Sinusoidal Parameter Estimation

- A Bayesian Approach -

Master's Thesis

Jesper Kjær Nielsen

10th semester
Department of Electronic Systems
Aalborg University 2009

This thesis has been printed with Computer Modern 10pt and been typeset using LaTeX 2_ε on a computer running the GNU/Linux operating system. All figures have been created using GNUPLOT, PGF and the macro packages Ti*k*Z and PGFPLOTS. Simulations have been run in Matlab™.

AALBORG UNIVERSITY

Department of Electronic Systems
Fredrik Bajers Vej 7
DK-9220 Aalborg Ø
http://es.aau.dk

Title:

Sinusoidal Parameter Estimation - A Bayesian Approach

Master Programme:

Wireless Communication

Project period:

E9, autumn semester 2008
E10, spring semester 2009

Participant:

Jesper Kjær Nielsen

Supervisors:

Mads Græsbøll Christensen
Søren Holdt Jensen
Ali Taylan Cemgil
Simon J. Godsill
Torben Larsen

Copies: 10

Page numbers: 138

Date of completion: June 3, 2009

Abstract:

Sinusoidal parameter estimation is an important problem in a wide range of signal processing applications such as audio coding, compression, signal enhancement and restoration. In this thesis, sinusoidal parameter estimation is treated from a Bayesian perspective which is an emerging signal processing field. In this connection, we give in the first part of the thesis an introduction to the fundamentals of Bayesian thinking and compare it against traditional signal processing methods. In the second part of the thesis, we propose and develop a new Bayesian inference scheme for the sinusoidal model parameters of the dynamic sinusoidal model. This model can be used for modelling non-stationary signals and it is thus more flexible than the more popular static sinusoidal model.

The developed inference scheme is evaluated through simulations on synthetic signals as well as on a real audio signal. These simulations show that the developed inference scheme works very well for making inference about unknown model parameters as well as for restoration. The major drawback of the inference scheme is that it suffers from a high computational complexity which renders it infeasible for most real-time applications.

AALBORG UNIVERSITY

Institut for Elektroniske Systemer
Fredrik Bajers Vej 7
DK-9220 Aalborg Ø
http://es.aau.dk

Titel:

Estimering af sinusparametre
- En bayesiansk tilgang

Specialeretning:

Trådløs kommunikation

Projektperiode:

E9, efterårssemestret 2008
E10, forårssemestret 2009

Deltager:

Jesper Kjær Nielsen

Vejledere:

Mads Græsbøll Christensen
Søren Holdt Jensen
Ali Taylan Cemgil
Simon J. Godsill
Torben Larsen

Eksemplarer: 10

Sideantal: 138

Afleveringsdato: 3. juni 2009

Resumé:

Estimering af sinusparametre er et vigtigt problem inden for en lang række af signalbehandlingsapplikationer. Det drejer sig for eksempel om audiokodning, komprimering, signalforbedring og -genoprettelse. I dette speciale er estimering af sinusparametre behandlet fra et Bayesiansk synspunkt, der er et voksende område inden for signalbehandling. I den første del af specialet gives der en introduktion til den fundamentale Bayesianske tankegang, og den sammenlignes med traditionelle signalbehandlingsmetoder. I den anden del af specialet fremsættes og udvikles en ny Bayesiansk metode til at drage statistiske slutninger for sinusparametre i en dynamisk signalmodel. Denne model kan bruges til at modellere ikke-stationære signaler og er derfor mere fleksibel end den mere populære statiske signalmodel.
Den udviklede Bayesianske metode er evalueret ved hjælp af simuleringer på syntetiske signaler og på et rigtigt audiosignal. Simuleringerne viser, at den udviklede Bayesianske metode med succes kan bruges til at drage slutninger om de ukendte sinusparametre og til signalgenopretning. Den største ulempe ved metoden er, at den lider af en så høj beregningsmæssig kompleksitet, at den ikke ville kunne bruges i de fleste realtidsapplikationer.

Contents

Preface

This master's thesis is written by me, Jesper Kjær Nielsen, at the Department of Electronic Systems on Aalborg University during the 9th and 10th semesters in the project period spanning from September 1, 2008 to June 3, 2009. During the project period, I was affiliated with the Multimedia Information and Signal Processing (MISP) Group at Aalborg University. The thesis is concerned with making inference about sinusoidal parameters. This is a very important problem in a wide range of application, and it has been an active field of research in recent years in the MISP group. The developed approaches so far for solving the inference problem have been based on tools from classical statistics. In this thesis, however, a different approach is taken based on tools from Bayesian statistics. This approach offers some conceptual advantages over the methods from classical statistics.

The purpose of the master project has been twofold: First of all, it has been to analyse a well known problem to the MISP group, but from a different perspective by using Bayesian signal processing. In this way, new knowledge of Bayesian signal processing has been added into the MISP group and this may be important for future research work. Secondly, a purpose has been to develop new methods for sinusoidal parameter inference based on tools from Bayesian statistics.

The contents of the thesis reflects the two purposes. After a short introduction in chapter 1, part I of the thesis is concerned with the fundamentals of Bayesian thinking in chapter 2 as well as the fundamental numerical Bayesian methods in chapter 3. Chapter 4 concludes part I by applying Bayesian signal processing to the sinusoidal parameter inference problem. This also serves the purpose of demonstrating the state-of-the-art Bayesian inference scheme for the sinusoidal frequency. Part II of the thesis is concerned with the proposal and development of a new Bayesian inference scheme for the sinusoidal parameters of the dynamic sinusoidal model. In chapter 5, the dynamic sinusoidal model is defined and the inference scheme for the parameters of it is developed and evaluated in chapter 6 and chapter 7. The thesis in concluded in chapter 8. In the appendices, important background information is provided in terms of a list of several probability distributions and their properties in appendix A, Bayesian inference for the parameters of the Gaussian distribution in appendix B, and the Kalman filter and

smoother in appendix C. Appendix D includes a submitted paper for WASPAA 2009 based on the proposed inference scheme in part II of the thesis.

The reader should pay attention to the following typographical conventions on perusal of this thesis:

- The main part of the thesis is divided into numbered chapters whereas the appendices are arranged alphabetically.
- Figures, tables, equations, examples and algorithms are numbered consecutively according to the chapter symbol. Hence, the first figure in chapter one is named figure 1.1, the second figure figure 1.2 and so on.
- Two types of figures are used. Closed figures are used for displaying results of simulations whereas open figures are used for displaying conceptual and illustrative figures.
- In the first part of the thesis, several examples are presented in order to demonstrate important points. These examples are marked with a vertical grey line in the left margin. Several algorithms are given in this thesis and they are all presented in a framed box.

I would like to take the opportunity to thank my main supervisors Prof. Søren Holdt Jensen and especially Ass. Prof. Mads Græsbøll Christensen who have carefully guided me through this long term master project as well as the projects conducted on the 7th and 8th semester. They have been an integral part of what I have achieved during the learning process that I have underwent in the last couple of years. They have also played a major part in setting up a three month visit to Prof. Simon J. Godsill and his signal processing and communications group at University of Cambridge, UK in January-March 2009. In that connection, I would like to thank Prof. Godsill for having me and being an inspirational source of information and ideas as well as for spending some of his precious time on supervising me.

Last but not least, I would also like to thank Ass. Prof. Ali Taylan Cemgil previously affiliated with University of Cambridge but now with Boğaziçi University, TR. He initially proposed the idea behind the developed inference scheme in this thesis. I would also like to thank Ass. Prof. Cemgil for reading my numerous e-mails and answering some of my many questions.

Aalborg University June 3, 2009

Jesper Kjær Nielsen
<jkjaer@es.aau.dk>

List of Symbols

x	scalar
$\boldsymbol{x}$	vector
$\boldsymbol{X}$	matrix
$\boldsymbol{I}_N$	the $N \times N$ identity matrix
$\lvert \cdot \rvert$	determinant
$(\cdot)^T$	matrix transpose
$\mathrm{tr}(\cdot)$	the trace of a matrix
$\mathrm{diag}(\cdot)$	the diagonal of a matrix
$\boldsymbol{x}_{N_1:N_2}$	set of time ordered scalars or vectors $\boldsymbol{x}_n$ for $n = N_1, \cdots, N_2$
$\mathrm{vec}(\cdot)$	vectorisation of matrix, time ordered scalars or time ordered vectors into a single column vector
$(\cdot)^c$	complex conjugate
j	imaginary unit
$p(\cdot)$	marginal probability distribution
$p(\cdot, \cdot)$	joint probability distribution
$p(\cdot \vert \cdot)$	conditional probability distribution
Inv-$\mathcal{G}(x; \alpha, \beta)$	x has an inverse gamma distribution with shape parameter α and scale parameter β
Inv-$\mathcal{W}(\boldsymbol{X}; \nu, \boldsymbol{\Psi})$	$\boldsymbol{X}$ has an inverse Wishart distribution with ν degrees of freedom and scale matrix $\boldsymbol{\Psi}$
$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$	$\boldsymbol{x}$ has a Gaussian (or normal) distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$

$\mathrm{St}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$	$\boldsymbol{x}$ has a student's t-distribution with mean vector $\boldsymbol{\mu}$, covariance matrix $\boldsymbol{\Sigma}$ and ν degrees of freedom
$\mathcal{U}(x; a, b)$	x has a uniform distribution with lower boundaries a and upper boundary b
$\mathcal{VM}(x; \kappa, \mu)$	x has a von Mises distribution with concentration parameter κ and location parameter μ
$E\{\cdot\}$	expected value
$\arg\max_{\boldsymbol{x}} f(\boldsymbol{x})$	the value of $\boldsymbol{x}$ which maximises $f(\boldsymbol{x})$
$\hat{(\cdot)}$	estimator or estimate
$\boldsymbol{x}^{[\tau]}$	the value of $\boldsymbol{x}$ at the τ'th iteration
$\mathbb{R}$	the set of real numbers

List of Abbreviations

AIC	Akaike information criterion
AR	autoregressive
ARMA	autoregressive moving average
BIC	Bayesian information criterion
ESPRIT	estimation of signal parameters by rotational invariance techniques
EVD	eigenvalue decomposition
FFT	fast Fourier transform
i.i.d.	independent, identically distributed
LS	least squares
MA	moving average
MAP	maximum a posteriori
MCMC	Markov chain Monte Carlo
MH	Metropolis-Hastings
ML	maximum likelihood
MLE	maximum likelihood estimate
MMSE	minimum mean squared error
p.d.	positive definite
RJMCMC	reversible jump Markov chain Monte Carlo
SNR	signal-to-noise ratio
sym.	symmetric

WASPAA	workshop on applications of signal processing to audio and acoustics

Chapter 1

Introduction

1.1 Background

A fundamental problem encountered in a wide range of applications is the ability to extract characteristic *features* from some physical process based on a set of observations. Some examples of such applications are:

- *Speech recognition* in which a computer converts spoken words into text or commands. This, e.g., enables a human being to interact with his or her electronic equipment by means of his or her voice. A related field is that of *speaker recognition* in which a computer uniquely identifies the speaker.
- *Compression* in which a computer discards irrelevant and/or redundant parts of an auditory or visual signal thus decreasing the total amount of data. Audio coding and video streaming are well known and widespread examples of this.
- *Source separation* in which a computer separates a mixture of sound sources. This enables, e.g., hearing aids to focus the listening attention to a single speaker within a mixture of speakers and background noise.
- *Automatic music transcription* in which a computer translates an acoustic signal from a musical instrument into musical notation, i.e., notes in a stave.
- *Restoration* in which a computer approximately recovers a distorted or corrupted signal. Examples of this are restoration of old gramophone recordings or reconstruction of lost packages on a package based network.

Due to the wide range of applications in which this feature extraction problem shows up, extensive research has been conducted throughout the years. This has resulted

in several proposed solutions which can be divided into two groups based on whether they assume an underlying parametric structure, i.e., a model, for the physical process or not. If a parametric model is assumed, the method is referred to as a *parametric* method whereas it is referred to as a *non-parametric* method in the absence of the model assumption. There exists an almost endless number of parametric models with some of the more popular being the *autoregressive (AR)*, the *moving average (MA)*, the *autoregressive moving average (ARMA)* and the *sinusoidal* model.

In this thesis, we initially restrict our attention to the real *static*[1] sinusoidal model given by

$$x_n = \sum_{l=1}^{L} \alpha_l e^{-\gamma_l n} \cos(\omega_l n + \varphi_l) + w_n \ , \quad \text{for } n = 1, \cdots, N \tag{1.1}$$

where $\alpha_l > 0$, $\varphi_l \in [-\pi, \pi]$, $\omega_l \in [0, \pi]$, $\gamma_l > 0$ are the amplitude, phase, (angular) frequency and log-damping coefficient of the l'th sinusoid, respectively. The observed signal x_n is at time index n the sum of L of these sinusoids and the stochastic noise term w_n. Unless stated otherwise, we always assume the noise to be white and Gaussian distributed with variance σ_w^2. This model accurately describes segments of observations from physical processes such as voiced speech, a wide range of musical instruments and the direction of arrival of a received wireless signal. Since the signal is completely specified once the parameters of the model are known, the task is to acquire the values of these parameters based on the observations. This process of drawing conclusions from the observations about the characteristics features of the underlying signal is referred to as *statistical inference*.

Due to the applicability of the sinusoidal model, the estimation of the sinusoidal parameters has received a lot of attention. Especially the frequency parameter has been subject to extensive research since it enters the signal model in a non-linear fashion. A thorough overview of some of the fundamental frequency estimators is given in [Stoica and Moses, 2005, ch. 4]. Almost all research has been based on the widespread *classical* approach to statistics. The research based on the other major statistical approach, *Bayesian* statistics, has been much more sparse primarily due to practical problems such as evaluation of complicated high-dimensional integrals. In recent years, however, many computational algorithms such as *Markov chain Monte Carlo (MCMC)* sampling have been embraced and developed by the Bayesian community. This has to a large extend overcome many of the practical problems and led to various developments in Bayesian frequency estimation (see, e.g., [Bretthorst, 1988], [Andrieu and Doucet, 1999], [Davy et al., 2006] and the references therein). In this thesis, we also use the Bayesian approach, and we extend this work further by proposing an inference scheme for the

[1] In this thesis, static refers to that the amplitudes of the sinusoids do not change with time while the amplitudes in a dynamic model do. Strictly speaking, the model in Eq. (1.1) is only static for zero damping, i.e., $\gamma_l = 0$, and dynamic for non-zero damping, i.e., $\gamma_l > 0$. In this thesis, however, we refer to the model in Eq. (1.1) as the static model for both zero and non-zero damping.

parameters of the sinusoidal model in which the amplitude and phase are allowed to vary as a function of time. We refer to this model as the *dynamic* sinusoidal model. Before proposing this new inference scheme in part II, however, we first briefly outline the basic difference between the two schools of statistical inference in the rest of the introduction, and give an introduction to the fundamentals of analytical and numerical Bayesian inference in part I.

1.2 Classical versus Bayesian Statistics

As mentioned above, the field of statistics is divided into two schools: The school of classical statistics (also known as sampling theory, orthodox and frequentist statistics) and the school of Bayesian statistics. The main difference between these two statistical approaches is the way they use and interpret probability. Classical statisticians interpret probability as the *frequency of outcomes of repeated random experiments*, i.e., the probability of a particular outcome is the limiting ratio between the number of observations of the outcome to the total number of trials. For example, if we want to discover the probability of heads for a tossed coin, we simply toss the coin N times while counting the number of heads N_h. The probability of heads is then the fraction $p = N_h/N$ for $N \to \infty$. This interpretation of probability has two consequences [Bolstad, 2007, p. 5]:

- The underlying parameters of the experiments are unknown and *deterministic* variables, i.e., they are fixed for every (hypothetical) repetition of the experiment. For example, we would not be able to compute the probability of heads as the fraction $p = N_h/N$ if we did not assume p to be fixed for all of the N tosses.
- The statistical inference scheme is evaluated from the long-run average performance over an infinite number of (hypothetical) repetitions of the experiment. For example, the fraction N_h/N for a finite number of tosses is only an estimate of the probability of heads. In order to evaluate this estimate, we have to imagine that we have access to an infinity number of similar observation sets and then investigate the statistical properties of this estimate.

Bayesian statisticians treat probability in a more general way as the *degree of belief* [Bolstad, 2007, p. 6]. That is, before an experiment is conducted the possible outcomes x are given a *prior probability* reflecting the subjective anticipation of the outcome. After the experiment has been conducted, the prior probabilities are updated with the new knowledge obtained from a set of observations y. The distribution summarising the prior belief and the knowledge from the observations is called the *posterior probability*. The relationships between these quantities are given by Bayes' theorem [Gelman et al., 2003, p. 8]

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} \tag{1.2}$$

which is the cornerstone of Bayesian statistics. For the coin tossing, for example, we would expect the coin to be approximately fair for which reason we would assign a prior probability $p(x)$ that would reflect that. After observing some tosses, we then, by using Bayes' theorem, compute the posterior probability $p(x|y)$ which now summarises our current state of knowledge on the probability of heads. If we at some later point are given another set of observations, we could again update our knowledge by using Bayes' theorem. Due to the notion of prior probabilities, the underlying parameters from a Bayesian point of view are random variables.

It is beyond the scope of this thesis to perform a comprehensive comparison between the two approaches to statistics. However, we devote the rest of the introduction to illustrate some of the key differences between the two approaches. This also motivates the use of the Bayesian methods for the frequency estimation problem treated in chapter 4 and part II of this thesis.

1.2.1 Key Differences

The Bayesian approach to statistics has been struggling at two major points:

- From a *philosophical* point of view, the classical statisticians have been criticising the use of prior probabilities in the Bayesian approach. Since the prior is designed and chosen before any observations are made, they inject too much subjectivity into the analysis. Also, the use of a prior leads to an interpretation of an unknown parameter as a random variable, and the classical statisticians argue that this is inconsistent with what a parameter really is; a fixed quantity.

- From a *practical* point of view, the Bayesian approach requires, in most cases, evaluation of high-dimensional and complex integrals. This requirement arise from *marginalisation* of uninteresting parameters (known as *nuisance* parameters) and from computation of moments and intervals of the posterior distribution. In many cases, even simple ones, analytical solutions may not exist and numerical integration might be infeasible. In recent years, many computational algorithms based on analytical approximations or stochastic sampling techniques have highly remedied for this practical obstacle. However, the computational complexity of these algorithms is typically still much higher than their classical counterparts.

From an engineering point of view, the philosophical concerns are rarely an issue. The practical concerns are still the major drawback of the methods offered by the Bayesian approach. This is unfortunate since Bayesian statistics offer some important analytical and conceptual advantages over classical statistics.

The results of statistical analysis is often reported in the form of point estimates, interval estimates or hypothesis tests. In the classical approach, terms such as biased/unbiased estimators, confidence intervals, significance levels and p-values describe the basic set of tools for performing the statistical inference. In the Bayesian approach,

all statistical inference is based on the posterior distribution and this entails some key advantages as we describe next.

Point Estimation

Point estimation is an important subject since results are most often reported in this form. In the classical approach, point estimates are computed by using various estimators which loosely can be defined as mappings of the observations into a single point: the estimate. However, it is not always clear which estimator to use for a particular problem. Classical statisticians often struggle with the problem of deriving optimal estimators for the inference problem at hand. This is in general a very difficult task so the search is often constrained to that of finding optimal unbiased estimators. Even in this case, an optimal estimator does not always exist and, if it does, it may be hard to find and/or infeasible to compute [Kay, 1993, p. 12+19-22]. In some cases, unbiased estimators, despite having good statistical properties, also lead to non-sensible answers. A prominent example of this is that the use of an unbiased covariance matrix estimate in a power spectral estimation problem can lead to negative power spectral density estimates which, for physical reasons, have zero probability. The estimators are often derived from the *likelihood* function (also known as the sampling distribution) whose parametric form is the same as the conditional distribution $p(y|x)$ of Bayes' theorem[2]. These estimators are referred to as maximum likelihood estimators, and their popularity stems from the fact that they are optimal unbiased and efficient estimators in the limit of an infinite number of observations [Kay, 1993, p. 157]. An inspection of the scientific literature confirms the importance and difficulty of point estimation within the classical framework. The large majority of publications are concerned with deriving new estimators and benchmarking their statistical performance against other estimators and bounds, and entire books have been written with the purpose of finding good estimators (see, e.g., [Stoica and Moses, 2005] and [Kay, 1993]) for different problems.

In the Bayesian approach, there is no need to search for good estimators since the point estimates are derived from the posterior distribution and thus uniquely defined. Or, as stated by David MacKay, an advocate of the Bayesian approach [MacKay, 2002, p. 50]

> There is no need to invent 'estimators'; nor do we need to invent criteria for comparing alternative estimators with each other. Whereas orthodox statisticians offer twenty ways of solving a problem, and another twenty different criteria for deciding which of these solutions is the best, Bayesian statistics only offers one answer to a well-posed problem.

[2]In the classical approach, the likelihood function is just a deterministic function and not a probability distribution as in the Bayesian approach. Strictly speaking, it is therefore misleading to refer to the conditional distribution as the likelihood function when using Bayesian inference methods. However, in accordance with most of the literature, we adopt this naming convention anyway.

Interval Estimation

In the classical approach, confidence intervals are used for describing intervals likely to contain the value of an unknown parameter. The confidence level is used for quantifying how likely the confidence interval is to contain the parameter. The confidence level is often misinterpreted as the probability that the confidence interval contains the true value of the unknown parameter [Bolstad, 2007, p. xxi]. This interpretation is only valid when using the Bayesian approach to statistics in which case the confidence interval is called the credible or posterior interval. The correct interpretation of the confidence interval is that, for an infinite number of repetitions, a proportion (equal to the confidence level) of the calculated confidence intervals contain the true value of the parameter. In many typical cases, the numerical outcome of the confidence interval and credible interval is the same; however, their interpretation is very different. The latter is demonstrated in example 1.1.

Example 1.1 (Confidence Interval for a Constant in Uniform Noise)
Consider the case of an unknown constant A in zero-mean uniform noise, i.e.,

$$x_n = A + w_n \ , \quad w_n \sim \mathcal{U}(-\Delta/2, \Delta/2) \tag{1.3}$$

where the length of the interval Δ is assumed known. We want to find a confidence interval for the parameter A from N observations at some confidence level $1 - \alpha$.

We write the confidence interval as $(u(\boldsymbol{x}), v(\boldsymbol{x}))$ where $\boldsymbol{x}$ is a vector containing the N observations, and $u(\boldsymbol{x})$ and $v(\boldsymbol{x})$ are statistics which map the observations into the lower and upper endpoints, respectively, of the confidence interval. These statistics are not uniquely defined, but are often selected as points on the distribution of the estimate of A at the significance level α. Using this approach, we first have to select an estimate of A. One candidate could be the sample mean, but it can be shown to have a larger variance than the estimator [Grachev, 1977]

$$\hat{A} = \frac{\min(\boldsymbol{x}) + \max(\boldsymbol{x})}{2} \tag{1.4}$$

which we therefore select as our estimator for A. From an analysis of the distribution of $\hat{A}$ at the significance level α, the confidence interval can be found to [Grachev, 1977]

$$\left(\hat{A} - \frac{\Delta}{2}(1 - \sqrt[N]{\alpha}), \hat{A} + \frac{\Delta}{2}(1 - \sqrt[N]{\alpha}) \right) \ . \tag{1.5}$$

Now, suppose we were given the $N = 2$ observations 0.4 and 0.9 for $\Delta = 1$ and were asked to compute the 75 % confidence interval for A. Inserting these observations into Eq. (1.5) yields the confidence interval $(0.4, 0.9)$ for A with a confidence level of 75 %. By logic, however, this confidence interval must contain the true value of A with probability one since the two observations are separated by at least $\Delta/2$ (for these two

observations, the separation is exactly equal to $\Delta/2$). This clearly underlines the fact that we cannot interpret the confidence level as a probability interval.

For the case of $N = 2$, we could also compute a confidence interval in a more heuristic way. Since, in the long run, the interval defined by the two observations contains the true value of A with probability 0.5, the confidence interval $(\min(\boldsymbol{x}), \max(\boldsymbol{x}))$ has a confidence level of 50 %. If we observe the same two points as before, the confidence interval would still contain the true value of A with probability one but even so, according to classical statistics, we should report it as a 50 % confidence interval. The numerical values of this interval are the same as in the previous case, but since our choice of statistics, $u(\boldsymbol{x})$ and $v(\boldsymbol{x})$, is different, the confidence levels differ. This highlights the importance of selecting appropriate estimators but also the fact that a confidence interval is not fully specified if the statistics are not provided.

The Bayesian credible interval is derived from the posterior distribution and can be interpreted as the probability of containing the true value of the parameter. This illustrates once more the conceptual simplicity and power of the Bayesian approach; the current knowledge is encapsulated in the posterior distribution, and statistical inference is entirely based on computing statistical quantities from it.

Hypothesis Testing

Hypothesis testing is important for, e.g., model comparison. Given some observations we could propose several hypothesis of the underlying model structure responsible for the generation of the observations. We could, for example, propose sinusoidal models of different model orders or an alternative structure such as an AR-model. In the classical approach, the hypothesis are accepted or rejected at some significance level. As in the case of interval estimation, the significance level cannot be interpreted as the probability of the hypothesis being true. Instead it is the proportion of possible set of observations which would be at least as extreme as the observed data set. Due to the fact that hypothesis are accepted or rejected at some significance level, it is hard to answer how much more probable one hypothesis is compared against one or more alternative hypothesis [MacKay, 2002, p. 460]. Using the Bayesian approach, we can answer this question by using a posterior probability distribution over the alternative hypothesis. In this way, we can select the most probable model as well as access how much more probable it is compared against the other models.

1.3 Concluding Remarks

Although the comparison of the classical and Bayesian approaches to statistical inference is an important and interesting research area, it is not the topic of this thesis. The short

comparison above barely scratched the surface of this subject[3], and it only served as a motivation for using the Bayesian approach to the problem of performing statistical inference in the sinusoidal model. By doing so, we obtain some attractive advantages:

- There is no need to search for optimal estimators and compare them against various bounds since the Bayesian approach offers the complete and optimal solution in terms of the posterior distribution. This distribution is always obtained using a single tool: Bayes' theorem.
- The Bayesian approach allows us to interpret the reported statistics as probability statements about the parameters using, e.g., credible intervals.
- In the case of model selection, we can derive probabilities for alternative hypothesised models. Also, we avoid the problem of over-fitting which is a severe drawback of the maximum likelihood method of the classical approach.

In the remaining of this thesis, we therefore adopt the Bayesian approach and apply it to inference for the parameters of the dynamic sinusoidal model. This model formulation is a generalisation of the static sinusoidal model in equation Eq. (1.1), and it allows the amplitudes and phases to evolve according to an autoregressive process. Part I of this thesis provides an introduction to the fundamentals of Bayesian thinking and how the practical problems previously outlined are overcome by using various numerical techniques. Equipped with these tools, we describe in part II how Bayesian inference is performed for the parameters of the dynamic sinusoidal model.

[3] A more thorough discussion on the different approaches to statistics can be found in, e.g., [Jaynes, 2003, ch. 16-17], [Bolstad, 2007, ch. 9+12], [Bernardo and Smith, 1994, ap. B] and [MacKay, 2002, ch. 37].

Part I

Fundamentals

Chapter 2

Bayesian Inference

In this chapter, we give an introduction to the Bayesian approach to statistical inference. We begin with a presentation of Bayes' theorem on which Bayesian inference is based, and we give a description of the types of variables and probability distributions constituting it. One of these probability distributions is the prior distribution which, as discussed in the introduction, constitutes one of the key differences between the classical and Bayesian approaches. We describe various strategies for selecting it and outline the basic terminology pertaining to it. A unique feature of the Bayesian approach is that all inference is based on the posterior distribution. Therefore, we consider it and derive some important point estimates from it. We also introduce the Bayesian pendant to confidence intervals, the credible interval. Finally, we introduce Bayesian model selection.

2.1 Bayes' Theorem and Bayesian Terminology

As alluded in the introduction, Bayesian inference is based on a single fundamental tool: Bayes' Theorem. Before introducing it, however, we restate the two fundamental rules of probability [Bishop, 2006, p. 14]

$$\text{Sum rule:} \qquad p(x) = \int p(x, y)dy \tag{2.1}$$

$$\text{Product rule:} \qquad p(x, y) = p(x|y)p(y) \tag{2.2}$$

where $p(\cdot, \cdot)$, $p(\cdot)$ and $p(\cdot|\cdot)$ denote the *joint*, *marginal* and *conditional* distribution, respectively. These two rules constitute the theoretical basis on which probability distributions are manipulated. Since $p(x, y) = p(y, x)$, we obtain from the product rule

Bayes' Theorem in the general form as

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} \ . \tag{2.3}$$

In the Bayesian framework, we know the value of y and wish to infer the value of x. Assuming a probabilistic model for y's dependence on x as well as independent probabilistic models for x and y, Bayes' Theorem is used for combining these terms into a probabilistic model summarising the uncertainty about x given the known value y.

2.1.1 The General Data Model

In this thesis, we consider a specific setup in which we observe some data $\mathcal{D}$ which originate from one of K models $\mathcal{M}_1, \cdots, \mathcal{M}_K$. For each of these K models, our setup also includes several types of unknown variables and parameters which we for the k'th model collectively denote as ϑ_k. Given the k'th model we therefore consider Bayes' theorem as

$$p(\vartheta_k|\mathcal{D}, \mathcal{M}_k) = \frac{p(\mathcal{D}|\vartheta_k, \mathcal{M}_k)p(\vartheta_k|\mathcal{M}_k)}{p(\mathcal{D}|\mathcal{M}_k)} \ . \tag{2.4}$$

Since this form is so fundamental to Bayesian inference, the distributions of Bayes' theorem are assigned special names:

$p(\vartheta_k|\mathcal{M}_k)$: *Prior Probability Distribution*
This term is one of the key differences between the classical and Bayesian approaches. It contains the knowledge of the unknown parameters ϑ_k before any data $\mathcal{D}$ have been observed.

