H}_i[\boldsymbol{B}_i + \rho\boldsymbol{P}_i] + \epsilon E_i \tag{6.29}$$

and the magnitude of the residual is controlled by ρ, another not yet known model parameter. The residual term can be thought of as an error term that describes the simplification of the underlying physics statistically. In other words, those real world contributions that are not covered by our model are treated as if they were random errors.

Because we focus on time stationary snapshots of the EMF, dating uncertainties are displaced into the residual term. Nonetheless, these errors are of importance if the temporal behaviour of the EMF is reconstructed. There already exist several approaches to accommodate dating errors for example Jack-knife [Korte et al., 2009, Licht et al., 2013],

MCMC sampling [Hellio et al., 2014] or transdimensional modelling [Livermore et al., 2018].

To proceed with the modelling concept outlined in Section 6.1.2, we combine the linearization and the Gaussian proxy for the directional errors. However, a POE for the linearization is still missing. The subsequent section covers this problem and also discusses the concrete incorporation of the error term and the residual. As the final ingredient of our modelling concept, the treatment of model parameters is the subject of Section 6.3.

6.2. Bayesian Update System

The need for linearization described in Section 6.1.3 requires a suitable POE. The linearization as a Taylor expansion performs better the less the POE deviates from the *truth*. Archeo- and palaeomagnetic directional and intensity data are determined from different laboratory experiments, and the majority of records report either one or two field components (incomplete records). The complete vector information (D, I, F) is only available in rare cases. Noting that it is easier to determine a POE from full vector records, we introduce a Bayesian update system to treat complete and incomplete records successively.

The posterior distribution is computed by a two step strategy only considering a subset of observations at a time. Records are partitioned into two disjoint groups $o_{\mathcal{I}}$ and $o_{\mathcal{C}}$ where subscripts are referring to incomplete and complete measurements. Making use of the conditional probability rule – that is $p(X|Y)p(Y) = p(X, Y)$ – and according to Bayes' law the posterior $\boldsymbol{B}$-field factorizes

$$p(\boldsymbol{B}|o) = p(\boldsymbol{B}|o_{\mathcal{C}}, o_{\mathcal{I}}) = \frac{p(o_{\mathcal{I}}|\boldsymbol{B}, o_{\mathcal{C}})}{p(o_{\mathcal{I}}|o_{\mathcal{C}})} p(\boldsymbol{B}|o_{\mathcal{C}}) \ , \tag{6.30}$$

that is the posterior EMF based on the complete observations $o_{\mathcal{C}}$ serves as prior for the Bayesian posterior based on $o_{\mathcal{I}}$. Not to incorporate the data all at once appears to be a promising strategy due to strong magnetic field correlations.

The complexity of the developed algorithm is growing through this and the following section. Figure 6.3 provides a schematic illustration so as not to loose the overview. The two step strategy is shown in the top panel of Figure 6.3.

6.2.1. Complete Records

In the initial step only complete records are taken into account. Triplets of declination, inclination and intensity are forming the set of complete records

$$o_{\mathcal{C}} = \left\{ \boldsymbol{o}_i = (D(\boldsymbol{z}_i), I(\boldsymbol{z}_i), F(\boldsymbol{z}_i))^\top \right\}_{i=1,\dots,n_{\mathcal{C}}} \ . \tag{6.31}$$

In order to apply the linearization, a POE is missing. The special case of knowing all three components allows the calculation of the inverse map (Eq. 6.22)

$$\tilde{\boldsymbol{B}}_{\mathcal{C}} = \left\{ \mathbf{H}^{-1}[\boldsymbol{o}_i] \right\}_{i=1,\dots,n_{\mathcal{C}}} \tag{6.32}$$

which will serve as the point to linearize about. That point is reasonable as long as the prior $\boldsymbol{B}$-field variance is large in comparison with the measurement errors. If measurement

errors were not negligible w.r.t. the a priori distribution, then the POE is no longer known with confidence and we have to propagate uncertainties. To ensure this, Section 6.3 is devoted to choosing an uninformative a priori field. A Gaussian approximation in this way is also known as Laplace's method [Murphy, 2012, section 8.4.1].

In order to apply the modelling scheme introduced in Section 6.1.2 we use the Gaussian proxy error model. The diagonal error covariance matrix is denoted by

$$\Sigma_{E,\mathcal{C}} = \operatorname{diag}\left(\sigma_1^2, \ldots, \sigma_{n_\mathcal{C}}^2\right) \ , \tag{6.33}$$

where σ_i refers to individual standard errors w.r.t. D, I and F. The linearised observation functionals translate between (D, I, F) and $\boldsymbol{B}$. To keep equations concise, the dipole basis and the Jacobi matrices are collected in big matrices

$$\acute{Y}_{1,\mathcal{C}} = \{\nabla \boldsymbol{Y}_1(\boldsymbol{z}_i)\}_{i=1,\ldots,n_\mathcal{C}} \qquad (3n_\mathcal{C} \times 3) \tag{6.34}$$

$$\acute{\mathrm{H}}_\mathcal{C} = \left\{\delta_{ij} \nabla \mathbf{H}[\tilde{\boldsymbol{B}}_i]\right\}_{i,j=1,\ldots,n_\mathcal{C}} \qquad (3n_\mathcal{C} \times 3n_\mathcal{C}) \tag{6.35}$$

where δ_{ij} refers to the Kronecker delta and $\acute{\mathrm{H}}_\mathcal{C}$ is 3×3 block-diagonal. For $o_\mathcal{C}$, the approximative prior mean vector is given by

$$\bar{O}_\mathcal{C} \approx o_\mathcal{C} + \acute{\mathrm{H}}_\mathcal{C}^\top \left(-\acute{Y}_{1,\mathcal{C}} \bar{\boldsymbol{g}}_1 - \tilde{\boldsymbol{B}}_\mathcal{C}\right) \tag{6.36}$$

where $\boldsymbol{B}_\mathcal{C}$ means evaluated at all the locations of observation. Due to assumed independence of error model and EMF, the linearised covariance matrix for complete records reads

$$\Sigma_\mathcal{C} = \mathbb{V}[O_\mathcal{C}] \approx \acute{\mathrm{H}}_\mathcal{C}^\top \left(\mathbb{V}[\boldsymbol{B}_\mathcal{C}] + \rho^2 \mathbb{I}\right) \acute{\mathrm{H}}_\mathcal{C} + \epsilon^2 \Sigma_{E,\mathcal{C}} \ , \tag{6.37}$$

where $\mathbb{V}[\boldsymbol{B}_\mathcal{C}]$ is constructed from the kernel (Eq. 6.9) at any two locations of observation. Due to bi-linearity of the covariance $\acute{\mathrm{H}}_\mathcal{C}$ is factored out and all constant terms are stripped off. For arbitrary design points, the linearised cross covariance amongst EMF and measurements is given by

$$\Sigma_{\boldsymbol{B}\mathcal{C}} = \operatorname{Cov}[\boldsymbol{B}, O_\mathcal{C}] \approx \operatorname{Cov}[\boldsymbol{B}, \boldsymbol{B}_\mathcal{C}] \acute{\mathrm{H}}_\mathcal{C} \ . \tag{6.38}$$

According to Equations 6.14 and 6.15, a Gaussian proxy of $\boldsymbol{B}$'s posterior distribution is determined through the conditional mean $\mathbb{E}[\boldsymbol{B}|o_\mathcal{C}]$ and conditional covariance $\mathbb{V}[\boldsymbol{B}|o_\mathcal{C}]$. This first step of our modelling scheme is illustrated in the top left panel of Figure 6.3.

Analogous to Eq. 6.17, the linearised cross covariance amongst Gauss coefficients and observations is given by

$$\Sigma_{\ell\mathcal{C}} = \operatorname{Cov}[\boldsymbol{g}_\ell, O_\mathcal{C}] \approx \operatorname{Cov}[\boldsymbol{g}_\ell, \boldsymbol{B}_\mathcal{C}] \acute{\mathrm{H}}_\mathcal{C} = -\Sigma_\ell \acute{Y}_{\ell,\mathcal{C}} \acute{\mathrm{H}}_\mathcal{C} \ . \tag{6.39}$$

where $\acute{Y}_{\ell,\mathcal{C}}$ extends Eq. 6.34 up to SH degree ℓ. The approximations of the conditional mean $\mathbb{E}[\boldsymbol{g}_\ell|o_\mathcal{C}]$ and conditional covariance $\mathbb{V}[\boldsymbol{g}_\ell|o_\mathcal{C}]$ are given through Equations 6.18 and 6.19.

Although the subset of complete records is comparatively small, as a first *guess*, we anticipate a reconstruction of the EMF's dominating features due to strong magnetic field correlations.

6.2.2. Incomplete Records

Inference with the incomplete records only implicitly depends on the a priori mean and covariance function through the first step. The distribution of the EMF posterior to $o_{\mathcal{C}}$ may be understood as the prior to incorporate the remaining measurements $o_{\mathcal{I}}$. To carry out the linearization, the mean conditional on $o_{\mathcal{C}}$ will serve as POE

$$\tilde{\boldsymbol{B}}_{|\mathcal{C}} = \mathbb{E}[\boldsymbol{B}|o_{\mathcal{C}}] \ . \tag{6.40}$$

From a theoretical point of view this is a function that is going to be evaluated by the observational functionals. We expect those points to be well suited, as in a Gaussian model the mean is the most likely solution. However, this is an arbitrary choice and does not necessary imply any optimality.

Because the EMF's proxy posterior to $o_{\mathcal{C}}$ is normal we are going to use the same modelling concept as we have already done (Sec. 6.1.2). To facilitate the second step it is necessary not only to predict on design points but also on *auxiliary* quantities in the first step.

Incomplete records are treated individually even though a certain location may report more than one observable. The approximative mean is given by

$$\bar{O}_{\mathcal{I}|\mathcal{C}} \approx \left\{ \mathrm{H}_i\big[\tilde{\boldsymbol{B}}_{|\mathcal{C}}\big] \right\}_{i=1,\dots,n_{\mathcal{I}}} \tag{6.41}$$

and we use the subscript i to indicate both the location $\boldsymbol{z}_i$ and the type of record in H_i, either D, I or F. If we collect all gradients within one big matrix

$$\acute{H}_{\mathcal{I}|\mathcal{C}} = \left\{ \delta_{ij} \nabla \mathrm{H}_i[\tilde{\boldsymbol{B}}_{|\mathcal{C}}] \right\}_{ij=1,\dots,n_{\mathcal{I}}} , \tag{6.42}$$

then the covariance matrix for incomplete records reads

$$\Sigma_{\mathcal{I}|\mathcal{C}} \approx \acute{H}^{\top}_{\mathcal{I}|\mathcal{C}} \left(\mathbb{V}[\boldsymbol{B}_{\mathcal{I}}|o_{\mathcal{C}}] + \rho^2 \mathbb{I} \right) \acute{H}_{\mathcal{I}|\mathcal{C}} + \epsilon^2 \Sigma_{E,\mathcal{I}} \, . \tag{6.43}$$

The error covariance matrix is analogous to Eq. 6.33. The auxiliary quantities we have to carry along are conditional mean vector and covariance matrix at points of observation.

To model the posterior $\boldsymbol{B}$-field we again have to calculate linearised cross correlations amongst $\boldsymbol{B}|o_{\mathcal{C}}$ and $O_{\mathcal{I}}$. If we store the matrix $\mathrm{Cov}[\boldsymbol{B}, \boldsymbol{B}_{\mathcal{I}}|o_{\mathcal{C}}]$ within the first step, linearised cross correlations amongst design points and incomplete records are given by

$$\Sigma_{\boldsymbol{B}\mathcal{I}|\mathcal{C}} = \mathrm{Cov}[\boldsymbol{B}, O_{\mathcal{I}}|o_{\mathcal{C}}] \approx \mathrm{Cov}[\boldsymbol{B}, \boldsymbol{B}_{\mathcal{I}}|o_{\mathcal{C}}] \, \acute{H}_{\mathcal{I}} \ . \tag{6.44}$$

The Gaussian proxy for the EMF's posterior distribution is again determined through Equations 6.14 and 6.15. Conditional mean and covariance read

$$\mathbb{E}[\boldsymbol{B}|o] = \mathbb{E}[\boldsymbol{B}|o_{\mathcal{C}}] + \Sigma_{\boldsymbol{B}\mathcal{I}|\mathcal{C}} \Sigma^{-1}_{\mathcal{I}|\mathcal{C}} (o_{\mathcal{I}} - \bar{O}_{\mathcal{I}|\mathcal{C}}) \tag{6.45}$$

$$\mathbb{V}[\boldsymbol{B}|o] = \mathbb{V}[\boldsymbol{B}|o_{\mathcal{C}}] - \Sigma_{\boldsymbol{B}\mathcal{I}|\mathcal{C}} \Sigma^{-1}_{\mathcal{I}|\mathcal{C}} \Sigma^{\top}_{\boldsymbol{B}\mathcal{I}|\mathcal{C}} \ . \tag{6.46}$$

The top panel of Figure 6.3 illustrates how the posterior EMF is built within two steps.

Gauss coefficients are estimated analogously. If we store the matrix $\mathrm{Cov}[\boldsymbol{g}_\ell, \boldsymbol{B}_{\mathcal{I}}|\mathcal{C}]$ while performing the first step, then, the linearised cross correlations are given by

$$\Sigma_{\ell\mathcal{I}|\mathcal{C}} = \mathrm{Cov}[\boldsymbol{g}_\ell, O_{\mathcal{I}}|o_{\mathcal{C}}] \approx \mathrm{Cov}[\boldsymbol{g}_\ell, \boldsymbol{B}_{\mathcal{I}}|o_{\mathcal{C}}] \, \acute{H}_{\mathcal{I}|\mathcal{C}} \ . \tag{6.47}$$

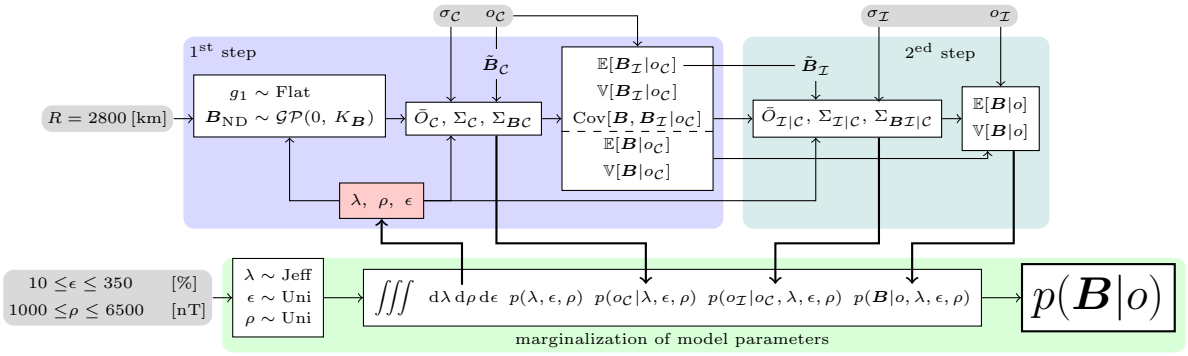

Figure 6.3.: Illustration of how the two-step strategy and a marginalization intertwine to compute the posterior compound distribution. Invariants – such as observations – are shaded in grey. The upper part refers to the update system where model parameters are highlighted in red. Top left-hand panel shows the initial step (Sec. 6.2.1) whereas top right-hand illustrates the update (Sec. 6.2.2). The marginalization is highlighted in green (Sec. 6.3.2). Arrows indicate how information is passed.

Approximations of conditional mean and covariance translate according to Equations 6.18 and 6.19.

Unfortunately, we cannot yet directly apply that algorithm since we do not know specific values for the model parameters $\bar{\boldsymbol{g}}_1$, Σ_1, λ, ρ and ϵ. Before we deal with abandoning these parameters (see Sec. 6.3) we carry out tests with synthetic data to validate the proposed framework.

6.2.3. Synthetic Tests

The current IGRF coefficients are used as a realistic reference field. Synthetic data is generated from the reference field and is corrupted by artificial noise. A gamma distribution is used to corrupt intensity records whereas directional data is randomly drawn from a vMF distribution. To check for robustness, the measurement errors are chosen on purpose not to coincide with the proposed Gaussian proxy error model. For the tests we use several data sets that differ in the error level, the complete/incomplete ratio and the spatial distribution. Since the reference model is known, we use the mean absolute error (MAE)

$$\mathrm{MAE} = \frac{1}{n}\sum_{j}^{n} \left\| \boldsymbol{B}(\boldsymbol{z}_j) - \mathbb{E}[\boldsymbol{B}(\boldsymbol{z}_j)|o] \right\|_1 \tag{6.48}$$

as a test characteristic. Reconstruction and reference field are compared at the Earth's surface, sampled at a rate that accounts for the length scales present in the reference field. Our modelling strategy is able to recover the reference field for all data sets considered.

Furthermore the influence of the POE on the linearization is examined. We compare the proposed strategy to linearization about an axial dipole of $g_1^0 = -23\,\mu\mathrm{T}$ [1]. For all data sets under consideration, the MAE of our modelling strategy falls below linearization about the axial dipole. Especially if the data set mimics reality, the proposed strategy performs better.

[1] An arbitrarily chosen value that roughly fits in terms of magnitude and sign.

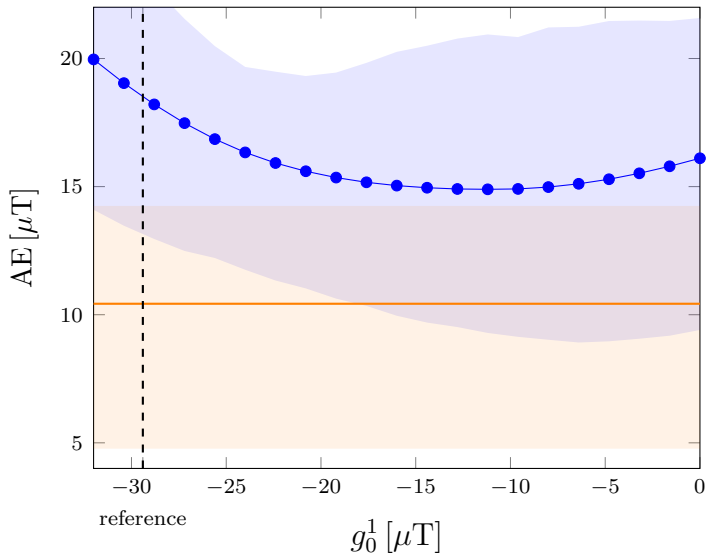

Figure 6.4.: Linearization about an axial dipole compared to the strategy proposed. The horizontal red line indicates the MAE of our approach whereas the blue dots refer to the MAE by varying the axial dipole strength. The shaded areas extend from the 25 to 75 per cent quantile to illustrates the spread of absolute errors. The dashed vertical line refers to the axial dipole of the reference field.

To make sure that $g_1^0 = -23\,\mu\mathrm{T}$ is not a coincidence, we also vary the axial dipole strength. For the synthetic data that mimic reality, Figure 6.4 compares the MAE of the proposed strategy with linearizing about a range of axial dipoles. Even the axial dipole featuring the smallest MAE has a value that is still above the proposed POE built from the complete data. Throughout the synthetic data sets considered, the proposed strategy outperforms the linearization about the *best* suited axial dipole. A detailed description and the full test results are available together with the source code by Schanner and Mauerberger [2019].

6.3. Model Parameters

Before applying the outlined modelling strategy we need to address several model parameters. As the posterior EMF must not depend on a certain choice of model parameters, the a priori mean $\bar{\boldsymbol{g}}_1$ and covariance Σ_1 of the dipole and the scaling factors λ, ρ and ϵ are considered so-called *nuisance* parameters, that is parameters that are not of primary interest. To abandon those quantities from the EMF's posterior distribution we are going to use two different techniques: Section 6.3.1 analytically eliminates $\bar{\boldsymbol{g}}_1$ and Σ_1 by exploring the limit of a flat dipole prior. Section 6.3.2 addresses a hierarchical Bayesian approach to marginalize the scaling factors λ, ρ and ϵ.

6.3.1. Uninformative Dipole

To make as few assumptions as possible we aim to explore the limit of an uninformative prior dipole. This approach is beneficial in two different ways: 1) It prevents us from

accidentally choosing an overly confident prior. 2) We abandon all of the nine dipole-parameters. We expect the data to be strong enough to estimate $\boldsymbol{g}_1$ with confidence.

Loosely speaking, a flat prior may be understood as the limit of a Gaussian with variance sent to infinity. Since an unbounded covariance is not well defined the standard approach is to express formulae w.r.t. the precision and explore the limiting case of vanishing dipole precision that is $\Sigma_1^{-1} \to 0$. In the following, we closely follow Rasmussen and Williams [2006, section 2.7], additionally taking our update system and the linearization into account.

Even though the prior dipole precision is sent to zero, the posterior distribution remains normal. The second step in our update system depends only implicitly on the a priori assumptions made, thus, only the first step needs modifications. This means that the overall modelling concept persists, although we have to modify the conditional mean and covariance (Eqs. 6.14 and 6.15).

First of all, we have to partition the covariance matrix into dipole and non-dipole contributions

$$\Sigma_{\mathcal{C}} = \Sigma_{\mathcal{C},\mathrm{DP}} + \Sigma_{\mathcal{C},\mathrm{ND}} = \acute{H}_{\mathcal{C}} \acute{Y}_{1,\mathcal{C}} \Sigma_1 \acute{Y}_{1,\mathcal{C}}^\top \acute{H}_{\mathcal{C}}^\top + \Sigma_{\mathcal{C},\mathrm{ND}} \tag{6.49}$$

where $\Sigma_{\mathcal{C},\mathrm{ND}}$ is constructed from $K_{\boldsymbol{B},\mathrm{ND}}$ also containing measurement errors and residuals. To keep equations concise, it is beneficial to predict the dipole coefficients first. Making use of the matrix inversion lemma [Rasmussen and Williams, 2006, A.9], conditional mean and inverse of the covariance result in

$$\mathbb{V}[\boldsymbol{g}_1|o_{\mathcal{C}}]^{-1} = \Sigma_1^{-1} + \acute{Y}_{1,\mathcal{C}}^\top \acute{H}_{\mathcal{C}}^\top \Sigma_{\mathcal{C},\mathrm{ND}}^{-1} \acute{H}_{\mathcal{C}} \acute{Y}_{1,\mathcal{C}} \tag{6.50}$$

$$\mathbb{E}[\boldsymbol{g}_1|o_{\mathcal{C}}] = \mathbb{V}[\boldsymbol{g}_1|o_{\mathcal{C}}]^{-1} \left(\Sigma_1^{-1} \bar{\boldsymbol{g}}_1 - \acute{Y}_{1,\mathcal{C}}^\top \acute{H}_{\mathcal{C}}^\top \Sigma_{\mathcal{C},\mathrm{ND}}^{-1} \acute{H}_{\mathcal{C}} \tilde{B}_{\mathcal{C}} \right) \tag{6.51}$$

and the data $o_{\mathcal{C}}$ enters through the POE (see Eq. 6.36). Considering the limit of the uninformative dipole yields

$$\Sigma_{1|\mathcal{C}}^{-1} := \lim_{\Sigma_1^{-1} \to 0} \mathbb{V}[\boldsymbol{g}_1|o_{\mathcal{C}}]^{-1} = \acute{Y}_{1,\mathcal{C}}^\top \acute{H}_{\mathcal{C}}^\top \Sigma_{\mathcal{C},\mathrm{ND}}^{-1} \acute{H}_{\mathcal{C}} \acute{Y}_{1,\mathcal{C}} \tag{6.52}$$

$$\bar{g}_{1|\mathcal{C}} := \lim_{\Sigma_1^{-1} \to 0} \mathbb{E}[\boldsymbol{g}_1|o_{\mathcal{C}}] \quad = -\Sigma_{1|\mathcal{C}}^{-1} \acute{Y}_{1,\mathcal{C}}^\top \acute{H}_{\mathcal{C}}^\top \Sigma_{\mathcal{C},\mathrm{ND}}^{-1} \acute{H}_{\mathcal{C}} \tilde{B}_{\mathcal{C}} \ . \tag{6.53}$$

Interestingly, $\boldsymbol{g}_1|o_{\mathcal{C}}$ does not depend on $\bar{\boldsymbol{g}}_1$, rendering the dipole's prior mean irrelevant.

Predicting Gauss coefficients of higher SH degree is straightforward since a priori dipole and non-dipole contributions are assumed independent. For $\ell > 1$ cross correlations are of no concern as they do not depend on $\boldsymbol{g}_1$. Analogous to Equation 6.39, conditional mean and covariance result in

$$\mathbb{E}[\boldsymbol{g}_{2:\ell}|o_{\mathcal{C}}] = \quad - \Sigma_{2:\ell} \acute{Y}_{2:l,\mathcal{C}}^\top \acute{H}_{\mathcal{C}}^\top \Omega_{\mathcal{C}} \acute{H}_{\mathcal{C}} \tilde{B}_{\mathcal{C}} \tag{6.54}$$

$$\mathbb{V}[\boldsymbol{g}_{2:\ell}|o_{\mathcal{C}}] = \Sigma_{2:\ell} - \Sigma_{2:\ell} \acute{Y}_{2:l,\mathcal{C}}^\top \acute{H}_{\mathcal{C}}^\top \Omega_{\mathcal{C}} \acute{H}_{\mathcal{C}} \acute{Y}_{2:l,\mathcal{C}} \Sigma_{2:\ell} \tag{6.55}$$

where $\Omega_{\mathcal{C}}$ refers to the limiting precision matrix

$$\Omega_{\mathcal{C}} = \Sigma_{\mathcal{C},\mathrm{ND}}^{-1} - \Sigma_{\mathcal{C},\mathrm{ND}}^{-1} \acute{H}_{\mathcal{C}} \acute{Y}_{1,\mathcal{C}} \Sigma_{1|\mathcal{C}} \acute{Y}_{1,\mathcal{C}}^\top \acute{H}_{\mathcal{C}}^\top \Sigma_{\mathcal{C},\mathrm{ND}}^{-1} \tag{6.56}$$

and we again made use of the matrix inversion lemma. To perform the second step, the whole posterior covariance matrix is necessary. However, the posterior cross covariance

amongst dipole and non-dipole coefficients is missing. The difficulty is to find the limiting case. According to Equation 6.19 we have

$$\mathrm{Cov}[\boldsymbol{g}_1, \boldsymbol{g}_{2:\ell}|o_\mathcal{C}] = -\mathrm{Cov}[\boldsymbol{g}_1, O_\mathcal{C}]\, \Sigma_\mathcal{C}^{-1} \mathrm{Cov}[O_C, \boldsymbol{g}_{2:\ell}] \tag{6.57}$$

since a priori dipole and non-dipole are assumed independent. Plugging in the linearization we end up with

$$\mathrm{Cov}[\boldsymbol{g}_1, \boldsymbol{g}_{2:\ell}|o_\mathcal{C}] = -\Sigma_{1|\mathcal{C}} \acute{Y}_{1,\mathcal{C}} \acute{H}_\mathcal{C} \Sigma_{\mathcal{C},\mathrm{ND}}^{-1} \acute{H}_\mathcal{C}^\top \acute{Y}_{2:l,\mathcal{C}}^\top \Sigma_{2:\ell} \,, \tag{6.58}$$

which no longer depends on Σ_1. To see this, use the matrix inversion lemma to express the precision matrix, factor in the left hand side and expand by $\Sigma_{1|\mathcal{C}}^{-1}\Sigma_{1|\mathcal{C}}$. Using Equation 6.50 and re-arranging terms yields Equation 6.58.