$p(\mathcal{D}|\vartheta_k, \mathcal{M}_k)$: *Likelihood Function* or *Sampling Distribution*
This term is frequently used in the classical approach where it is often denoted as $p(\mathcal{D}; \vartheta_k)$ or $L(\mathcal{D}; \vartheta_k)$. It is a key element of the *maximum likelihood* method where it is treated as a function of ϑ_k and not the data $\mathcal{D}$ [Box and Tiao, 1973, p. 10]. In the Bayesian framework, it expresses the probability that a certain set of parameters ϑ_k would have generated the observed data $\mathcal{D}$.

$p(\vartheta_k|\mathcal{D}, \mathcal{M}_k)$: *Posterior Probability Distribution*
This term is a measure of how probable a particular set of parameters ϑ_k is given the observation $\mathcal{D}$. The posterior probability combines the prior information about the parameters and the obtained information about these from observing the data in an optimal way [Godsill and Rayner, 1998, p. 75]. In the Bayesian approach, all statistical inference such as point and interval estimates is based on the posterior distribution. This is in contrast to the classical approach in which the optimal estimator depends on the inference problem.

$p(\mathcal{D}|\mathcal{M}_k)$: *Evidence*
This term is a scale factor independent of the parameters ϑ_k. It is a measure of the probability that we would observe a particular realisation $\mathcal{D}$, and its primary function is to ensure that the posterior probability integrates to one [Duda et al., 2000, p. 23]. It is usually evaluated by marginalisation, i.e.,

$$p(\mathcal{D}|\mathcal{M}_k) = \int p(\mathcal{D}|\vartheta_k, \mathcal{M}_k) p(\vartheta_k|\mathcal{M}_k) d\vartheta_k \ , \tag{2.5}$$

and is in many situations infeasible to compute. Since the evidence is independent of ϑ_k, it is used for model selection. From Bayes' theorem we have

$$p(\mathcal{M}_k|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{M}_k)p(\mathcal{M}_k)}{p(\mathcal{D})} \tag{2.6}$$

for every candidate model. Thus, the evidence acts as the likelihood for the model selection inference problem.

In many cases, we wish to make inference about the parameters ϑ_k. Since the evidence is independent of these parameters, it is a mere scale factor which only complicates the inference problem. For this reason, Bayes' theorem is often written as

$$p(\vartheta_k|\mathcal{D}, \mathcal{M}_k) \propto p(\mathcal{D}|\vartheta_k, \mathcal{M}_k) p(\vartheta_k, \mathcal{M}_k) \tag{2.7}$$

where $\propto$ indicates that the posterior distribution is *proportional to* the product of the prior distribution and the likelihood. Example 2.1 makes use of this trick to simplify Bayesian inference in a regression model.

Example 2.1 (Bayesian Regression in the Sinusoidal Model)
In this and the next chapter, we illustrate some of the key concept of analytical and numerical Bayesian inference by use of simple but useful examples. In these examples, we consider a special case of the sinusoidal model in Eq. (1.1) in which the phases and log-damping coefficients are zero. Further, we assume the signal to be harmonic so that the value of the l'th frequency ω_l is l times the fundamental frequency ω, which we assume to be known, i.e.,

$$x_n = \sum_{l=1}^{L} \alpha_l \cos(l\omega n) + w_n \ , \quad \text{for } n = 1, \cdots, N \tag{2.8}$$

This can be written in matrix-vector notation as

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} \cos(\omega) & \cdots & \cos(L\omega) \\ \vdots & \ddots & \vdots \\ \cos(\omega N) & \cdots & \cos(L\omega N) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_L \end{bmatrix} + \begin{bmatrix} w_1 \\ \vdots \\ w_N \end{bmatrix} \triangleq \boldsymbol{A}\boldsymbol{\alpha} + \boldsymbol{w} \ . \tag{2.9}$$

In this example, we assume the noise variance σ_w^2 and the model order L to be known. The observed data $\mathcal{D}$ is the vector $\boldsymbol{x}$ and we wish to find the posterior distribution for the unknown parameters ϑ which is the L amplitudes in the parameter vector $\boldsymbol{\alpha}$.

To solve this inference problem, we first select a prior distribution $p(\boldsymbol{\alpha})$ for the amplitudes. We select the Gaussian prior $\mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}_{\boldsymbol{\alpha}}, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}})$ with known mean and covariance. The reason for selecting this prior is discussed in section 2.2. From Eq. (2.9), we see that the likelihood of the parameters given the data, $p(\boldsymbol{x}|\boldsymbol{\alpha})$, is also Gaussian distributed and given by $\mathcal{N}(\boldsymbol{x}; \boldsymbol{A}\boldsymbol{\alpha}, \sigma_w^2 \boldsymbol{I}_N)$. Using Bayes' theorem in the form of Eq. (2.7), we thus have

$$p(\boldsymbol{\alpha}|\boldsymbol{x}) \propto p(\boldsymbol{x}|\boldsymbol{\alpha})p(\boldsymbol{\alpha}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{A}\boldsymbol{\alpha}, \sigma_w^2 \boldsymbol{I}_N)\mathcal{N}(\boldsymbol{\alpha}; \boldsymbol{\mu}_{\boldsymbol{\alpha}}, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}}) \tag{2.10}$$

which by result B.1 yields the Gaussian posterior distribution $\mathcal{N}(\boldsymbol{\alpha}; \boldsymbol{\mu}_{\boldsymbol{\alpha}|\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}|\boldsymbol{x}})$ with mean and covariance given by

$$\begin{aligned}
\boldsymbol{\mu}_{\boldsymbol{\alpha}|\boldsymbol{x}} &= \boldsymbol{\Sigma}_{\boldsymbol{\alpha}|\boldsymbol{x}}(\boldsymbol{A}^T \sigma_w^{-2} \boldsymbol{x} + \boldsymbol{\Sigma}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\alpha}}) && (2.11)\\
&= \boldsymbol{\Sigma}_{\boldsymbol{\alpha}} \left[\sigma_w^2 (\boldsymbol{A}^T \boldsymbol{A})^{-1} + \boldsymbol{\Sigma}_{\boldsymbol{\alpha}}\right]^{-1} (\boldsymbol{A}^T \boldsymbol{A})^{-1} \boldsymbol{A}^T \boldsymbol{x} \\
&\quad + \sigma_w^2 (\boldsymbol{A}^T \boldsymbol{A})^{-1} \left[\sigma_w^2 (\boldsymbol{A}^T \boldsymbol{A})^{-1} + \boldsymbol{\Sigma}_{\boldsymbol{\alpha}}\right]^{-1} \boldsymbol{\mu}_{\boldsymbol{\alpha}} && (2.12)\\
\boldsymbol{\Sigma}_{\boldsymbol{\alpha}|\boldsymbol{x}} &= (\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}^{-1} + \sigma_w^{-2} \boldsymbol{A}^T \boldsymbol{A})^{-1} \, . && (2.13)
\end{aligned}$$

In Eq. (2.12) and Eq. (2.13), we have expanded the expression for the posterior mean and highlighted the terms pertaining to the prior and likelihood in order to enable an interpretation of the moments of the posterior distribution. From Eq. (2.12), we see that the posterior mean is a weighted linear combination of the prior mean $\boldsymbol{\mu}_{\boldsymbol{\alpha}}$ and the *least-squares* estimate $\hat{\boldsymbol{\alpha}}_{\mathrm{LS}} = (\boldsymbol{A}^T \boldsymbol{A})^{-1} \boldsymbol{A}^T \boldsymbol{x}$ of the parameter vector. The weights are determined by the covariance matrix of the prior and a covariance term that depends on the noise variance as well as the particular structure of the data. If we factor the prior covariances matrix as $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}} = \sigma_{\boldsymbol{\alpha}}^2 \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha}}$, we obtain

$$\lim_{\sigma_{\boldsymbol{\alpha}}^2 \to \infty} \boldsymbol{\mu}_{\boldsymbol{\alpha}|\boldsymbol{x}} = \hat{\boldsymbol{\alpha}}_{\mathrm{LS}} \, . \tag{2.14}$$

Thus, in the absence of any prior knowledge the posterior mean asymptotically equals the least-squares estimate. The posterior precision matrix, i.e., the inverse posterior covariance matrix, is from Eq. (2.13) seen to be a simple sum of the prior precision matrix and the precision term of the noise variance as well as the particular structure of the data. Again, if we increase the prior noise covariance, i.e.,

$$\lim_{\sigma_{\boldsymbol{\alpha}}^2 \to \infty} \boldsymbol{\Sigma}_{\boldsymbol{\alpha}|\boldsymbol{x}} = \sigma_w^2 (\boldsymbol{A}^T \boldsymbol{A})^{-1} \, , \tag{2.15}$$

the posterior covariance is entirely determined from terms pertaining to the likelihood.

As in example 2.1, we usually omit the explicit conditioning on the k'th model in order to keep the notation as simple as possible. Instead, the conditioning on the k'th model is implicit and follows from the context.

2.1.2 Types of Variables and Parameters

In the previous section, we collectively denoted all variables and parameters, which our inference problem depends on, as ϑ[1]. This symbol represents several types of variables and parameters such as *latent* (hidden) variables, *model* parameters, *hyperparameters* and *nuisance* parameters. One variable can easily be a representative of several types of variables at the same time and the classification of a variable into one or more of these different types is not unique. However, it is still useful to define and use these types of variables and assign special symbols to them.

Latent variables
A latent variable is a hidden or unobservable variable which is used in the description of the probability distributions for one or more observable variables [MacKay, 2002, p. 436]. Examples of latent variables are the noise terms in the sinusoidal model, states of a state-space model or a hidden Markov model, and missing samples in a set of observations. Model parameters such as the mixing coefficients of a Gaussian mixture model are also latent variables [Bishop, 2006, p. 430].

Model parameters
Although model parameters are latent variables, they are so important that we use a separate symbol $\boldsymbol{\theta}$ for denoting them. Model parameters are quantities that constitute the variables of a certain model. For example, the model parameters of the sinusoidal model in Eq. (1.1) are the L amplitudes, phases, log-damping coefficients and frequencies as well as the noise variance. Most statistical inference is concerned with deriving point estimates for the unknown model parameters. In this thesis, we use the symbol $\boldsymbol{\theta}$ for denoting *unknown* model parameters. For example, $\boldsymbol{\theta} = \boldsymbol{\alpha}$ in example 2.1 since the phases, log-damping coefficients, frequencies and noise variance are known.

Hyperparameters
In a Bayesian framework, we assign prior distributions to the unknown model parameters. The parameters of these prior distributions are referred to as hyperparameters and we denote them as $\boldsymbol{\phi}$. An example of hyperparameters are the mean and variance of the Gaussian prior distribution of the amplitudes in example 2.1. The choice of hyperparameters is an important subject in Bayesian inference, especially in the case of few observations. Therefore, hyperpriors are

[1] We have omitted the subscript k in order to simplify notation. The conditioning on the k'th model is thus implicit.

sometimes assigned to the hyperparameters. We consider the selection of hyperparameters in more detail in section 2.2.

Nuisance parameters
In some situations, we are only interested in estimating a subset of desired parameters, say ϑ_d, of the unknown parameters. The uninteresting parameters are referred to as nuisance parameters and denoted as ϑ_u. Thus, the total set of unknown parameters is written as $\vartheta = \{\vartheta_d, \vartheta_u\}$. Since we are only interested in ϑ_d, the marginal posterior distribution $p(\vartheta_d|\mathcal{D})$ describes all we need to know in order to make inferences about ϑ_d. This marginal posterior distribution can be obtained from the complete posterior distribution $p(\vartheta_d, \vartheta_u|\mathcal{D})$ through marginalisation. Thus, we simply use the sum rule in Eq. (2.1) for integrating the nuisance parameters out, i.e.,

$$p(\vartheta_d|\mathcal{D}) = \int p(\vartheta_d, \vartheta_u|\mathcal{D}) d\vartheta_u \ . \tag{2.16}$$

The ability to remove nuisance parameters by integration is one the advantages of the Bayesian approach. In the classical approach, no general method exists for dealing with nuisance parameters [Kay, 1993, p. 329]. An example of performing Bayesian inference in the presence of nuisance parameters is given in example 2.2.

Example 2.2 (Prediction in the Sinusoidal Model)
Consider the same signal model as in example 2.1. Instead of computing the posterior distribution of the unknown model parameters $\boldsymbol{\theta}$, we wish to find the posterior distribution over an unobserved sample, say, at time index m. Since we are performing prediction, the posterior distribution over the future sample is sometimes called the posterior predictive distribution [Gelman et al., 2003, p. 8]. We denote this future sample as z_m. As in example 2.1, we do not know the amplitudes of the sinusoids. Since we are only interested in making a prediction is this example, the amplitudes are nuisance parameters for this particular inference problem. The total set of variables is thus $\vartheta = \{z_m, \boldsymbol{\theta}, \boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\}$ where z_m is a latent variable constituting the set of desired parameters ϑ_d, $\boldsymbol{\theta}$ contains the unknown model parameters and constitutes the set of nuisance parameters ϑ_u, and $\boldsymbol{\phi} = \{\boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\}$ is the set of known hyperparameters.

By using Eq. (2.16), we obtain

$$p(z_m|\boldsymbol{x}) = \int p(z_m, \boldsymbol{\theta}|\boldsymbol{x}) d\boldsymbol{\theta} = \int p(z_m|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{x}) d\boldsymbol{\theta} \tag{2.17}$$

where the last equality follows from the product rule in Eq. (2.2) and the fact that z_m is independent of $\boldsymbol{x}$ given $\boldsymbol{\theta}$. The future sample z_m obeys the same model as our observations $\boldsymbol{x}$ so $z_m \sim \mathcal{N}(z_m; \boldsymbol{a}_m^T \boldsymbol{\alpha}, \sigma_w^2)$ with $\boldsymbol{a}_m^T = \begin{bmatrix} \cos(\omega m) & \cdots & \cos(L\omega m) \end{bmatrix}$. The distribution $p(\boldsymbol{\theta}|\boldsymbol{x})$ is the posterior distribution found in example 2.1 since $\boldsymbol{\theta} = \boldsymbol{\alpha}$.

Therefore, by using result B.1, the posterior predictive distribution is

$$p(z_m|\boldsymbol{x}) = \int \mathcal{N}(z_m; \boldsymbol{a}_m^T\boldsymbol{\alpha}, \sigma_w^2)\mathcal{N}(\boldsymbol{\alpha}; \boldsymbol{\mu}_{\boldsymbol{\alpha}|\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}|\boldsymbol{x}})d\boldsymbol{\alpha} \tag{2.18}$$

$$= \mathcal{N}(z_m; \boldsymbol{a}_m^T\boldsymbol{\mu}_{\boldsymbol{\alpha}|\boldsymbol{x}}, \sigma_w^2 + \boldsymbol{a}_m^T\boldsymbol{\Sigma}_{\boldsymbol{\alpha}|\boldsymbol{x}}\boldsymbol{a}_m) \ . \tag{2.19}$$

Thus, by using the Bayesian approach, we have enabled ourselves to perform prediction without explicitly having to estimate the values of the nuisance parameter $\boldsymbol{\alpha}$.

2.2 The Prior Distribution

The use of a prior distribution for the unknown variables in the Bayesian approach is one of the most fundamental differences as compared against the classical approach. In example 2.1 and 2.2, we used a Gaussian prior for the unknown model parameters $\boldsymbol{\alpha}$ without motivating or justifying why we choose it or how we selected the value of the hyperparameters. In most real life examples, we do not know neither the functional form of the prior distribution nor the value of the corresponding hyperparameters for which reason we have to base our choice on assumptions. These assumptions are based on various factors such as physical considerations, degree of knowledge and, more controversially, mathematical convenience. Since the posterior distribution depends on these subjective assumptions, the Bayesian approach is often criticised as being too biased [Bishop, 2006, p. 23]. On the other hand, the use of priors enables us to incorporate actual prior knowledge in the inference process which would otherwise be impossible or hard to do in the classical framework.

2.2.1 Informative and Conjugate Priors

An informative prior distribution is used when prior knowledge is available. For example in the case of an unknown noise variance, then, for physical reasons, we would a priori know that the noise variance is positive, and we should therefore choose a prior distribution reflecting that. An important subclass of the informative priors is the conjugate priors. If the posterior distribution has the same functional form as the prior distribution, then the prior distribution is said to be conjugate to the likelihood. As we saw in example 2.1, the likelihood as well as the posterior and prior distributions were Gaussian distributions so the conjugate prior for the Gaussian distribution with known variance is also a Gaussian distribution. In example 2.3, we consider the case in which the noise variance of the sinusoidal model is unknown and the amplitudes are known. The conjugate prior in this case, is the inverse gamma distribution. One of the advantages of using a conjugate prior is that it is easy to update the posterior distribution if additional data are observed. In this case, the old posterior distribution, which is now

the prior, is the conjugate to the likelihood, and the new posterior distribution thus keeps the same functional form.

2.2.2 Non-informative Priors

In cases where there is no or vague prior knowledge, non-informative priors can be used. Non-informative priors, also called diffuse priors, reflects this uncertainty by being very flat so that they play a minimal role. As we saw in example 2.1, we would obtain a non-informative prior by selecting the prior covariance matrix to be very large. In the limit of infinite covariance, the prior distribution did not affect the moments of the posterior distribution at all. In this limit, however, the prior distribution is not a valid probability distribution since it does not integrate to one. These priors are called improper priors and are often used in practice since, in many cases, the posterior distribution is still proper [Gelman et al., 2003, p. 62], i.e., it integrates to one. However, care must be taken when using improper priors. This is also demonstrated in example 2.3.

Example 2.3 (Bayesian Regression for an Unknown Noise Variance)
Consider again the sinusoidal model in example 2.1 but this time with known amplitudes and unknown noise variance. We wish to find the posterior distribution for the noise variance. For physical reasons, the noise variance must be positive so we should select a prior distribution with positive support. A convenient prior fulfilling this property is the inverse gamma prior Inv-$\mathcal{G}(\sigma_w^2; a, b)$ since it is the conjugate prior for the Gaussian distribution with known mean and unknown variance. Using this prior, we obtain

$$p(\sigma_w^2|\boldsymbol{x}) \propto p(\boldsymbol{x}|\sigma_w^2)p(\sigma_w^2) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{A\alpha}, \sigma_w^2 \boldsymbol{I}_N)\text{Inv-}\mathcal{G}(\sigma_w^2; a, b) \tag{2.20}$$

which from result B.3 yields an inverse gamma distribution Inv-$\mathcal{G}(\sigma^2; a_{\sigma^2|\boldsymbol{x}}, b_{\sigma^2|\boldsymbol{x}})$ with parameters

$$a_{\sigma^2|\boldsymbol{x}} = a + N/2 \tag{2.21}$$

$$b_{\sigma^2|\boldsymbol{x}} = b + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{A\alpha})^T(\boldsymbol{x} - \boldsymbol{A\alpha}) \ . \tag{2.22}$$

If we let $a \to 0$ and $b \to 0$, the inverse gamma prior converges to the uniform distribution $\mathcal{U}(0, \infty)$ which is clearly improper. The posterior distribution, however, is still proper in this limit. On the other hand, if we wish to compute the evidence, then from result B.3 we have that

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}|\sigma_w^2)p(\sigma_w^2)d\sigma_w^2 = \text{St}(\boldsymbol{x}; \boldsymbol{A\alpha}, \frac{b}{a}\boldsymbol{I}_N, 2a) \tag{2.23}$$

which is improper for an improper prior. In general, the evidence is always improper for an improper prior [MacKay, 2002, p. 354].

2.3 Summarising Inference

As we have stated several times and illustrated in a few examples, the posterior distribution optimally combines the prior information about unknown parameters with the information about these gained from observing the data. Thus, the posterior distribution is the complete answer to the inference problem. In practice, however, the statistical inference is often summarised and reported in terms of point estimates, interval estimates or posterior odds. A point estimate is often needed in signal processing applications, and it is an estimate of the unknown parameter vector ϑ and is often denoted as $\hat{\vartheta}$. The point estimate is typically derived by using a *cost-function* $C(\mathcal{E})$, also called a loss-function, which expresses the cost (or loss) incurred in selecting a particular set of parameters as an estimate. That is,

$$C(\mathcal{E}) = C(\vartheta - \hat{\vartheta}) \ . \tag{2.24}$$

Since the parameters are random variables, the error $\mathcal{E}$ is a random variable too, and that leads to a cost-function with the intractable property that is depends on the particular realisation of the parameters ϑ[2] and observed data $\mathcal{D}$. The remedy for this is the *Bayes' risk* which measures the average cost and is defined as [Kay, 1993, p.343]

$$R(\hat{\vartheta}) = E\{C(\mathcal{E})\} = \int\int C(\mathcal{E})p(\vartheta, \mathcal{D})d\vartheta d\mathcal{D} \tag{2.25}$$

where the expectation is with respect to the joint distribution of ϑ and $\mathcal{D}$.

2.3.1 The Minimum Mean Square Error and Maximum a Posteriori Point Estimates

Two widely used Bayesian point estimators, the *minimum mean square error* (MMSE) and the *maximum a posteriori* (MAP), can be derived by minimising the Bayes' risk in Eq. (2.25) for two particular choices of the cost-function. These choices are the *quadratic* cost-function defined by

$$C(\mathcal{E}) = |\mathcal{E}|^2 \tag{2.26}$$

and the *uniform* cost-function defined by

$$C(\mathcal{E}) = \begin{cases} 1 & |\mathcal{E}| \geq \epsilon \\ 0 & |\mathcal{E}| < \epsilon \end{cases} \tag{2.27}$$

[2]Recall that the parameters are interpreted as random variables in the Bayesian framework

which lead to the MMSE estimator and the MAP estimator, respectively. In order to find these estimators, we first rewrite Eq. (2.25) by using the product rule in Eq. (2.2). This yields

$$R(\hat{\vartheta}) = \int \left[\int C(\mathcal{E}) p(\vartheta|\mathcal{D}) d\vartheta \right] p(\mathcal{D}) d\mathcal{D} \tag{2.28}$$

from which it is seen that the dependence on ϑ is confined to the inner integral. Since $p(\mathcal{D}) \geq 0$, we can thus minimise the Bayes' risk if the inner integral is minimised. Performing this minimisation for the quadratic cost-function yields the MMSE estimator as given by [Godsill and Rayner, 1998, p. 76]

$$\hat{\vartheta}_{\text{MMSE}} = E\{\vartheta|\mathcal{D}\} = \int \vartheta p(\vartheta|\mathcal{D}) d\vartheta \tag{2.29}$$

and for the uniform cost-function the MAP estimator as given by [Godsill and Rayner, 1998, p. 76]

$$\hat{\vartheta}_{\text{MAP}} = \arg\max_{\vartheta} p(\vartheta|\mathcal{D}) \ . \tag{2.30}$$

Thus, the MMSE estimator is the mean of the posterior distribution whereas the MAP estimator is the mode, i.e., the maximum, of the posterior distribution as illustrated in figure 2.1. Notice that the MAP estimator resembles the maximum likelihood (ML) estimator in that it is a point estimate corresponding to the argument of the maximum of a function that relates the unknown parameters and the observed data. In fact, it can be shown that for an improper uniform prior distribution or an infinite large sample size the MAP and the ML estimators yield the same estimate [Godsill and Rayner, 1998, p. 77].

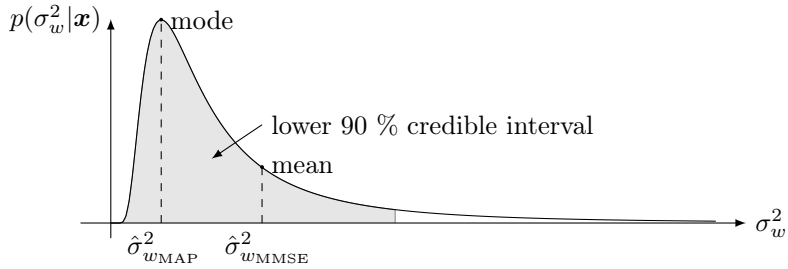

Figure 2.1: Illustration of the posterior inverse gamma distribution for the noise variance σ_w^2 of example 2.3. The figure also shows the MMSE and MAP estimate as well as the lower 90 % credible interval.

2.3.2 Credible Intervals

The Bayesian pendant to confidence intervals of classical statistics is the credible or posterior interval. It is often useful to use this for expressing the uncertainty associated with a point estimate. As opposed to a confidence interval, a credible interval can be interpreted as the probability that the true parameter is contained in the interval. The credible interval it typically two-sided with equal tail areas, but it can also be one-sided or even separated into two or more intervals [Gelman et al., 2003, pp. 38-39]. Figure 2.1 shows the lower 90 % probability interval of an inverse gamma distribution and example 2.4 derives the credible interval for the constant in uniform noise previously considered from a classical viewpoint in example 1.1.

Example 2.4 (Credible Interval for a Constant in Uniform Noise)
We return to the problem of deriving an interval estimate for an unknown constant A in uniform noise previously considered in example 1.1. There, we used the classical confidence interval for the interval estimation, but the results were counter-intuitive and more or less useless. In this example, we derive the Bayesian credible interval for A.

We select the prior distribution to be a uniform distribution centred at zero and with a width of δ, i.e., $p(A) = \mathcal{U}(A; -\delta/2, \delta/2)$. By selecting δ to be very large, we can make the uniform distribution non-informative so it plays a minimal role. The likelihood for the n'th observation is also uniform and given by

$$p(x_n|A) = \mathcal{U}(x_n; A - \Delta/2, A + \Delta/2) = \begin{cases} \Delta^{-1} & \text{for } A - \Delta/2 < x_n < A + \Delta/2 \\ 0 & \text{otherwise} \end{cases} . \tag{2.31}$$

By interchanging the roles of x_n and A in this equation, we see that every observation x_n is the centre of an interval of non-zero probability in which A must lie. This is illustrated in figure 2.2 for a few observations and for the prior. From Bayes' theorem, we have

$$p(A|\boldsymbol{x}) \propto p(\boldsymbol{x}|A)p(A) = \left[\prod_{n=1}^{N} \mathcal{U}(x_n; A - \Delta/2, A + \Delta/2)\right] \mathcal{U}(A; -\delta/2, \delta/2) \tag{2.32}$$

so the posterior probability equals zero if any of the uniform distribution equals zero. In terms of figure 2.2, this means that the posterior probability is only non-zero in the interval which is included in all of the uniform intervals. Thus, the posterior distribution must also be a uniform distribution $\mathcal{U}(A; a, b)$ with

$$a = \max(\max(\boldsymbol{x}) - \Delta/2, -\delta/2) = \max(\boldsymbol{x}, \Delta/2 - \delta/2) - \Delta/2 \tag{2.33}$$

$$b = \min(\min(\boldsymbol{x}) + \Delta/2, \delta/2) = \min(\boldsymbol{x}, -\Delta/2 + \delta/2) + \Delta/2 \ . \tag{2.34}$$

We see that the posterior distribution is proper for an improper prior so we let $\delta \to \infty$. The mean of the resulting posterior distribution, i.e., the MMSE estimate, is $\hat{A}_{\text{MMSE}} = (a+b)/2 = (\min(\boldsymbol{x}) + \max(\boldsymbol{x}))/2$ which is the same as the mean estimate used in example 1.1. Thus, the Bayesian approach directly provides us with the optimal estimator! The 75 % credible interval for the $N = 2$ observations, 0.4 and 0.9, can be derived directly from the posterior distribution. Thus, the 75 % central, two-sided credible interval for A is $(0.4625, 0.8375)$. Notice, that the 100 % credible interval for A is $(0.4, 0.9)$ whereas the confidence level for the confidence interval with the same numerical values was only 75 % for the optimal mean estimator in example 1.1. This underlines ones more that credible intervals can be interpreted according to our intuition as true probability intervals whereas the interpretation of the confidence intervals is much more subtle and based on long run frequencies of hypothetical observations.

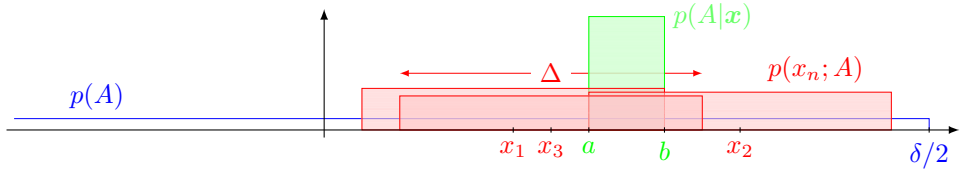

Figure 2.2: Illustration of the prior distribution, likelihood function and posterior distribution for $N = 3$ observations. The height of the individual likelihood functions are shifted a little around the true height for illustrative purposes.

2.3.3 Posterior Odds

If we take two points from the posterior distribution, say, $p(\vartheta_1|\mathcal{D})$ and $p(\vartheta_2|\mathcal{D})$, the ratio between them is the posterior odds. When using posterior odds, Bayes' Theorem takes a particular form given as

$$\underbrace{\frac{p(\vartheta_1|\mathcal{D})}{p(\vartheta_2|\mathcal{D})}}_{\text{Posterior odds}} = \underbrace{\frac{p(\mathcal{D}|\vartheta_1)}{p(\mathcal{D}|\vartheta_2)}}_{\text{Bayes' factor}} \underbrace{\frac{p(\vartheta_1)}{p(\vartheta_2)}}_{\text{Prior odds}} \tag{2.35}$$

where Bayes' factor is sometimes referred to as the likelihood ratio [Gelman et al., 2003, p. 9]. Posterior odds are used in cases of discrete distributions and in particular for hypothesis testing. An example of this is model selection in which the odds for alternative candidate models with respect to a reference model are compared. This is described in greater detail in the next section.

2.4 Model Selection

So far, we have assumed the model $\mathcal{M}_k$ as known and focused on estimating the unknown parameters ϑ_k of this model. In practical problems, however, the model is not always known so it is necessary to perform inference about it. One example of this is the problem of making prediction for unobserved observations $\mathcal{Z}$ which we considered in example 2.2 for a known model. In the case of an unknown model, the posterior predictive distribution is given by

$$p(\mathcal{Z}|\mathcal{D}) = \sum_{k=1}^{K} \int p(\mathcal{Z}, \vartheta_k, \mathcal{M}_k|\mathcal{D}) d\vartheta_k \tag{2.36}$$

$$= \sum_{k=1}^{K} p(\mathcal{M}_k|\mathcal{D}) \int p(\mathcal{Z}|\vartheta_k, \mathcal{M}_k, \mathcal{D}) p(\vartheta_k|\mathcal{M}_k, \mathcal{D}) d\vartheta_k \tag{2.37}$$

$$= \sum_{k=1}^{K} p(\mathcal{M}_k|\mathcal{D}) p(\mathcal{Z}|\mathcal{M}_k, \mathcal{D}) \tag{2.38}$$

which is a mixture distribution where the posterior predictive distribution given the k'th model $p(\mathcal{Z}|\mathcal{M}_k, \mathcal{D})$ is weighted with the posterior probability $p(\mathcal{M}_k|\mathcal{D})$ for this model. In many cases, the full posterior predictive distribution is approximated by making predictions from the single most probable predictive distribution. The problem of selecting the most probable model, which is equivalent to the problem of finding the largest $p(\mathcal{M}_k|\mathcal{D})$, is known as model selection [Bishop, 2006, p. 162].

The Bayesian approach provides a unified method for doing so by using Bayes' theorem. The posterior probability for the k'th model is given by Eq. (2.6) and restated here for easy reference

$$p(\mathcal{M}_k|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{M}_k)p(\mathcal{M}_k)}{p(\mathcal{D})} \tag{2.39}$$

where $p(\mathcal{M}_k)$ and $p(\mathcal{D}|\mathcal{M}_k)$ are the prior model probability and model evidence of Eq. (2.4), respectively. Since the latter acts as the likelihood in Eq. (2.39) and as a marginal distribution in Eq. (2.4), it is often referred to as the *marginal likelihood*. In the case of a uniform prior distribution for the models, the marginal likelihood is proportional to the posterior probability for the models and is thus all we need to know in order to select the most probable model. Alternative models are often compared using posterior odds considered in the previous section. From Eq. (2.35), we see that the posterior odds are equal to Bayes' factor for a uniform prior. Therefore, Bayes' factor is often used for the model selection task.

As we have seen so far, the marginal likelihood plays a key role for model selection. The marginal likelihood favours the simplest possible model that can explain the data reasonably well. This is in contrast to the maximum likelihood method of classical

statistics which suffers from *over-fitting*, i.e., that the model is fitted to the particular noise realisation contaminating the observations. To understand why the marginal likelihood does not suffer from over-fitting, write it as

$$p(\mathcal{D}|\mathcal{M}_k) = \int p(\mathcal{D}|\vartheta_k, \mathcal{M}_k) p(\vartheta_k|\mathcal{M}_k) d\vartheta_k \tag{2.40}$$

which is the integral over the numerator of Eq. (2.4), i.e., the product of the likelihood and prior for the parameters ϑ_k, which is proportional to the posterior distribution of the parameters. Now, assume that the prior and the posterior distributions are approximately uniform distributions with width $\Delta_{\vartheta_k}^{\text{prior}}$ and $\Delta_{\vartheta_k}^{\text{post}}$, respectively, with the posterior distribution centred on the MAP estimate of the parameters. Then, the integral in Eq. (2.40) can be written as the approximation

$$p(\mathcal{D}|\mathcal{M}_k) \approx p(\mathcal{D}|\hat{\vartheta}_{k_{\text{MAP}}}, \mathcal{M}_k) \frac{\Delta_{\vartheta_k}^{\text{post}}}{\Delta_{\vartheta_k}^{\text{prior}}} \tag{2.41}$$

and taking the logarithm yields

$$\ln p(\mathcal{D}|\mathcal{M}_k) \approx \ln p(\mathcal{D}|\hat{\vartheta}_{k_{\text{MAP}}}, \mathcal{M}_k) + \ln \frac{\Delta_{\vartheta_k}^{\text{post}}}{\Delta_{\vartheta_k}^{\text{prior}}} \ . \tag{2.42}$$

The two terms on the right hand side of the latter equation are the maximum value of the log-likelihood function and a penalty term. The value of the log-likelihood function increases for increasing model complexity whereas the width $\Delta_{\vartheta_k}^{\text{post}}$ of the posterior distribution decreases for increasing model complexity [Bishop, 2006, p. 163]. Thus, the largest marginal likelihood is a trade-off between these to competing terms. Due to the build in penalty term in the Bayesian model selection inference scheme, over-fitting is not an issue. The model selection methods of classical statistics often mimics the Bayesian counter part by introducing a similar penalty term as demonstrated in example 2.5.

Example 2.5 (Maximum Likelihood and Bayesian Model Selection)
Consider the sinusoidal model first considered in example 2.1 but this time with unknown amplitudes and noise variance. Suppose the number of sinusoids L is also unknown and we are asked to make inference about it based on an observed data set. Figure 2.3 shows an example for four alternative models $\mathcal{M}_k$ for $k = 1, \cdots, 4$ with k specifying the number of sinusoids L. Using result B.4, the marginal posterior distributions for the amplitudes and noise variance are calculated based on $N = 20$ observed data points. Every plot in the top row shows these data points, which span four periods of the fundamental frequency, as well as a fit to the data based on the mean of the posterior distribution for the amplitudes. The plots in the bottom row

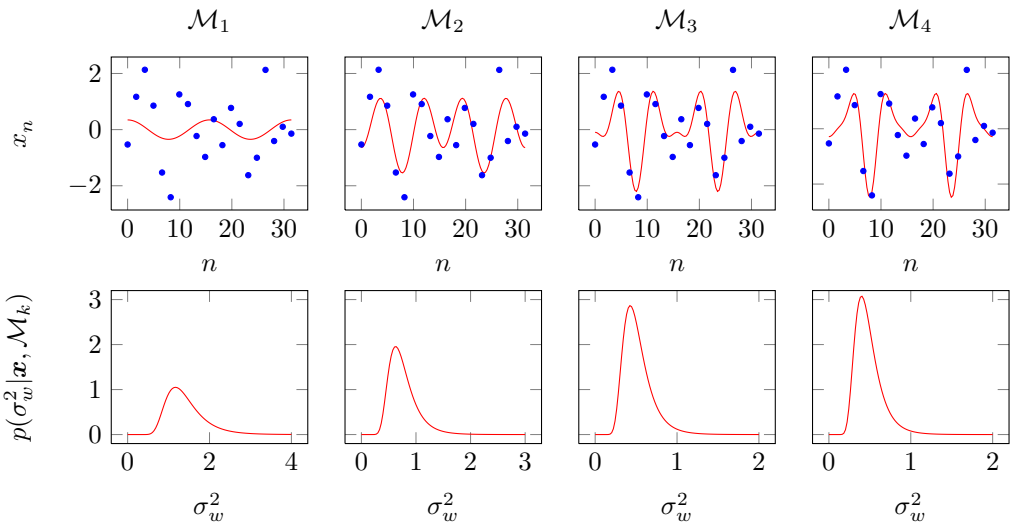

Figure 2.3: Bayesian fitting of sinusoidal model with unknown amplitudes and noise variance for four alternative candidate models.

show the posterior distribution for the observation noise variance pertaining to the four models. Clearly, the mode of the posterior distribution for the noise variance decreases with an increasing number of sinusoids. It is not clear from the plots which model best explains the observations. Therefore, we use the Bayesian approach for comparing the alternative models. Before discussing this, however, we review the maximum likelihood method for solving this problem.

Maximum Likelihood Method

In classical statistics, the fitting is often based on the maximum likelihood solution (which for Gaussian noise is the same as the least squares solution given in example 2.1). This solution is obtained by a maximisation of the likelihood function given by

$$\hat{\boldsymbol{\theta}}_{k_{\mathrm{ML}}} = \arg\max_{\boldsymbol{\theta}_k} p(\boldsymbol{x}; \boldsymbol{\theta}_k) \tag{2.43}$$

where $\boldsymbol{\theta}_k$ is a $k+1$ dimensional vector denoting the k amplitudes and the noise variance for the k'th model. The maximum value of the likelihood function $p(\boldsymbol{x}; \hat{\boldsymbol{\theta}}_{k_{\mathrm{ML}}})$ increases as a function of increasing model order so we cannot simply obtain an estimate of the number of sinusoids by jointly maximising $p(\boldsymbol{x}; \boldsymbol{\theta}_k)$ with respect to both $\boldsymbol{\theta}_k$ and k. If we did so, the model with the largest number of free parameters would be selected due to the over-fitting property of the maximum likelihood method. Therefore, a penalty

term is often introduced which compensates for the over-fitting, i.e., [Stoica and Moses, 2005, p. 417]

$$J(k) = -2 \ln p(\boldsymbol{x}; \hat{\boldsymbol{\theta}}_{k_{\text{ML}}}) + (k+1)\eta(k, N) \tag{2.44}$$

where $\eta(k, N)$ is the penalty coefficient. Two simple and popular choices for the this coefficient are $\eta(k, N) = 2$ and $\eta(k, N) = \ln N$ which are known as the *Akaike information criterion* (AIC) and *Bayesian information criterion* (BIC), respectively. Figure 2.4 shows the value of $J(k)$ for the same data points as in figure 2.3. The figure also shows the case where the two criteria, AIC and BIC, are used. Both of these criteria favours the third model which corresponds to the model with three sinusoids. However, neither of these two criteria indicate how much more probable $\mathcal{M}_3$ is compared against the other models. In general, the two criteria do not yield the same answer which leads to the problem of choosing the best penalty term.

Bayesian Method

We assume the joint prior distribution for the amplitudes and noise variance to be given by

$$p(\boldsymbol{\theta}_k) = p(\boldsymbol{\alpha}_k, \sigma^2_{w,k}) = p(\boldsymbol{\alpha}_k | \sigma^2_{w,k}) p(\sigma^2_{w,k}) \tag{2.45}$$

$$= \mathcal{N}(\boldsymbol{\alpha}_k; \boldsymbol{\mu}_{\boldsymbol{\alpha}_k}, \sigma^2_{w,k} \boldsymbol{C}_{\boldsymbol{\alpha}_k}) \text{Inv-}\mathcal{G}(\sigma^2_{w,k}, a_k, b_k) \tag{2.46}$$

since it is the conjugate prior for the likelihood. From result B.4, we know that the evidence (or marginal likelihood) is a student's t-distribution given by $p(\boldsymbol{x}|k) = \text{St}(\boldsymbol{x}; \boldsymbol{A}_k \boldsymbol{\mu}_{\boldsymbol{\alpha}_k}, \frac{b_k}{a_k}(\boldsymbol{I}_N + \boldsymbol{A}_k \boldsymbol{C}_{\boldsymbol{\alpha}_k} \boldsymbol{A}_k^T), 2a_k)$ with $\boldsymbol{\phi}_k = \{a_k, b_k, \boldsymbol{\mu}_{\boldsymbol{\alpha}_k}, \boldsymbol{C}_{\boldsymbol{\alpha}_k}\}$ being the known hyperparameters for the noise variance and amplitudes of the k'th model. If we assume a uniform prior over the four models and assign zero probability to all other models, the posterior distribution for the number of sinusoids is given by Eq. (2.39). Figure 2.5 shows this distribution for the data points of figure 2.3 as well as for $N = 30$ and $N = 40$ observed data points. For $N = 20$ data points, the model with three sinusoids is the most probable, but it is not much more probable than the model with two sinusoids. Increasing the number observations also increases the posterior probability for the model with three sinusoids.

If we were not willing to assign zero probability to all other models but the four of figure 2.3 a priori, we would not be able to derive the posterior probabilities for the four models directly. Instead, we could in this case use posterior odds for comparing the four models.

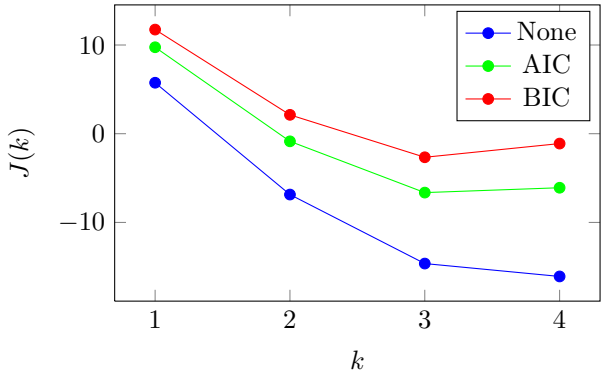

Figure 2.4: Different criteria for selecting between the alternative models in figure 2.3 using tools from classical statistics.

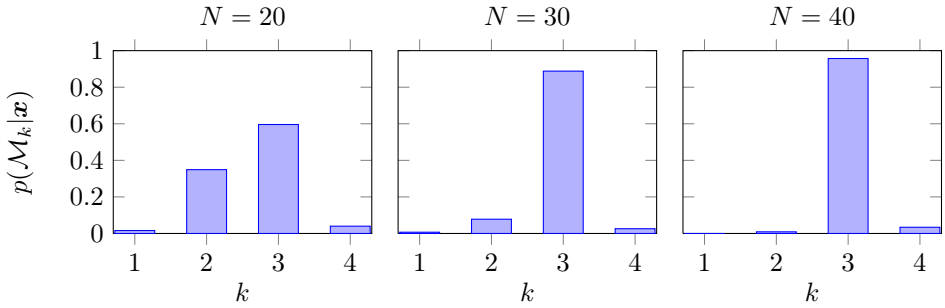

Figure 2.5: Posterior distributions for the number of sinusoids in the case of the $N = 20$ observations in figure 2.3 as well as for $N = 30$ and $N = 40$ observations.

2.5 Summary

In this chapter, we have revised the fundamentals of Bayesian inference and thinking. As demonstrated through several examples, the Bayesian method basically boils down to applying Bayes' theorem on the inference problem whose answer is given in terms of a posterior probability distribution. This distribution encapsulates and combines in an optimal way the information gained from observing some data with the prior knowledge of the unknown quantities, and it can be used for deriving statistics such as point and interval estimates. There is no reason to search for other estimators or to use performance bounds for the statistics since the Bayesian method offers the optimal answer to the inference problem. This is the key strength of Bayesian statistics. The major drawback of Bayesian statistics is that, except for the simple cases such as the problems considered in the examples of this chapter, it is very hard or even impossible to find the posterior distribution or its moments in closed form. The remedy for this problem is to use various numerical techniques. We investigate some of these in the next chapter.

Chapter 3

Numerical Bayesian Inference

In chapter 2, we gave an introduction to the theoretical basis of Bayesian inference and signal processing and demonstrated its applicability through several examples. In all of these examples, we used Bayes' theorem for deriving an analytical expression for the posterior distribution of the interesting parameters or models. Unfortunately, it is not possible in general to derive an analytical expression for the posterior distribution and, as already discussed in the introduction, this has been the main drawback of the Bayesian approach to statistical inference. This chapter discusses some of the solutions to the practical problems that arise when analytical Bayesian inference is not feasible or impossible. This is caused by the requirement to evaluate integrals needed for, e.g., computing expectations and for marginalisation of nuisance parameters. The main difficulties, that prohibit analytical evaluation of these integrals, are highly complex distributions, lack of closed-form solutions, and high dimensionality of the parameter space [Bishop, 2006, pp. 461-462]. Therefore, alternative *approximate* methods have been proposed and investigated in recent years, and they are generally partitioned into two groups:

1. *Deterministic methods* which are based on analytical approximations by assuming a particular parametric form or factorisation of the true distribution. For example, a complex posterior distribution can be approximated by fitting another distribution to it from a subset of simpler distributions by using various optimisation techniques. This is known as *variational Bayesian inference.*

2. *Stochastics methods* in which samples are drawn directly or indirectly from the true posterior distribution. These samples then form a histogram which converges to the true posterior distribution in the limit of infinitely many samples.

The deterministic methods never generate exact results but only approximate results. This is in contrast to stochastic methods which converge to the exact result for an

increasing amount of samples. On the other hand, the stochastic methods entail a high computational complexity whereas the deterministic methods typically are more efficient - especially for large and complex problems. In this thesis, we primarily focus on the stochastic methods.

There exist several stochastic methods for approximate Bayesian inference with the most dominating being the *Monte Carlo* methods and in particular an important subclass thereof known as *Markov chain Monte Carlo* (MCMC) methods. These methods draw, say, T numerical samples $\boldsymbol{y}^{[\tau]}$, for $\tau = 1, \cdots, T$, from a distribution $p(\boldsymbol{y})$ in order to evaluate the integral of the form [Bishop, 2006, p. 524]

$$E\{f(\boldsymbol{y})\} = \int f(\boldsymbol{y})p(\boldsymbol{y})d\boldsymbol{y} \tag{3.1}$$

by the approximation

$$\hat{f}(\boldsymbol{y}) = \frac{1}{T}\sum_{\tau=1}^{T} f(\boldsymbol{y}^{[\tau]}) \ . \tag{3.2}$$

Notice, that the form of Eq. (3.1) is very general and can represent marginalisation or expectation problems which are often encountered in Bayesian inference as already alluded in the introduction of this chapter. For example, if $p(\boldsymbol{y})$ is the posterior distribution and $f(\boldsymbol{y}) = \boldsymbol{y}$, then $\hat{f}(\boldsymbol{y}) = \hat{\boldsymbol{y}}$ is the MMSE estimator. The approximation is a classic unbiased and consistent estimator. Thus, the expected value of the estimator is the true value, and the variance of the estimator decreases and approaches 0 as the sample size T increases. If we therefore took an infinite sample size, there would not be anything approximate about Monte Carlo methods. In practice, however, we have to work with a finite sample size due to computational limitations.

3.1 Basic Sampling Techniques

A significant challenge of the Monte Carlo methods is that of drawing samples from some known distribution. First of all, the samples must be truly random which is impossible on a digital computer that produces pseudo-random numbers. Secondly, the samples must be drawn according to some known distribution which can take on a large number of forms, and we cannot in general assume that there exist an algorithm for that on the computer. In our treatment of Markov chain Monte Carlo inference, we will therefore begin with an investigation of two basic methods for drawing samples from some known distribution.

Our discussion on the basic sampling techniques will be based on the assumptions that we have access to an algorithm that can produce a random variate u sampled from a uniform distributed on the interval $[0, 1]$. We also restrict ourselves to the univariate case leaving the more complicated multivariate case to the Markov chain Monte Carlo

sampling. This is justified by the fact that the basic sampling techniques are not suitable for performing sampling in high dimensional problems [Bishop, 2006, p. 537].

3.1.1 Inverse Transform Sampling

In inverse transform sampling, the uniformly distributed random variable U is mapped to the random variable Y through the relation $Y = g(U)$. The mapping is such that Y has the desired distribution $p(y)$. In order to establish this mapping between the random variables U and Y, we use the probability integral transform theorem [Devroye, 1986, p. 28]

$$U = F(Y) = \int_{-\infty}^{Y} p(\eta) d\eta \tag{3.3}$$

where η is an integration variable. The theorem states that if Y is distributed according to the distribution $p(y)$ and has the cumulative distribution function $F(y)$, then $F(Y)$ is uniformly distributed on the interval $[0, 1]$. Thus, the inverse of the cumulative distribution function constitutes the desired mapping and we have that $Y = F^{-1}(U) = g(U)$[1]. Unfortunately, it is not always possible to find a closed-form solution for $g(U)$ from Eq. (3.3). Therefore, we have to resort to alternative techniques which we describe next.

3.1.2 Rejection Sampling

We assume that the desired distribution $p(y)$ is so complicated that we cannot use inverse transform sampling for drawing samples from it. Instead, we use a simpler *proposal* distribution $q(y)$ which satisfies $kq(y) \geq p(y)$ for all random variates y of non-zero probability and which is feasible to sample from using, e.g., inverse transform sampling. The algorithm is simple and summarised in algorithm 3.1.

Algorithm 3.1 (Rejection Sampling)

1. Select a proposal distribution $q(y)$, which is feasible to sample from, satisfying $kq(y) \geq p(y)$ for every y in the support for the desired distribution $p(y)$.
2. Repeat for $\tau = 1, 2, 3, \cdots, T$
   (a) Repeat until sample is accepted
      i. Draw a candidate sample y^* from the proposal distribution $q(y)$.
      ii. Draw a random variate u from the uniform distribution $\mathcal{U}(0, 1)$.

[1] A table of various functions $g(\cdot)$ for different desired distributions can be found in [Devroye, 1986, p. 29].

iii. Accept y^* as a sample from $p(y)$ if

$$u < \frac{p(y^*)}{kq(y^*)} \ .$$

Otherwise reject it.

(b) Set $y^{[\tau]} = y^*$.

Figure 3.1 illustrates the idea of rejection sampling. We sample from the upper curve $q(y)$ and if that sample falls within the shaded area, we reject it. Since the purpose is to generate accepted samples, we want to decrease the probability of rejecting a sample. Intuitively, this is done by selecting a $q(y)$ such that the shaded area is as small as possible. The *acceptance ratio* is a measure of the expected proportion of samples that are accepted, i.e.,

$$\eta_{\mathrm{A}} = \int \frac{p(y)}{kq(y)} q(y) dy = \frac{1}{k} \int p(y) dy = \frac{1}{k} \ . \tag{3.5}$$

Thus, since $k > 1$, maximising the acceptance ratio is equivalent to minimising k subject to the constraint that $kq(y) \geq p(y)$.

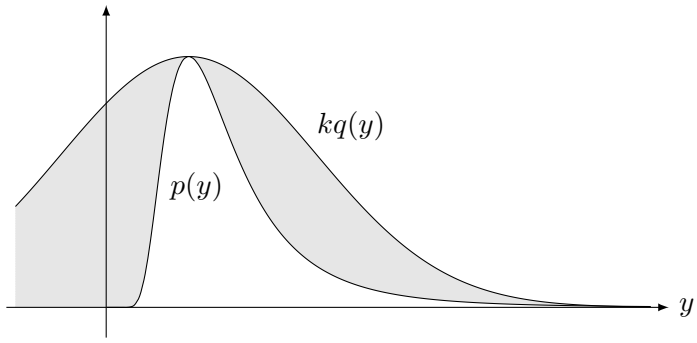

Figure 3.1: Illustration of rejection sampling which is used for drawing samples from the distribution $p(y)$ by sampling from the proposal distribution $q(y)$ satisfying $kq(y) \geq p(y)$. The shaded area to the total area under $kq(y)$ indicates the proportion of rejected samples generated from $q(y)$.

3.2 Markov Chain Monte Carlo Sampling

The basic sampling methods described above are very useful for problems involving distributions of low dimension, but for higher dimensional distributions they suffer from

severe limitations [Bishop, 2006, p. 537]. Markov chain Monte Carlo (MCMC) sampling, however, has proven to be very versatile regardless of the dimension of the problem, and it is therefore considered here. In MCMC sampling, samples are drawn from the desired distribution $p(\boldsymbol{y})$ by forming a Markov chain whose stationary distribution converges to $p(\boldsymbol{y})$. The most fundamental MCMC sampling technique is the *Metropolis-Hastings (MH)* algorithm which resembles rejection sampling since samples are drawn from a proposal distribution and accepted or rejected. An important special case of the MH-algorithm is *Gibbs* sampling which we also consider. First, however, we give a short review of Markov chains.

3.2.1 Markov Chains

A Markov random sequence is a random sequence whose samples satisfy the property that, given the current state $\boldsymbol{y}^{[\tau]}$, the future states are independent of the past states, i.e., [Stark and Woods, 2001, p. 362]

$$p(\boldsymbol{y}^{[\tau+k]}|\boldsymbol{y}^{[\tau]}, \boldsymbol{y}^{[\tau-1]}, \cdots, \boldsymbol{y}^{[1]}) = p(\boldsymbol{y}^{[\tau+k]}|\boldsymbol{y}^{[\tau]}) \tag{3.6}$$

for some integers $k > 0$ and $\tau > 0$. This property is called the Markov property. A special case is the one-step case where $k = 1$ which we restrict ourselves to. If the states $\{\boldsymbol{y}^{[1]}, \cdots, \boldsymbol{y}^{[\tau+1]}\}$ are discrete-valued, then the Markov random sequence is referred to as a Markov chain [Kay, 2005, p. 739]; however, the term is also sometimes used for describing a Markov random process with a continuous state space.

The conditional distribution $p(\boldsymbol{y}^{[\tau+1]}|\boldsymbol{y}^{[\tau]})$ on the right hand side of Eq. (3.6) is for a continuous state space called the *transition probability kernel* and denoted as $T(\boldsymbol{y}^{[\tau+1]}|\boldsymbol{y}^{[\tau]})$. If the transition kernel does not change as a function of τ, the Markov chain is *homogeneous* [Bishop, 2006, p. 540]. In the rest of this section, we assume that the Markov chain is homogeneous and use the simpler notation $T(\boldsymbol{y}'; \boldsymbol{y})$ for the transition kernel where $\boldsymbol{y}'$ and $\boldsymbol{y}$ are generic variables representing the future state and the current state, respectively. The marginal distribution of the current state $p^{[\tau]}(\boldsymbol{y})$ is called the *state probability* and is a function of τ. A homogeneous Markov chain is fully specified by its initial state distribution and transition kernel.

The transition kernel relates two successive state distributions by

$$p^{[\tau+1]}(\boldsymbol{y}') = \int T(\boldsymbol{y}'; \boldsymbol{y}) p^{[\tau]}(\boldsymbol{y}) d\boldsymbol{y} \tag{3.7}$$

which is known as the *Chapman-Kolmogorov* equation [Stark and Woods, 2001, pp. 429-430]. Using this equation, we can find the state distribution at any time τ. In many applications involving Markov chains, the problem is to determine the *stationary* (or invariant) distribution $\pi(\boldsymbol{y})$, if it exists, which $p^{[\tau]}(\boldsymbol{y})$ converges towards as τ increases, i.e.,

$$\pi(\boldsymbol{y}) = \lim_{\tau \to \infty} p^{[\tau]}(\boldsymbol{y}) \ . \tag{3.8}$$

For this distribution, the Chapman-Kolmogorov equation yields

$$\pi(\boldsymbol{y}') = \int T(\boldsymbol{y}';\boldsymbol{y})\pi(\boldsymbol{y})d\boldsymbol{y} \tag{3.9}$$

This equation may have zero, one or more solutions. If the Markov chain is *reversible*, it satisfies the *detailed balance* property given by

$$p^{[\tau]}(\boldsymbol{y})T(\boldsymbol{y}';\boldsymbol{y}) = p^{[\tau]}(\boldsymbol{y}')T(\boldsymbol{y};\boldsymbol{y}') \ . \tag{3.10}$$

The detailed balance property is illustrated in figure 3.2 for a discrete state space and univariate variable y. In the figure, y and y' represents any two states that the samples $y^{[\tau]}$ and $y^{[\tau+1]}$ can be in. Reversibility of a Markov chain is a sufficient, but not necessary, condition guaranteeing that $p^{[\tau]}(\boldsymbol{y}) = \pi(\boldsymbol{y})$ is a stationary distribution for the Markov chain [Bishop, 2006, p. 540]. Finally, a Markov chain is *ergodic* if Eq. (3.8) is satisfied for any initial state distribution. This means, that the Markov chain has only one stationary distribution which the chain converges to from any initial state distribution.

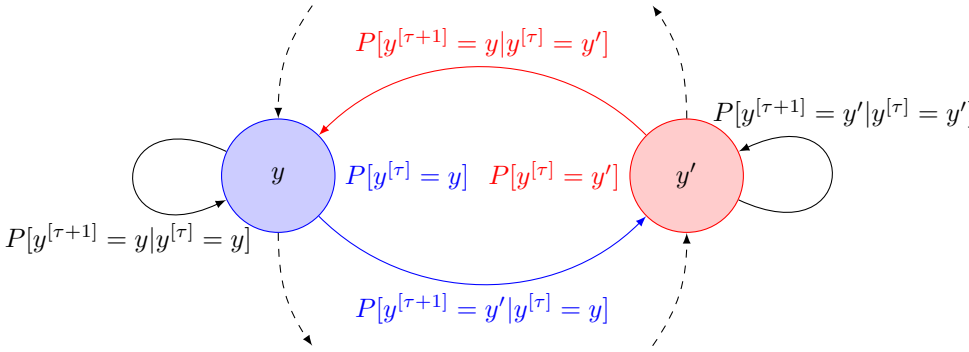

Figure 3.2: The property of detailed balance in a Markov chain with a finite number of discrete states. When the product of the red probabilities equals the product of the blue probabilities for every pair of states, the detailed balance property is fulfilled and stationarity of the Markov chain is ensured. The dotted lines indicate that the shown states are a part of a larger set of states.

3.2.2 The Metropolis-Hastings Algorithm

The Metropolis-Hastings (MH) algorithm is the most general algorithm for performing MCMC sampling, and other MCMC methods can be considered as a special case thereof [Liu, 2002, p. 105]. It requires that we know the posterior distribution $p(\boldsymbol{y})$, that we wish to draw samples from, up to some unknown *normalisation* constant Z_p, i.e., $p(\boldsymbol{y}) = \tilde{p}(\boldsymbol{y})/Z_p$, and that we have some proposal transition kernel $Q(\boldsymbol{y}';\boldsymbol{y})$ from which

we can easily draw samples. The MH-algorithm constructs, in a simple manner, a Markov chain whose stationary state distribution is $p(\boldsymbol{y})$. In order to explain why the algorithm works, we use figure 3.2 even though it only illustrate the case for a univariate variable y and a discrete state space. However, this special case is much easier to illustrate.