To predict on the EMF, we divide the cross covariance into dipole and non-dipole contributions. Since a priori the dipole term is assumed independent of the non-dipole contribution we have

$$\Sigma_{\boldsymbol{B}\mathcal{C}} = \Sigma_{\boldsymbol{B}\mathcal{C},\mathrm{DP}} + \Sigma_{\boldsymbol{B}\mathcal{C},\mathrm{ND}} = \acute{Y}_1^\top \Sigma_1 \acute{Y}_{1,\mathcal{C}} \acute{H}_\mathcal{C} + \Sigma_{\boldsymbol{B}\mathcal{C},\mathrm{ND}} \tag{6.59}$$

and $\Sigma_{\boldsymbol{B}\mathcal{C},\mathrm{ND}}$ is constructed from $K_{\boldsymbol{B},\mathrm{ND}}$. Using the same strategy as we did with the Gauss coefficients, the posterior mean and covariance arise to

$$\mathbb{E}[\boldsymbol{B}|o_\mathcal{C}] = -\acute{Y}_1^\top \bar{g}_{1|\mathcal{C}} + \Sigma_{\boldsymbol{B}\mathcal{C},\mathrm{ND}} \Sigma_{\mathcal{C},\mathrm{ND}}^{-1} \big(\tilde{B}_\mathcal{C} - \acute{Y}_{1,\mathcal{C}}^\top \bar{g}_{1|\mathcal{C}}\big) \tag{6.60}$$

$$\mathbb{V}[\boldsymbol{B}|o_\mathcal{C}] = \acute{Y}_1^\top \Sigma_{1|\mathcal{C}} \acute{Y}_1 + K_{\boldsymbol{B},\mathrm{ND}} - \Sigma_{\boldsymbol{B}\mathcal{C},\mathrm{ND}} \Omega_\mathcal{C} \Sigma_{\boldsymbol{B}\mathcal{C},\mathrm{ND}}^\top \,. \tag{6.61}$$

From those equations we can obtain all quantities needed to proceed with the second step incorporating incomplete records.

The importance in the result is that conditional mean and covariance no longer depend on the choice of the a priori dipole, since we assumed zero precision.

6.3.2. Compound Distribution

Although we are not particularly interested in reconstructing the probability distribution of $\vartheta = (\lambda, \rho, \epsilon)$, their variabilities must be taken into account. The final result of the proceeding is the EMF's compound distribution

$$p(\boldsymbol{B}|o) = \int p(\boldsymbol{B}|o, \vartheta)\, p(\vartheta|o)\, \mathrm{d}\vartheta \tag{6.62}$$

which results from marginalizing the scaling factors. Figure 6.3 illustrates the interaction of the two-step strategy and the compound distribution, with the marginalization being depicted in the bottom part. This approach makes it possible to escape from Gaussianity. The integral will not be tractable analytically and must be evaluated by numerical methods.

Let us inspect the two PDFs we want to integrate over. For a certain choice of parameters, $p(\boldsymbol{B}|o, \vartheta)$ is calculated according to our two step strategy. To keep track of the scaling factor's posterior $p(\vartheta|o)$ we add a hierarchical stage. Applying Bayes' law, the posterior density is given by

$$p(\vartheta|o) \propto p(o|\vartheta) p(\vartheta) \tag{6.63}$$

where we neglected the normalizing constant. To suitably normalize we have to carry out another quadrature since $p(o) = \int p(o|\vartheta)p(\vartheta)\,\mathrm{d}\vartheta$ is unknown. Because of the flat dipole prior, calculating $p(o|\vartheta)$ needs special attention. Not to distract from our endeavour, this is the focus of Section 6.3.3.

A priori, all three parameters are assumed statistically independent. Both, residual and error level are considered uniformly distributed. As we are roughly aware of magnitudes, the chosen range is well-spaced (weakly informative). On the contrary, λ is a scale parameter bearing across orders of magnitude. The corresponding uninformative prior – representing the state of no prior information – is Jeffrey's prior

$$p(\lambda) \propto \frac{1}{\lambda}\,[\mathrm{nT}] \tag{6.64}$$

that is, values ten times larger are just as likely as values ten times smaller [Murphy, 2012, section 5.4.2]. However, there is a subtlety arising for the compound prior PDFs. The hierarchical approach has an impact because the a priori $\boldsymbol{B}$-field depends on λ. The scale invariance is passed on rendering the compound prior distributions improper, that is the density can not be normalized. As an example, the compound PDF for the a priori non-dipole Gauss coefficients reads

$$p(g_\ell^m) = \frac{1}{\sqrt{2\pi}} \int_0^\infty \frac{1}{\lambda^2} \exp\left\{-\frac{1}{2}\frac{(g_\ell^m)^2}{\lambda^2}\right\} \mathrm{d}\lambda \propto \frac{1}{|g_\ell^m|}\,. \tag{6.65}$$

That density remains centred at zero whereas the variance does not exist. Nevertheless, the compound posterior PDFs are normalizable and well defined. Marginalizing λ results in an extremely weak a priori assumption.

Although the compound distribution is not normal, calculating higher-order moments is one of its potentials. Posterior mean and covariance are given by

$$\mathbb{E}[\boldsymbol{B}|o] = \int p(\vartheta|o)\ \mathbb{E}[\boldsymbol{B}|o,\vartheta]\ \mathrm{d}\vartheta \tag{6.66}$$

$$\mathbb{V}[\boldsymbol{B}|o] = \int p(\vartheta|o)\big(\mathbb{E}[\boldsymbol{B}|o,\vartheta]^2 + \mathbb{V}[\boldsymbol{B}|o,\vartheta]\,\big)\,\mathrm{d}\vartheta - \mathbb{E}[\boldsymbol{B}|o]^2 \tag{6.67}$$

and the mean and covariance we are integrating over are explicitly given. For modelling Gauss coefficients the above formulae translate analogously.

An actual implementation requires to perform three numeric integrations, in total. Since numeric quadrature in three dimensions is feasible, both the fully Bayesian posterior density and also the Gaussian moment matching proxy are right at hand.

6.3.3. Marginal Likelihood

We postponed the calculation of the *marginal likelihood* until here. The terms *marginal* and *likelihood* refer to the marginalization over the EMF as a function in ϑ. To actually discretize and integrate the compound distribution, $p(o|\vartheta)$ is still missing. We again separate into complete and incomplete records. The marginal likelihood factorizes

$$p(o|\vartheta) = p(o_\mathcal{I}|o_\mathcal{C},\vartheta)p(o_\mathcal{C}|\vartheta) \tag{6.68}$$

where we used the conditional probability rule. The major benefit is that we already know the quantities needed to evaluate the PDFs on the right hand side of Eq. 6.68 through our two step strategy. Figure 6.3 illustrates where Eq. 6.68 enters the modelling scheme and how quantities are passed between marginalization and two-step strategy.

For the sake of simplicity we analyse incomplete records, first. For a certain choice of parameters, $\boldsymbol{B}|o_\mathcal{C}, \vartheta$ and $O_\mathcal{I}|\vartheta$ are jointly normal due to linearization and error approximation. By taking advantage of the joint normality, we can directly observe the prior predictive distribution of the incomplete records. The PDF follows to read

$$p(o_\mathcal{I}|o_\mathcal{C}, \vartheta) = \frac{\exp\left\{-\frac{1}{2}(o_\mathcal{I} - \bar{O}_{\mathcal{I}|\mathcal{C}})^\top \Sigma_{\mathcal{I}|\mathcal{C}}^{-1}(o_\mathcal{I} - \bar{O}_{\mathcal{I}|\mathcal{C}})\right\}}{\sqrt{(2\pi)^{n_\mathcal{I}}|\Sigma_{\mathcal{I}|\mathcal{C}}|}} \tag{6.69}$$

and the mean and covariance are given by Eqs. 6.41 and 6.43, implicitly depending on λ, ρ and ϵ through the first step in our update system. Regarded as a function in ϑ, the marginal likelihood is certainly not normal.

Because of the flat dipole prior, the marginal likelihood for complete records needs special attention. We proceed analogously to the incomplete data, however, having to bear the limiting case

$$p(o_\mathcal{C}|\vartheta) = \lim_{\Sigma_1^{-1}\to 0} \frac{\exp\left\{-\frac{1}{2}(o_\mathcal{C} - \bar{O}_\mathcal{C})^\top \Sigma_\mathcal{C}^{-1}(o_\mathcal{C} - \bar{O}_\mathcal{C})\right\}}{\sqrt{(2\pi)^{n_\mathcal{C}}|\Sigma_\mathcal{C}|}} \,. \tag{6.70}$$

We closely follow Rasmussen and Williams [2006, section 2.7] and again split into dipole and non-dipole contributions. As already mentioned in Section 6.3.1, in the limit $\bar{\boldsymbol{g}}_1$ is irrelevant and we set $\bar{O}_\mathcal{C} = o_\mathcal{C} - \acute{H}_\mathcal{C}\tilde{B}_\mathcal{C}$ (see Eq. 6.36). The limiting precision matrix is of no concern and given by Equation 6.56. The big concern, however, is the determinant as the dipole variance tends to infinity. According to Rasmussen and Williams [2006, Eq. 2.45], the marginal likelihood for the complete records results in

$$p(o_\mathcal{C}|\vartheta) = \frac{\exp\left\{-\frac{1}{2}\tilde{B}_\mathcal{C}^\top \acute{H}_\mathcal{C}^\top \Omega_\mathcal{C} \acute{H}_\mathcal{C}\tilde{B}_\mathcal{C}\right\}}{\sqrt{(2\pi)^{n_\mathcal{C}-3}|\Sigma_{\mathcal{C},\mathrm{ND}}|\,|\Sigma_{1|\mathcal{C}}^{-1}|}} \tag{6.71}$$

and we already computed all relevant quantities in the first step of our update system. Although not explicitly indicated, the non-dipole covariance, the limiting precision and the dipole covariance depend on ϑ.

6.4. Application

As a practical proof of concept we apply the suggested method to the archeomagnetic and volcanic data offered by the GEOMAGIA50.v3 database [Brown et al., 2015]. We used all directional and intensity records between 753 AD and 1950 AD that were included in GEOMAGIA50 version 3.3 in November 2019. To simplify matters, the Earth is assumed a sphere of radius $R_E = 6\,371.2$ km and coordinates are treated as if they were spherical. We estimate the committed error to be less than $\frac{1}{2}\,\mu$T which is small compared to modelling uncertainties.

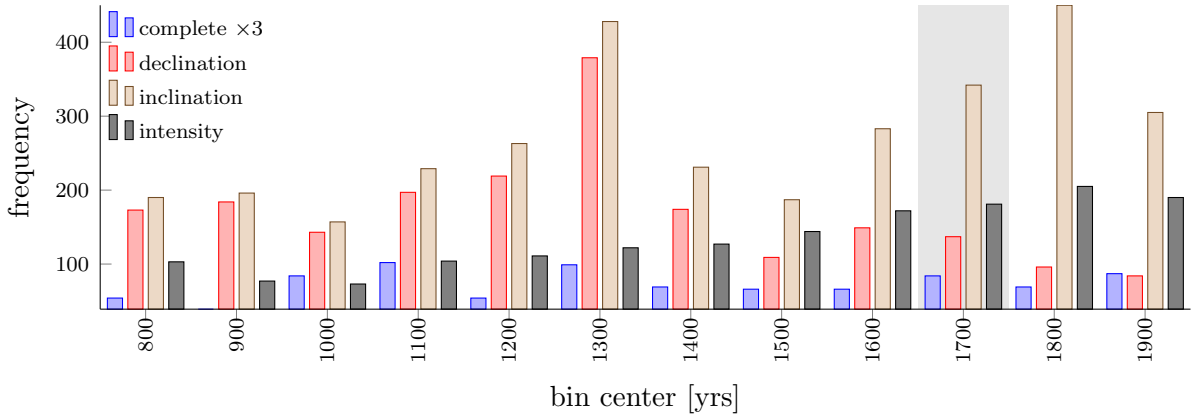

Figure 6.5.: Number of data per 100 yr bin (D, I and F counts are without complete records). For comparability each complete record is counted as three data points. The bin that we use as example in the discussion is shaded in grey. The horizontal line refers to the criterion of how the window length is chosen.

We use the individual, originally reported error estimates. If uncertainties are not available (ca. 8.4 per cent of the data), we assign $\alpha_{95} = 4.5°$ as directional errors and $\sigma_F = 8.25\,\mu\mathrm{T}$ as intensity errors [Licht et al., 2013, section. 2.2]. Single unpaired declination records are not used.

As our model does not yet account for time dependence we group the data into disjoint bins of 100 yr. The decisive factor for the window length is the number of complete records per bin. Our two step strategy and the linearization are the basic rationale behind this choice. The number of complete records has to determine a reasonable POE for incorporating the incomplete records. Let us assume that at the surface the EMF is dipole dominated (about 90%) for the timespan under consideration. As a rule of thumb, a Taylor expansion performs reasonably well when deviating less than 10% from the *truth*. Having a parametric view back in mind, the field's dominating features may be described by only nine parameters ($\boldsymbol{g}_1$ and Σ_1). With a minimum of ten complete records per bin we anticipate a coarse field estimate suited as POE. This consideration leads to a window length of 100 yr. Figure 6.5 shows the temporal data distribution.

To demonstrate the potential of our modelling strategy in recovering the stationary field we use the interval $[1650, 1750]$ as an example. In total, this bin summarizes 744 observations acquired at 480 sites. Figure 6.6 depicts the highly irregular data coverage which is dominated by the northern hemisphere.

We are going to compare our findings to three previous, continuous magnetic field models. The models ARCH10k.1 [Constable et al., 2016], arhimag1k [Senftleben, 2019] and COV-ARCH [Hellio and Gillet, 2018] are considered eligible competitors as they stem from a similar data basis, although arhimag1k additionally includes direct historical observations. All three models report Gauss coefficients up to SH degree $\ell = 10$. ARCH10k.1 and arhimag1k are continuous in time, but do not report modelling errors. In contrast, COV-ARCH provides an ensemble of 50 realizations but uses the same coarse time steps as we do. To compare with COV-ARCH's ensemble we calculate sample mean and sample variance. Histograms are computed according to *Scott's rule of thumb* [Scott, 1979].

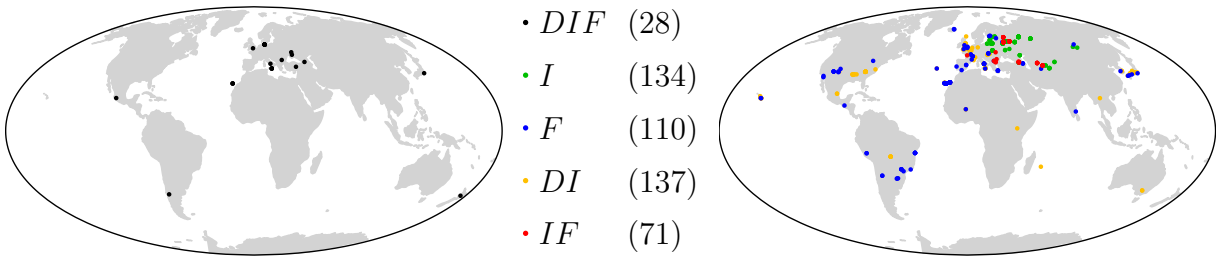

Figure 6.6.: Illustration of the very irregular data coverage for the 1700 epoch. The left-hand panel shows complete records whereas the right-hand panel illustrates combinations of D, I and F measurements.

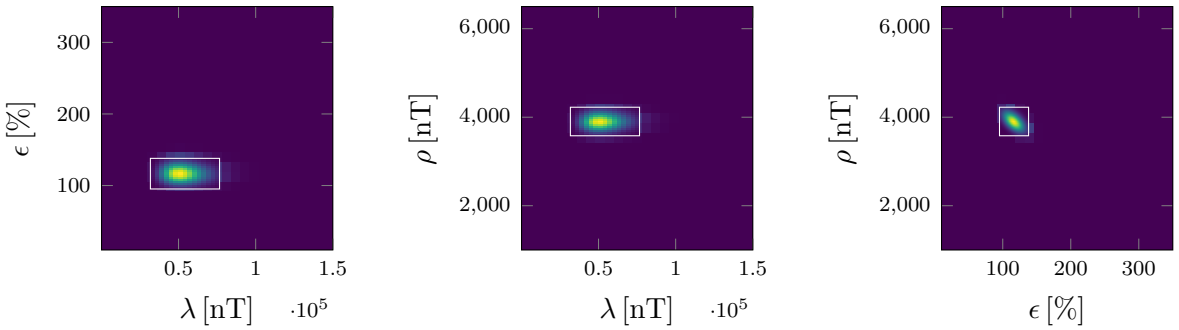

Figure 6.7.: Discretisation of the parameter space for the epoch 1700. The accurately sampled marginal densities for the integration phase (white framed box) are drawn atop of the coarse exploratory grid.

6.4.1. Numeric Integration

To evaluate the compound distribution $p(\boldsymbol{B}|o)$ we discretize the integral in Equation 6.62. The posterior PDF is approximated by a simple Riemann sum

$$p(\boldsymbol{B}|o) \approx \sum_i p(\boldsymbol{B}|o, \vartheta_i)\, p(\vartheta_i|o) \Delta\vartheta_i \ . \tag{6.72}$$

Although not highly accurate, we favour using a Riemann sum because of its simplicity. Since by construction the mixture components $p(\boldsymbol{B}|o, \vartheta_i)$ are normally distributed, the r.h.s. of Equation 6.72 gives a (finite) Gaussian mixture distribution [Murphy, 2012, section 11.2.1], for which moments are readily available by translating Equations 6.66 and 6.67 accordingly.

To actually discretize the parameter space, we use a regular and equally spaced grid. Thus we have the posterior

$$p(\boldsymbol{B}|o) \approx \sum_{i,j,k=0}^{N} p(\boldsymbol{B}|o, \lambda_i, \epsilon_j, \rho_k)\, p(\lambda_i, \epsilon_j, \rho_k|o)\, \Delta\lambda\Delta\epsilon\Delta\rho \ , \tag{6.73}$$

where N^3 is the number of gridpoints for the Riemann sum, $\Delta\lambda$ is the step width, λ_0 and λ_N specify the bounds of the interval we integrate over and $\lambda_i = \lambda_0 + i\Delta\lambda$ (Similar for ϵ and ρ). As mentioned above, we carry out a second quadrature to normalize the

parameter posterior $p(\lambda_i, \epsilon_j, \rho_k|o)$. This also serves the purpose of having a well defined Gaussian mixture density. We normalize it w.r.t. the Riemann sum, that is such that

$$\sum_{i,j,k=0}^{N} p(\lambda_i, \epsilon_j, \rho_k|o)\Delta\lambda\Delta\epsilon\Delta\rho = 1 \ . \tag{6.74}$$

Under these considerations Equation 6.73 can be seen as the discrete version of Equation 6.62, where all terms are replaced by their respective discrete equivalent.

We consider the approximation (Eq. 6.72) reasonable, since we found the probability mass of the parameter posterior $p(\vartheta|o)$ to be unimodal and localized in a finite region. Thus to calculate the Gaussian mixture proxy (Eq. 6.73) for the compound distribution, we perform two steps:

Exploration of the parameter space We need to identify the region in which the parameter posterior is localized. Hence we span a coarse grid over all values we believe to be (physically) reasonable and calculate the posterior $p(\vartheta|o)$ on this grid. We choose $N = 25$ gridpoints along each of the three dimensions and choose the bounds as

$$\begin{aligned} \lambda_0 &= 100 \text{ nT} & \lambda_N &= 150000 \text{ nT} \\ \epsilon_0 &= 10 \ \% & \epsilon_N &= 350 \ \% \\ \rho_0 &= 1000 \text{ nT} & \rho_N &= 6500 \text{ nT} \ . \end{aligned} \tag{6.75}$$

Although in principle a wider extent of the grid may be "physically reasonable", by trial and error we found these bounds to be sufficient (i.e. outside of the bounds the probability mass was approximately zero for all considered cases). Finally, from the posterior we calculate the marginal distribution for each parameter λ, ϵ, ρ via another Riemann sum. We then calculate each (empirical) mean and standard deviation from these coarse marginals. The region where the probability mass is located is then covered by the cuboid centred at the empirical mean with edge-lengths given by two empirical standard deviations.

Calculation The actual numerical integration only takes place within the cuboid that is derived in the exploration step. Inside that cuboid a refined grid is spanned with $N = 15$ nodes per dimension. By calculating not only the posterior on this new grid, but running for each gridpoint an inversion for the EMF, the Gauss coefficients and other quantities of interest, we can calculate a proxy for the full compound posterior for each of these quantities, using Equation 6.73. This proxy, which is a Gaussian mixture distribution, is the final result of our modelling strategy.

For all epochs the parameter posterior is of good-nature only featuring a single mode. For epoch 1700, Figure 6.7 depicts the posterior parameter PDF. Presented are all combinations of 2-D marginal parameter posterior at the coarse and refined grid. The white rectangle refers to the edges of the refined grid that is used for numeric integration. The maximum of the error level ϵ is slightly above 100% at a rather small spread. This means that measurement errors are considered a little larger than described in the dataset. Taking the approximations into account, a shift toward higher values seems reasonable. The residual term ρ has its maximum at about $4\,\mu$T and is also relatively sharp. It is interesting to note that the residual is of the same order of magnitude as the equatorial

dipole. This need not only be due to unexplained sources, it can also be an effect of the linearization or not taking time dependencies into account. The scaling parameter λ features the widest distribution. In comparison with the IGRF, the magnitude and range appear reasonable. However, the posterior PDF of the model parameters is not sharp enough to justify a point estimate. Therefore we integrate out model parameters and thus the uncertainties from the parameter posteriors are translated into the posterior of the quantities of interest. This way the posterior variance, which is easily available due to the Gaussian mixture structure, does not only reflect uncertainties arising from the data, but also reproduces the model uncertainties.

6.4.2. Vector field predictions

The discretized versions of $\boldsymbol{B}$'s posterior mean and variance translate according to Equations 6.66 and 6.67 and are given by

$$\mathbb{E}[\boldsymbol{B}|o] = \sum p(\vartheta_i|o)\mathbb{E}[\boldsymbol{B}|o,\vartheta_i]\,\Delta\vartheta \tag{6.76}$$

$$\mathbb{V}[\boldsymbol{B}|o] = \sum p(\vartheta_i|o)\left(\mathbb{E}[\boldsymbol{B}|o,\vartheta_i]^2 + \mathbb{V}[\boldsymbol{B}|o,\vartheta_i]\right)\Delta\vartheta - \mathbb{E}[\boldsymbol{B}|o]^2\ . \tag{6.77}$$

The pointwise posterior standard deviation serves for realistic location dependent uncertainty estimates. Again, $\boldsymbol{B}$-field predictions are non-parametric and do not depend on Gauss coefficients.

The top row of Figure 6.8 depicts both posterior mean and standard deviation of the down component at the Earth's surface. The field is evaluated at 2000 design points, equally distributed over the sphere [Deserno, 2004]. We are able to quantify what previous studies are suggesting: The EMF is reconstructed with confidence within areas of dense data coverage, for example in Europe. Structures of large parts of the southern hemisphere, however, remain vague. We see quite similar patterns across all epochs under investigation.

6.4.3. Declination, Inclination and Intensity

Because the EMF shares a non-linear relation with declination, inclination and intensity we do not see the possibility of analytically deriving properties such as the posterior mean for D, I and F. For a moderate number of design points we can obtain the distribution by sampling strategies. However, trying to draw from a high dimensional normal mixture distribution appears absurd. For the 2000 design points we are interested in, storing all the covariance matrices of the entire parameter grid is hardly possible, let alone drawing samples from the resulting mixture.

The idea behind uncovering the posterior mean and variance of D, I and F is once more a linearization. To do so the same strategy as pointed out in Section 6.1.3 is used. The linearization of D, I and F are given by Equations 6.24, 6.25 and 6.26. We build upon the EMF estimate from the previous section and the posterior mean $\mathbb{E}[\boldsymbol{B}|\boldsymbol{o}]$ serves as POE. Because of the linearization approximations for mean and (co)variance are explicit.

To provide an example we consider the intensity, only. Utilizing Equation 6.26, the

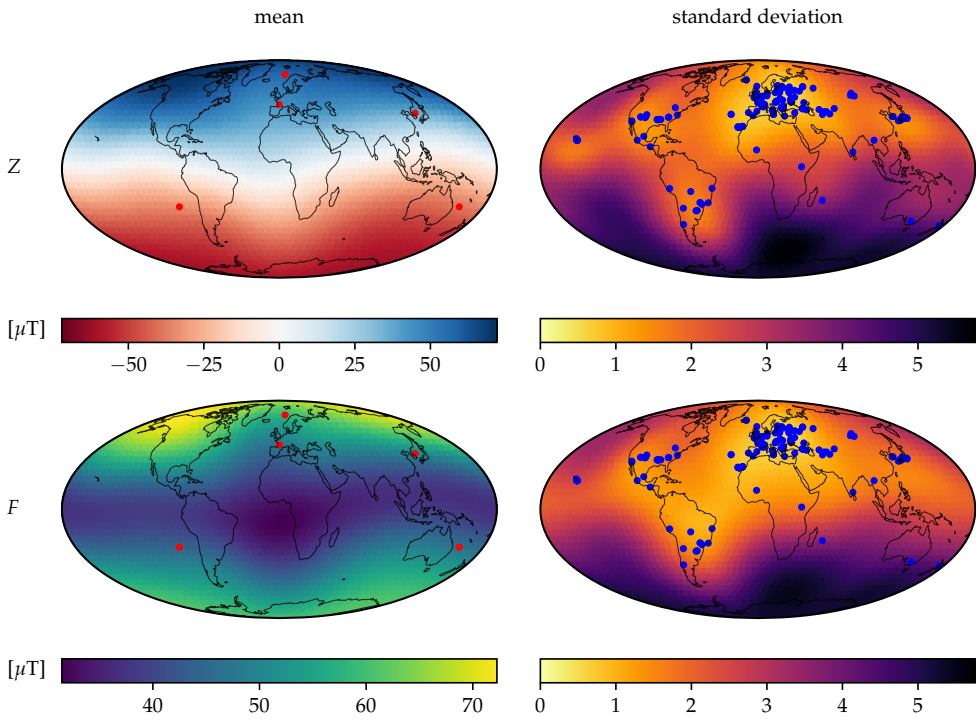

Figure 6.8.: Posterior mean (left-hand panel) and standard deviation (right-hand panel) of the EMF at the Earth's surface for the 1700 epoch. The top row shows the down component B_Z. The locations at which we perform normality tests are indicated in red (see Tab. 6.1). The bottom row shows the field intensity F. Blue dots refer to the underlying data sites.

approximate posterior mean and variance are given by

$$\mathbb{E}[F|o] \approx \| \mathbb{E}[\boldsymbol{B}|\boldsymbol{o}] \| \tag{6.78}$$

$$\mathbb{V}[F|o] \approx \frac{\mathbb{E}[\boldsymbol{B}|\boldsymbol{o}]^\top \mathbb{V}[\boldsymbol{B}|\boldsymbol{o}] \, \mathbb{E}[\boldsymbol{B}|\boldsymbol{o}]}{\| \mathbb{E}[\boldsymbol{B}|\boldsymbol{o}] \|^2} \, . \tag{6.79}$$

We pursue a non-parametric approach but the concept is similar to that of Hellio et al. [2014, Appendix A]. For the 1700 epoch, the bottom row of Figure 6.8 depicts mean and standard deviation of the intensity at the Earth's surface. Although not directly observed by a single record, the evolving area of weak field known as the South Atlantic Anomaly [e.g., Mandea et al., 2007, Hartmann and Pacca, 2009] is certainly significant within one standard deviation. Looking at the other epochs, the westward drift is also visible. This feature is constrained here by only a few points around the actual anomaly, which shows that our modelling approach is capable of uncovering features that are known from models with stronger data basis. This ability stems from long ranging spatial correlations of the kernel and is controlled by the reference radius R (see Figure 6.1).