The challenge in Markov chain theory is typically to determine, given the transition probability kernel, the stationary distribution that the sequence of state distributions converges to in case of convergence. Here, we face the opposite problem; we want to determine the transition probabilities such that the sequence of state distributions converges to the desired distribution $p(\boldsymbol{y})$ that we know up to some normalisation constant. From the detailed balance property, we know that if we can find a transition kernel that fulfils the detailed balance property in Eq. (3.10), then convergence to the desired distribution is ensured, i.e., in terms of figure 3.2, the probability of being in state $\boldsymbol{y}$ and then moving from state $\boldsymbol{y}$ to $\boldsymbol{y}'$ is the same as being in state $\boldsymbol{y}'$ and then moving from state $\boldsymbol{y}'$ to $\boldsymbol{y}$. If we propose some random proposal transition kernel $Q(\boldsymbol{y}';\boldsymbol{y})$, that we draw samples from[2], then the detailed balance property is most likely not fulfilled, and we observe that we are, e.g., moving more often from state $\boldsymbol{y}$ to $\boldsymbol{y}'$. This can be written as

$$p(\boldsymbol{y})Q(\boldsymbol{y}';\boldsymbol{y}) > p(\boldsymbol{y}')Q(\boldsymbol{y};\boldsymbol{y}') \ . \tag{3.11}$$

In order to satisfy the detailed balance property, we are therefore forced to reduce the number of moves from state $\boldsymbol{y}$ to $\boldsymbol{y}'$. This can be obtained by introducing the new bivariate distribution $\alpha(\boldsymbol{y}', \boldsymbol{y})$ yielding

$$\alpha(\boldsymbol{y}', \boldsymbol{y})p(\boldsymbol{y})Q(\boldsymbol{y}';\boldsymbol{y}) = p(\boldsymbol{y}')Q(\boldsymbol{y};\boldsymbol{y}') \tag{3.12}$$

and solving for $\alpha(\boldsymbol{y}', \boldsymbol{y})$ yields

$$\alpha(\boldsymbol{y}, \boldsymbol{y}') = \frac{p(\boldsymbol{y}')Q(\boldsymbol{y};\boldsymbol{y}')}{p(\boldsymbol{y})Q(\boldsymbol{y}';\boldsymbol{y})} \ . \tag{3.13}$$

In general, we cannot rely on that we select $Q(\boldsymbol{y}';\boldsymbol{y})$ such that we move more often from state $\boldsymbol{y}$ to $\boldsymbol{y}'$ as we assumed in the inequality given by Eq. (3.11). Nevertheless, we evaluate $\alpha(\boldsymbol{y}', \boldsymbol{y})$ in Eq. (3.13) when we are trying to move from state $\boldsymbol{y}$ to $\boldsymbol{y}'$ and take the following action:

- If $\alpha(\boldsymbol{y}', \boldsymbol{y})$ turns out to be *greater than or equal to* one, we make the move with probability one since we are in a situation where it is more probable to move from state $\boldsymbol{y}'$ to $\boldsymbol{y}$ than from $\boldsymbol{y}$ to $\boldsymbol{y}'$.

[2]Recall that the transition kernel is just a conditional distribution dependent on the current state. We therefore assume that we can draw samples from it using one of the basic sampling algorithm described in section 3.1.

- If $\alpha(\boldsymbol{y}', \boldsymbol{y})$ turns out to be *less than* one, we make the move with probability $\alpha(\boldsymbol{y}', \boldsymbol{y})$ since we are in a situation where it is more probable to move from state $\boldsymbol{y}$ to $\boldsymbol{y}'$ than from $\boldsymbol{y}'$ to $\boldsymbol{y}$.

Thus, the general expression for $\alpha(\boldsymbol{y}', \boldsymbol{y})$ can be summarised as

$$\alpha(\boldsymbol{y}', \boldsymbol{y}) = \min\left[1, \frac{\tilde{p}(\boldsymbol{y}')Q(\boldsymbol{y}; \boldsymbol{y}')}{\tilde{p}(\boldsymbol{y})Q(\boldsymbol{y}'; \boldsymbol{y})}\right] . \tag{3.14}$$

where we have used the fact that we only know the desired distribution up to some normalisation constant.

This concludes the derivation of the Metropolis-Hastings algorithm. A more thorough introduction to it is given in [Chib and Greenberg, 1995]. The algorithm for the Metropolis-Hastings algorithm is outlined in algorithm 3.2.

Algorithm 3.2 (Metropolis-Hastings Algorithm)

1. Initialise $\boldsymbol{y}^{[0]}$.
2. Repeat for $\tau = 0, 1, 2, \cdots, T$
   (a) Draw a candidate sample $\boldsymbol{y}^*$ from the proposal transition kernel $Q(\boldsymbol{y}; \boldsymbol{y}^{[\tau]})$.
   (b) Evaluate the probability of move given by
   $$\alpha(\boldsymbol{y}^*, \boldsymbol{y}^{[\tau]}) = \min\left[1, \frac{\tilde{p}(\boldsymbol{y}^*)Q(\boldsymbol{y}^{[\tau]}; \boldsymbol{y}^*)}{\tilde{p}(\boldsymbol{y}^{[\tau]})Q(\boldsymbol{y}^*; \boldsymbol{y}^{[\tau]})}\right]$$
   where $\tilde{p}(\boldsymbol{y})$ is the unnormalised desired distribution.
   (c) Draw a random variate $u^{[\tau]}$ from the univariate uniform distribution $\mathcal{U}(0, 1)$.
   (d) If $u^{[\tau]} \leq \alpha(\boldsymbol{y}^*, \boldsymbol{y}^{[\tau]})$, then accept the sample $\boldsymbol{y}^*$ as a sample from $p(\boldsymbol{y})$ and set $\boldsymbol{y}^{[\tau+1]} = \boldsymbol{y}^*$. Otherwise reject $\boldsymbol{y}^*$ as a sample from $p(\boldsymbol{y})$ and set $\boldsymbol{y}^{[\tau+1]} = \boldsymbol{y}^{[\tau]}$.

The outlined algorithm does not put any restrictions on the choice of proposal transition kernel $Q(\boldsymbol{y}'; \boldsymbol{y})$. No matter the choice of proposal kernel, the MH algorithm guarantees that the desired distribution $p(\boldsymbol{y})$ is the stationary distribution of the Markov chain. The convergence rate, however, is affected by the choice of proposal distribution. A common choice is a multivariate Gaussian centred on the current state and with variance selected as a trade-off between correlation time and rejection rate [Bishop, 2006,

pp. 541-542]. Due to the convergence time, the first samples cannot be seen as samples from the stationary distribution of the underlying Markov chain and are thus discarded. These initial samples are referred to as burn-in samples and the convergence time is referred to as burn-in time. The MH-algorithm is demonstrated in example 3.1.

Example 3.1 (Bayesian Inference using the Metropolis-Hastings Algorithm) In this example, we again consider the sinusoidal model introduced in example 2.1. For illustrative purposes, the signal in this example only consists of one sinusoid with known frequency and unknown amplitude. The noise variance is also unknown so the unknown model parameters are $\boldsymbol{\theta} = \begin{bmatrix}\alpha & \sigma_w^2\end{bmatrix}^T$. The prior distributions for these model parameters are the same as in example 2.5, i.e., $p(\boldsymbol{\theta}) = p(\alpha, \sigma_w^2) = \mathcal{N}(\alpha; \mu_\alpha, \sigma_w^2 c_\alpha)\text{Inv-}\mathcal{G}(\sigma_w^2, a, b)$. We consider the problem of making inferences for the joint posterior distributions for the amplitude and noise variance, i.e., $p(\boldsymbol{\theta}|\boldsymbol{x})$. It is given by the factorisation

$$p(\boldsymbol{\theta}|\boldsymbol{x}) \propto p(\boldsymbol{x}, \boldsymbol{\theta}) = p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{3.15}$$

$$= \mathcal{N}(\boldsymbol{x}; \boldsymbol{A}\alpha, \sigma_w^2)\mathcal{N}(\alpha; \mu_\alpha, \sigma_w^2 c_\alpha)\text{Inv-}\mathcal{G}(\sigma_w^2; a, b) \ . \tag{3.16}$$

Therefore, we can easily evaluate $p(\boldsymbol{\theta}|\boldsymbol{x})$ up to a normalisation factor as required by the MH-algorithm.

Suppose we are given a data set of $N = 20$ observations with a true, but unknown amplitude and noise variance of $\alpha = 1$ and $\sigma_w^2 = 0.5$, respectively. In order to apply the MH-algorithm on these data, we have to specify a proposal transition kernel $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{[\tau]})$, from which we can easily draw samples, and we have to make and initial guess of the model parameters $\boldsymbol{\theta}$. The proposal transition kernel is selected to be a bivariate Gaussian distribution centred on the current state with isotropic covariance with the variance term $\sigma^2 = 0.1$, i.e.,

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{[\tau]}) = \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\theta}^{[\tau]}, \sigma^2 \boldsymbol{I}_2) \ . \tag{3.17}$$

The initial state of the Markov chain was set to $\boldsymbol{\theta}^{[0]} = \begin{bmatrix}2 & 0.1\end{bmatrix}^T$. $T = 100,000$ samples were drawn using the MH-algorithm in algorithm 3.2, and the results are shown in figure 3.3.

The middle plot shows three contours of the Gaussian-inverted gamma distribution and the trace of the first 100 samples generated by the MH-algorithm. The proposed samples, which were rejected, are shown as red dotted lines. Starting from the initial state of the Markov chain $\boldsymbol{\theta}^{[0]}$, a proposed move to $\boldsymbol{\theta}^* = \begin{bmatrix}1.98 & 0.04\end{bmatrix}^T$ is rejected. Therefore, the state $\boldsymbol{\theta}^{[1]}$ of the Markov chain equals the initial state $\boldsymbol{\theta}^{[0]}$. The next proposed move is accepted and the Markov chain jumps to the state $\boldsymbol{\theta}^{[2]}$. In the figure, we see the trace of moves until the state given by $\boldsymbol{\theta}^{[100]}$. The green circles centred on this state show the 0.1 and 0.9 probability regions of the proposal transition kernel

$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{[100]})$. Thus, with a probability of 0.1 the next proposed move lies inside the inner circle and inside the outer circle with a probability of 0.9. This illustrates that the successive samples generated by the MH-algorithm are correlated and that the degree of correlation depends on the variance of the proposal distribution. However, if we increase the variance of the proposal, a larger fraction of proposed moves are rejected. The key property of a good proposal distribution is therefore to find a distribution which results in a good trade-off between acceptance ratio and correlation time.

The top and right plots of figure 3.3 show the (analytical) marginal distributions and histograms for the amplitude and noise variance, respectively. The histograms are computed from all of the 100,000 samples except for the burn-in samples. The burn-in length was set to 100. We see that the histograms coincide with the (analytical) marginal distributions so the samples obtained using the MH-algorithm are indeed samples from the desired joint and marginal posterior distributions.

3.2.3 The Gibbs Sampler

Gibbs sampler is the most widely used MCMC method, and it can be seen as a special case of the Metropolis-Hastings algorithm. It partitions the multivariate sampling from the desired distribution $p(\boldsymbol{y}) = p(\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_k)$ into lower dimensional sampling of the k conditional distributions given by

$$
\begin{aligned}
&p(\boldsymbol{y}_1|\boldsymbol{y}_2, \cdots, \boldsymbol{y}_k) \\
&p(\boldsymbol{y}_2|\boldsymbol{y}_1, \boldsymbol{y}_3 \cdots, \boldsymbol{y}_k) \\
&\quad \vdots \\
&p(\boldsymbol{y}_i|\boldsymbol{y}_1, \cdots, \boldsymbol{y}_{i-1}, \boldsymbol{y}_{i+1}, \cdots, \boldsymbol{y}_k) \\
&\quad \vdots \\
&p(\boldsymbol{y}_k|\boldsymbol{y}_1, \cdots, \boldsymbol{y}_{k-1}) \ .
\end{aligned}
$$

The conversion from a distribution of high dimensionality into a series of distribution of lower dimensionality enables the use of the basic sampling methods described in section 3.1.

At each iteration of the Gibbs sampler, we cycle through the set of conditional distributions and draw one sample from each. When a sample is drawn from one conditional distribution, the succeeding distributions are updated with the new value of that sample.

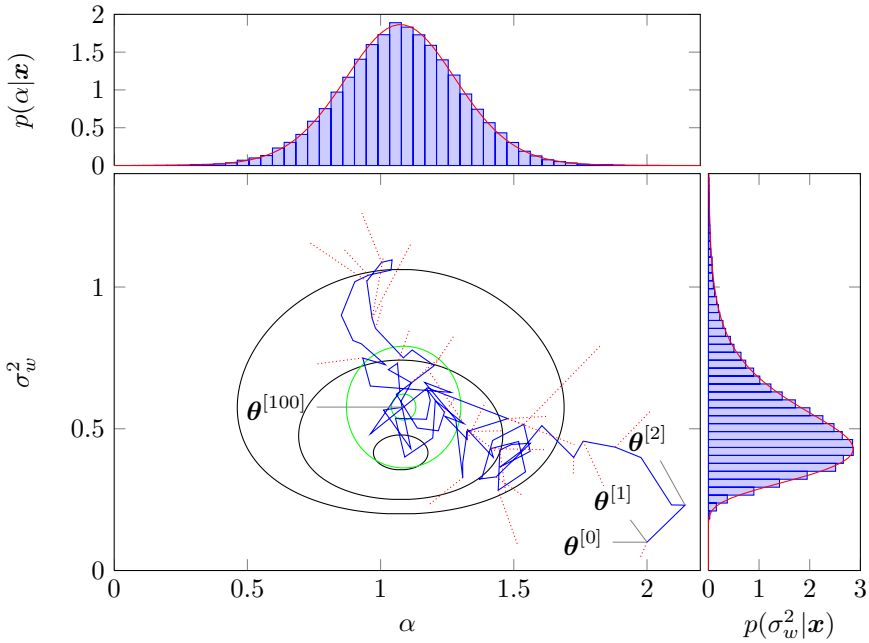

Figure 3.3: Illustration of the MH-algorithm for drawing samples from a Gaussian-inverted gamma distribution. The trace of the first 100 samples (blue curve) are shown on top of contour plots (black) of the Gaussian-inverted gamma distribution. The red dotted lines indicate the rejected samples. The green circles centred on $\boldsymbol{\theta}^{[100]}$ are the boundaries of the 0.1 and 0.9 probability regions of the proposal transition kernel. The plots in the margin show the analytical marginal distributions as well as the histograms based on 100,000 samples.

This yields the following drawings at the τ'th iteration

$$
\begin{aligned}
\boldsymbol{y}_1^{[\tau+1]} &\sim p(\boldsymbol{y}_1|\boldsymbol{y}_2^{[\tau]}, \cdots, \boldsymbol{y}_k^{[\tau]}) \\
\boldsymbol{y}_2^{[\tau+1]} &\sim p(\boldsymbol{y}_2|\boldsymbol{y}_1^{[\tau+1]}, \boldsymbol{y}_3^{[\tau]}, \cdots, \boldsymbol{y}_k^{[\tau]}) \\
&\vdots \\
\boldsymbol{y}_i^{[\tau+1]} &\sim p(\boldsymbol{y}_i|\boldsymbol{y}_1^{[\tau+1]}, \cdots, \boldsymbol{y}_{i-1}^{[\tau+1]}, \boldsymbol{y}_{i+1}^{[\tau]}, \cdots, \boldsymbol{y}_k^{[\tau]}) \\
&\vdots \\
\boldsymbol{y}_k^{[\tau+1]} &\sim p(\boldsymbol{y}_k|\boldsymbol{y}_1^{[\tau+1]}, \cdots, \boldsymbol{y}_{k-1}^{[\tau+1]}) \ .
\end{aligned}
$$

Since we sample from the desired distribution, the detailed balance property is fulfilled and we do not have to reject any samples. Like the Metropolis-Hastings algorithm,

however, the Markov chain of the Gibbs sampler needs some initial transient period to converge to the desired stationary distribution. The Gibbs sampler is outlined in algorithm 3.3 and demonstrated in example 3.2.

Algorithm 3.3 (Gibbs Sampler)

1. Determine expressions for the conditional distributions

$$p(\boldsymbol{y}_i|\boldsymbol{y}_1, \cdots, \boldsymbol{y}_{i-1}, \boldsymbol{y}_{i+1}, \cdots, \boldsymbol{y}_k)$$

   of the unknown variables for $i = 1, \cdots, k$.

2. Initialise $\boldsymbol{y}_i^{[0]}$ for $i = 2, \cdots, k$.

3. Repeat for $\tau = 0, 1, 2, \cdots, T$

   (a) Draw samples from the conditional distributions in an alternating pattern. Once a new sample is drawn, it is immediately substituted into the future conditional distributions

$$\begin{aligned}
\boldsymbol{y}_1^{[\tau+1]} &\sim p(\boldsymbol{y}_1|\boldsymbol{y}_2^{[\tau]}, \cdots, \boldsymbol{y}_k^{[\tau]}) \\
\boldsymbol{y}_2^{[\tau+1]} &\sim p(\boldsymbol{y}_2|\boldsymbol{y}_1^{[\tau+1]}, \boldsymbol{y}_3^{[\tau]}, \cdots, \boldsymbol{y}_k^{[\tau]}) \\
&\vdots \\
\boldsymbol{y}_k^{[\tau+1]} &\sim p(\boldsymbol{y}_k|\boldsymbol{y}_1^{[\tau+1]}, \cdots, \boldsymbol{y}_{k-1}^{[\tau+1]}) \ .
\end{aligned}$$

Example 3.2 (Bayesian Inference using the Gibbs Sampler)
Consider the same setup as in example 3.1. In this example, we use the Gibbs sampler for drawing samples from the joint posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{x})$. Unlike the MH-algorithm, we do not have to specify a proposal distribution. Instead, we have to derive expressions for the conditional distributions $p(\alpha|\sigma_w^2, \boldsymbol{x})$ and $p(\sigma_w^2|\alpha, \boldsymbol{x})$. Using result B.1, B.3 and B.4, we obtain from Eq. (3.16) that

$$p(\alpha|\sigma_w^2, \boldsymbol{x}) = \mathcal{N}\Big(\alpha; (c_\alpha^{-1} + \boldsymbol{A}^T\boldsymbol{A})^{-1}(\boldsymbol{A}^T\boldsymbol{x} + c_\alpha^{-1}\mu_\alpha), \sigma_w^2(c_\alpha^{-1} + \boldsymbol{A}^T\boldsymbol{A})^{-1}\Big) \tag{3.19}$$

$$p(\sigma_w^2|\alpha, \boldsymbol{x}) = \text{Inv-}\mathcal{G}\Big(\sigma_w^2; a + \frac{N+1}{2}, b + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{A}\alpha)^T(\boldsymbol{x} - \boldsymbol{A}\alpha) + \frac{(\alpha - \mu_\alpha)^2}{2c_\alpha}\Big) \tag{3.20}$$

which we are able to sample from. Each iteration of the Gibbs sampler in algorithm 3.3 involves taking a sample from these two distributions given the previous samples. If

we use the same initial state $\boldsymbol{\theta}^{[0]}$ as in example 3.1, we therefore first draw a sample from $p(\alpha|\sigma_w^{2\,[0]}, \boldsymbol{x})$. This is shown in figure 3.4 as the horizontal move from $\boldsymbol{\theta}^{[0]}$ to the intermediate state $\begin{bmatrix}\alpha^{[1]} & \sigma_w^{2\,[0]}\end{bmatrix}^T$. The second step of the first iteration of Gibbs sampler is to draw a sample from $p(\sigma_w^2|\alpha^{[1]}, \boldsymbol{x})$. This corresponds to the vertical move from the intermediate state to $\boldsymbol{\theta}^{[1]}$. The second iteration also consists of a horizontal and vertical move and ends in $\boldsymbol{\theta}^{[2]}$. The figure shows the trace of the first 25 of such moves along with the histograms based on 100,000 samples with the first 100 samples removed as burn-in samples.

The clear advantages of the Gibbs sampler as compared against the MH-algorithm is that no proposal distribution is required, that no samples are rejected, and that we can break the sampling process down to a series of lower dimensional samplings processes. The latter often reduces the computational complexity significantly. Unfortunately, the conditional distribution used in the Gibbs sampler are not always easy to sample from. In this case, the MH-algorithm can be used within the Gibbs sampler for drawing samples from the intractable distribution [Gelman et al., 2003, p. 292]. Later in this thesis, we make use of this hybrid sampling scheme.

3.3 Model Selection

A significant challenge in Bayesian inference is to compute the value of the normalisation constant Z_p relating the normalised distribution $p(\boldsymbol{y})$ by the unnormalised distribution $\tilde{p}(\boldsymbol{y})$ through

$$p(\boldsymbol{y}) = \frac{1}{Z_p}\tilde{p}(\boldsymbol{y}) \ . \tag{3.21}$$

Integrating both sides of this equation with respect to $\boldsymbol{y}$ readily yields that the normalisation constant is given by

$$Z_p = \int \tilde{p}(\boldsymbol{y}) d\boldsymbol{y} \ . \tag{3.22}$$

In applications such as model selection, as already encountered in section 2.4, it is important to compute the value of Z_p since it represents the model evidence $p(\mathcal{D}|\mathcal{M}_k)$. The unnormalised distribution $\tilde{p}(\boldsymbol{y})$ is in this case the product of the likelihood and the prior distribution for the parameters. Thus, Eq. (3.22) is for the case of model selection given by

$$p(\mathcal{D}|\mathcal{M}_k) = \int p(\mathcal{D}|\vartheta_k, \mathcal{M}_k) p(\vartheta_k|\mathcal{M}_k) d\vartheta_k \ . \tag{3.23}$$

When it is not possible to compute the normalisation constant using analytical tools, the numerical techniques discussed in this chapter can be used. This can be seen by comparing Eq. (3.23) with Eq. (3.1). These two equations have the same form so the basic

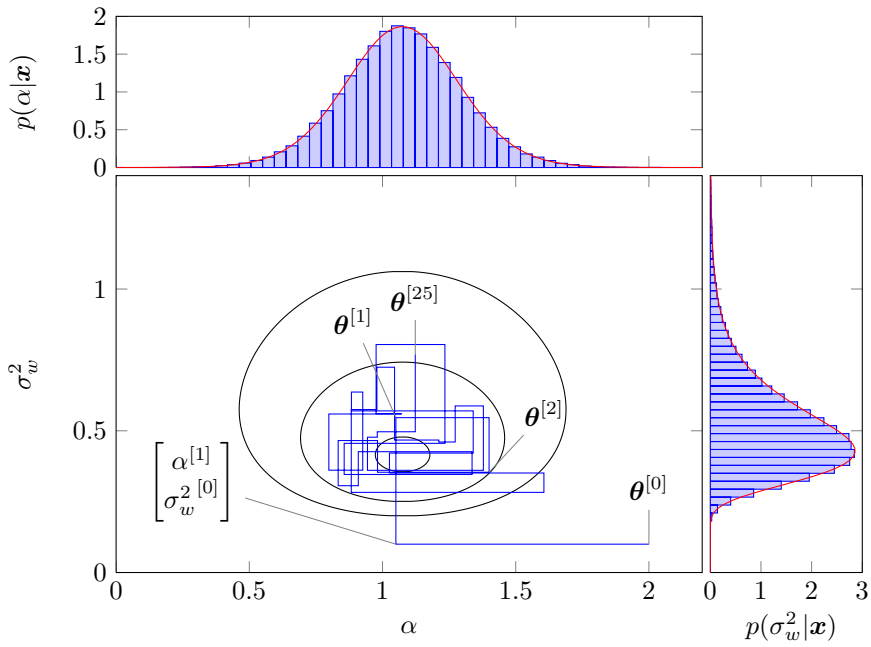

Figure 3.4: Illustration of drawing samples from a Gaussian-inverted gamma distribution using the Gibbs sampler. The trace of the first 25 samples (blue curve) are shown on top of contour plots (black) of the Gaussian-inverted gamma distribution. The plots in the margin show the analytical marginal distributions as well as the histograms based on 100,000 samples.

sampling techniques as well as MCMC-based sampling techniques can be used for the computation of the model evidence $p(\mathcal{D}|\mathcal{M}_k)$. Special MCMC-based algorithms have also been introduced for computing the model evidence. Two of these are the Chib's algorithm [Chib, 1995], which is based on the Gibbs sampler, and the reversible jump MCMC (RJMCMC) [Green, 1995]. It is beyond the scope of this thesis to present these methods. Instead, we consider two simple approximate ways of computing the normalisation constant: *The Laplace approximation* and the *Bayesian information criterion* (BIC).

3.3.1 The Laplace Approximation

The Laplace approximation is based on approximating $p(\boldsymbol{y})$ by a Gaussian distribution $q(\boldsymbol{y})$ centred on a mode of $p(\boldsymbol{y})$. We denote this mode as $\boldsymbol{y}_0$. The covariance of the Gaussian distribution is found by using a second order Taylor expansion of the logarithm of the unnormalised distribution $\tilde{p}(\boldsymbol{y})$. This yields [Petersen and Pedersen, 2008, p. 58]

$$\ln \tilde{p}(\boldsymbol{y}) \approx \ln \tilde{p}(\boldsymbol{y}_0) + \boldsymbol{g}^T(\boldsymbol{y}_0)(\boldsymbol{y} - \boldsymbol{y}_0) + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{y}_0)^T \boldsymbol{H}(\boldsymbol{y}_0)(\boldsymbol{y} - \boldsymbol{y}_0) \tag{3.24}$$

where

$$\boldsymbol{g}(\boldsymbol{y}_0) = \left. \frac{\partial \ln \tilde{p}(\boldsymbol{y})}{\partial \boldsymbol{y}} \right|_{\boldsymbol{y}=\boldsymbol{y}_0} \tag{3.25}$$

$$\boldsymbol{H}(\boldsymbol{y}_0) = \left. \frac{\partial^2 \ln \tilde{p}(\boldsymbol{y})}{\partial \boldsymbol{y} \partial \boldsymbol{y}^T} \right|_{\boldsymbol{y}=\boldsymbol{y}_0} \tag{3.26}$$

are the gradient and Hessian of $\ln \tilde{p}(\boldsymbol{y})$, respectively. Since $\boldsymbol{y}_0$ is the mode, the gradient is a zero vector at $\boldsymbol{y}_0$. Thus, if we define $\boldsymbol{\Lambda} = -\boldsymbol{H}(\boldsymbol{y}_0)$, then Eq. (3.24) can be written as

$$\ln \tilde{p}(\boldsymbol{y}) \approx \ln \tilde{p}(\boldsymbol{y}_0) - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{y}_0)^T \boldsymbol{\Lambda}(\boldsymbol{y} - \boldsymbol{y}_0) \tag{3.27}$$

The second term of the right hand side of this equation has the same form as the exponent of the exponential in the Gaussian distribution. Thus, if we take the exponential on both sides, we obtain

$$\tilde{p}(\boldsymbol{y}) \approx \tilde{p}(\boldsymbol{y}_0) \exp\left\{ -\frac{1}{2}(\boldsymbol{y} - \boldsymbol{y}_0)^T \boldsymbol{\Lambda}(\boldsymbol{y} - \boldsymbol{y}_0) \right\} \propto q(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{y_0}, \boldsymbol{\Lambda}^{-1}) \ . \tag{3.28}$$

The inverse covariance matrix $\boldsymbol{\Lambda}$ is referred to as the precision matrix. Using the approximation in Eq. (3.28), we can easily compute the normalisation factor in Eq. (3.22) as

$$Z_p \approx \int \tilde{p}(\boldsymbol{y}_0) \exp\left\{ -\frac{1}{2}(\boldsymbol{y} - \boldsymbol{y}_0)^T \boldsymbol{\Lambda}(\boldsymbol{y} - \boldsymbol{y}_0) \right\} d\boldsymbol{y} = \tilde{p}(\boldsymbol{y}_0)(2\pi)^{D/2} |\boldsymbol{\Lambda}|^{-1/2} \tag{3.29}$$

where D is the dimensionality of $\boldsymbol{y}$.

For model selection, $\tilde{p}(\boldsymbol{y}_0) = p(\mathcal{D}, \hat{\vartheta}_{k_{\mathrm{MAP}}} | \mathcal{M}_k) = p(\mathcal{D} | \hat{\vartheta}_{k_{\mathrm{MAP}}} \mathcal{M}_k) p(\hat{\vartheta}_{k_{\mathrm{MAP}}} | \mathcal{M}_k)$ and the Laplace approximation of the model evidence in Eq. (3.23) is therefore

$$\ln p(\mathcal{D}|\mathcal{M}_k) \approx \ln p(\mathcal{D}|\hat{\vartheta}_{k_{\mathrm{MAP}}}, \mathcal{M}_k) + \ln p(\hat{\vartheta}_{k_{\mathrm{MAP}}}|\mathcal{M}_k) + \frac{D}{2} \ln 2\pi - \frac{1}{2} \ln |\boldsymbol{\Lambda}| \ . \tag{3.30}$$

The first term of this equation is the log-likelihood, which grows with increasing model complexity, whereas the last three terms are penalty terms that decrease with increasing model complexity.

3.3.2 The Bayesian Information Criterion

The Bayesian information criterion (BIC) is a further simplification of the Laplace approximation. If the number of samples N is very large, the Hessian matrix $\mathbf{\Lambda}$ grows as $N\mathbf{\Lambda}_0$ for some fixed $\mathbf{\Lambda}_0$. We can therefore write

$$\ln|\mathbf{\Lambda}| \approx \ln|N\mathbf{\Lambda}_0| = \ln N^D|\mathbf{\Lambda}_0| = D\ln N + \ln|\mathbf{\Lambda}_0| \ . \tag{3.31}$$

If we assume the prior distribution to be flat, and only keep the terms that depend on N, we obtain the BIC from Eq. (3.30) as

$$\ln p(\mathcal{D}|\mathcal{M}_k) \approx \ln p(\mathcal{D}|\hat{\vartheta}_{k_{\mathrm{MAP}}}, \mathcal{M}_k) - \frac{D}{2}\ln N \ . \tag{3.32}$$

Since the BIC assumes a flat prior, the MAP estimate coincides with the maximum likelihood estimate from classical statistics. As we have already seen in Eq. (2.44) in example 2.5, the BIC is also used for model selection in the classical framework.

3.4 Summary

In this chapter, we have presented the most fundamental numerical techniques for drawing samples from arbitrary probability distributions. This is a very important topic in Bayesian statistics since analytical solutions to inference problems cannot always be computed. In general there exist two numerical inference techniques: the deterministic methods and the stochastic methods, and we focused on the latter. We introduced inverse transform sampling and rejection sampling as two simple ways of sampling from low-dimensional distributions. For sampling in higher dimensions, these methods suffer from severe limitations for which reason we looked into Markov chain Monte Carlo techniques. In this connection, we presented the Metropolis-Hastings algorithm and the Gibbs sampler, which are the most popular MCMC-based sampling techniques, and we demonstrated their applicability to Bayesian inference by using two small-scale examples. Finally, we looked at methods for computing the normalisation constant of an unnormalised distribution. This is an important problem for model selection, and we presented two popular approximate techniques: the Laplace approximation and the Bayesian information criterion. In the next chapter, we apply some of the introduced techniques on sinusoidal frequency estimation. This is an example of an important real world application in which analytical inference is impossible in general.

Chapter 4

Case Study: Bayesian Inference for the Frequency

In the last two chapters, we have presented the basic methods for making analytical and numerical Bayesian inference. We demonstrated these methods through several examples on a simplified version of the static sinusoidal model first introduced in Eq. (1.1). In all of these examples, we assumed the frequency as known. Although this assumption does not reflect the state of knowledge in most real world applications, it was necessary in order to enable analytical inference. In this chapter, we assume the much more realistic (and interesting) scenario in which the frequency is unknown, and we focus on making Bayesian inference about it. Specifically, we consider the model

$$x_n = \alpha \cos(\omega n + \varphi) + w_n \ , \quad \text{for } n = 1, \cdots, N \tag{4.1}$$

where the amplitude α, phase φ and frequency ω are all unknown and the white Gaussian noise term has unknown variance σ_w^2. Notice, that we are restricting ourselves to the case of a single sinusoid for illustrative purposes. As compared to the general static sinusoidal model in Eq. (1.1), we have also assumed the log-damping coefficient to be zero. This is a quite common assumption and this is also assumed in most Bayesian inference schemes for the frequency. In this chapter, we present two of these inference schemes which are based on the Gibbs sampler [Dou and Hodgson, 1995] and the Metropolis-Hastings algorithm [Andrieu and Doucet, 1999], respectively. Both of these schemes are based on a reformulated version of Eq. (4.1) given by

$$x_n = \beta_1 \cos(\omega n) + \beta_2 \sin(\omega n) + w_n \ , \quad \text{for } n = 1, \cdots, N \tag{4.2}$$

where $\alpha = \sqrt{\beta_1^2 + \beta_2^2}$ and $\varphi = -\arctan(\beta_2/\beta_1)$. In vector notation, we write this as

$$\boldsymbol{x} = \boldsymbol{D}\boldsymbol{\beta} + \boldsymbol{w} \tag{4.3}$$

where we define

$$\boldsymbol{x} \triangleq \begin{bmatrix} x_1 & \cdots & x_N \end{bmatrix}^T \tag{4.4}$$

$$\boldsymbol{D} \triangleq \begin{bmatrix} \boldsymbol{d}_1 & \boldsymbol{d}_2 \end{bmatrix} \tag{4.5}$$

$$\boldsymbol{d}_1 \triangleq \begin{bmatrix} \cos(\omega) & \cdots \cos(\omega N) \end{bmatrix}^T \tag{4.6}$$

$$\boldsymbol{d}_2 \triangleq \begin{bmatrix} \sin(\omega) & \cdots \sin(\omega N) \end{bmatrix}^T \tag{4.7}$$

$$\boldsymbol{\beta} \triangleq \begin{bmatrix} \beta_1 & \beta_2 \end{bmatrix}^T \tag{4.8}$$

$$\boldsymbol{w} \triangleq \begin{bmatrix} w_1 & \cdots & w_N \end{bmatrix}^T \ . \tag{4.9}$$

Since we focus on the inference for the frequency ω, we treat the amplitude $\boldsymbol{\beta}$ and noise variance σ_w^2 as nuisance parameters.

4.1 Inference based on a Gibbs Sampler

We begin by presenting the approach based on Gibbs sampler given in [Dou and Hodgson, 1995]. The full joint posterior distribution for the unknown model parameters is by Bayes' theorem given by

$$p(\boldsymbol{\beta}, \omega, \sigma_w^2 | \boldsymbol{x}) \propto p(\boldsymbol{x} | \boldsymbol{\beta}, \omega, \sigma_w^2) p(\boldsymbol{\beta}, \omega, \sigma_w^2) \ . \tag{4.10}$$

In [Dou and Hodgson, 1995], the prior distribution for the unknown parameters are assumed to be an improper prior proportional to the inverse noise variance, i.e., that $p(\boldsymbol{\beta}, \omega, \sigma_w^2) \propto p(\sigma_w^2) \propto \sigma_w^{-2}$, which we also assume. Since the noise is Gaussian distributed, we can therefore write the full joint posterior distribution as

$$p(\boldsymbol{\beta}, \omega, \sigma_w^2 | \boldsymbol{x}) \propto (\sigma_w^2)^{-N/2} \exp\left\{ \frac{-1}{2\sigma_w^2} (\boldsymbol{x} - \boldsymbol{D}\boldsymbol{\beta})^T (\boldsymbol{x} - \boldsymbol{D}\boldsymbol{\beta}) \right\} \sigma_w^{-2} \ . \tag{4.11}$$

The marginal posterior distribution for the frequency can be found by marginalising the full joint posterior distribution, i.e.,

$$p(\omega | \boldsymbol{x}) = \int p(\boldsymbol{\beta}, \omega, \sigma_w^2 | \boldsymbol{x}) d\boldsymbol{\beta} d\sigma_w^2 \ . \tag{4.12}$$

Although this marginalisation can be performed analytically for a single sinusoid, it cannot be done for the more general case of multiple sinusoids where we wish to compute, say, $p(\omega_l | \boldsymbol{x})$. In [Dou and Hodgson, 1995] a solution to this problem is proposed based on the Gibbs sampler consisting of three conditional distributions from which samples are drawn. These three conditional distributions are

Amplitude:	$p(\boldsymbol{\beta} \mid \omega, \sigma_w^2, \boldsymbol{x})$
Noise variance:	$p(\sigma_w^2 \mid \omega, \boldsymbol{\beta}, \boldsymbol{x})$
Frequency:	$p(\omega \mid \boldsymbol{\beta}, \sigma_w^2, \boldsymbol{x})$

and derived below.

Marginal Distribution for the Amplitude

If the frequency and noise variance are assumed known, the conditional distribution for the amplitude is

$$p(\boldsymbol{\beta}|\omega, \sigma_w^2, \boldsymbol{x}) \propto p(\boldsymbol{\beta}, \boldsymbol{x}|\omega, \sigma_w^2) = p(\boldsymbol{x}|\boldsymbol{\beta}, \sigma_w^2, \omega)p(\boldsymbol{\beta}|\omega, \sigma_w^2) \tag{4.13}$$

$$\propto p(\boldsymbol{x}|\boldsymbol{\beta}, \sigma_w^2, \omega) \propto \mathcal{N}\left(\boldsymbol{\beta}; \hat{\boldsymbol{\beta}}, \sigma_w^2(\boldsymbol{D}^T\boldsymbol{D})^{-1}\right) \tag{4.14}$$

where the last step follows from result B.1 and $\hat{\boldsymbol{\beta}} = (\boldsymbol{D}^T\boldsymbol{D})^{-1}\boldsymbol{D}^T\boldsymbol{x}$. Thus, we can obtain samples for the amplitude conditioned on the frequency and the noise variance by sampling the bivariate Gaussian distribution[1].

Marginal Distribution for the Noise Variance

If the amplitude and frequency are assumed known, the conditional distribution for the noise variance is

$$p(\sigma_w^2|\omega, \boldsymbol{\beta}, \boldsymbol{x}) \propto p(\sigma_w^2, \boldsymbol{x}|\omega, \boldsymbol{\beta}) = p(\boldsymbol{x}|\omega, \boldsymbol{\beta}, \sigma_w^2)p(\sigma_w^2) \propto \mathcal{N}(\boldsymbol{x}; \boldsymbol{D}\boldsymbol{\beta}, \sigma_w^2\boldsymbol{I}_N)\sigma_w^{-2} \tag{4.15}$$

$$\propto \text{Inv-}\mathcal{G}\left(\sigma_w^2; N/2, \frac{1}{2}(\boldsymbol{x} - \boldsymbol{D}\boldsymbol{\beta})^T(\boldsymbol{x} - \boldsymbol{D}\boldsymbol{\beta})\right) \tag{4.16}$$

where the last step follows from result B.3.

Marginal Distribution for the Frequency

We cannot derive the conditional distribution for the frequency by following the same procedure as above since this results in a non-standard conditional distribution for the frequency

$$p(\omega|\boldsymbol{\beta}, \sigma_w^2, \boldsymbol{x}) \propto \exp\left\{\frac{-1}{2\sigma_w^2}(\boldsymbol{x} - \boldsymbol{D}\boldsymbol{\beta})^T(\boldsymbol{x} - \boldsymbol{D}\boldsymbol{\beta})\right\} \tag{4.17}$$

which we cannot sample from directly. To overcome this problem, [Dou and Hodgson, 1995] propose making the Laplace approximation to this distribution as described in section 3.3.1, i.e., we approximate $p(\omega|\boldsymbol{\beta}, \sigma_w^2, \boldsymbol{x})$ by a Gaussian distribution. Thus, we first find the MAP estimate of the frequency $\hat{\omega}_{\text{MAP}}$ as the frequency ω minimising

$$J(\omega) = (\boldsymbol{x} - \boldsymbol{D}\boldsymbol{\beta})^T(\boldsymbol{x} - \boldsymbol{D}\boldsymbol{\beta}) \ . \tag{4.18}$$

[1] This bivariate conditional distribution can also be broken down into two univariate Gaussian distributions or marginalised w.r.t the noise variance yielding a student's t-distribution. See [Dou and Hodgson, 1995] for details.

This is a nonlinear least-squares problem, and the MAP estimate is the mean of the Gaussian we use for the approximation. Secondly, we find the Hessian of $\ln p(\omega|\boldsymbol{\beta}, \sigma_w^2, \boldsymbol{x})$ evaluated at $\hat{\omega}_{\text{MAP}}$ as [Dou and Hodgson, 1995]

$$H(\hat{\omega}_{\text{MAP}}) = -\frac{\boldsymbol{\beta}^T \dot{\boldsymbol{D}}^T \dot{\boldsymbol{D}} \boldsymbol{\beta}}{\sigma_w^2}\bigg|_{\omega=\hat{\omega}_{\text{MAP}}} \triangleq -\frac{1}{\sigma^2} \tag{4.19}$$

where $\dot{\boldsymbol{D}} = \frac{\partial}{\partial \omega}\boldsymbol{D}$ and σ^2 is the variance of the Gaussian we use for the approximation. The Laplace approximation of the conditional distribution of the frequency is therefore given by

$$p(\omega|\boldsymbol{\beta}, \sigma_w^2, \boldsymbol{x}) \approx \mathcal{N}(\omega; \hat{\omega}_{\text{MAP}}, \sigma^2) \tag{4.20}$$

from which we can easily draw samples.

The Gibbs sampling algorithm as proposed in [Dou and Hodgson, 1995] is summarised in algorithm 4.1.

Algorithm 4.1 (Gibbs Sampler for Frequency Estimation)

1. Initialise ${\sigma_w^2}^{[0]}$ and $\omega^{[0]}$
2. Repeat for $\tau = 0, 1, 2, \cdots, T$
   (a) Draw a sample for the amplitude
   $$\boldsymbol{\beta}^{[\tau+1]} \sim \mathcal{N}\left(\boldsymbol{\beta}; \hat{\boldsymbol{\beta}}^{[\tau]}, {\sigma_w^2}^{[\tau]} ({\boldsymbol{D}^{[\tau]}}^T \boldsymbol{D}^{[\tau]})^{-1}\right)$$
   where $\hat{\boldsymbol{\beta}}^{[\tau]} = ({\boldsymbol{D}^{[\tau]}}^T \boldsymbol{D}^{[\tau]})^{-1} {\boldsymbol{D}^{[\tau]}}^T \boldsymbol{x}$ and $\boldsymbol{D}^{[\tau]}$ denotes that $\boldsymbol{D}$ is evaluated using $\omega^{[\tau]}$.
   (b) Draw a sample for the noise variance
   $${\sigma_w^2}^{[\tau+1]} \sim \text{Inv-}\mathcal{G}\left(\sigma_w^2; N/2, \frac{1}{2}(\boldsymbol{x} - \boldsymbol{D}^{[\tau]}\boldsymbol{\beta}^{[\tau+1]})^T(\boldsymbol{x} - \boldsymbol{D}^{[\tau]}\boldsymbol{\beta}^{[\tau+1]})\right) .$$
   (c) Find the MAP estimate of the frequency $\hat{\omega}_{\text{MAP}}^{[\tau+1]}$ by minimising
   $$J(\omega) = (\boldsymbol{x} - \boldsymbol{D}^{[\tau]}\boldsymbol{\beta}^{[\tau+1]})^T(\boldsymbol{x} - \boldsymbol{D}^{[\tau]}\boldsymbol{\beta}^{[\tau+1]})$$
   w.r.t. ω.
   (d) Draw sample for the frequency
   $$\omega^{[\tau+1]} \sim \mathcal{N}(\omega; \hat{\omega}_{\text{MAP}}^{[\tau+1]}, {\sigma^2}^{[\tau+1]})$$
   where σ^2 is given by Eq. (4.19).

There are two serious drawbacks of this algorithm. First of all, it is only an approximate Bayesian inference scheme since we use the Laplace approximation in order to enable sampling for the frequency parameter. Secondly, step c) of the algorithm is highly intractable since it involves minimisation of a sharply peaked multimodel cost function which cannot be computed in closed form [Stoica and Moses, 2005, p. 159]. In part II of this thesis, we develop a similar inference scheme based on the Gibbs sampler but avoiding both of these drawbacks.

4.2 Inference Based on the Metropolis-Hastings Algorithm

As we saw in the approach based on the Gibbs sampler, it was only approximate due to the use of the Laplace approximation to the conditional distribution for the frequency. In the more recent paper [Andrieu and Doucet, 1999], this is avoided by using the Metropolis-Hastings inference scheme instead of the Gibbs sampler. In the paper, the full joint posterior distribution for the unknown frequency, amplitude and noise variance is again given by Eq. (4.10). The prior distribution, however, is assumed to factor as

$$p(\boldsymbol{\beta}, \omega, \sigma_w^2) = p(\boldsymbol{\beta}|\sigma_w^2, \omega)p(\sigma_w^2)p(\omega) \tag{4.25}$$

$$= \mathcal{N}(\boldsymbol{\beta}; \mathbf{0}, \sigma_w^2 \boldsymbol{\Sigma}_{\boldsymbol{\beta}})\text{Inv-}\mathcal{G}(\sigma_w^2; a, b)\mathcal{U}(\omega; 0, \pi) \tag{4.26}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = g(\boldsymbol{D}^T\boldsymbol{D})^{-1}$. The scalar g can be interpreted as the expected signal-to-noise ratio (SNR) [Andrieu and Doucet, 1999]. The particular form of the prior for $p(\boldsymbol{\beta}, \sigma_w^2|\omega) = p(\boldsymbol{\beta}|\sigma_w^2, \omega)p(\sigma_w^2)$ corresponds to the *Zellner's g-prior* [Zellner, 1986] which is a special case of the prior distribution considered in example 2.5 and section B.3. The assumed factorisation of $p(\boldsymbol{\beta}, \omega, \sigma_w^2)$ can be obtained using a maximum entropy method [Andrieu and Doucet, 1999], i.e., the assumed factorisation is the least subjective prior distribution in the case of little or no prior knowledge.

A convenient property of this prior is that it allows us to integrate the nuisance parameters $\boldsymbol{\beta}$ and σ_w^2 out of the full joint posterior distribution. To see this, first consider the full joint posterior distribution

$$p(\boldsymbol{\beta}, \omega, \sigma_w^2|\boldsymbol{x}) \propto p(\boldsymbol{x}|\boldsymbol{\beta}, \omega, \sigma_w^2)p(\boldsymbol{\beta}|\sigma_w^2, \omega)p(\sigma_w^2)p(\omega) \tag{4.27}$$

$$= \mathcal{N}(\boldsymbol{x}; \boldsymbol{D}\boldsymbol{\beta}, \sigma_w^2)\mathcal{N}(\boldsymbol{\beta}; \mathbf{0}, \sigma_w^2\boldsymbol{\Sigma}_{\boldsymbol{\beta}})\text{Inv-}\mathcal{G}(\sigma_w^2; a, b)\mathcal{U}(\omega; 0, \pi) \tag{4.28}$$

$$\begin{aligned} &\propto (\sigma_w^2)^{-N/2} \exp\left\{\frac{-1}{2\sigma_w^2}(\boldsymbol{x} - \boldsymbol{D}\boldsymbol{\beta})^T(\boldsymbol{x} - \boldsymbol{D}\boldsymbol{\beta})\right\} \\ &\quad \times (\sigma_w^2)^{-1}|\boldsymbol{\Sigma}_{\boldsymbol{\beta}}|^{-1/2} \exp\left\{\frac{-1}{2\sigma_w^2}\boldsymbol{\beta}^T\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\beta}\right\} \\ &\quad \times (\sigma_w^2)^{-(a+1)} \exp\left\{\frac{-b}{\sigma_w^2}\right\}\mathbb{I}(\omega) \end{aligned} \tag{4.29}$$

where

$$\mathbb{I}(\omega) = \begin{cases} 1 & \text{for } 0 < \omega < \pi \\ 0 & \text{otherwise} \end{cases} \tag{4.30}$$

is the indicator function. The terms can be rearranged and grouped in a clever way as

$$\begin{aligned} p(\boldsymbol{\beta}, \omega, \sigma_w^2|\boldsymbol{x}) &\propto (\sigma_w^2)^{-1}|\boldsymbol{\Sigma}_{\boldsymbol{\beta}|\boldsymbol{x}}|^{-1/2} \exp\left\{ \frac{-1}{2\sigma_w^2}(\boldsymbol{\beta} - \boldsymbol{\mu}_{\boldsymbol{\beta}|\boldsymbol{x}})^T \boldsymbol{\Sigma}_{\boldsymbol{\beta}|\boldsymbol{x}}^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_{\boldsymbol{\beta}|\boldsymbol{x}}) \right\} \\ &\times \left[b + \frac{1}{2}\boldsymbol{x}^T(\boldsymbol{I}_N + \boldsymbol{D}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{D}^T)^{-1}\boldsymbol{x} \right]^{\frac{N}{2}+a} (\sigma_w^2)^{-(\frac{N}{2}+a+1)} \exp\left\{ -\frac{b + \frac{1}{2}\boldsymbol{x}^T(\boldsymbol{I}_N + \boldsymbol{D}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{D}^T)^{-1}\boldsymbol{x}}{\sigma_w^2} \right\} \\ &\times \mathbb{I}(\omega)|\boldsymbol{\Sigma}_{\boldsymbol{\beta}}|^{-1/2}|\boldsymbol{\Sigma}_{\boldsymbol{\beta}|\boldsymbol{x}}|^{1/2} \left[b + \frac{1}{2}\boldsymbol{x}^T(\boldsymbol{I}_N + \boldsymbol{D}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{D}^T)^{-1}\boldsymbol{x} \right]^{-(\frac{N}{2}+a)} \end{aligned} \tag{4.31}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\beta}|\boldsymbol{x}} = (\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} + \boldsymbol{D}^T\boldsymbol{D})^{-1}$ and $\boldsymbol{\mu}_{\boldsymbol{\beta}|\boldsymbol{x}} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}|\boldsymbol{x}}\boldsymbol{D}^T\boldsymbol{x}$. The first line of Eq. (4.31) contains all factors dependent on $\boldsymbol{\beta}$ and is proportional to a Gaussian distribution. Thus, integrating Eq. (4.31) w.r.t. $\boldsymbol{\beta}$, is equivalent to integrating the Gaussian in line one w.r.t. $\boldsymbol{\beta}$. The integration is easily performed and gives a scale factor independent of σ_w^2 and ω. After the integration w.r.t. $\boldsymbol{\beta}$, the remaining parts dependent on σ_w^2 constitute an inverse gamma distribution (line two of Eq. (4.31)). Thus, integrating w.r.t. σ_w^2 also gives a simple scale factor independent of ω. Hence, the marginal posterior distribution for the frequency is proportional to the third line of Eq. (4.31), i.e.,

$$p(\omega|\boldsymbol{x}) \propto \mathbb{I}(\omega)|\boldsymbol{\Sigma}_{\boldsymbol{\beta}}|^{-1/2}|\boldsymbol{\Sigma}_{\boldsymbol{\beta}|\boldsymbol{x}}|^{1/2} \left[b + \frac{1}{2}\boldsymbol{x}^T(\boldsymbol{I}_N + \boldsymbol{D}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{D}^T)^{-1}\boldsymbol{x} \right]^{-(N/2+a)} \tag{4.32}$$

$$\propto \mathbb{I}(\omega) \left[2b + \boldsymbol{x}^T(\boldsymbol{I}_N + \boldsymbol{D}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{D}^T)^{-1}\boldsymbol{x} \right]^{-(N/2+a)} \triangleq \tilde{p}(\omega|\boldsymbol{x}) \tag{4.33}$$

where we have used the fact that

$$|\boldsymbol{\Sigma}_{\boldsymbol{\beta}}|^{-1/2}|\boldsymbol{\Sigma}_{\boldsymbol{\beta}|\boldsymbol{x}}|^{1/2} = |(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} + \boldsymbol{D}^T\boldsymbol{D})\boldsymbol{\Sigma}_{\boldsymbol{\beta}}|^{-1/2} = |\boldsymbol{I}_2 + \boldsymbol{D}^T\boldsymbol{D}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}|^{-1/2} = (1+g)^{-1} \,. \tag{4.34}$$

The marginal posterior distribution $p(\omega|\boldsymbol{x})$ is clearly a non-standard distribution and non-linear in the frequency parameter. Thus, we cannot find, e.g., the mean of this distribution using analytical tools and we have to resort to numerical methods. In [Andrieu and Doucet, 1999], the Metropolis-Hastings algorithm is used for drawing samples from $p(\omega|\boldsymbol{x})$. For the MH-algorithm, we have to specify a proposal transition kernel $Q(\omega'; \omega)$ which, as discussed in section 3.2.2, should be selected as a trade-off between correlation time and acceptance ratio. In order to achieve this, [Andrieu and Doucet, 1999] suggest using the mixture proposal transition kernel given by

$$Q(\omega'; \omega) = \lambda Q_1(\omega'; \omega) + (1 - \lambda)Q_2(\omega'; \omega) \tag{4.35}$$

where $Q_1(\omega';\omega)$ is a mixture distribution of uniform distributions with mixing coefficients ψ calculated from the periodogram, i.e.,

$$Q_1(\omega';\omega) = \sum_{k=0}^{K-1} \psi(\omega_k)\mathcal{U}(\omega;\omega_k,\omega_{k+1}) \ . \tag{4.36}$$

The value of K can be selected to be equal to the number of samples N or larger to improve the interpolation. The transition kernel $Q_2(\omega';\omega)$ is a Gaussian distribution with variance σ^2 and centred on the current state. The advantage of using this mixture kernel is that $Q_1(\omega';\omega)$ is independent of the current state whereas $Q_2(\omega';\omega)$, for a small variance σ^2, ensures a high acceptance ratio. Thus, $Q_1(\omega';\omega)$ is used for making independent jumps on $p(\omega|\boldsymbol{x})$ whereas $Q_2(\omega';\omega)$ makes a local exploration on $p(\omega|\boldsymbol{x})$ dependent on the current state. The MH-algorithm as proposed in [Andrieu and Doucet, 1999] is summarised in algorithm 4.2.

Algorithm 4.2 (MH-Algorithm for Frequency Estimation)

1. Compute the periodogram from the (possible zero-padded) vector of observations $\boldsymbol{x}$ and normalise it so it integrates to one. Denote the resulting distribution as $Q_1(\omega';\omega)$.
2. Select the mixing coefficient $0 \leq \lambda \leq 1$ and the variance σ^2 of the Gaussian transition kernel given by $Q_2(\omega';\omega) = \mathcal{N}(\omega';\omega,\sigma^2)$.
3. Initialise $\omega^{[0]}$.
4. Repeat for $\tau = 0, 1, 2, \cdots, T$
   (a) Draw a sample $u_1^{[\tau]}$ from the uniform distribution $\mathcal{U}(0,1)$.
   (b) If $u_1^{[\tau]} < \lambda$, set $i = 1$. Else set $i = 2$.
   (c) Draw a candidate sample ω^* from the proposal transition kernel $Q_i(\omega;\omega^{[\tau]})$.
   (d) Evaluate the probability of move given by

$$\alpha(\omega^*,\omega^{[\tau]}) = \min\left[1, \frac{\tilde{p}(\omega^*|\boldsymbol{x})Q_i(\omega^{[\tau]};\omega^*)}{\tilde{p}(\omega^{[\tau]}|\boldsymbol{x})Q_i(\omega^*;\omega^{[\tau]})}\right]$$

   where $\tilde{p}(\omega|\boldsymbol{x})$ is the unnormalised desired distribution given by Eq. (4.33).
   (e) Draw a random variate $u_2^{[\tau]}$ from the univariate uniform distribution $\mathcal{U}(0,1)$.

(f) If $u_2^{[\tau]} \leq \alpha(\omega^*, \omega^{[\tau]})$, then accept the sample ω^* as a sample from $p(\omega|\boldsymbol{x})$ and set $\omega^{[\tau+1]} = \omega^*$. Otherwise reject ω^* as a sample from $p(\omega|\boldsymbol{x})$ and set $\omega^{[\tau+1]} = \omega^{[\tau]}$.

The algorithm requires that we select the mixing coefficient λ and the variance σ^2 of the Gaussian distribution. Although these user defined parameters do not affect the posterior distribution, which we draw samples from, they influence the convergence time, correlation time and acceptance ratio.

4.3 Simulations

We demonstrate the applicability of the inference schemes in algorithm 4.1 and 4.2 on a small-scale example. In the example, we generated $N = 96$ samples from the signal model in Eq. (4.2) with a true frequency of $\omega = 0.4$, true amplitudes of $\boldsymbol{\beta} = \frac{1}{2}\begin{bmatrix}\sqrt{3} & 1\end{bmatrix}^T$, and a true noise variance $\sigma_w^2 = 1$, i.e., the SNR was 0 dB. All of these values were assumed unknown.

The Gibbs sampling inference scheme in algorithm 4.1 requires that we find the MAP estimate of the frequency for every iteration by solving a non-linear optimisation problem. This is very hard in general (see, e.g., [Stoica and Moses, 2005, pp. 457-462]), but [Dou and Hodgson, 1995] do not mention how they solve this problem. We have therefore selected to find the MAP estimate in the following way:

- For the first iteration, we sample the cost-function in Eq. (4.18) marginalised w.r.t. to $\boldsymbol{\beta}$ on a fine grid and select the first MAP-estimate as the argument of the minimum value.
- For all subsequent iterations, we find the new MAP estimate by performing one Newton step given by

$$\hat{\omega}_{\text{MAP}}^{[\tau+1]} = \hat{\omega}_{\text{MAP}}^{[\tau]} - \left.\frac{\frac{\partial}{\partial\omega}J(\omega)}{\frac{\partial^2}{\partial\omega^2}J(\omega)}\right|_{\omega=\hat{\omega}_{\text{MAP}}^{[\tau]}}. \tag{4.37}$$

Figure 4.1 shows the trace of samples generated from the marginal posterior distribution of the frequency by using the Gibbs sampling inference scheme. A total of 100,000 samples were drawn and the right margin of figure 4.1 shows the 50-bin histogram of these samples. The red curve shows the true marginal posterior distribution for the frequency evaluated on a grid. The true marginal posterior distribution was derived analytically from Eq. (4.12). Since the histogram and true distribution are approximately equal, the Laplace approximation in the inference scheme must be a good approximation - at least for the data of this simulation.

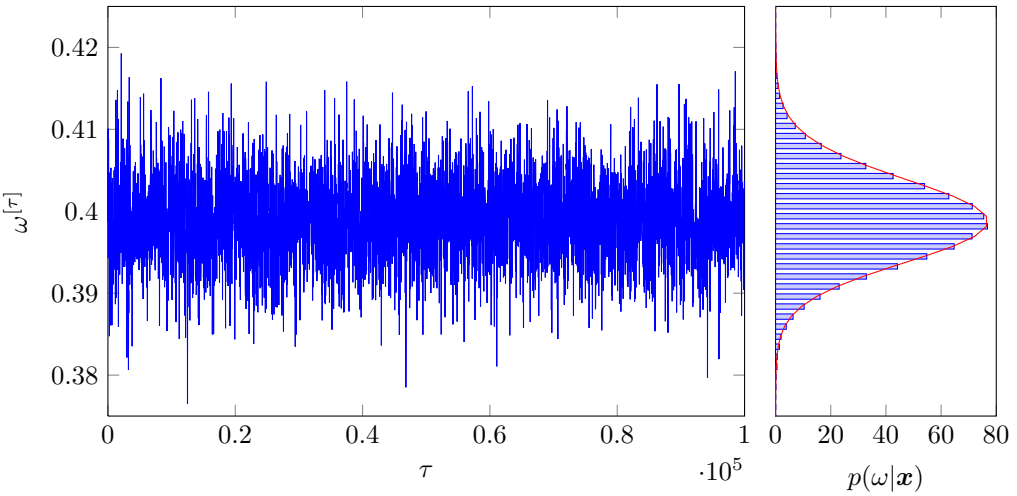

Figure 4.1: Trace of the $T = 100,000$ samples for the frequency generated by the algorithm based on Gibbs sampler. The plot in the margin shows the true distribution (solid red line) and the 50-bin histogram of the generated samples with the first 100 samples removed as burn-in samples.

Before running the Metropolis-Hastings inference scheme in algorithm 4.2, the user-defined parameters must be specified. For the prior distribution of $\boldsymbol{\beta}$, we must select the value of g. Although g can also be treated as a random variable and incorporated into the sampling scheme (see [Andrieu and Doucet, 1999] for details), we selected it to be large so that the prior covariance of the prior distribution of $\boldsymbol{\beta}$ was large. The mixing coefficient and variance of the proposal Gaussian distribution were selected to have the same values as in [Andrieu and Doucet, 1999], i.e., $\lambda = 0.2$ and $\sigma^2 = (5N)^{-1}$. For the calculation of the periodogram, we selected an FFT-length of $2N$.

Figure 4.2 shows the trace of samples generated from the marginal posterior distribution of the frequency by using the MH inference scheme. As in the case of the Gibbs sampler, a total of 100,000 samples were drawn and the right margin of figure 4.1 shows the 50-bin histogram of these samples. The red curve shows the true marginal posterior distribution for the frequency evaluated on a grid. The true marginal posterior distribution is given by Eq. (4.33). The histogram has again the same envelope as the true distribution. The acceptance ratio in the simulation was 71.3 %.