When comparing epoch 800 with figure 11(a) in Hellio and Gillet [2018] two things stand out. The magnitudes are similar but a low intensity patch is found in the Pacific rather than the Atlantic. In the standard deviation we also see a rather sharp transition between north and south. Since this is a proof of concept a detailed comparison or even an interpretation would be premature.

Since our posterior distributions are certainly not normal it is questionable if the standard deviation is adequate to describe uncertainties. If the posterior is highly skewed or

	Lat	-26°	45°	71°	39°	-26°
	Lon	-88°	0°	9°	131°	159°
D	Δ_{16}	2.5	0.7	0.3	-0.2	1.5
	Δ_{84}	-0.3	2.9	3.9	1.8	0.4
I	Δ_{16}	3.6	-0.3	-0.4	-1.5	3.8
	Δ_{84}	-3.0	1.0	3.0	6.7	-2.3
F	Δ_{16}	4.6	-2.4	0.7	1.5	7.1
	Δ_{84}	-1.7	0.5	-0.1	-0.3	0.6
Z	Δ_{16}	2.9	-2.6	-1.1	-0.4	1.2
	Δ_{84}	1.6	0.1	0.6	0.6	1.9

Table 6.1.: Estimation of the quality of the proxy standard deviation as a measure of error, compared to the 16-/84-percentiles. If Δ is larger (smaller) than 0, the standard deviation overestimates (underestimates) the error. Values are given in percent of one standard deviation.

even multimodal, the standard deviation would not be well suited to quantify modelling errors. One would preferably use percentiles which, however, cannot be derived analytically. As long as the number of design points is moderate, one brute-force method is to calculate percentiles by sampling. This works well for percentiles that are not far from the bulk of the probability mass.

A simplistic approach to obtain samples from a mixture distribution is the following algorithm. First, by chance the k^{th} random variable is selected from the mixing distribution $p(\vartheta_k|o)$ (categorical). Then the value of the selected random variable $p(\boldsymbol{B}|o,\vartheta_k)$ is realized (multivariate normal). Repeat until the desired amount of samples is achieved.

Although we have to radically reduce the number of design points we can gain insight into how strong the proxy standard deviation and percentiles deviate. We compare 16- and 84-percentiles to mean ± standard deviation via

$$\Delta_{16} = 1 - \frac{\mu - \alpha_{16}}{\sigma} \qquad \text{and} \qquad \Delta_{84} = 1 - \frac{\alpha_{84} - \mu}{\sigma} \tag{6.80}$$

where α refers to percentiles, μ and σ indicate mean and standard deviation. The interpretation of Δ is as follows: If Δ is larger (smaller) than zero, the standard deviation overestimates (underestimates) the uncertainty given by the percentile, by $|\Delta|$ standard deviations. Note, that to keep this interpretation consistent for both Δ_{16} and Δ_{84} the sign of the second term in (6.80) changes.

At 5 randomly selected locations we apply this check to the down component B_Z and to the three commonly used archeomagnetic observables D, I and F. As can be seen in Table 6.1, at all locations the deviation is well below 10%. Thus we believe the proxy standard deviation is qualified to describe uncertainties. Furthermore, the proxy standard deviation obtained by linearization is computationally feasible and easy to visualize and interpret.

6.4.4. Predictions at the Core-Mantle Boundary

Until here predictions were carried out at the Earth's surface. It is straightforward to predict the EMF at arbitrarily chosen design points outside of the reference sphere of radius R. Figure 6.9 presents mean and standard deviation of the down component B_Z at the CMB. We again use 2000 equidistributed design points but at radius $R_{\text{CMB}} =$

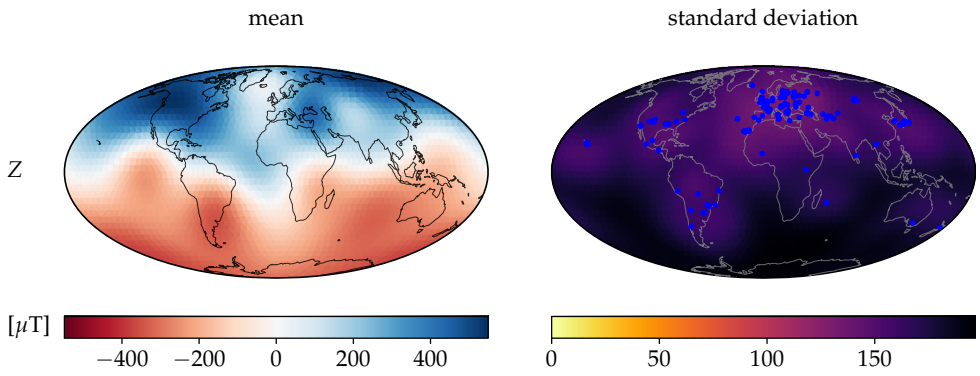

Figure 6.9.: Posterior mean (left-hand panel) and standard deviation (right-hand panel) of the EMF down component B_Z at the CMB for the 1700 epoch. Blue dots refer to the sites where records are acquired.

3480 km. One observes that uncertainties are greater than at the Earth's surface. Roughly speaking, the standard deviation is about 40% compared to the mean, while at the Earth's surface relative errors only amount up to 10% (see right-hand column of Figure 6.8). The modelling error strongly depends on how far design points are from the reference sphere.

This behaviour is best understood thinking in terms of Gauss coefficients. In Equation 6.2 the term that is raised to the power of $\ell + 1$ causes this effect. Since design points lie outside the reference sphere, the ratio $\frac{R}{|x|}$ is smaller than one. Thus, Gauss coefficients are more *penalized* the larger the SH degree, and the ratio $\frac{R}{|x|}$ determines the rate of decline. In turn, the closer to the reference radius the more impact higher SH degrees have. Related to smaller structures, the higher the SH degree the less certain we are. When scaling down – for example to the CMB – higher SH degrees increase and so do the respective uncertainties. This idea translates to our non-parametric model.

Compared to what Hellio and Gillet [2018, figure 8c] found, our reconstruction of epoch 1400 looks very similar. Because of the large variabilities, a detailed comparison does not make much sense.

6.4.5. Gauss Coefficients

We want to stress again that our model is inherently non-parametric. The fundamental quantity of the inference is the geomagnetic potential. Nevertheless, our approach allows to infer Gauss coefficients since geomagnetic potential and Gauss coefficients are linearly related (see Eq. 6.16). The procedure is similar to inferring the EMF. The discretized versions of the compound mean (Eq. 6.66) and (co)variance (Eq. 6.67) for the Gauss coefficients g_ℓ^m read

$$\mathbb{E}[g_\ell^m|o] = \sum p(\vartheta_i|o)\mathbb{E}[g_\ell^m|o,\vartheta_i]\,\Delta\vartheta \tag{6.81}$$

$$\mathbb{V}[g_\ell^m|o] = \sum p(\vartheta_i|o)\left(\mathbb{E}[g_\ell^m|o,\vartheta_i]^2 + \mathbb{V}[g_\ell^m|o,\vartheta_i]\right)\Delta\vartheta - \mathbb{E}[g_\ell^m|o]^2 \quad . \tag{6.82}$$

In principle we can predict up to arbitrary SH degree. However, storing the component's mean and covariances for the whole parameter grid becomes memory intensive. If we restrict our selves to a moderate SH degree – for example $\ell \leq 10$ – we are able to calculate the full mixture PDF, sample from the posterior and calculate percentiles.

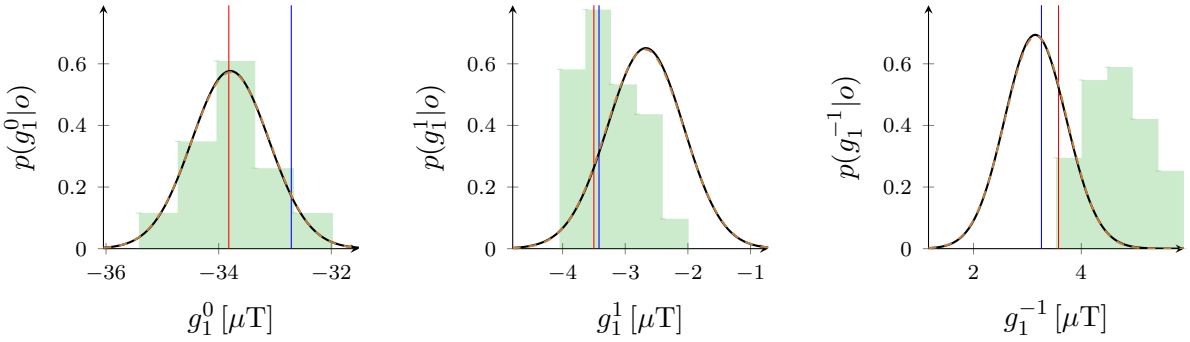

Figure 6.10.: Distribution of the dipole coefficients at the Earth's surface for epoch 1700. The resulting mixture PDF is indicated by the solid black lines whereas the, nearly identical, dashed lines refers to the moment matching Gaussian proxies. The green histograms illustrates the COV-ARCH ensemble coefficients. The blue vertical lines refer to arhimag1k and the red ones to ARCH10k.1 .

Although not explicitly indicated, Gauss coefficients depend on the reference radius R. Given the g_ℓ^m at the reference radius, we can scale them to a radius $\tilde{R} > R$ by

$$\tilde{g}_\ell^m = g_\ell^m \left(\frac{R}{\tilde{R}}\right)^{\ell+2} . \tag{6.83}$$

Mean and covariances translate accordingly.

For comparison, we scale Gauss coefficients w.r.t. the Earth's surface. The mixture PDFs for the dipole coefficients are shown in Figure 6.10. As can be seen, the mixture is quite close to the moment matching normal proxy. Across all epochs the dipole coefficients are close to being normally distributed, except for the bin $[850, 950]$, for which the mixture distribution deviates slightly from the normal proxy. However, the higher the SH degree the less normal the posterior is and the more the scale invariant prior dominates (see Eq. 6.65). We further see that for the epoch under consideration, the histogram built from the COV-ARCH ensemble shows a considerable overlap with the mixture's PDF and the mean aligns with the ARCH10k.1 estimate. It is not too much of a surprise that the arhimag1k prediction deviates since it additionally incorporates historical records, which are not included in our data set.

Although well suited to compare with existing models, care has to be taken when interpreting Gauss coefficients. A finite set of Gauss coefficients does not represent the full information contained in our non-parametric modelling approach. Areas of dense data coverage may feature a resolution than can not be captured by an expansion, for example until degree 10. However, this is less relevant in the context of geomagnetic core field modelling, where SH degrees larger than around 14 are dominated by lithospheric field signals, which cannot be resolved by a sparse data distribution. Figure 6.11 compares the predicted Gauss coefficients to the selected reference models until SH degree 5, which is considered the approximate global resolution of the spherical harmonics based models [Korte et al., 2009, Licht et al., 2013, Sanchez et al., 2016, Constable et al., 2016, Hellio and Gillet, 2018]. With few exceptions, our findings are on a par with existing models within one standard deviation.

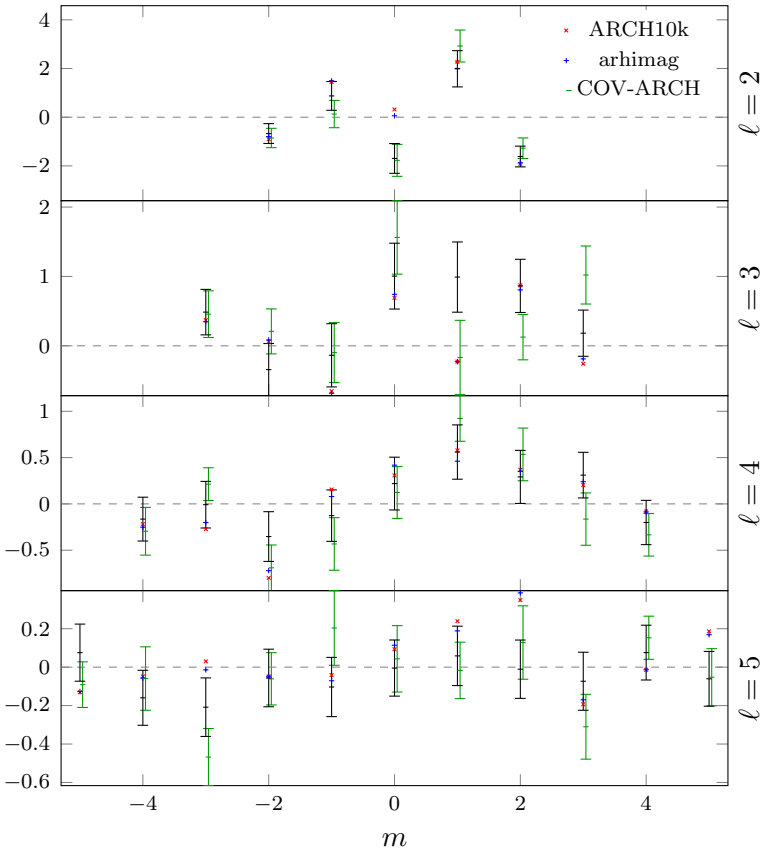

Figure 6.11.: Comparison of Gauss coefficients for SH degrees 2 - 5 at Earth's surface for epoch 1700. Results from this study are shown in black. Error bars indicate one standard deviation.

6.4.6. Spatial Power Spectrum

It is hard to digest all the information contained in a collection of Gauss coefficients. Therefore, it has become common to consider the geomagnetic power spectrum [Backus et al., 1996, section 4.4.2], that reflects the contributions of different spatial wavelengths in the SHs. For degree ℓ the corresponding wavelength is $\lambda_\ell \approx \frac{4\pi R}{2\ell+1}$ [Langel and Hinze, 1998, section 4.3.5]. The EMF is apportioned such that

$$\boldsymbol{B} = \sum_\ell \boldsymbol{B}_\ell = -\nabla \sum_\ell \Phi_\ell \tag{6.84}$$

where

$$\Phi_\ell(\boldsymbol{x}) = R \left(\frac{R}{|\boldsymbol{x}|}\right)^{\ell+1} \sum_{-\ell \le m \le \ell} g_\ell^m \, Y_\ell^m(\hat{\boldsymbol{x}}) \ . \tag{6.85}$$

The components Φ_ℓ are certainly orthogonal, since SHs form an orthogonal system. A characteristic that describes $\boldsymbol{B}_\ell$'s variations is the so-called *average square value*. Since $\boldsymbol{B}$ is divergence free, the net total flux through a closed surface is zero. Thus, the *average* over the sphere vanishes

$$\langle \boldsymbol{B}_\ell \rangle_R = 0 \ , \tag{6.86}$$

where the angle brackets are an abbreviation for the surface integral (see Eq. 6.16). The definition of the average can be extended to the (centred) average square value

$$\left\langle (\boldsymbol{B}_\ell - \langle \boldsymbol{B}_\ell \rangle_R)^2 \right\rangle_R = \left\langle \boldsymbol{B}_\ell^2 \right\rangle_R \ . \tag{6.87}$$

Due to Parseval's Theorem [Backus et al., 1996, Eq. 4.4.21], the average square can be expressed in terms of Gauss coefficients. Regarded as a function in ℓ, the quantity

$$r_\ell := \left\langle \mathbf{B}_\ell^2 \right\rangle_R = (\ell+1) \sum_{-\ell \le m \le \ell} (g_\ell^m)^2 \tag{6.88}$$

is called the *geomagnetic power spectrum* [Lowes, 1974]. Again, Gauss coefficients g_ℓ^m depend on the reference radius and so does r_ℓ.

Within the setting of statistical inversions, calculating the power spectrum requires special attention since uncertainties have an appreciable effect. Squaring and summing normal distributed Gauss coefficients gives a random variable (RV) that is distributed as the sum of weighted non-central χ^2 RVs. Unfortunately, for the PDF of a linear combination of non-central chi-square RVs no closed, analytic expression is known [Bausch, 2013]. Nonetheless, using the algebraic formula for the variance $\mathbb{E}[XY] = \mathrm{Cov}[X,Y] + \mathbb{E}[X]\,\mathbb{E}[Y]$ we obtain an expression for the expectation of the power spectrum

$$\mathbb{E}[r_\ell | o] = (\ell+1) \sum_m \left(\mathbb{E}[g_\ell^m | o]^2 + \mathbb{V}[g_\ell^m | o] \right) \tag{6.89}$$

and it is obvious that variances play an important role. In other words, the larger the uncertainties the bigger the impact on the spectrum's mean.

Although the second moment may be accessible, the standard deviation is not suited to quantify errors. Roughly speaking, if standard deviations of the g_ℓ^ms dominate over the mean, the PDF of r_ℓ is highly skewed with wide tails. In turn, a moment matching

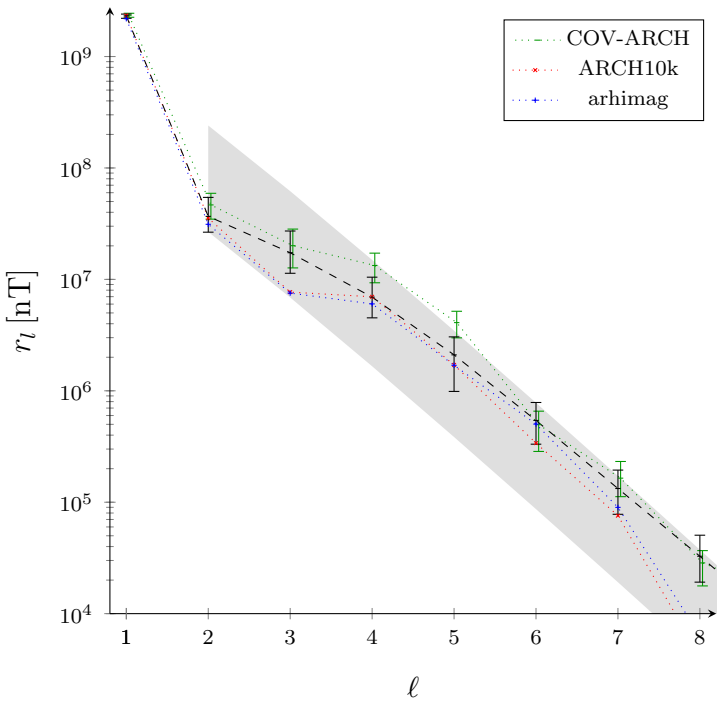

Figure 6.12.: Power spectrum at the Earth' surface for epoch 1700. The mean of the posterior power spectrum is indicated by the dashed line. The top and bottom of the error bars refer to 16 and 84 percentiles, respectively. Except for a constant offset, the grey shaded area indicates the slope of the a priori power spectrum.

Gaussian proxy would violate the positivity constraint of r_ℓ. For an increasing SH degree this is certainly the case. Percentiles are better suited to estimate the error level, but impossible to access analytically. We again calculate the distribution and percentiles empirically by brute-force sampling Gauss coefficients from the posterior. The sampling strategy is described in Section 6.4.3.

For epoch 1700 the resulting power spectrum is shown in Figure 6.12. Within the error margins our findings and the COV-ARCH model are in good agreement. Differences arise comparing with ARCH10k.1 and arhimag1k. Both models report less power at degree $\ell = 3$ and feature a rapid loss for degrees $\ell \geq 8$. While the degree 3 deviation might be due to differences in the underlying data basis, the latter likely is caused by the influence of the global regularization in the spherical harmonic models. Nonetheless, for $\ell \leq 8$ ARCH10k.1 and arhimag1k are potential realizations from what we find.

6.4.7. Dipole Moment

To a first approximation, the EMF is dipolar. This is, its shape is similar to that of a hypothetical bar magnet placed at the centre of the Earth. With respect to a Cartesian coordinate frame, the corresponding vector dipole moment is given by

$$\boldsymbol{m} = \frac{4\pi R_E^3}{\mu_0}(g_1^1\hat{\boldsymbol{x}} + g_1^{-1}\hat{\boldsymbol{y}} + g_1^0\hat{\boldsymbol{z}}) \ , \tag{6.90}$$

where $\mu_0 \approx 4\pi \cdot 10^{-7}\,[\mathrm{Tm/A}]$ refers to the permeability of free space [Backus et al., 1996, Eq. 4.4.17]. Vector components are indicated by subscripts x, y and z, for example

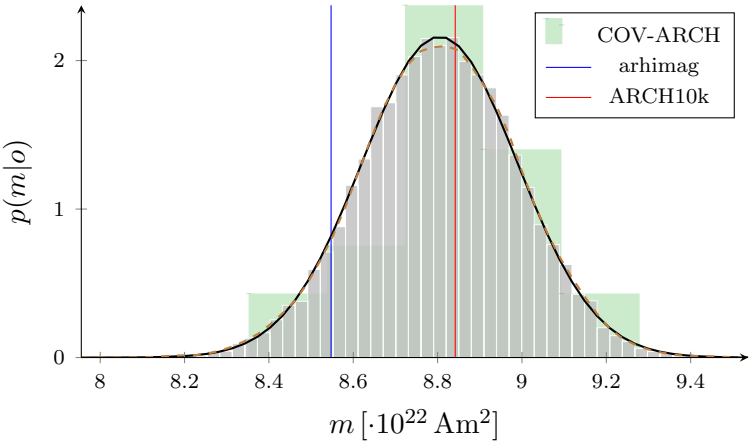

Figure 6.13.: Distribution of the dipole moment for epoch 1700 based on 10000 drawn samples. The kernel density estimate is indicated by the red dashed line, that is nearly identical to the Gaussian proxy (black line).

$m_x \propto g_1^1$. Considering the discretization, the vector dipole moment $\boldsymbol{m}$ is a Gaussian mixture borrowing its statistical properties from the dipole coefficients. The dipole's magnitude can be computed directly via

$$\tau = \|\boldsymbol{m}\| = \frac{4\pi R_E^3}{\mu_0}\sqrt{\left(g_1^1\right)^2 + \left(g_1^{-1}\right)^2 + \left(g_1^0\right)^2} \tag{6.91}$$

and is proportional to the square root of the power r_1. We are interested in statistical properties of τ. Because of the square root, we are not able to derive an analytic expression for $\mathbb{E}[\tau|o]$. Nonetheless, if we knew $\mathbb{E}[\tau|o]$, the variance is right at hand

$$\mathbb{V}[\tau|o] = \mathbb{E}\left[\tau^2|o\right] - \mathbb{E}[\tau|o]^2 \ , \tag{6.92}$$

since $\mathbb{E}[\tau^2] \propto \mathbb{E}[r_1]$, which is given by Eq. 6.89. To obtain a proxy of $\mathbb{E}[\tau|o]$ and to calculate the empiric distribution we again use sampling. An ordinary Gaussian kernel density estimate is used to smooth the histogram [Murphy, 2012, section 14.7.2]. The bandwidth is selected by *Scott's rule of thumb* which – due to its simplicity – strongly influences the estimate.

For the epoch of choice, Figure 6.13 compares the density estimate, the according histogram and the Gaussian proxy. The density estimate looks rather normal and is approximated well by the moment matching Gaussian. This is also the case for all other epochs in our study. At least for the epoch 1700 our findings agree with ARCH10k.1 and COV-ARCH. Presumably, due to its stronger data basis arhimag1k deviates. However, the agreement varies through epochs.

As the dipole moment is known to change with time, Figure 6.14 displays the time-series of all epochs under consideration. Compared to COV-ARCH, ARCH10k.1 and arhimag1k, we find a similar temporal evolution of the dipole moment. Presumably the most recent epoch deviates from the other models, since we do not include historical information in our model. Especially for the earlier epochs we see slightly higher intensities than reported by existing models. The strong deviation of epochs 1100 and 1500 may be caused by outliers, which we did not test for so far, but it is also possible that the variations are real and earlier models underestimate strong variations due to the treatment of dating

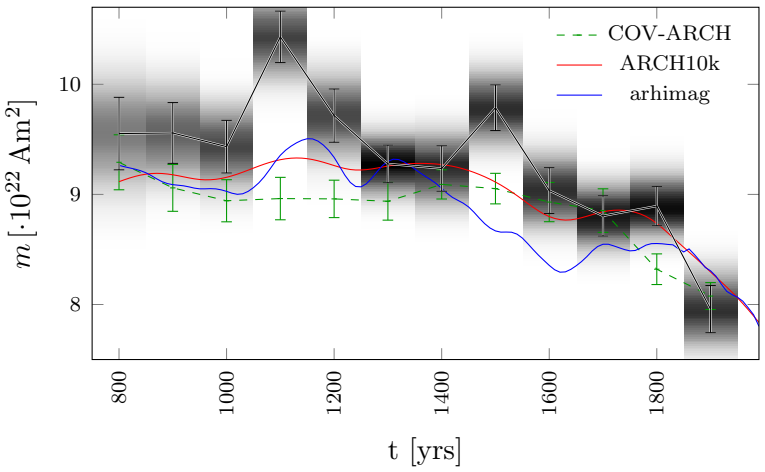

Figure 6.14.: Temporal variation of the dipole moment magnitude based on 10000 drawn samples. The posterior means are connected by black lines and the error bars indicate one standard deviation. The gray shaded background refers to the kernel density estimate. Results from models COV-ARCH, ARCH10k and arhimag are shown for comparison.

uncertainties and temporal regularization. It is beyond the scope of our proof-of-concept model to resolve this question. For both epochs the deviation is caused by g_1^0. The parameter distributions $p(\vartheta|o)$ for these epochs do not show any noticeable problems. However, epoch 1500 features a rather weak data basis.

We are further interested in the distribution of the dipole's north pole a.k.a. *geomagnetic north*. The geomagnetic north pole is given as the antipode of the projection of the vector dipole moment onto the sphere. In other words, the intersection of the axis of the hypothetical bar magnet with the Earth's surface. The location w.r.t. spherical coordinates is given by

$$\theta_m = \arccos\left(-\frac{m_z}{\tau}\right) , \qquad \phi_m = \arctan\left(\frac{-m_y}{-m_x}\right) \tag{6.93}$$

where θ refers to colatitude and ϕ to longitude. We proceed with a similar approach as presented in Khokhlov et al. [2006] and translate the dipole's PDF that is interpreted w.r.t. a Cartesian reference frame into spherical coordinates. The PDF of the vector dipole moment transformed to spherical coordinates is given by

$$p\left(\begin{pmatrix}\tau\\ \theta_m\\ \phi_m\end{pmatrix}\middle| o\right) = p\left(\begin{pmatrix}m_x = \sin\theta_m\cos\phi_m\\ m_y = \sin\theta_m\sin\phi_m\\ m_z = \cos\theta_m\end{pmatrix}\middle| o\right)\tau^2\sin\theta_m \ , \tag{6.94}$$

where we made use of the change of variables theorem [Murphy, 2012, section 2.6]. To obtain the distribution of the location we have to marginalize the magnitude from the mixture distribution

$$p\left(\begin{pmatrix}\theta_m\\ \phi_m\end{pmatrix}\middle| o\right) = \sum_i p(\vartheta_i|o)\int_0^\infty p\left(\begin{pmatrix}\tau\\ \theta_m\\ \phi_m\end{pmatrix}\middle| o, \vartheta_i\right)\mathrm{d}\tau \ . \tag{6.95}$$

Since individual dipole coefficients are normally distributed, we can analytically solve the integral. To do so, factor out m, complete the square and marginalize via the standard

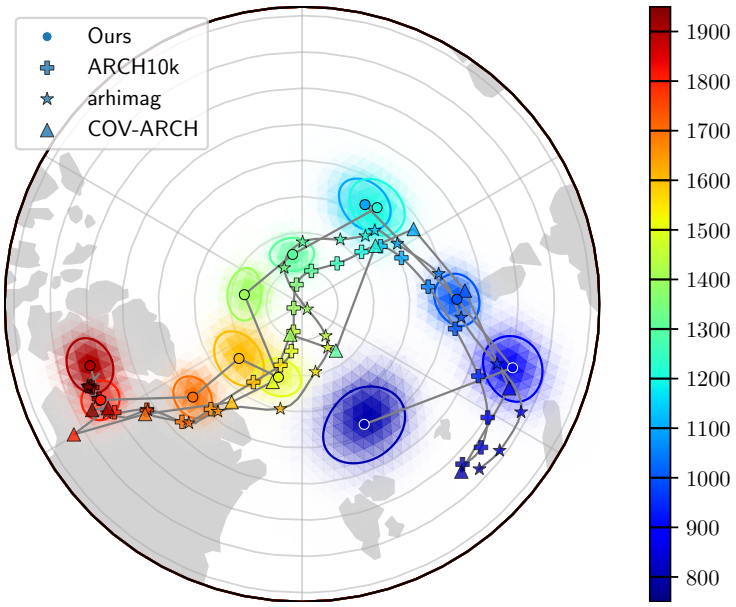

Figure 6.15.: Wander of the geomagnetic north pole. For each epoch in the study, the probability distribution of the location is shown with the mean marked by dots and one sigma marked by an ellipse. The colour varies according to time. For comparison, wander paths for the reference models are shown, without uncertainties in order not to clutter the figure.