4.4 Summary

In this chapter, we have demonstrated the applicability of the numerical Bayesian inference to a real world problem. This was accomplished by demonstrating how the

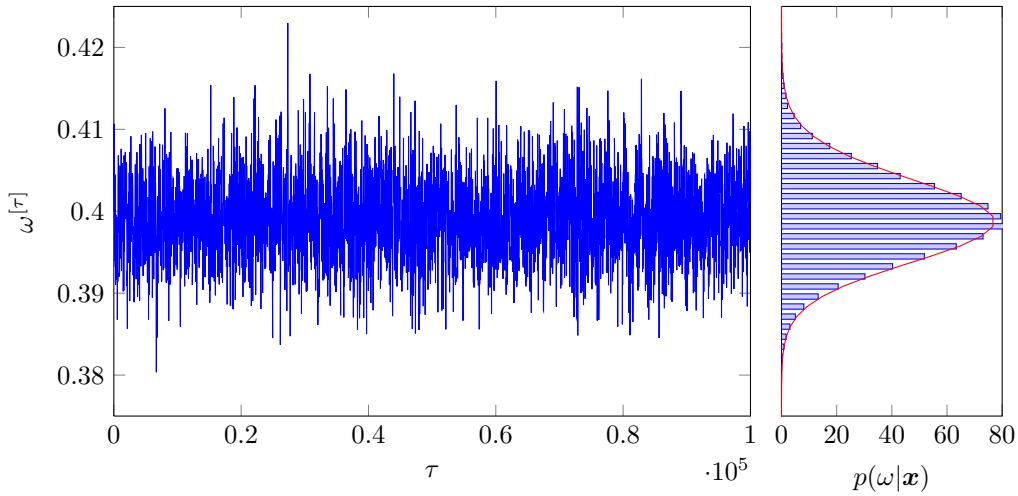

Figure 4.2: Trace of the $T = 100,000$ samples for the frequency generated by the MH-algorithm. The plot in the margin shows the true distribution (solid red line) and the 50-bin histogram of the generated samples with the first 100 samples removed as burn-in samples.

state-of-the-art Bayesian inference schemes for the frequency in the static sinusoidal model work. Although we have only demonstrated the applicability of the state-of-the-art sinusoidal frequency Bayesian inference schemes on a very simple example, it should be clear what the strength and weaknesses of the algorithms are. The algorithm proposed in [Dou and Hodgson, 1995] is based on a Gibbs sampler, but the inference scheme suffers from that the conditional distribution for the frequency is a non-standard distribution which is not easy to sample from. The proposed remedy for this is to use the Laplace approximation which, as demonstrated in the simulations, is a very good approximation. The major drawback of using the Laplace approximation is that we have to solve a non-linear least squares problem in order to compute the mean of the Laplace approximation. The algorithm proposed in [Andrieu and Doucet, 1999] is based on the Metropolis-Hastings algorithm and does not employ any approximations. The algorithm, however, requires careful tuning of the proposal distribution in order to obtain a good trade-off between acceptance ratio and correlation time. In the next part of this thesis, we propose and develop an inference scheme based on the more flexible dynamic sinusoidal model. By using this model, the conditional distribution for the frequency parameters in the Gibbs sampler turns out to be a standard distribution which we can easily and efficiently draw samples from. Thus, we do not have to tune any proposal distribution and we avoid sampling from an approximate distribution and solving non-linear least squares problem.

Part II

Bayesian Inference for the Dynamic Sinusoidal Model

Chapter 5

The Dynamic Sinusoidal Signal Model

In the first part of this thesis, we introduced the basic analytical and numerical tools for performing Bayesian inference. We also discussed the state-of-the-art Bayesian methods for deriving the marginal posterior distribution for the frequency of a static sinusoidal model. Although the static model given by Eq. (1.1), and its special cases, is by far the most popular sinusoidal model, there exist other sinusoidal models. In the second part of this thesis, we consider Bayesian inference for a dynamic sinusoidal model which is based on a state space formulation of the static sinusoidal model. This dynamic model has the static model as a special case and is thus more flexible. It allows the amplitudes to evolve as a first-order autoregressive (AR(1)) model. In this chapter, we derive the dynamic model and discuss its properties and relation to the static model. In the next chapter, we develop the inference scheme for it.

5.1 State-Space Formulation of the Sinusoidal Model

So far we have only encountered the real static sinusoidal model in Eq. (1.1) which we have restated here for easy reference

$$x_n = \sum_{l=1}^{L} \alpha_l e^{-\gamma_l n} \cos(\omega_l n + \varphi_l) + w_n \ , \quad \text{for } n = 1, \cdots, N \ . \tag{5.1}$$

The model parameters are the noise variance σ_w^2 of the white Gaussian noise w_n and the l'th amplitude $\alpha_l > 0$, the l'th phase $\varphi_l \in [-\pi, \pi]$, the l'th (angular) frequency $\omega_l \in [0, \pi]$, the l'th log-damping coefficient $\gamma_l > 0$ for each of the L sinusoids. Using

complex notation, we can rewrite this model as

$$x_n = \sum_{l=1}^{L} (c_l z_l^n + c_l^c (z_l^c)^n) + w_n \tag{5.2}$$

where $(\cdot)^c$ denotes complex conjugation, $c_l = (\alpha_l/2)e^{j\varphi_l}$, $z_l = e^{-\gamma_l} e^{j\omega_l}$ and $j = \sqrt{-1}$ is the imaginary unit. In matrix notation, we can write this as

$$x_n = \begin{bmatrix} z_1^n & (z_1^c)^n & \cdots & z_L^n & (z_L^c)^n \end{bmatrix} \begin{bmatrix} c_1 \\ c_1^c \\ \vdots \\ c_L \\ c_L^c \end{bmatrix} + w_n \tag{5.3}$$

$$= \begin{bmatrix} 1 & 1 & \cdots 1 & 1 \end{bmatrix} \begin{bmatrix} z_1 & 0 & \cdots & 0 & 0 \\ 0 & z_1^c & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & z_L & 0 \\ 0 & 0 & \cdots & 0 & z_L^c \end{bmatrix}^n \begin{bmatrix} c_1 \\ c_1^c \\ \vdots \\ c_L \\ c_L^c \end{bmatrix} + w_n \tag{5.4}$$

$$\triangleq \tilde{\boldsymbol{b}}^T \tilde{\boldsymbol{A}}^n \tilde{\boldsymbol{c}} + w_n \ . \tag{5.5}$$

This formulation constraints the log-damping coefficients and the frequency parameters to be completely determined (in a deterministic way) by the time-invariant and complex diagonal matrix $\tilde{\boldsymbol{A}}$. The diagonal representation of x_n is not unique since we for any invertible matrix $\boldsymbol{T}$ have that

$$x_n = \tilde{\boldsymbol{b}}^T \left(\boldsymbol{T}^{-1} \boldsymbol{T} \tilde{\boldsymbol{A}} \boldsymbol{T}^{-1} \boldsymbol{T} \right)^n \tilde{\boldsymbol{c}} + w_n = \tilde{\boldsymbol{b}}^T \boldsymbol{T}^{-1} \left(\boldsymbol{T} \tilde{\boldsymbol{A}} \boldsymbol{T}^{-1} \right)^n \boldsymbol{T} \tilde{\boldsymbol{c}} + w_n \tag{5.6}$$

$$= \boldsymbol{b}^T \boldsymbol{A}^n \boldsymbol{c} + w_n \tag{5.7}$$

where $\boldsymbol{b}$, $\boldsymbol{A}$ and $\boldsymbol{c}$ are given by

$$\boldsymbol{b} \triangleq \left(\tilde{\boldsymbol{b}}^T \boldsymbol{T}^{-1} \right)^T \tag{5.8}$$

$$\boldsymbol{A} \triangleq \boldsymbol{T} \tilde{\boldsymbol{A}} \boldsymbol{T}^{-1} \tag{5.9}$$

$$\boldsymbol{c} \triangleq \boldsymbol{T} \tilde{\boldsymbol{c}} \ . \tag{5.10}$$

Notice that Eq. (5.9) can be interpreted as an eigenvalue decomposition (EVD) of $\boldsymbol{A}$ with $\tilde{\boldsymbol{A}}$ containing the eigenvalues and $\boldsymbol{T}$ the eigenvectors. Thus from any matrix $\boldsymbol{A}$, we can always recover the diagonal representation of x_n by computing the eigenvalue decomposition of $\boldsymbol{A}$. The matrix $\tilde{\boldsymbol{A}}$ and vector $\tilde{\boldsymbol{c}}$ are in general complex valued. In

order to avoid the complex terms, we define the Hermitian complex block diagonal matrix $\boldsymbol{T} = \text{diag}(\boldsymbol{T}_1, \cdots, \boldsymbol{T}_l, \cdots \boldsymbol{T}_L)$ with

$$\boldsymbol{T}_l = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ j & -j \end{bmatrix} \ . \tag{5.11}$$

By using this particular choice of $\boldsymbol{T}$, we now obtain

$$\boldsymbol{b} = \sqrt{2} \begin{bmatrix} 1 & 0 & \cdots & 1 & 0 \end{bmatrix}^T \tag{5.12}$$

$$\boldsymbol{A} = \text{diag}(\boldsymbol{A}_1, \cdots, \boldsymbol{A}_l, \cdots, \boldsymbol{A}_L) \tag{5.13}$$

$$\boldsymbol{A}_l = e^{-\gamma_l} \begin{bmatrix} \cos\omega_l & \sin\omega_l \\ -\sin\omega_l & \cos\omega_l \end{bmatrix} \tag{5.14}$$

$$\boldsymbol{c} = \frac{1}{\sqrt{2}} \begin{bmatrix} \alpha_1 \cos\varphi_1 & -\alpha_1 \sin\varphi_1 & \cdots & \alpha_L \cos\varphi_L & -\alpha_L \sin\varphi_L \end{bmatrix}^T \ . \tag{5.15}$$

which are all real. In the rest of this thesis, we use this real representation of x_n.

Returning to Eq. (5.7), we can expand it into

$$x_n = \boldsymbol{b}^T \underbrace{\boldsymbol{A} \cdots \boldsymbol{A}}_{n \text{ times}} \boldsymbol{c} + w_n \tag{5.16}$$

from which we see that the amplitudes in $\boldsymbol{c}$ are rotated by the same matrix $\boldsymbol{A}$ for every time index n. This update of the amplitude can be interpreted as a state $\boldsymbol{s}_n$ of the amplitude at time index n and can be done recursively by separating Eq. (5.16) into an observation and a state equation. If we also introduce noise into the state equation, we can write the sinusoidal model in Eq. (5.1) as a linear Gaussian time-invariant state space model given by

$$\begin{aligned} y_n &= \boldsymbol{b}^T \boldsymbol{s}_n + w_n \qquad \text{(observation equation)} \\ \boldsymbol{s}_{n+1} &= \boldsymbol{A}\boldsymbol{s}_n + \boldsymbol{v}_n \qquad \text{(state equation)} \end{aligned} \tag{5.17}$$

where $\boldsymbol{v}_n$ is white Gaussian state noise with covariance matrix $\boldsymbol{Q}$. We also assume a Gaussian prior for the initial state vector $\boldsymbol{s}_1$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{P}$. The dynamic model in Eq. (5.17) is slightly different (hence the use of y_n instead of x_n for denoting the observations) than the original model in Eq. (5.1) for state noise having non-zero covariance. This allows the amplitudes in $\boldsymbol{c}$ to develop as an AR(1) process. In the case of initial state vector equal to $\boldsymbol{s}_1 = \boldsymbol{A}\boldsymbol{c}$ and zero state-noise, the two models are identical, i.e., $y_n = x_n$. In the next section, we establish a general relationship between the static and dynamic models by examining the state noise $\boldsymbol{v}_n$.

5.2 Relationship Between the Static and Dynamic Models

In order to establish a connection between the static and dynamic sinusoidal models, we want to see how the state noise shows up in the original sinusoidal model in Eq. (5.1). To do this, we simply merge the state and observation equation of the state space model in Eq. (5.17) as

$$\begin{aligned} y_n &= \boldsymbol{b}^T \boldsymbol{s}_n + w_n && (5.18)\\ &= \boldsymbol{b}^T (\boldsymbol{A}\boldsymbol{s}_{n-1} + \boldsymbol{v}_{n-1}) + w_n && (5.19)\\ &= \boldsymbol{b}^T \left(\boldsymbol{A}(\boldsymbol{A}\boldsymbol{s}_{n-2} + \boldsymbol{v}_{n-2}) + \boldsymbol{v}_{n-1}\right) + w_n && (5.20)\\ &\;\;\vdots \\ &= \boldsymbol{b}^T \boldsymbol{A}^{n-1} \boldsymbol{s}_1 + \sum_{k=1}^{n-1} \boldsymbol{b}^T \boldsymbol{A}^{k-1} \boldsymbol{v}_{n-k} + w_n && (5.21) \end{aligned}$$

which for $\boldsymbol{s}_1 = \boldsymbol{A}\boldsymbol{c}$ can be written as

$$\begin{aligned} y_n &= \sum_{l=1}^{L} \alpha_l e^{-\gamma_l n} \cos(\omega_l n + \varphi_l) + e_n + w_n && (5.22)\\ &= x_n + e_n && (5.23) \end{aligned}$$

where we have defined the noise term e_n as

$$e_n \triangleq \sum_{k=1}^{n-1} \boldsymbol{b}^T \boldsymbol{A}^{k-1} \boldsymbol{v}_{n-k} \; . \tag{5.24}$$

Thus, the observations x_n of the static model in Eq. (5.1) and observations y_n of the dynamic model in Eq. (5.17) are related by an additive noise term e_n which is a stochastic process dependent on the log-damping coefficients as well as the frequency parameters. The expression for e_n can be simplified to

$$e_n = \sum_{k=1}^{n-1} \boldsymbol{d}_{k-1}^T \boldsymbol{v}_{n-k} \tag{5.25}$$

$$\boldsymbol{d}_i \triangleq \sqrt{2} \begin{bmatrix} e^{-\gamma_1 i} \begin{bmatrix} \cos \omega_1 i \\ \sin \omega_1 i \end{bmatrix}^T & \cdots & e^{-\gamma_L i} \begin{bmatrix} \cos \omega_L i \\ \sin \omega_L i \end{bmatrix}^T \end{bmatrix}^T \; . \tag{5.26}$$

The variance of e_n can now be calculated as

$$\sigma_{e_n}^2 = E\left\{\left(\sum_{k=1}^{n-1} \boldsymbol{d}_{k-1}^T \boldsymbol{v}_{n-k}\right)\left(\sum_{k=1}^{n-1} \boldsymbol{d}_{k-1}^T \boldsymbol{v}_{n-k}\right)^T\right\} \tag{5.27}$$

$$= \sum_{k=1}^{n-1}\sum_{r=1}^{n-1} \boldsymbol{d}_{k-1}^T E\left\{\boldsymbol{v}_{n-k}\boldsymbol{v}_{n-r}^T\right\}\boldsymbol{d}_{r-1} \tag{5.28}$$

$$= \sum_{k=1}^{n-1} \boldsymbol{d}_{k-1}^T \boldsymbol{Q}\boldsymbol{d}_{k-1} \tag{5.29}$$

where the last equality follows from the fact that

$$E\left\{\boldsymbol{v}_{n-k}\boldsymbol{v}_{n-r}^T\right\} = \begin{cases} \boldsymbol{Q} & \text{for } k = r \\ \boldsymbol{0} & \text{for } k \neq r \end{cases} . \tag{5.30}$$

Since the variance is a function of the time index n, the noise term e_n is in general non-stationary. Except for the trivial case of zero state-covariance, the dynamic model can thus be used for modelling non-stationary observations as opposed to the static model. If we assume that the covariance matrix of the state noise has the structure $\boldsymbol{Q} = \mathrm{diag}(q_1, q_1, \cdots q_l, q_l, \cdots, q_L, q_L)$, the variance of e_n reduces to

$$\sigma_{e_n}^2 = 2\sum_{l=1}^{L} q_l \sum_{k=1}^{n-1} e^{-2\gamma_l(k-1)} = 2\sum_{l=1}^{L} q_l \frac{1 - e^{-2\gamma_l(n-1)}}{1 - e^{-2\gamma_l}} . \tag{5.31}$$

In the simplest case, we assume zero damping and $\boldsymbol{Q}$ to be isotropic, i.e, $\boldsymbol{Q} = q\boldsymbol{I}_{2L}$. This yields

$$\sigma_{e_n}^2 = 2qL(n-1). \tag{5.32}$$

Thus for these approximations, the variance increases linearly with time. This should come as no surprise since we allow the amplitudes to develop as an AR(1) process in the dynamic model.

5.3 Summary

In this chapter, we have rewritten the static sinusoidal model into the dynamic sinusoidal model, and we have established a connection between the two models. The static sinusoidal model is the most popular sinusoidal model, but the dynamic sinusoidal model is more flexible and can be seen as a generalisation of the static model. In this generalisation, the amplitudes are allowed to evolve as a first-order autoregressive process, and the dynamic sinusoidal model is therefore able to model non-stationary signals which

are often encountered in practice. We can therefore expect the dynamic model to be able to accurately model a much broader class of signals as compared against the static model. The cost of the increased flexibility is that the number of unknown parameters is much larger in the dynamic model than in the static model. In the next chapter, we propose and develop an inference scheme for the unknown parameters of the dynamic model. Although this inference scheme suffers from a high computational complexity, it has some clear advantages over the inference schemes outlined in chapter 4.

Chapter 6

Derivation of Inference Scheme

The dynamic sinusoidal model derived in the previous chapter has previously been treated in a Bayesian framework for the application of music transcription (see, e.g., [Cemgil, 2004], [Cemgil and Godsill, 2005] and [Cemgil et al., 2006]). In these papers, however, the sinusoidal frequency is defined as a discrete random variable corresponding to the pitch of a musical note. The inference schemes suggested in the papers are based on analytical approximations including variational inference. In this thesis, we do not restrict ourselves to the application of music transcription and we therefore treat the frequency as a continuous random variable. Also, we base our inference scheme on stochastic methods such as the Gibbs sampler since it turns out to be a feasible inference scheme for the dynamic model. In this chapter, we develop this inference scheme.

6.1 Definitions and Problem Formulation

Having obtained the dynamic sinusoidal formulation in Eq. (5.17), we are now able to derive a Bayesian inference scheme for the latent states and model parameters of it. The state space model can also be defined in terms of three Gaussian distributions given by

$$p(y_n|\boldsymbol{s}_n, \sigma_w^2) = \mathcal{N}(y_n; \boldsymbol{b}^T\boldsymbol{s}_n, \sigma_w^2) \tag{6.1}$$

$$p(\boldsymbol{s}_{n+1}|\boldsymbol{s}_n, \boldsymbol{A}, \boldsymbol{Q}) = \mathcal{N}(\boldsymbol{s}_{n+1}; \boldsymbol{A}\boldsymbol{s}_n, \boldsymbol{Q}) \tag{6.2}$$

$$p(\boldsymbol{s}_1) = \mathcal{N}(\boldsymbol{s}_1; \boldsymbol{\mu}, \boldsymbol{P}) \ , \tag{6.3}$$

for $n = 1, \cdots, N$, where y_n is the scalar observation at time index n, $\boldsymbol{s}_n$ is the $2L$-dimensional state vector at time index n, $\boldsymbol{b}$ is the $2L$-dimensional system observation vector, $\boldsymbol{A}$ is the $2L \times 2L$ sinusoidal block-diagonal matrix, σ_w^2 is the variance of the observation noise, $\boldsymbol{Q}$ is the $2L \times 2L$ state noise covariance matrix, and the $2L$-dimensional vector $\boldsymbol{\mu}$ and the $2L \times 2L$ matrix $\boldsymbol{P}$ are the mean vector and covariance matrix, respectively, of the initial state $\boldsymbol{s}_1$. The sinusoidal matrix $\boldsymbol{A}$ is dependent on the L-dimensional

vectors of frequency parameters $\boldsymbol{\omega}$ and log-damping coefficient $\boldsymbol{\gamma}$ in a non-linear but deterministic way given by Eq. (5.13) and Eq. (5.14).

The model contains several random variables which we, according to our definitions in chapter 2, divide in three groups and denote as

$$\begin{aligned}
&\text{Observations:} && y_{1:N} = \{y_1, y_2, \cdots, y_N\} \\
&\text{Latent variables:} && \boldsymbol{s}_{1:N} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_N\} \\
&\text{Model parameters:} && \boldsymbol{\theta} = \{\boldsymbol{Q}, \sigma_w^2, \boldsymbol{\omega}, \boldsymbol{\gamma}\} \\
& && \boldsymbol{\omega} = \begin{bmatrix}\omega_1 & \omega_2 & \cdots & \omega_L\end{bmatrix}^T \\
& && \boldsymbol{\gamma} = \begin{bmatrix}\gamma_1 & \gamma_2 & \cdots & \gamma_L\end{bmatrix}^T
\end{aligned}$$

The model parameters $\boldsymbol{\theta}$ can be further divided into desired parameters and nuisance parameters depending on the specific inference problem. The prior distributions for the unknown parameters are parametrised by hyperparameters $\boldsymbol{\phi}$. We base our inference scheme on conjugate priors since they are, in most cases, general enough and they are also convenient to work with from a mathematical perspective. We assume the initial state statistics $\boldsymbol{\mu}$ and $\boldsymbol{P}$ to be known.

With these definitions and assumptions, we can now state the inference problem as the construction of the full joint posterior distribution for the unknown variables given the observations $y_{1:N}$, i.e.,

$$p(\boldsymbol{s}_{1:N}, \boldsymbol{Q}, \sigma_w^2, \boldsymbol{\omega}, \boldsymbol{\gamma} | y_{1:N}) = p(\boldsymbol{s}_{1:N}, \boldsymbol{\theta} | y_{1:N}) \ , \tag{6.4}$$

or as the construction of the marginal posterior distributions such as

$$p(\boldsymbol{s}_{1:N} | y_{1:N}) = \int p(\boldsymbol{s}_{1:N}, \boldsymbol{\theta} | y_{1:N}) d\boldsymbol{Q} d\sigma_w^2 d\boldsymbol{\omega} d\boldsymbol{\gamma} \tag{6.5a}$$

$$p(\boldsymbol{Q} | y_{1:N}) = \int p(\boldsymbol{s}_{1:N}, \boldsymbol{\theta} | y_{1:N}) d\boldsymbol{s}_{1:N} d\sigma_w^2 d\boldsymbol{\omega} d\boldsymbol{\gamma} \tag{6.5b}$$

$$p(\sigma_w^2 | y_{1:N}) = \int p(\boldsymbol{s}_{1:N}, \boldsymbol{\theta} | y_{1:N}) d\boldsymbol{s}_{1:N} d\boldsymbol{Q} d\boldsymbol{\omega} d\boldsymbol{\gamma} \tag{6.5c}$$

$$p(\boldsymbol{\omega} | y_{1:N}) = \int p(\boldsymbol{s}_{1:N}, \boldsymbol{\theta} | y_{1:N}) d\boldsymbol{s}_{1:N} d\boldsymbol{Q} d\sigma_w^2 d\boldsymbol{\gamma} \tag{6.5d}$$

$$p(\boldsymbol{\gamma} | y_{1:N}) = \int p(\boldsymbol{s}_{1:N}, \boldsymbol{\theta} | y_{1:N}) d\boldsymbol{s}_{1:N} d\boldsymbol{Q} d\sigma_w^2 d\boldsymbol{\omega} \ . \tag{6.5e}$$

The joint as well as the marginal posterior distributions are non-standard distribution. Thus, we have to resort to numerical techniques in order to solve the inference problem. In this chapter, we develop a numerical inference scheme based on a Gibbs sampler.

6.2 Bayesian Inference using a Gibbs Sampler

The inference problem can be solved using a Gibbs sampler which we discussed in section 3.2.3. In an alternating pattern, it draws samples from the desired marginal distributions by drawing samples from the distributions of the individual unknown variables conditioned on the remaining unknown variables, the observations and the prior information. Thus, we have to derive expressions for the conditional distributions given by

State:	$p(\boldsymbol{s}_{1:N}\|\boldsymbol{Q}, \sigma_w^2, \boldsymbol{\omega}, \boldsymbol{\gamma}, y_{1:N})$
State covariance matrix:	$p(\boldsymbol{Q}\|\boldsymbol{s}_{1:N}, \sigma_w^2, \boldsymbol{\omega}, \boldsymbol{\gamma}, y_{1:N})$
Observation variance:	$p(\sigma_w^2\|\boldsymbol{s}_{1:N}, \boldsymbol{Q}, \boldsymbol{\omega}, \boldsymbol{\gamma}, y_{1:N})$
Frequency:	$p(\boldsymbol{\omega}\|\boldsymbol{s}_{1:N}, \boldsymbol{Q}, \sigma_w^2, \boldsymbol{\gamma}, y_{1:N})$
Log-damping:	$p(\boldsymbol{\gamma}\|\boldsymbol{s}_{1:N}, \boldsymbol{Q}, \sigma_w^2, \boldsymbol{\omega}, y_{1:N})$

in the case where nothing but the model structure and the observations are given. If some of the parameters are known, the steps in which samples are generated for them are simply skipped. The derivation of these five conditional distributions is presented in the next five sections.

6.2.1 Conditional Distribution for the States

By using Bayes' theorem, the conditional state distribution can be factored as

$$p(\boldsymbol{s}_{1:N}|\boldsymbol{\theta}, y_{1:N}) \propto p(y_{1:N}|\boldsymbol{s}_{1:N}, \boldsymbol{\theta}) p(\boldsymbol{s}_{1:N}|\boldsymbol{\theta}) \ . \tag{6.6}$$

For known model parameters $\boldsymbol{\theta}$, the conditional state distribution $p(\boldsymbol{s}_{1:N}|\boldsymbol{\theta}, y_{1:N})$ thus factors into a conditional distribution $p(y_{1:N}|\boldsymbol{s}_{1:N}, \boldsymbol{\theta})$ and a marginal distribution $p(\boldsymbol{s}_{1:N}|\boldsymbol{\theta})$. By using the Markov property, the latter can be factored as

$$\begin{aligned} p(\boldsymbol{s}_{1:N}|\boldsymbol{\theta}) &= \left[\prod_{n=1}^{N-1} p(\boldsymbol{s}_{n+1}|\boldsymbol{s}_n, \boldsymbol{Q}, \boldsymbol{\omega}, \boldsymbol{\gamma})\right] p(\boldsymbol{s}_1) \\ &= \left[\prod_{n=1}^{N-1} \mathcal{N}(\boldsymbol{s}_{n+1}; \boldsymbol{A}\boldsymbol{s}_n, \boldsymbol{Q})\right] \mathcal{N}(\boldsymbol{s}_1; \boldsymbol{\mu}, \boldsymbol{P}) \end{aligned} \tag{6.7}$$

where the last equality follows from the state space formulation in Eq. (6.2) and Eq. (6.3). We can now write the marginal distribution as

$$\begin{aligned} p(\boldsymbol{s}_{1:N}|\boldsymbol{\theta}) &\propto \exp\left\{\frac{-1}{2}\left[\sum_{n=1}^{N-1}(\boldsymbol{s}_{n+1}-\boldsymbol{A}\boldsymbol{s}_n)^T\boldsymbol{Q}^{-1}(\boldsymbol{s}_{n+1}-\boldsymbol{A}\boldsymbol{s}_n)+(\boldsymbol{s}_1-\boldsymbol{\mu})^T\boldsymbol{P}^{-1}(\boldsymbol{s}_1-\boldsymbol{\mu})\right]\right\} \\ &= \exp\Bigg\{\frac{-1}{2}\Bigg[\boldsymbol{s}_N^T\boldsymbol{Q}^{-1}\boldsymbol{s}_N+\sum_{n=2}^{N-1}\boldsymbol{s}_n^T(\boldsymbol{Q}^{-1}+\boldsymbol{A}^T\boldsymbol{Q}^{-1}\boldsymbol{A})\boldsymbol{s}_n+\boldsymbol{s}_1^T(\boldsymbol{P}^{-1}+\boldsymbol{A}^T\boldsymbol{Q}^{-1}\boldsymbol{A})\boldsymbol{s}_1 \\ &\quad -\sum_{n=1}^{N-1}\boldsymbol{s}_{n+1}^T\boldsymbol{Q}^{-1}\boldsymbol{A}\boldsymbol{s}_n-\sum_{n=1}^{N-1}\boldsymbol{s}_n^T\boldsymbol{A}^T\boldsymbol{Q}^{-1}\boldsymbol{s}_{n+1}-2\boldsymbol{s}_1^T\boldsymbol{P}^{-1}\boldsymbol{\mu}+\boldsymbol{\mu}^T\boldsymbol{P}^{-1}\boldsymbol{\mu}\Bigg]\Bigg\}. \end{aligned} \tag{6.8}$$

Using the method of *completing the squares* (see, e.g., the proof of result A.1), we can construct the multivariate Gaussian distribution for $\boldsymbol{s} = \text{vec}(\boldsymbol{s}_{1:N}) = \begin{bmatrix}\boldsymbol{s}_1^T & \boldsymbol{s}_2^T & \cdots & \boldsymbol{s}_N^T\end{bmatrix}^T$ as

$$p(\boldsymbol{s}|\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{s}; \boldsymbol{\mu}_{\boldsymbol{s}|\boldsymbol{\theta}}, \boldsymbol{C}_{\boldsymbol{s}|\boldsymbol{\theta}}) \tag{6.9}$$

where

$$\boldsymbol{\mu}_{\boldsymbol{s}|\boldsymbol{\theta}} = \begin{bmatrix}\boldsymbol{\mu}^T & (\boldsymbol{A}\boldsymbol{\mu})^T & (\boldsymbol{A}^2\boldsymbol{\mu})^T & \cdots & (\boldsymbol{A}^{N-1}\boldsymbol{\mu})^T\end{bmatrix}^T \tag{6.10}$$

$$\boldsymbol{C}_{\boldsymbol{s}|\boldsymbol{\theta}}^{-1} = \begin{bmatrix} \boldsymbol{P}^{-1}+\boldsymbol{A}^T\boldsymbol{Q}^{-1}\boldsymbol{A} & -\boldsymbol{A}^T\boldsymbol{Q}^{-1} & \cdots & \boldsymbol{0} & \boldsymbol{0} \\ -\boldsymbol{Q}^{-1}\boldsymbol{A} & \boldsymbol{Q}^{-1}+\boldsymbol{A}^T\boldsymbol{Q}^{-1}\boldsymbol{A} & \cdots & \boldsymbol{0} & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{Q}^{-1}+\boldsymbol{A}^T\boldsymbol{Q}^{-1}\boldsymbol{A} & -\boldsymbol{A}^T\boldsymbol{Q}^{-1} \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & -\boldsymbol{Q}^{-1}\boldsymbol{A} & \boldsymbol{Q}^{-1} \end{bmatrix} \tag{6.11}$$

Notice that the inverse covariance matrix is block triagonal.

The construction of the conditional distribution $p(y_{1:N}|\boldsymbol{s}_{1:N}, \boldsymbol{\theta})$ is easier. We again use the Markov property and obtain

$$p(y_{1:N}|\boldsymbol{s}_{1:N}, \boldsymbol{\theta}) = \prod_{n=1}^{N} p(y_n|\boldsymbol{s}_n, \sigma_w^2) = \prod_{n=1}^{N}\mathcal{N}(y_n; \boldsymbol{b}^T\boldsymbol{s}_n, \sigma_w^2) \tag{6.12}$$

$$\propto \exp\left\{\frac{-1}{2\sigma_w^2}\sum_{n=1}^{N}(y_n-\boldsymbol{b}^T\boldsymbol{s}_n)^T(y_n-\boldsymbol{b}^T\boldsymbol{s}_n)\right\} \tag{6.13}$$

$$\propto \mathcal{N}(\boldsymbol{y}; \boldsymbol{B}\boldsymbol{s}, \sigma_w^2\boldsymbol{I}_N) \tag{6.14}$$

where $\boldsymbol{I}_N$ is the $N \times N$ identity matrix and

$$\boldsymbol{y} = \text{vec}(y_{1:N}) = \begin{bmatrix}y_1 & y_2 & \cdots & y_N\end{bmatrix}^T \tag{6.15}$$

$$\boldsymbol{B} = \boldsymbol{I}_N \otimes \boldsymbol{b}^T\ . \tag{6.16}$$

The symbol $\otimes$ denotes the Kronecker product.

From these derivations, it is now clear that we can rewrite Eq. (6.6) as

$$\begin{aligned} p(\boldsymbol{s}_{1:N}|\boldsymbol{\theta}, y_{1:N}) = p(\boldsymbol{s}|\boldsymbol{y}, \boldsymbol{\theta}) &\propto p(\boldsymbol{y}|\boldsymbol{s}, \boldsymbol{\theta}) p(\boldsymbol{s}|\boldsymbol{\theta}) \\ &= \mathcal{N}(\boldsymbol{y}; \boldsymbol{B}\boldsymbol{s}, \sigma_w^2 \boldsymbol{I}_N) \mathcal{N}(\boldsymbol{s}; \boldsymbol{\mu}_{\boldsymbol{s}|\boldsymbol{\theta}}, \boldsymbol{C}_{\boldsymbol{s}|\boldsymbol{\theta}}) \qquad (6.17) \\ &\propto \mathcal{N}(\boldsymbol{s}; \boldsymbol{\mu}_{\boldsymbol{s}|\boldsymbol{y},\boldsymbol{\theta}}, \boldsymbol{C}_{\boldsymbol{s}|\boldsymbol{y},\boldsymbol{\theta}}) \qquad (6.18) \end{aligned}$$

where the last proportional sign follows from result B.1 and

$$\boldsymbol{\mu}_{\boldsymbol{s}|\boldsymbol{y},\boldsymbol{\theta}} = \boldsymbol{C}_{\boldsymbol{s}|\boldsymbol{y},\boldsymbol{\theta}} (\sigma_w^{-2} \boldsymbol{B}^T \boldsymbol{y} + \boldsymbol{C}_{\boldsymbol{s}|\boldsymbol{\theta}}^{-1} \boldsymbol{\mu}_{\boldsymbol{s}|\boldsymbol{\theta}}) \qquad (6.19a)$$

$$\boldsymbol{C}_{\boldsymbol{s}|\boldsymbol{y},\boldsymbol{\theta}} = (\boldsymbol{C}_{\boldsymbol{s}|\boldsymbol{\theta}}^{-1} + \sigma_w^{-2} \boldsymbol{B}^T \boldsymbol{B})^{-1} \ . \qquad (6.19b)$$

Thus, we can obtain a sample for the states $\boldsymbol{s}_{1:N}$ by sampling from a multivariate Gaussian distribution of dimension $2LN \times 1$. In most cases, this dimension leads to a very high computational complexity. The remedy for this is to use the simulation smoother which is a more efficient sampling scheme for drawing one or more samples for $\boldsymbol{s}_{1:N}$.

Simulation Smoothing

Samples for the states conditioned on the observations can be obtained in a more efficient way by using a recursive technique based on the Kalman filter and smoother (see appendix C). One of the first algorithms employing this technique was proposed in [Carter and Kohn, 1994] and it was based on the factorisation[1]

$$\begin{aligned} p(\boldsymbol{s}_{1:N}|y_{1:N}) &= p(\boldsymbol{s}_1|\boldsymbol{s}_{2:N}, y_{1:N}) p(\boldsymbol{s}_2|\boldsymbol{s}_{3:N}, y_{1:N}) \cdots p(\boldsymbol{s}_{N-1}|\boldsymbol{s}_N, y_{1:N}) p(\boldsymbol{s}_N|y_{1:N}) \qquad (6.20) \\ &= p(\boldsymbol{s}_N|y_{1:N}) \prod_{n=1}^{N-1} p(\boldsymbol{s}_n|\boldsymbol{s}_{n+1:N}, y_{1:N}) \qquad (6.21) \\ &= p(\boldsymbol{s}_N|y_{1:N}) \prod_{n=1}^{N-1} p(\boldsymbol{s}_n|\boldsymbol{s}_{n+1}, y_{1:n}) \ . \qquad (6.22) \end{aligned}$$

where the last equality follows from the fact that $\boldsymbol{s}_n$ is independent of $y_{n+1:N}$ and $\boldsymbol{s}_{n+2:N}$ given $\boldsymbol{s}_{n+1}$ and $\boldsymbol{y}_{1:n}$[2]. The distribution $p(\boldsymbol{s}_n|\boldsymbol{s}_{n+1}, y_{1:n})$ is a Gaussian distribution and given by Eq. (C.31). Its mean and variance are calculated by running the Kalman filter which computes the means and variances of $p(\boldsymbol{s}_n|y_{1:n})$ and $p(\boldsymbol{s}_{n+1}|y_{1:n})$ for $n =$

[1] We have omitted the conditioning on the model parameters $\boldsymbol{\theta}$ in order to keep the notation uncluttered.

[2] It can sometimes be hard to determine whether two random variables are conditional independent. A systematic way of determining this is to use graphical models. For an introduction to graphical models see [Bishop, 2006, ch. 8].

$1, \cdots, N$. Thus, the last mean and variance computed by the Kalman filter are the moments of $p(\boldsymbol{s}_N|y_{1:N})$ from which we draw a sample $\boldsymbol{s}_N^*$. Inserting this sample as well as the means and variances of $p(\boldsymbol{s}_{N-1}|y_{1:N-1})$ and $p(\boldsymbol{s}_N|y_{1:n-1})$, which are computed in the $N-1$'th iteration of the Kalman filter, into Eq. (C.31) enables us hereafter to draw a sample $\boldsymbol{s}_{N-1}^*$ from $p(\boldsymbol{s}_{N-1}|\boldsymbol{s}_N^*, y_{1:N-1})$. Continuing in this way enables us to draw a sample for all states. The computational complexity of this algorithm was decreased later by [De Jong and Shephard, 1995] and even further by [Durbin and Koopman, 2002]. This was achieved by using the disturbance smoother [Koopman, 1993] which computes the posterior distribution for the observation and state noise. The disturbance smoother thus acts in an analogue way to the Kalman smoother and a thorough treatment of it is given in [Durbin and Koopman, 2001]. The simulation smoother based on the disturbance smoother as proposed in [Durbin and Koopman, 2002] is outlined in algorithm 6.1. The algorithm basically consists of three steps. The first step is a Kalman filtering step which is run in the forward time direction from $n = 1, \cdots, N$. The second step is a disturbance smoothing step in which smoothed values for the state noise is computed from a simple recursion in the reverse time direction from $n = N, \cdots, 1$. In the third and final step, the state equation of Eq. (5.17) is run in forward time direction from $n = 1, \cdots, N$.

Algorithm 6.1 (Simulation Smoother)
A set of samples $\boldsymbol{s}_n^*$ for $n = 1, \cdots, N$ from the posterior distribution $p(\boldsymbol{s}_{1:N}|y_{1:N})$ for the states given the observation can be drawn by

1. Draw a sample $\boldsymbol{s}_1^+$ from $p(\boldsymbol{s}_1) = \mathcal{N}(\boldsymbol{s}_1; \boldsymbol{\mu}, \boldsymbol{P})$ given by Eq. (6.3).
2. For $n = 1, \cdots, N$ do
   (a) Draw a sample w_n^+ from $p(w_n) = \mathcal{N}(w_n; 0, \sigma_w^2)$.
   (b) Compute a simulated observation y_n^+ from the observation equation of Eq. (5.17) by inserting $\boldsymbol{s}_n^+$ and w_n^+, i.e.,
   $$y_n^+ = \boldsymbol{b}^T \boldsymbol{s}_n^+ + w_n^+ \ .$$
   (c) Draw a sample $\boldsymbol{v}_n^+$ from $p(\boldsymbol{v}_n) = \mathcal{N}(\boldsymbol{v}_n; \boldsymbol{0}, \boldsymbol{Q})$.
   (d) Compute a simulated state $\boldsymbol{s}_{n+1}^+$ from the state equation of Eq. (5.17) by inserting $\boldsymbol{s}_n^+$ and $\boldsymbol{v}_n^+$, i.e.,
   $$\boldsymbol{s}_{n+1}^+ = \boldsymbol{A}\boldsymbol{s}_n^+ + \boldsymbol{v}_n^+ \ .$$

3. Define the new observation $y'_n = y_n - y_n^+$ for $n = 1, \cdots, N$. Using these observations, run the Kalman filter by means of the recursion given by Eq. (C.24).

4. Compute the means $\hat{\boldsymbol{v}}'_n$ of the smoothed posterior distributions for the state noise $p(\boldsymbol{v}'_n|y'_{1:N})$ for $n = N, \cdots, 1$ by running the recursion

$$\begin{aligned}\hat{\boldsymbol{v}}'_n &= \boldsymbol{Q}\boldsymbol{r}_n \\ \boldsymbol{r}_{n-1} &= \boldsymbol{b}\boldsymbol{F}_n^{-1}\boldsymbol{v}_n + \boldsymbol{L}_n^T\boldsymbol{r}_n\end{aligned}$$

   where $\boldsymbol{F}_n$ and $\boldsymbol{L}_n$ were computed by the Kalman filter in the previous step, and $\boldsymbol{r}_N = \boldsymbol{0}$.

5. Compute the means $\hat{\boldsymbol{s}}'_n$ of the smoothed posterior distributions for the state $p(\boldsymbol{s}'_n|y'_{1:N})$ for $n = 1, \cdots, N$ by running the state equation of Eq. (5.17) using the smoothed state noise computed in the previous step, i.e.,

$$\hat{\boldsymbol{s}}'_{n+1} = \boldsymbol{A}\hat{\boldsymbol{s}}'_n + \hat{\boldsymbol{v}}'_n$$

   where $\hat{\boldsymbol{s}}'_1 = \boldsymbol{P}\boldsymbol{r}_0$.

6. The samples $\boldsymbol{s}_n^*$ for $n = 1, \cdots, N$ from $p(\boldsymbol{s}_{1:N}|y_{1:N})$ can now be computed from

$$\boldsymbol{s}_n^* = \hat{\boldsymbol{s}}'_n + \boldsymbol{s}_n^+ \ . \tag{6.26}$$

This concludes the derivation of the sampling scheme for the conditional distribution $p(\boldsymbol{s}_{1:N}|\boldsymbol{Q}, \sigma_w^2, \boldsymbol{\omega}, \boldsymbol{\gamma}, y_{1:N})$ for the states. Although the simulation smoother reduces the computational complexity as compared against direct sampling from the multivariate normal distribution given by Eq. (6.18), the computational complexity of the algorithm is still very high.

6.2.2 Conditional Distribution for the State Noise Covariance

By exploiting conditional independence as well as using Bayes' theorem, we can rewrite the conditional state noise covariance matrix distribution as

$$p(\boldsymbol{Q}|\boldsymbol{s}_{1:N}, \sigma_w^2, \boldsymbol{\omega}, \boldsymbol{\gamma}, y_{1:N}) = p(\boldsymbol{Q}|\boldsymbol{s}_{1:N}, \boldsymbol{\omega}, \boldsymbol{\gamma}) \propto p(\boldsymbol{s}_{1:N}|\boldsymbol{Q}, \boldsymbol{\omega}, \boldsymbol{\gamma})p(\boldsymbol{Q}|\boldsymbol{\phi}_{\boldsymbol{Q}}) \tag{6.27}$$

where $\boldsymbol{\phi}_{\boldsymbol{Q}}$ denotes the hyperparameters for $\boldsymbol{Q}$. The conditional state noise covariance matrix distribution can thus be factored into the prior distribution $p(\boldsymbol{Q}|\boldsymbol{\phi}_{\boldsymbol{Q}})$ for $\boldsymbol{Q}$ and $p(\boldsymbol{s}_{1:N}|\boldsymbol{Q}, \boldsymbol{\omega}, \boldsymbol{\gamma})$. From Eq. (6.7), we already know that the parametric form of the latter

distribution. Here, however, the states are known and the state noise covariance matrix is unknown and we obtain

$$p(\boldsymbol{s}_{1:N}|\boldsymbol{Q},\boldsymbol{\omega},\boldsymbol{\gamma}) = p(\boldsymbol{s}_1)\left[\frac{1}{\sqrt{(2\pi)^{2L}|\boldsymbol{Q}|}}\right]^{N-1}\exp\left\{\frac{-1}{2}\sum_{n=1}^{N-1}\boldsymbol{v}_n^T\boldsymbol{Q}^{-1}\boldsymbol{v}_n\right\} \tag{6.28}$$

$$= p(\boldsymbol{s}_1)(2\pi)^{-(N-1)L}|\boldsymbol{Q}|^{-(N-1)/2}\exp\left\{\frac{-1}{2}\mathrm{tr}\left(\sum_{n=1}^{N-1}\boldsymbol{v}_n\boldsymbol{v}_n^T\boldsymbol{Q}^{-1}\right)\right\} \tag{6.29}$$

where $|\cdot|$ and $\mathrm{tr}(\cdot)$ denote the determinant and trace, respectively. Notice that the initial state distribution $p(\boldsymbol{s}_1)$ is independent of $\boldsymbol{Q}$ for which reason it is ignored at a later stage. The conjugate prior for this distribution is the inverse Wishart distribution Inv-$\mathcal{W}(\nu,\boldsymbol{\Psi})$ (see appendix A) with the hyperparameters $\boldsymbol{\phi}_{\boldsymbol{Q}} = \{\nu,\boldsymbol{\Psi}\}$ and given by

$$p(\boldsymbol{Q}|\boldsymbol{\phi}_{\boldsymbol{Q}}) = B(\nu,\boldsymbol{\Psi})|\boldsymbol{Q}|^{-(\nu+2L+1)/2}\exp\left\{-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Psi}\boldsymbol{Q}^{-1})\right\} \tag{6.30}$$

where

$$B(\nu,\boldsymbol{\Psi}) = \frac{|\boldsymbol{\Psi}|^{\nu/2}}{2^{\nu L}\pi^{L(2L-1)/2}\prod_{i=1}^{2L}\Gamma(\frac{\nu+1-i}{2})} \tag{6.31}$$

and $\Gamma(\cdot)$ is the gamma function. Since the inverse Wishart distribution is the conjugate prior for $\boldsymbol{Q}$ in $p(\boldsymbol{s}_{1:N}|\boldsymbol{Q},\boldsymbol{\omega},\boldsymbol{\gamma})$, the posterior distribution is also an inverse Wishart distribution. Inserting the inverse Wishart distribution for the prior in Eq. (6.27) and combining terms readily yields

$$p(\boldsymbol{Q}|\boldsymbol{s}_{1:N},\boldsymbol{\omega},\boldsymbol{\gamma},\boldsymbol{\phi}_{\boldsymbol{Q}}) = \text{Inv-}\mathcal{W}\left(\boldsymbol{Q};\nu+N-1,\boldsymbol{\Psi}+\sum_{n=1}^{N-1}\boldsymbol{v}_n\boldsymbol{v}_n^T\right)\ . \tag{6.32}$$

The hyperparameters ν and $\boldsymbol{\Psi}$ of the prior distribution are referred to as the degrees of freedom and the scale matrix, respectively, and they can be selected such that the prior is diffuse or concentrated. Although it is possible to draw samples from the inverse Wishart distribution, there is a couple of reasons for avoiding it. First of all, the number of elements of $\boldsymbol{Q}$ grows quadratically with the number of sinusoids and the computational complexity associated with sampling from the inverse Wishart distribution is high. Secondly, it turns out that assuming a diagonal structure for $\boldsymbol{Q}$ greatly simplifies the inference task for the frequency parameters and log-damping coefficients as well as reduces the computational complexity of the simulation smoother. In the rest of this thesis, we therefore restrict ourselves to two different structures of the state covariance: the isotropic structure given by $\boldsymbol{Q} = q\boldsymbol{I}_{2L}$ and the diagonal structure given by

$$\boldsymbol{Q} = \mathrm{diag}(q_1,q_1,\cdots q_l,q_l,\cdots,q_L,q_L)\ . \tag{6.33}$$

Assume that $\boldsymbol{Q}$ is isotropic. If we insert this into Eq. (6.30), we obtain the univariate inverse Wishart distribution which is the same as the inverse gamma distribution Inv-$\mathcal{G}(a_v, b_v)$ and given by

$$p(q|\boldsymbol{\phi}_q) = \frac{1}{\Gamma(a_v)} b_v^{a_v} q^{-(a_v+1)} \exp\left\{\frac{-b_v}{q}\right\} \tag{6.34}$$

where the hyperparameters $\boldsymbol{\phi}_q = \{a_v, b_v\}$ of the inverse gamma distribution are related to the hyperparameters of the univariate Wishart distribution by $a_v = \nu/2$ and $b_v = \Psi/2$. Inserting the isotropic covariance matrix into Eq. (6.29) and combining it with the inverse gamma prior distribution yield

$$p(q|\boldsymbol{s}_{1:N}, \boldsymbol{\omega}, \boldsymbol{\gamma}, \boldsymbol{\phi}_q) = \text{Inv-}\mathcal{G}\left(q; a_v + (N-1)L, b_v + \frac{1}{2}\sum_{n=1}^{N-1} \boldsymbol{v}_n^T \boldsymbol{v}_n\right) . \tag{6.35}$$

If we assume that $\boldsymbol{Q}$ has a diagonal structure, then, by the same arguments as above, we can decouple the sampling problem into sampling from L inverse Gamma distributions given by

$$\begin{aligned}
&p(q_1, \cdots, q_L|\boldsymbol{s}_{1:N}, \boldsymbol{\omega}, \boldsymbol{\gamma}, \boldsymbol{\phi}_{q_1}, \cdots, \boldsymbol{\phi}_{q_L}) \\
&\quad\propto \left[\prod_{l=1}^{L} q_l^2\right]^{-(N-1)/2} \prod_{l=1}^{L} \exp\left\{\frac{-1}{2q_l}\sum_{n=1}^{N-1} \boldsymbol{v}_{n,l}^T \boldsymbol{v}_{n,l}\right\} \prod_{l=1}^{L} p(q_l|a_{v,l}, b_{v,l}) && (6.36)\\
&\quad= \prod_{l=1}^{L}\left[q_l^{-(N-1)} \exp\left\{\frac{-1}{2q_l}\sum_{n=1}^{N-1} \boldsymbol{v}_{n,l}^T \boldsymbol{v}_{n,l}\right\} p(q_l|a_{v,l}, b_{v,l})\right] && (6.37)\\
&\quad\propto \prod_{l=1}^{L} \text{Inv-}\mathcal{G}\left(q_l; a_{v,l} + N - 1, b_{v,l} + \frac{1}{2}\sum_{n=1}^{N-1} \boldsymbol{v}_{n,l}^T \boldsymbol{v}_{n,l}\right) && (6.38)
\end{aligned}$$

where the subvector $\boldsymbol{v}_{n,l}$ contains the $(2l-1)$'th and $(2l)$'th element of $\boldsymbol{v}_n$.

There exist straightforward and efficient methods for drawing samples from the inverse gamma distribution. By assuming an isotropic or diagonal structure of the state noise covariance matrix, the number of unknown parameters is independent of the model order and grows linearly with it, respectively, which is a clear advantage from a computational complexity point of view as compared against the quadratic increase in the number of unknown parameters associated with the use of an inverse Wishart prior distribution.

6.2.3 Conditional Distribution for the Observation Variance

The conditional observation variance distribution can be derived by following the same procedure as for the conditional state covariance matrix distribution. Thus, we first

factorise it as

$$p(\sigma_w^2|\boldsymbol{s}_{1:N}, \boldsymbol{Q}, \boldsymbol{\omega}, \boldsymbol{\gamma}, y_{1:N}) = p(\sigma_w^2|\boldsymbol{s}_{1:N}, y_{1:N}) \propto p(y_{1:N}|\boldsymbol{s}_{1:N}, \sigma_w^2) p(\sigma_w^2|\boldsymbol{\phi}_{\sigma_w^2}) \ . \tag{6.39}$$

Then, we recall from Eq. (6.1) that $p(y_{1:N}|\boldsymbol{s}_{1:N}, \sigma_w^2)$ is given by

$$p(y_{1:N}|\boldsymbol{s}_{1:N}, \sigma_w^2) = \left(2\pi\sigma_w^2\right)^{-N/2} \exp\left\{\frac{-1}{2\sigma_w^2} \sum_{n=1}^{N} w_n^T w_n\right\} \tag{6.40}$$

so that, if we propose an inverse Gamma prior Inv-$\mathcal{G}(a_w, b_w)$ for the observation noise variance, we obtain

$$p(\sigma_w^2|\boldsymbol{s}_{1:N}, y_{1:N}, \boldsymbol{\phi}_{\sigma_w^2}) = \text{Inv-}\mathcal{G}\left(\sigma_w^2; a_w + N/2, b_w + \frac{1}{2}\sum_{n=1}^{N} w_n^T w_n\right) \ . \tag{6.41}$$

6.2.4 Conditional Distribution for the Frequency Parameters

For a diagonal state covariance matrix $\boldsymbol{Q}$, it turns out that we are able to decouple the L-dimensional posterior distribution $p(\boldsymbol{\omega}|\boldsymbol{s}_{1:N}, \sigma_w^2, \boldsymbol{Q}, \boldsymbol{\gamma}, y_{1:N})$ into L univariate distributions

$$p(\omega_l|\boldsymbol{s}_{1:N}, \sigma_w^2, \boldsymbol{Q}, \boldsymbol{\omega}_{\setminus l}, \boldsymbol{\gamma}, y_{1:N}) \tag{6.42}$$

for $l = 1, \cdots, L$ where $(\cdot)_{\setminus l}$ denotes 'without element l'. The l'th distribution can by using conditional independence and Bayes' theorem be factored as

$$p(\omega_l|\boldsymbol{s}_{1:N}, \boldsymbol{Q}, \sigma_w^2, \boldsymbol{\omega}_{\setminus l}, \boldsymbol{\gamma}, y_{1:N}) = p(\omega_l|\boldsymbol{s}_{1:N}, \boldsymbol{Q}, \boldsymbol{\omega}_{\setminus l}, \boldsymbol{\gamma}) \propto p(\boldsymbol{s}_{1:N}|\boldsymbol{Q}, \boldsymbol{\omega}, \boldsymbol{\gamma}) p(\omega_l) \tag{6.43}$$

where $p(\omega_l)$ is the prior distribution for ω_l left undefined for the moment. Since $\boldsymbol{A}$ is 2×2 block diagonal and $\boldsymbol{Q}$ is diagonal, we can factor the likelihood of the state equation into L bivariate likelihoods with the l'th given by

$$p(\boldsymbol{s}_{1:N,l}|q_l, \omega_l, \gamma_l) = (2\pi q_l)^{-(N-1)} \exp\left\{\frac{-1}{2q_l} \sum_{n=1}^{N-1} (\boldsymbol{s}_{n+1,l} - \boldsymbol{A}_l \boldsymbol{s}_{n,l})^T (\boldsymbol{s}_{n+1,l} - \boldsymbol{A}_l \boldsymbol{s}_{n,l})\right\} \tag{6.44}$$

$$= \tilde{Z}_l^{-1} \exp\left\{\frac{-1}{2q_l} \sum_{n=1}^{N-1} \boldsymbol{s}_{n,l}^T \boldsymbol{A}_l^T \boldsymbol{A}_l \boldsymbol{s}_{n,l} + \frac{1}{q_l} \sum_{n=1}^{N-1} \boldsymbol{s}_{n+1,l}^T \boldsymbol{A}_l \boldsymbol{s}_{n,l}\right\} \tag{6.45}$$

where $\boldsymbol{A}_l$ is given by Eq. (5.14), $\boldsymbol{s}_{(\cdot),l}$ is the l'th 2×1 subvector of $\boldsymbol{s}_{(\cdot)}$ and $\tilde{Z}_l$ is a normalisation constant containing the terms independent of $\boldsymbol{A}_l$. Inserting the expression

for $\boldsymbol{A}_l$ and evaluating the two terms of the exponent in the exponential yield

$$\begin{aligned}\frac{-1}{2q_l}\sum_{n=1}^{N-1}\boldsymbol{s}_{n,l}^T\boldsymbol{A}_l^T\boldsymbol{A}_l\boldsymbol{s}_{n,l} &= \frac{-1}{2q_l}\sum_{n=1}^{N-1}\boldsymbol{s}_{n,l}^T e^{-2\gamma_l}\begin{bmatrix}\cos\omega_l & -\sin\omega_l\\ \sin\omega_l & \cos\omega_l\end{bmatrix}\begin{bmatrix}\cos\omega_l & \sin\omega_l\\ -\sin\omega_l & \cos\omega_l\end{bmatrix}\boldsymbol{s}_{n,l}\\ &= \frac{-1}{2q_l}\sum_{n=1}^{N-1}\boldsymbol{s}_{n,l}^T e^{-2\gamma_l}\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\boldsymbol{s}_{n,l} = -e^{-2\gamma_l}\frac{1}{2q_l}\sum_{n=1}^{N-1}\boldsymbol{s}_{n,l}^T\boldsymbol{s}_{n,l}\\ &= -e^{-2\gamma_l}\alpha_{1,l} \end{aligned} \tag{6.46}$$

where we have defined

$$\alpha_{1,l} \triangleq \frac{1}{2q_l}\sum_{n=1}^{N-1}\boldsymbol{s}_{n,l}^T\boldsymbol{s}_{n,l}\ , \tag{6.47}$$

and

$$\frac{1}{q_l}\sum_{n=1}^{N-1}\boldsymbol{s}_{n+1,l}^T\boldsymbol{A}_l\boldsymbol{s}_{n,l} = e^{-\gamma_l}(d_{1,l}\cos\omega_l + d_{2,l}\sin\omega_l) \tag{6.48}$$

where we have defined

$$\begin{bmatrix}d_{1,l}\\ d_{2,l}\end{bmatrix} \triangleq \frac{1}{q_l}\sum_{n=1}^{N-1}\begin{bmatrix}\boldsymbol{s}_{n+1,l}^T\boldsymbol{s}_{n,l}\\ \Big|\begin{bmatrix}\boldsymbol{s}_{n+1,l} & \boldsymbol{s}_{n,l}\end{bmatrix}\Big|\end{bmatrix}\ . \tag{6.49}$$

The notation $|\cdot|$ denotes the determinant. Thus, the first term of the exponent in Eq. (6.45) is independent of ω_l whereas the second term depends on it. We can therefore write the likelihood in Eq. (6.45) as

$$p(\boldsymbol{s}_{1:N,l}|q_l,\omega_l,\gamma_l) = \tilde{Z}_l^{-1}\exp\left\{-e^{-2\gamma_l}\alpha_{1,l} + e^{-\gamma_l}(d_{1,l}\cos\omega_l + d_{2,l}\sin\omega_l)\right\} \tag{6.50}$$

$$= Z_l^{-1}\exp\left\{e^{-\gamma_l}d_{1,l}\cos\omega_l + e^{-\gamma_l}d_{2,l}\sin\omega_l\right\}\ . \tag{6.51}$$

The exponent now consists of a superposition of two sinusoids of the same frequency. Thus, if we define

$$\kappa_l \triangleq e^{-\gamma_l}\sqrt{d_{1,l}^2 + d_{2,l}^2} \tag{6.52a}$$

$$\psi_l \triangleq \arctan\frac{d_{2,l}}{d_{1,l}}\ , \tag{6.52b}$$

Eq. (6.45) can be rewritten as

$$p(\boldsymbol{s}_{1:N,l}|q_l,\omega_l,\gamma_l) = Z_l^{-1}\exp\left\{\kappa_l\cos(\psi_l - \omega_l)\right\} \tag{6.53}$$

which has the same parametric form as the von Mises distribution $\mathcal{VM}(\psi_l; \kappa_l, \omega_l)$ (see appendix A). Notice, that ω_l acts as the location parameter of this likelihood which is therefore unknown. The concentration parameter κ_l is known.