Gaussian integral equations [Owen, 1980, Eqs. 10, 11 and 12]. We deliberately skip the resulting expression as it is lengthy and of no particular interest.

However, to visualize the wander of geomagnetic north it is useful to numerically evaluate the resulting PDF. As can be seen in Figure 6.15 the general movement is similar to the models we are comparing with. Again the 1500 epoch deviates and is ahead of the comparison path, which may be caused by outliers as stated previously. For comparison we only show COV-ARCH's sample mean because a scatter plot of the complete ensemble leads to an overstuffed picture. The ensemble covariance features similar magnitudes compared to what we find.

6.5. Conclusions and Perspectives

The extensive theory of Sections 6.1 - 6.3 build the foundation of a new modelling strategy for archeo- and palaeomagnetic field models. The key advantage of this probabilistic approach is that realistic modelling uncertainty estimates are obtained, for example via the standard deviation. The a priori distributions we choose have pros and cons. Taking the least subjective choice is an advantage as we a priori do not specify any preferred direction and, thus, our method is even well suited for time periods featuring reversals. However, a shortcoming – in particular concerning the prior dipole – is that we could not visualize a comparison of prior and posterior uncertainties.

Besides the a priori covariance structure and the weakly informative parameter priors, our modelling strategy depends only on a single parameter, the reference radius R. Conceptually, it is no problem to also integrate out this remaining parameter. Only a technicality arises if we want to integrate out R, Gauss coefficients require scaling to a common radius – for example the Earth's surface – as they depend on R. The limiting

factor is a 4-D parameter space with excessive computational demands.

To save the effort of implementing the exhaustive theory, a ready-to-use software suite called CORBASS [Schanner and Mauerberger, 2019] was developed as part of the project. CORBASS is written in python and licensed under the GPLv3. A public GIT repository serves for development, maintenance and support. To facilitate first steps we provide usage examples in the form of a web based interactive environment (Jupyter notebooks), that further illustrate the modelling concept and the algorithm. To lighten system requirements we make use of the package and environment manager *conda*.

We carried out a case study to demonstrate the potential of our statistical model. All results presented in Section 6.4 are produced using CORBASS. Even though we use a rather small data set, computational costs are not negligible. At the time of writing, processing all epochs under consideration took about 30 hr on an ordinary workstation. The computational complexity is set by the number of observations and the parameter grid chosen. For a certain choice of model parameters, the computational complexity of GP regression is cubic w.r.t. number of observations. Performing a Riemann sum along one dimension scales quadratically according to the number of collocation points. Thus the complexity of the quadrature scheme for all three model parameters grows with the sixth power. Although numeric integration offers room for optimizations, the computational complexity cannot be lightened.

Although a proof of concept rather than a fully featured EMF model, our case study already supports the findings of existing studies. In comparison with models using traditional methods this is useful since it is another source that quantifies what was described qualitatively. As an example, early studies questioned the reconstruction of the EMF in the southern hemisphere from archeomagnetic and volcanic data only due to poor data coverage e.g. Korte et al. [2009], Constable and Korte [2015]. That statement is quantified by our findings: The EMF's posterior standard deviation is small in areas of good data coverage, such as Europe, while uncertainties are large on the southern hemisphere (see top panel of Fig. 6.8). Even though our uninformative prior assumptions are significantly weaker, we find uncertainties similar to for example Licht et al. [2013], while noting that bootstrap ensemble methods tend to underestimate uncertainties in regions where there are no data to draw from. Surprisingly, our uncertainty estimates are on a par with more elaborate modelling concepts [e.g. Hellio and Gillet, 2018] whereas our approach does not yet account for the temporal evolution. Future work will show if this is a coincidence, and if taking temporal dynamics into account yields different results.

In a general context, caution is advised when performing any further processing with posterior Gauss coefficients. Whenever possible, use the posterior EMF instead. At densely covered regions a vast amount of Gauss coefficients are necessary to represent all the information that is contained in the posterior EMF at the surface. However, this is irrelevant when studying the core field which cannot be retrieved beyond degrees around 14 anyway, due to the distance from the source and dominance of the lithospheric field at higher degrees.

The spatial correlation structure we employ makes truncating the SH decomposition obsolete and the model resolves according to the availability of data. This raises the obvious question of how our global, non-parametric model compares to higher resolving studies. To investigate this question one would have to include other sources of data such as historical logs and observatory data for recent times. Although the modelling strategy inherently works with magnetic field components B_N, B_E and B_Z we deliberately

left out recent observatory data to put attention on non-linear observables. If we focus on areas of small modelling errors, precise locations must not be neglected. Considering elevation and the coordinate conversion from geodetic to geocentric is straightforward. The difficulty with historical logs is that the direction of travel is affected by large inaccuracies [Jackson et al., 2000, Jonkers et al., 2003]. To account for location uncertainties, our data model requires an extension. Another difficulty is computational costs related to the large amount of historical and observatory data. Therefore one has to introduce a data selection and reduction process as the interest is in time spans of years but not days or less. Furthermore, our proxy Gaussian error model is intolerant of outliers. With only a few records that strongly deviate, the Gaussian error model causes a highly distorted reconstruction. Therefore it is important to perform outlier analysis and select data with care. The work by Khokhlov et al. [2006] appears to be well suited to discriminate data that are incompatible with our modelling approach.

For the time increments of interest – $\Delta t \geq 1\,\mathrm{yr}$ – we know with confidence when observatory data and historical records were acquired. For volcanic and archeomagnetic records, the average dating uncertainties amount to about a hundred years [Licht et al., 2013, section 2.2], and they tend to increase further back in time. For this study, we assume that our rather long bin width of 100 yr balances temporal errors, although it results in poor temporal resolution. Arguably, our model is still overly confident as we did not consider dating errors. In order to apply our method more generally and to longer times the inclusion of sediment records has to be implemented. Sediment records are affected by large dating uncertainties which require a data model that also accounts for temporal errors and preserves the stratification [e.g., Nilsson et al., 2014]. Moreover, a strategy to deal with the scaling or relative intensity has to be developed and our two-step approach might become a challenge if the number of available full vector records with absolute intensity information is small.

Nonetheless, the presented snapshot model should be considered a first step in the direction of a time continuous correlation based Holocene magnetic field model, and, more generally, a new modelling method for the palaeomagnetic field on various time-scales. We regard the time stationary binning an interim solution as it does not capture the dynamics of the EMF well. In our opinion a temporally continuous model also considering dating errors is needed to fulfil the needs of palaeomagnetic field modelling. For the extension of our modelling concept we are currently working on an empirical continuous time correlation kernel, similar to Bouligand et al. [2016], Hellio and Gillet [2018]. Combining a temporal and a spatial kernel to a space-time kernel will make any binning obsolete. Within this space-time GP setting, existing techniques may be used to address dating uncertainties [McHutchon and Rasmussen, 2011].

We conclude by emphasizing again that this is initial work towards a new palaeomagnetic field modelling strategy and an improved full Holocene model, and that it is destined to receive many improvements in the future. The open source modelling concept offers vast flexibility and a allows for a variety of refinements.

Bibliography

P. Abrahamsen. *A Review of Gaussian Random Fields and Correlation Functions*. Norsk Regnesentral/Norwegian Computing Center, 1997. ISBN 978-8-2539-0435-1.

K. Aki and P. G. Richards. *Quantitative seismology*. University Science Books, 2002. ISBN 0-935702-96-2.

T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley Series in Probability and Statistics. Wiley, 2003. ISBN 9780471360919.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950. doi: 10.2307/1990404.

G. Backus, R. Parker, and C. G. Constable. *Foundations of Geomagnetism*. Cambridge University Press, 1996. ISBN 0521410061.

Marco Ballesio, Joakim Beck, Anamika Pandey, Laura Parisi, Erik von Schwerin, and Raul Tempone. Multilevel Monte Carlo Acceleration of Seismic Wave Propagation under Uncertainty. *arXiv e-prints*, art. arXiv:1810.01710, 2018.

Johannes Bausch. On the efficient calculation of a linear combination of chi-square random variables with an application in counting string vacua. *Journal of Physics A: Mathematical and Theoretical*, 46(50):505202, nov 2013. doi: 10.1088/1751-8113/46/50/505202.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Space in Probability and Statistics*. Springer, Boston, MA, 2004. ISBN 978-1-4419-9096-9. doi: 10.1007/978-1-4419-9096-9.

Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.

W. M. Bolstad. *Introduction to Bayesian statistics*. Wiley, 2 edition, 2007. ISBN 9780470141151.

C. Bouligand, G. Hulot, A. Khokhlov, and GA Glatzmaier. Statistical paleomagnetic field modeling and dynamo numerical simulation. *Geophysical Journal International*, 161: 603–626, 06 2005. doi: 10.1111/j.1365-246X.2005.02613.x.

Claire Bouligand, Nicolas Gillet, Dominique Jault, Nathanaël Schaeffer, Alexandre Fournier, and Julien Aubert. Frequency spectrum of the geomagnetic field harmonic coefficients from dynamo simulations. *Geophysical Journal International*, 207:1142–1157, 11 2016. doi: 10.1093/gji/ggw326.

I. N. Bronstein, K. A. Semendjajew, G. Musiol, and H. Mühlig. *Taschenbuch der Mathematik*. Deutsch (Harri), 2000. ISBN 3817120052.

Maxwell C. Brown, Fabio Donadini, Andreas Nilsson, Sanja Panovska, Ute Frank, Kimmo Korhonen, Maximilian Schuberth, Monika Korte, and Catherine G. Constable. Geomagia50.v3: 2. a new paleomagnetic database for lake and marine sediments. *Earth, Planets and Space*, 67(1):70, May 2015. ISSN 1880-5981. doi: 10.1186/s40623-015-0233-z. URL `https://doi.org/10.1186/s40623-015-0233-z`.

T. Bui-Thanh. A gentle tutorial on statistical inversion using the bayesian paradigm. Technical report, Institute for Computational Engineering and Sciences, The University of Texas at Austin, 2012.

D. Calvetti and E. Somersalo. *Introduction to Bayesian scientific computing*. Springer-Verlag New York, 2007. ISBN 978-0-387-73393-7.

C. H. Chapman. *Fundamentals of Seismic Wave Propagation*. Cambridge University Press, 2004. ISBN 978-511-21037-2.

C. G. Constable and M. Korte. Centennial- to millennial-scale geomagnetic field variations. In G. Schubert, editor, *Treatise on Geophysics, 2nd ed.*, volume 5, pages 309–341. Elsevier, 2015.

C. G. Constable and R. L. Parker. Statistics of the geomagnetic secular variation for the past 5 m.y. *Journal of Geophysical Research: Solid Earth*, 93(B10):11569–11581, 1988. doi: 10.1029/JB093iB10p11569.

C. G. Constable, C. L. Johnson, and S. P. Lund. Global geomagnetic field models for the past 3000 years: transient or permanent flux lobes? *Phil. Trans. R. Soc. Lond. A*, 358: 991–1008, 2000.

Catherine G. Constable, Monika Korte, and Sanja Panovska. Persistent high paleosecular variation activity in southern hemisphere for at least 10 000 years. *Earth and Planetary Science Letters*, 453:78 – 86, 2016. ISSN 0012-821X. doi: 10.1016/j.epsl.2016.08.015.

I. Cook. *A Dictionary of Statistics*. OUP Oxford, New York, London, 2008. ISBN 978-0-199-54145-4.

R. A. Davis. *Gaussian Process: Theory*. John Wiley & Sons, Ltd, 2014. ISBN 978-1-118-44511-2. doi: 10.1002/9781118445112.stat07472.

Markus Deserno. How to generate equidistributed points on the surface of a sphere. *If Polymerforshung (Ed.)*, page 99, 2004.

S. Donner, M. Bernauer, and H. Igel. Inversion for seismic moment tensors combining translational and rotational ground motions. *Geophysical Journal International*, 207 (1):562–570, 2016. doi: 10.1093/gji/ggw298.

Colin G. Farquharson and Douglas W. Oldenburg. Non-linear inversion using general measures of data misfit and model structure. *Geophysical Journal International*, 134 (1):213–227, 07 1998. ISSN 0956-540X. doi: 10.1046/j.1365-246x.1998.00555.x.

A. Fichtner. *Full Seismic Waveform Modelling and Inversion*. Springer, Berlin, Heidelberg, 2011. ISBN 978-3-642-15807-0. doi: 10.1007/978-3-642-15807-0.

A. Fichtner, L. Stehly, L. Ermert, and C. Boehm. Generalized interferometry – i: theory for interstation correlations. *Geophysical Journal International*, 208(2):603–638, 2017. doi: 10.1093/gji/ggw420.

H. Fischer and H. Kaul. *Mathematik für Physiker*, volume 2. Springer Spektrum, Wiesbaden, 4th edition, 2014. ISBN 978-3-658-00477-4. doi: 10.1007/978-3-658-00477-4.

Jacob R. Gardner, Geoff Pleiss, David Bindel, Kilian Q. Weinberger, and Andrew Gordon Wilson. Gpytorch: Blackbox matrix-matrix gaussian process inference with GPU acceleration. *CoRR*, abs/1809.11165, 2018.

A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 3 edition, 2013. ISBN 9781439840955.

M. G. Genton. Classes of kernels for machine learning: A statistics perspective. *J. Mach. Learn. Res.*, 2:299–312, March 2002. ISSN 1532-4435.

Z. Ghahramani. Unsupervised learning. In O. Bousquet, U. von Luxburg, and G. Rätsch, editors, *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures*, pages 72–112. Springer Berlin Heidelberg, 2004. ISBN 978-3-540-28650-9. doi: 10.1007/978-3-540-28650-9_5.

Nicolas Gillet. Spatial And Temporal Changes Of The Geomagnetic Field: Insights From Forward And Inverse Core Field Models. In *Geomagnetism, aeronomy and space weather: a journey from the Earth's core to the sun*. 2019. URL `https://hal.archives-ouvertes.fr/hal-02042703`.

C. M. Grinstead and L. J. Snell. *Introduction to Probability*. American Mathematical Society, 2 edition, 1998. ISBN 9780821894149.

D. Gubbins. *Time series analysis and inverse theory for geophysicists*. Cambridge University Press, 2004. ISBN 978-0-521-52569-5.

D. Gubbins and J. Bloxham. Geomagnetic field analysis – iii. magnetic fields on the core-mantle boundary. *Geophysical Journal of the Royal Astronomical Society*, 80(3): 695–713, 1985. ISSN 1365-246X. doi: 10.1111/j.1365-246X.1985.tb05119.x.

Gelvam A Hartmann and Igor G Pacca. Time evolution of the south atlantic magnetic anomaly. *Anais da Academia Brasileira de Ciências*, 81(2):243–255, 2009.

Sebastian Heimann, Marius Isken, Daniela Kühn, Henriette Sudhaus, Andreas Steinberg, Simon Daout, Simone Cesca, Hannes Vasyura-Bathke, and Torsten Dahm. Grond – a probabilistic earthquake source inversion framework. *GFZ Data Services*, 2018. doi: 10.5880/GFZ.2.1.2018.003.

G. Hellio and N. Gillet. Time-correlation-based regression of the geomagnetic field from archeological and sediment records. *Geophysical Journal International*, 214(3):1585–1607, 2018. doi: 10.1093/gji/ggy214.

Gabrielle Hellio, Nicolas Gillet, Claire Bouligand, and Dominique Jault. Stochastic modelling of regional archaeomagnetic series. *Geophysical Journal International*, 199:931–943, 08 2014. doi: 10.1093/gji/ggu303.

M. Holschneider, V. Lesur, S. Mauerberger, and J. Baerenzung. Correlation-based modeling and separation of geomagnetic field components. *Journal of Geophysical Research: Solid Earth*, 121(5):3142–3160, 2016. ISSN 2169-9356. doi: 10.1002/2015JB012629.

A. Jackson, A. Jonkers, and M. Walker. Four centuries of geomagnetic secular variation from historical records. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 358(1768):957–990, 2000. ISSN 1364-503X. doi: 10.1098/rsta.2000.0569.

Art R. T. Jonkers, Andrew Jackson, and Anne Murray. Four centuries of geomagnetic data from historical records. *Reviews of Geophysics*, 41(2), 2003. doi: 10.1029/2002RG000115.

J. Kaipio and E. Somersalo. *Statistical and computational inverse problems*. Springer-Verlag New York, 2005. ISBN 0387220739.

B. L. N. Kennett. *Interpretation of seismograms on regional and global scales*. The Seismic Wavefield. Cambridge University Press, 2002. ISBN 978-052-100663-7.

A. Khokhlov, G. Hulot, and C. Bouligand. Testing statistical palaeomagnetic field models against directional data affected by measurement errors. *Geophysical Journal International*, 167(2):635–648, 11 2006. ISSN 0956-540X. doi: 10.1111/j.1365-246X.2006.03133.x.

M. Korte and C. G. Constable. Continuous global geomagnetic field models for the past 3000 years. *Phys. Earth Planet. Interiors*, 140:73–89, 2003.

M. Korte, F. Donadini, and C. G. Constable. Geomagnetic field for 0-3ka: 2. a new series of time-varying global models. *Geochem. Geophys. Geosys.*, 10, Q06008: doi:10.1029/2008GC002297, 2009.

R. Langel and W. Hinze. *The Magnetic Field of the Earth's Lithosphere: The Satellite Perspective*. Cambridge University Press, 1998. ISBN 9780521473330.

Alexis Licht, Gauthier Hulot, Yves Gallet, and Erwan Thébault. Ensembles of low degree archeomagnetic field models for the past three millennia. *Physics of the Earth and Planetary Interiors*, 224:38 – 67, 2013. ISSN 0031-9201. doi: 10.1016/j.pepi.2013.08.007.

G. Lindgren, H. Rootzén, and M. Sandsten. *Stationary stochastic processes for scientists and engineers*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2013. ISBN 9781466586185.

J. S. Liu. *Monte Carlo strategies in scientific computing*. Springer, 2008. ISBN 978-0-387-95230-7.

Philip W Livermore, Alexandre Fournier, Yves Gallet, and Thomas Bodin. Transdimensional inference of archeomagnetic intensity change. *Geophysical Journal International*, 215(3):2008–2034, 09 2018. ISSN 0956-540X. doi: 10.1093/gji/ggy383.

J. J. Love and C. G. Constable. Gaussian statistics for palaeomagnetic vectors. *Geophysical Journal International*, 152(3):515–565, 03 2003. ISSN 0956-540X. doi: 10.1046/j.1365-246X.2003.01858.x.

F. J. Lowes. Spatial power spectrum of the main geomagnetic field, and extrapolation to the core. *Geophysical Journal International*, 36(3):717–730, 1974.

A. Malinverno and R.L. Parker. Two ways to quantify uncertainty in geophysical inverse problems. *Geophysics*, 71(3):W15–W27, 2006. doi: 10.1190/1.2194516.

M. Mandea, M. Korte, D. Mozzoni, and P. Kotzé. The magnetic field changing over the Southern African continent - a unique behaviour. *S. Afr. J. Geol.*, 110:193–202, 2007.

Jonathan H. Manton and Pierre-Olivier Amblard. A primer on reproducing kernel hilbert spaces. *Foundations and Trends in Signal Processing*, 8(1–2):1–126, 2015. ISSN 1932-8346. doi: 10.1561/2000000050.

S. Mauerberger, M. Schanner, M. Korte, and M. Matthias. Correlation based snapshot models of the archeomagnetic field. *Geophysical Journal International*, 7 2020. doi: 10.1093/gji/ggaa336.

Andrew McHutchon and Carl E. Rasmussen. Gaussian process training with input noise. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 1341–1349. Curran Associates, Inc., 2011.

Thomas P. Minka. Expectation propagation for approximate bayesian inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, page 362–369, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558608001.

Robb J Muirhead. *Aspects of multivariate statistical theory*. John Wiley & Sons, 2 edition, 2005.

K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012. ISBN 978-0-262-01802-9.

Andreas Nilsson, Richard Holme, Monika Korte, Neil Suttie, and Mimi Hill. Reconstructing Holocene geomagnetic field variation: new methods, models and implications. *Geophysical Journal International*, 198(1):229–248, 05 2014. ISSN 0956-540X. doi: 10.1093/gji/ggu120.

Enzo Orsingher. Damped vibrations excited by white noise. *Advances in Applied Probability*, 16(3):562–584, 1984. doi: 10.2307/1427287.

D. B. Owen. A table of normal integrals. *Communications in Statistics - Simulation and Computation*, 9(4):389–419, 1980. doi: 10.1080/03610918008812164.

R.L. Parker. *Geophysical Inverse Theory*. Princeton series in geophysics. Princeton University Press, 1994. ISBN 9780691036342.

C.R. Rao. *Linear Statistical Inference and its Applications*. Wiley, 2 edition, 1973.

C.E. Rasmussen and C.K.I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006. ISBN 026218253X.

M. Reed and B. Simon. *I: Functional Analysis*. Methods of Modern Mathematical Physics. Elsevier Science, 1981. ISBN 9780080570488.

David Rios, Fabrizio Ruggeri, and Michael Wiper. *Bayesian Analysis of Stochastic Process Models*. Wiley Series in Probability and Statistics. Wiley, 04 2012. doi: 10.1002/9780470975916.

Christian Robert. Generalized inverse normal distributions. *Statistics & Probability Letters*, 11(1):37–41, 1991.

Havard Rue, Andrea Riebler, Sigrunn H. Sørbye, Janine B. Illian, Daniel P. Simpson, and Finn K. Lindgren. Bayesian computing with inla: a review. *Annual Review of Statistics and Its Application*, 4:395–421, 3 2017. ISSN 2326-8298. doi: 10.1146/annurev-statistics-060116-054045.

Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, 2009. doi: 10.1111/j.1467-9868.2008.00700.x.

H. Sadeghisorkhani, Ó. Gudmundsson, R. Roberts, and A. Tryggvason. Mapping the source distribution of microseisms using noise covariogram envelopes. *Geophysical Journal International*, 205(3):1473–1491, 2016. doi: 10.1093/gji/ggw092.

S. Sanchez, A. Fournier, J. Aubert, E. Cosme, and Y. Gallet. Modelling the archaeomagnetic field under spatial constraints from dynamo simulations: a resolution analysis. *Geophysical Journal International*, 207(2):983–1002, 08 2016. ISSN 0956-540X. doi: 10.1093/gji/ggw316.

Maximilian Schanner and Stefan Mauerberger. CORBASS: CORrelation Based Archeomagnetic SnapShot model v.1.1, 2019. URL `http://doi.org/10.5880/GFZ.2.3.2019.008`.

K.D. Schmidt. *Maß und Wahrscheinlichkeit*. Springer-Verlag Berlin Heidelberg, 2 edition, 2011. ISBN 9783642210266.

David W. Scott. On optimal and data-based histograms. *Biometrika*, 66(3):605–610, 12 1979. ISSN 0006-3444. doi: 10.1093/biomet/66.3.605.

M. Seeger. Gaussian processes for machine learning. *International Journal of Neural Systems*, 14(02):69–106, 2004. doi: 10.1142/S0129065704001899.

Matthias Seeger. *Bayesian Gaussian Process Models: PAC-Bayesian Generalisation Error Bounds and Sparse Approximations*. PhD thesis, 2003.

Robin Senftleben. *Earth's magnetic field over the last 1 000 years*. PhD thesis, University of Potsdam, 2019.

P. M. Shearer. *Introduction to seismology*. Cambridge University Press, 2009. ISBN 978-0-511-58010-9.

K. Siegrist. Random. Probability, Mathematical Statistics, Stochastic Processes, 2015. URL `http://www.randomservices.org/stat/`. Online; accessed 3-Jun.-2016.

M. L. Stein. *Interpolation of spatial data: some theory for kriging*. Springer series in statistics. Springer-Verlag New York, 1999. ISBN 0387986294.

S. Stein and M. Wysession. *An introduction to seismology, earthquakes, and earth structure*. Blackwell, 2003. ISBN 0-86542-078-5.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Publishing Company, Incorporated, 1 edition, 2008. ISBN 0387772413.

N. Suttie and A. Nilsson. Archaeomagnetic data: The propagation of an error. *Physics of the Earth and Planetary Interiors*, 289:73 – 74, 2019. ISSN 0031-9201. doi: 10.1016/j.pepi.2019.02.008.

Erwan Thébault, Christopher C. Finlay, Ciarán D. Beggan, Patrick Alken, Julien Aubert, Olivier Barrois, Francois Bertrand, Tatiana Bondar, Axel Boness, Laura Brocco, Elisabeth Canet, Aude Chambodut, Arnaud Chulliat, Pierdavide Coïsson, François Civet, Aimin Du, Alexandre Fournier, Isabelle Fratter, Nicolas Gillet, Brian Hamilton, Mohamed Hamoudi, Gauthier Hulot, Thomas Jager, Monika Korte, Weijia Kuang, Xavier Lalanne, Benoit Langlais, Jean-Michel Léger, Vincent Lesur, Frank J. Lowes, Susan Macmillan, Mioara Mandea, Chandrasekharan Manoj, Stefan Maus, Nils Olsen, Valeriy Petrov, Victoria Ridley, Martin Rother, Terence J. Sabaka, Diana Saturnino, Reyko Schachtschneider, Olivier Sirol, Andrew Tangborn, Alan Thomson, Lars Tøffner-Clausen, Pierre Vigneron, Ingo Wardinski, and Tatiana Zvereva. International geomagnetic reference field: the 12th generation. *Earth, Planets and Space*, 67(1):79, May 2015. ISSN 1880-5981. doi: 10.1186/s40623-015-0228-9.

Lloyd N. Trefethen and David Bau. *Numerical Linear Algebra*. SIAM, 1997. ISBN 0-89871-361-7.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake Vand erPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1. 0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 2020. doi: 10.1038/s41592-019-0686-2.

G. Wahba. *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics, 1990. ISBN 9780898712445.

Matthew Walker and Andrew Jackson. Robust modeling of the earth's magnetic field. *Geophysical Journal International*, 143:799 – 808, 09 2008. doi: 10.1046/j.1365-246X.2000.00274.x.

Ke Alexander Wang, Geoff Pleiss, Jacob R. Gardner, Stephen Tyree, Kilian Q. Weinberger, and Andrew Gordon Wilson. Exact gaussian processes on a million data points. *CoRR*, abs/1903.08114, 2019.

C. K. I. Williams. Prediction with gaussian processes: From linear regression to linear prediction and beyond. In *Learning and Inference in Graphical Models*, pages 599–621. Kluwer, 1997.