The conjugate prior for the von Mises likelihood with unknown location and known concentration is also the von Mises distribution [Guttorp and Lockhart, 1988]. Thus, if the prior distribution is given by $\mathcal{VM}(\omega_l; \kappa_0, \mu_0)$ with hyperparameters $\boldsymbol{\phi}_{\omega_l} = \{\kappa_0, \mu_0\}$, the posterior distribution in Eq. (6.43) is given by

$$p(\omega_l|\boldsymbol{s}_{1:N}, \boldsymbol{Q}, \boldsymbol{\omega}_{\setminus l}, \boldsymbol{\gamma}) = p(\omega_l|\boldsymbol{s}_{1:N,l}, q_l, \gamma_l) \propto p(\boldsymbol{s}_{1:N,l}|q_l, \omega_l, \gamma_l)p(\omega_l|\boldsymbol{\phi}_{\omega_l}) \tag{6.54}$$

$$\propto \mathcal{VM}(\psi_l; \kappa_l, \omega_l)\mathcal{VM}(\omega_l; \kappa_0, \mu_0) \tag{6.55}$$

$$\propto \mathcal{VM}\left(\omega_l; \sqrt{\delta_{1,l}^2 + \delta_{1,l}^2}, \arctan\frac{\delta_{2,l}}{\delta_{1,l}}\right) \tag{6.56}$$

where

$$\delta_{1,l} \triangleq \kappa_l \cos\psi_l + \kappa_0 \cos\mu_0 = e^{-\gamma_l} d_{1,l} + \kappa_0 \cos\mu_0 \tag{6.57a}$$

$$\delta_{2,l} \triangleq \kappa_l \sin\psi_l + \kappa_0 \sin\mu_0 = e^{-\gamma_l} d_{2,l} + \kappa_0 \sin\mu_0 \ . \tag{6.57b}$$

For $\kappa_0 \to 0$, we obtain the uniform prior in the limit on the interval $[-\pi, \pi]$ and the posterior distribution is proportional to the likelihood. Notice, that since the support of the von Mises prior is the interval $[-\pi, \pi]$, it is wider than the support $[0, \pi]$ for ω_l in the signal model given by Eq. (5.1). Our inference scheme has therefore the defect that is assigns non-zero probabilities to negative frequencies. This can be avoided by using another prior with the same support as for the frequencies. However, this leads to a more intractable sampling scheme so we retain the von Mises prior distribution.

Simulation From the von Mises Distribution

It is not possible to draw samples directly from the Von Mises distribution, but samples can be obtained by using an efficient rejection sampling scheme. The algorithm is given in [Best and Fisher, 1979] and restated in algorithm 6.2 for easy reference. The acceptance ratio depends on the value of κ, however, it can be shown that the minimum expected acceptance ratio is $\sqrt{e/2\pi} \approx 0.66$ [Best and Fisher, 1979].

Algorithm 6.2 (Best-Fisher Algorithm)
Given the concentration κ and location μ of the von Mises distribution $\mathcal{VM}(\omega; \kappa, \mu)$, a sample can be drawn by

1. Calculate

$$\tau = 1 + \sqrt{1 + 4\kappa^2} \ , \quad \rho = \frac{\tau - \sqrt{2\tau}}{2\kappa} \ , \quad r = \frac{1 + \rho^2}{2\rho} \ .$$

2. Repeat until sample is accepted
   (a) Draw two independent random variates, u_1 and u_2 from $\mathcal{U}(0,1)$.
   (b) Calculate
   $$z = \cos(\pi u_1) \ , \quad f = \frac{1+rz}{r+z} \ , \quad c = \kappa(r-f) \ .$$
   (c) Accept sample if $c(2-c) > u_2$ or $\ln \frac{c}{u_2} + 1 \geq c$.
3. Draw $u_3 \sim \mathcal{U}(0,1)$
4. The accepted sample is given by
   $$\omega = [(\text{sign}(u_3 - 0.5) \arccos f + \mu + \pi) \mod 2\pi] - \pi \ .$$

6.2.5 Conditional Distribution for the Log-Damping Coefficients

For a diagonal state covariance matrix $\boldsymbol{Q}$, we are again able to decouple the posterior distribution $p(\boldsymbol{\gamma}|\boldsymbol{s}_{1:N}, \sigma_w^2, \boldsymbol{Q}, \boldsymbol{\omega}, y_{1:N})$ into L univariate distributions and factor the l'th distribution as

$$p(\gamma_l|\boldsymbol{s}_{1:N}, \boldsymbol{Q}, \sigma_w^2, \boldsymbol{\omega}, \boldsymbol{\gamma}_{\setminus l}, y_{1:N}) = p(\gamma_l|\boldsymbol{s}_{1:N,l}, q_l, \omega_l) \propto p(\boldsymbol{s}_{1:N,l}|q_l, \omega_l, \gamma_l)p(\gamma_l) \ . \tag{6.61}$$

From Eq. (6.50), we know that the parametric form of the l'th likelihood is

$$p(\boldsymbol{s}_{1:N,l}|q_l, \omega_l, \gamma_l) = \tilde{Z}_l^{-1} \exp\left\{-e^{-2\gamma_l}\alpha_{1,l} + e^{-\gamma_l}\beta_{1,l}\right\} \tag{6.62}$$

where we have defined

$$\beta_{1,l} \triangleq d_{1,l} \cos\omega_l + d_{2,l} \sin\omega_l \ . \tag{6.63}$$

To our knowledge, this likelihood results in a non-standard posterior distribution, which we cannot sample directly from, no matter how we choose a continuous and non-trivial prior distribution for γ_l. Therefore, we propose generating samples from the posterior distribution in Eq. (6.61) based on the Laplace approximation described in section 3.3.1. This is motivated by the fact that the plausible values of the log-damping coefficients for physical reasons are confined to a small interval. For example, if $\gamma_l = 0.1$ for the l'th sinusoid $e^{-\gamma_l n}\cos(\omega_l n)$ with unit amplitude and zero phase, then the envelope of the sinusoid has decreased by a factor of e at time index $n = 10$. For larger values of γ_l, the envelope decays even faster, so except for the case of a very small number of observations

N, the contribution of this sinusoid is negligible. Since the Laplace approximation is a second order approximation of $e^{-\gamma_l}$ around the mode of the posterior distribution, we can therefore expect the approximation to be reasonable for the plausible interval of the values of γ_l.

For the prior distribution $p(\gamma_l)$, we assume a parametric form similar to the likelihood, i.e.,

$$p(\gamma_l|\boldsymbol{\phi}_{\gamma_l}) \propto \exp\left\{-e^{-2\gamma_l}\alpha_{0,l} + e^{-\gamma_l}\beta_{0,l}\right\} \tag{6.64}$$

where the hyperparameters $\boldsymbol{\phi}_{\gamma_l} = \{\alpha_{0,l}, \beta_{0,l}\}$ are selected such that large values of γ_l have very small probability. The posterior distribution is now readily found to be given by

$$\begin{aligned} p(\gamma_l|\boldsymbol{s}_{1:N,l}, q_l, \omega_l) &\propto p(\boldsymbol{s}_{1:N,l}|q_l, \omega_l, \gamma_l)p(\gamma_l|\boldsymbol{\phi}_{\gamma_l}) && (6.65)\\ &\propto \exp\left\{-e^{-2\gamma_l}(\alpha_{1,l} + \alpha_{0,l}) + e^{-\gamma_l}(\beta_{1,l} + \beta_{0,l})\right\} && (6.66)\\ &= \exp\left\{-e^{-2\gamma_l}\alpha_l + e^{-\gamma_l}\beta_l\right\} \triangleq \tilde{p}(\gamma_l|\boldsymbol{s}_{1:N,l}, q_l, \omega_l) && (6.67) \end{aligned}$$

The Laplace Approximation

The first order derivative of the log-posterior distribution is

$$\frac{\partial}{\partial\gamma_l}\ln\tilde{p}(\gamma_l|\boldsymbol{s}_{1:N,l}, q_l, \omega_l) = 2\alpha_l e^{-2\gamma_l} - \beta_l e^{-\gamma_l} = e^{-\gamma_l}(2\alpha_l e^{-\gamma_l} - \beta_l)\ . \tag{6.68}$$

The maximum value of the posterior distribution thus occurs at

$$\hat{\gamma}_{l_{\text{MAP}}} = -\ln\frac{\beta_l}{2\alpha_l} \tag{6.69}$$

since the second order derivative of the log-posterior distribution is negative at this point as we show next. The second order derivative at $\hat{\gamma}_{l_{\text{MAP}}}$ is

$$\begin{aligned} \frac{\partial^2}{\partial\gamma_l^2}\ln\tilde{p}(\gamma_l|\boldsymbol{s}_{1:N,l}, q_l, \omega_l)\bigg|_{\gamma_l=\hat{\gamma}_{l_{\text{MAP}}}} &= -4\alpha_l e^{-2\gamma_l} + \beta_l e^{-\gamma_l}\bigg|_{\gamma_l=\hat{\gamma}_{l_{\text{MAP}}}} && (6.70)\\ &= -4\alpha_l\frac{\beta_l^2}{4\alpha_l^2} + \beta_l\frac{\beta_l}{2\alpha_l} = -\frac{\beta_l^2}{2\alpha_l} && (6.71) \end{aligned}$$

so the variance of the Laplace approximation is

$$\sigma_q^2 = \frac{2\alpha_l}{\beta_l^2}\ . \tag{6.72}$$

The Laplace approximation $p(\gamma_l|\boldsymbol{s}_{1:N,l}, q_l, \omega_l)$ is therefore given by

$$p(\gamma_l|\boldsymbol{s}_{1:N,l}, q_l, \omega_l) \approx q(\gamma_l|\boldsymbol{s}_{1:N,l}, q_l, \omega_l) = \mathcal{N}(\gamma_l; \hat{\gamma}_{l_{\text{MAP}}}, \sigma_q^2)\ . \tag{6.73}$$

Figure 6.1 shows how well the Laplace approximation performs for a few values of α_l and β_l. Notice, that the approximation is in general very good and becomes better and better the more selective $p(\gamma_l|\boldsymbol{s}_{1:N,l}, q_l, \omega_l)$ is.

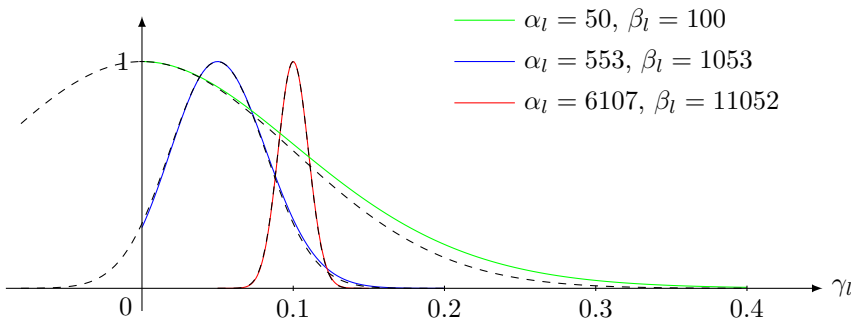

Figure 6.1: True probability distributions (solid colored lines) and their Laplace approximation (dashed lines) for some values of α_l and β_l. For illustrative purposes, the distributions are nomalised such that their maximum value equals 1. Notice, that in contrast to the true distributions, the support for the Laplace approximations also includes negative log-damping coefficients as indicated in the figure.

Simulation using the Laplace Approximation

One caveat of drawing samples from the Gaussian approximation $q(\gamma_l|\boldsymbol{s}_{1:N,l}, q_l, \omega_l)$ to the true posterior distribution $p(\gamma_l|\boldsymbol{s}_{1:N,l}, q_l, \omega_l)$ is that negative log-damping coefficients have non-zero probability of being generated. This could simply be ignored, or it could be remedied for by using a rejection sampling scheme which only accepts positive samples as in [Mazet et al., 2005]. The latter approach, however, entails a larger computational complexity.

The Gaussian approximation can also be used as a part of the proposal distribution of the Metropolis-Hastings algorithm. Since the Laplace approximation is a very good approximation to the true distribution, we expect a high acceptance ratio. Although this approach entails an increased computational complexity as compared against the approximate sampling schemes, it is a viable way of drawing samples from $p(\gamma_l|\boldsymbol{s}_{1:N,l}, q_l, \omega_l)$ without approximations. Algorithm 6.3 outlines this sampling scheme. Notice, that the mean of the Laplace approximation is being lower bounded by 0 in order to minimise the generated proportion of negative log-damping samples and thereby increase the acceptance ratio.

Algorithm 6.3 (MH-based Simulation of Log-damping Coefficients)
Given the previous sample $\gamma_l^{[\tau]}$ for the l'th log-damping coefficient, a new sample can be drawn from the conditional distribution $p(\gamma_l|\boldsymbol{s}_{1:N,l}, q_l, \omega_l)$ by

1. Compute α_l and β_l from

$$\alpha_l = \alpha_{1,l} + \alpha_{0,l}$$
$$\beta_l = \beta_{1,l} + \beta_{0,l}$$

where

$$\alpha_{1,l} = \frac{1}{2q_l} \sum_{n=1}^{N-1} \boldsymbol{s}_{n,l}^T \boldsymbol{s}_{n,l}$$
$$\beta_{1,l} = \frac{1}{q_l} \sum_{n=1}^{N-1} \begin{bmatrix} \boldsymbol{s}_{n+1,l}^T \boldsymbol{s}_{n,l} & \Big| \begin{bmatrix} \boldsymbol{s}_{n+1,l} & \boldsymbol{s}_{n,l} \end{bmatrix} \Big| \end{bmatrix} \begin{bmatrix} \cos\omega_l \\ \sin\omega_l \end{bmatrix}$$

and $\alpha_{0,l}$ and $\beta_{0,l}$ are the hyperparameters of the prior distribution $p(\gamma_l|\boldsymbol{\phi}_{\gamma_l})$ given by Eq. (6.64).

2. Calculate the mean and variance of the Gaussian proposal transistion kernel $Q(\gamma_l) = \mathcal{N}(\gamma_l; \mu_q, \sigma_q^2)$ as

$$\mu_q = \max(0, \hat{\gamma}_{l_{\mathrm{MAP}}})$$
$$\sigma_q^2 = \frac{2\alpha_l}{\beta_l^2}$$

where $\hat{\gamma}_{l_{\mathrm{MAP}}}$ is given by Eq. (6.69).

3. Draw a candidate sample γ_l^* from the proposal transition kernel $Q(\gamma_l)$.

4. Evaluate the probability of move given by

$$\alpha(\gamma_l^*, \gamma_l) = \min\left[1, \frac{\tilde{p}(\gamma_l^*|\boldsymbol{s}_{1:N,l}, q_l, \omega_l) Q(\gamma_l^{[\tau]})}{\tilde{p}(\gamma_l^{[\tau]}|\boldsymbol{s}_{1:N,l}, q_l, \omega_l) Q(\gamma_l^*)}\right]$$

where $\tilde{p}(\gamma_l|\boldsymbol{s}_{1:N,l}, q_l, \omega_l)$ is the unnormalised desired distribution given by Eq. (6.67).

5. Draw a random variate $u^{[\tau]}$ from the univariate uniform distribution $\mathcal{U}(0,1)$.

6. If $u^{[\tau]} \leq \alpha(\gamma_l^*, \gamma_l)$, then accept the sample γ_l^* as a sample from $p(\gamma_l|\boldsymbol{s}_{1:N,l}, q_l, \omega_l)$ and set $\gamma_l^{[\tau+1]} = \gamma_l^*$. Otherwise reject γ_l^* as a sample from $p(\gamma_l|\boldsymbol{s}_{1:N,l}, q_l, \omega_l)$ and set $\gamma_l^{[\tau+1]} = \gamma_l^{[\tau]}$.

6.3 Missing Observations

In for example audio restoration applications (see, e.g., [Godsill and Rayner, 1998] for an extensive treatment) some of the observations are corrupted or missing. This can be caused by a variety of phenomena such as scratches on a physical media, bad sectors in a storage media, noise bursts during recording, or package losses on a package based network. Audio restoration techniques aim at recovering the corrupted or missing samples by exploiting the correlation in the audio signal. For a set of observations $y_{1:N}$, partition it as

$$y_{1:N} = \{y_{1:K}, y_{K+1:K+R}, y_{R+1:N}\} \tag{6.74}$$

where the R observations $\boldsymbol{z} = y_{K+1:K+R}$ are corrupted or missing[3]. The problem is now to interpolate the samples in the gap $\boldsymbol{z}$ based on the valid observations $y_{1:K}$ and $y_{R+1:N}$. The interpolation is typically based on, in classical statistics, the maximum likelihood function $p(y_{1:K}, y_{R+1:N}|\boldsymbol{z})$ or, in Bayesian statistics, the marginal posterior distribution

$$p(\boldsymbol{z}|y_{1:K}, y_{R+1:N}) \ . \tag{6.75}$$

An intuitive way of recovering $\boldsymbol{z}$ is to use the ML estimate and MAP estimate for it by maximising the likelihood function and the posterior distribution, respectively. However, this has been shown to yield an atypical interpolant in the sense that it does not agree with the stochastic behaviour of the valid observations [**?**]. This is caused by the fact that ML/MAP minimises the variance of the stochastic part of the valid observations and thus acts as a kind of noise filter. In order to avoid this, the interpolant can be selected as a sample from the posterior distribution $p(\boldsymbol{z}|y_{1:K}, y_{R+1:N})$. We demonstrate the difference between these two ways of reconstruction the missing samples in the simulations in the next chapter.

The draw from $p(\boldsymbol{z}|y_{1:K}, y_{R+1:N})$ can easily be incorporated into our Gibbs sampling inference scheme. The conditional distribution for the missing observations $\boldsymbol{z}$ given all valid observation and all other latent and model parameters is

$$p(\boldsymbol{z}|\boldsymbol{s}_{1:N}, \boldsymbol{\theta}, y_{1:K}, y_{R+1:N}) = p(\boldsymbol{z}|\boldsymbol{s}_{K+1:K+R}, \sigma_w^2) \tag{6.76}$$

$$= \prod_{n=K+1}^{K+R} p(y_n|\boldsymbol{s}_n, \sigma_w^2) \ . \tag{6.77}$$

The R univariate distributions $p(y_n|\boldsymbol{s}_n, \sigma_w^2)$ are given by the observation equation in Eq. (6.1) and are thus Gaussian distributions which we can easily sample from. The implementation of the other steps of the Gibbs sampling inference scheme are unaffected by the inclusion of the interpolation stage and remain therefore the same.

[3] A significant problem in audio restoration is that of detecting and locating corrupt samples. In this thesis, however, we assume that we know which samples are corrupted/missing and which are not.

6.4 Summary of Inference Scheme

We have now derived the Gibbs sampling inference scheme for the latent states, model parameters and missing observations given the valid observation for the dynamic sinusoidal model. The inference scheme is summarised in algorithm 6.4. If some of the parameters are known a priori or all observations are valid, the sampling steps pertaining to these variables can be skipped in the algorithm.

Algorithm 6.4 (Gibbs Sampling Inference Scheme for Dynamic Model) For a diagonal or isotropic state noise covariance matrix, samples from the marginal posterior distributions given by Eq. (6.5) and Eq. (6.75) can be obtained by

1. Select hyperparameters of the prior distributions.
2. Initialise $\boldsymbol{z}^{[0]}$ and $\boldsymbol{\theta}^{[0]} = \{\sigma_w^{2\,[0]}, \boldsymbol{Q}^{[0]}, \boldsymbol{\omega}^{[0]}, \boldsymbol{\gamma}^{[0]}\}$ where $\boldsymbol{Q}^{[0]} = q^{[0]}\boldsymbol{I}_{2L}$ for the isotropic case and $\boldsymbol{Q}^{[0]} = \text{diag}(q_1^{[0]}, q_1^{[0]}, q_2^{[0]}, q_2^{[0]}, \cdots, q_L^{[0]}, q_L^{[0]})$ for the diagonal case.
3. Repeat for $\tau = 0, 1, 2, \cdots, T$
   (a) Draw a set of samples for the states
   $$\boldsymbol{s}_{1:N}^{[\tau+1]} \sim p(\boldsymbol{s}_{1:N}|\boldsymbol{\theta}^{[\tau]}, y_{1:K}, \boldsymbol{z}^{[\tau]}, y_{R+1:N})$$
   by using the simulation smoother in algorithm 6.1.
   (b) In the case of an isotropic state covariance matrix, draw a sample from
   $$q^{[\tau+1]} \sim \text{Inv-}\mathcal{G}\left(q; a_v + (N-1)L, b_v + \frac{1}{2}\sum_{n=1}^{N-1} \boldsymbol{v}_n^{[\tau]^T}\boldsymbol{v}_n^{[\tau]}\right)$$
   where $\boldsymbol{v}_n^{[\tau]} = \boldsymbol{s}_{n+1}^{[\tau]} - \boldsymbol{A}^{[\tau]}\boldsymbol{s}_n^{[\tau]}$, and $\boldsymbol{A}^{[\tau]}$ depends on $\boldsymbol{\omega}^{[\tau]}$ and $\boldsymbol{\gamma}^{[\tau]}$ through Eq. (5.13). In the case of a diagonal state covariance matrix, draw a sample from
   $$q_l^{[\tau+1]} \sim \text{Inv-}\mathcal{G}\left(q_l; a_{v,l} + (N-1), b_{v,l} + \frac{1}{2}\sum_{n=1}^{N-1} \boldsymbol{v}_{n,l}^{[\tau]^T}\boldsymbol{v}_{n,l}^{[\tau]}\right)$$
   for $l = 1, \cdots, L$ where $\boldsymbol{v}_{n,l}^{[\tau]} = \boldsymbol{s}_{n+1,l}^{[\tau]} - \boldsymbol{A}_l^{[\tau]}\boldsymbol{s}_{n,l}^{[\tau]}$, and $\boldsymbol{A}_l^{[\tau]}$ depends on $\omega_l^{[\tau]}$ and $\gamma_l^{[\tau]}$ through Eq. (5.14).

(c) Draw a sample for the observation variance

$$\sigma_w^{2\,[\tau+1]} \sim \text{Inv-}\mathcal{G}\left(\sigma_w^2; a_w + N/2, b_w + \frac{1}{2}\sum_{n=1}^{N} w_n^{[\tau]^T} w_n^{[\tau]}\right)$$

where $w_n^{[\tau]} = y_n - \boldsymbol{b}^T \boldsymbol{s}_n^{[\tau]}$. If y_n is a missing sample, the sample corresponding to it in $\boldsymbol{z}^{[\tau]}$ is used in its place.

(d) Draw a sample for the frequency parameters

$$\omega_l^{[\tau+1]} = \mathcal{VM}\left(\omega_l; \sqrt{\delta_{1,l}^{[\tau]^2} + \delta_{2,l}^{[\tau]^2}}, \arctan\frac{\delta_{2,l}^{[\tau]}}{\delta_{1,l}^{[\tau]}}\right)$$

for $l = 1, \cdots, L$ by using the Best-Fisher algorithm in algorithm 6.2. The values for $\delta_{1,l}^{[\tau]}$ and $\delta_{1,l}^{[\tau]}$ depend on $\boldsymbol{s}_{n+1,l}^{[\tau+1]}$, $\boldsymbol{s}_{n,l}^{[\tau+1]}$, $q_l^{[\tau+1]}$ and $\gamma_l^{[\tau]}$ through Eq. (6.57), Eq. (6.52) and Eq. (6.48).

(e) Draw a sample for the log-damping coefficients

$$\gamma_l^{[\tau+1]} \sim p(\gamma_l | \boldsymbol{s}_{n,l}^{[\tau+1]}, q_l^{[\tau+1]}, \omega_l^{[\tau+1]})$$

for $l = 1, \cdots, L$ by using the Laplace approximation given by Eq. (6.73) or the Metropolis-Hastings sampling scheme in algorithm 6.3.

(f) Draw an interpolant for the missing samples

$$\boldsymbol{z}^{[\tau+1]} \sim p(\boldsymbol{z} | \boldsymbol{s}_{K+1:K+R}^{[\tau+1]}, \sigma_w^{2\,[\tau+1]})$$

by drawing a sample from

$$p(y_n | \boldsymbol{s}_n^{[\tau+1]}, \sigma_w^{2\,[\tau+1]}) = \mathcal{N}(y_n; \boldsymbol{b}^T \boldsymbol{s}_n^{[\tau+1]}, \sigma_w^{2\,[\tau+1]})$$

for $n = K+1, \cdots, K+R$.

For a sufficiently large sample size, this algorithm can be used for determining the true joint as well as marginal posterior for the latent states, unknown variables and missing observations of the dynamical sinusoidal model. Based on these distributions, optimal point estimates as well as probability intervals can be computed. We do not focus on this post-processing step in this thesis; however, in the simulations in the next chapter, we give some examples of summarising the inference in terms of the MMSE estimate. The major disadvantage of the algorithm is, like most other Bayesian inference

schemes, the computational complexity which is very high.

Chapter 7

Simulation Study on Synthetic and Real Signals

After having derived the inference scheme for the latent variables and model parameters of the dynamic sinusoidal model in the previous chapter, this chapter demonstrates the application of the proposed method to analysis on synthetically generated signals as well as on a real audio signal. This is achieved through four simulations:

1. Inference for the unknown parameters in the static model constituted by a single damped sinusoid.
2. Inference in a reduced and simplified dynamic sinusoidal model where some parameters are assumed known.
3. Inference in a full dynamic sinusoidal model in which nothing but the model order is assumed known.
4. Inference for a frame of observations from a real piano audio signal.

First, however, we validate the implementation of the inference scheme by comparing the results obtained in individual sampling steps in the Gibbs sampler against the theoretical results.

7.1 Validation of the Individual Sampling Steps

We validated our implementation of the inference scheme of algorithm 6.4 by treating the sampling steps individually. This was necessary since we are unable to derive analytical expressions for the complete sampling scheme. The sampling steps for the states, the

frequency parameters and log-damping coefficients are not easily implemented whereas the sampling steps for the noise variances and missing samples are straightforward since they only involve sampling from inverse Gamma and Gaussian distributions, respectively. In this section, we therefore only focus on the validation of the sampling steps for the states, the frequency parameters and log-damping coefficients.

7.1.1 Simulation Smoothing for the States

We validated the implementation of the simulation smoother by using Monte Carlo simulations and compared the results against analytical results. From Eq. (6.18), we know that the posterior distribution for the states given the model parameters and observations is a multivariate Gaussian distribution whose mean and covariance are given by Eq. (6.19) and can be calculated analytically. For $N = 50$ observations from a dynamic sinusoidal model with a single sinusoid, i.e., $L = 1$, with the parameters $\boldsymbol{Q} = 0.01\boldsymbol{I}_2$, $\sigma_w^2 = 0.1$, $\omega = 0.5$, $\gamma = 0$, $\boldsymbol{\mu} = \sqrt{2}\begin{bmatrix}1 & 1\end{bmatrix}^T$ and $\boldsymbol{P} = 0.1\boldsymbol{I}_2$, we generated $T = 10,000$ set of state vectors $\boldsymbol{s}_{1:N}^{[\tau]}$ for $\tau = 1, \cdots, T$ by running the simulation smoother outlined in algorithm 6.1. From these T sets of state vectors, a sample mean and covariance were calculated by

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{s}|\boldsymbol{\theta},\boldsymbol{y}} = \frac{1}{T}\sum_{\tau=1}^{T} \boldsymbol{s}^{[\tau]} \tag{7.1a}$$

$$\hat{\boldsymbol{C}}_{\boldsymbol{s}|\boldsymbol{\theta},\boldsymbol{y}} = \frac{1}{T}\sum_{\tau=1}^{T} \boldsymbol{s}^{[\tau]}\boldsymbol{s}^{[\tau]T} \tag{7.1b}$$

where $\boldsymbol{s} = \text{vec}(\boldsymbol{s}_{1:N})$ and $\boldsymbol{y} = \text{vec}(y_{1:N})$. Since the dynamic model only consists of a single sinusoid, each state vector $\boldsymbol{s}_n$ is vector with two elements. We denote these two elements as $\boldsymbol{s}_{n,1}$ and $\boldsymbol{s}_{n,2}$, respectively.

Figure 7.1 and figure 7.2 show the results of the simulation for $\boldsymbol{s}_{n,1}$ and $\boldsymbol{s}_{n,2}$, respectively. In each of the two figures, two plots are shown; the left one comparing the true, analytical and estimated smoothed means, and the right one comparing the diagonal elements of the analytical and estimated covariance matrices. It is seen that the analytical and estimated smoothed means appear to completely coincide[1]. For the diagonal elements in the analytical and estimated covariance matrices, a small random deviation between the two curves are observed. However, they clearly follow the same trend.

Although we have not compared the off-diagonal elements of the analytical and estimated covariance matrices, we believe that the results in figure 7.1 and figure 7.2 have provided sufficient evidence for that the implementation of the simulation smoother produce samples from the posterior distribution for the states given by Eq. (6.18).

[1]Notice, that the thickness of the estimated mean has been increased in order to make it visible. It would have been completely hidden behind the true mean if this was not done.

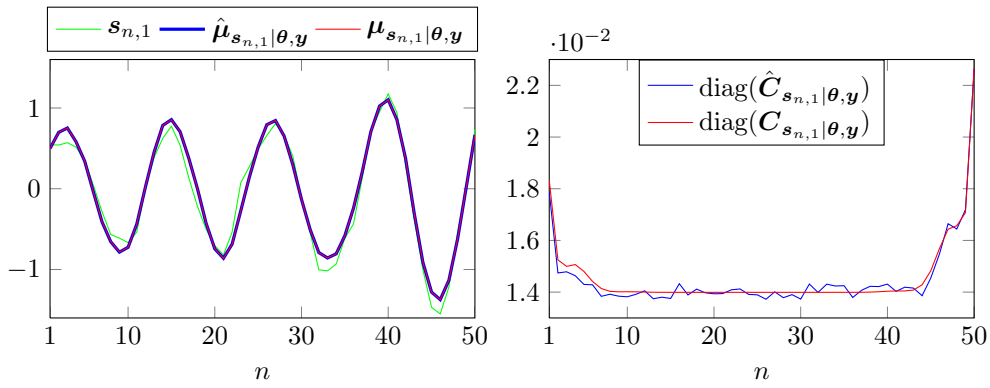

Figure 7.1: Comparison of the true, analytical and numerically estimated mean (left plot) for the first element of the 2-dimensional state vector. The right plot shows the diagonal elements of the analytical and numerically estimated covariance matrices.

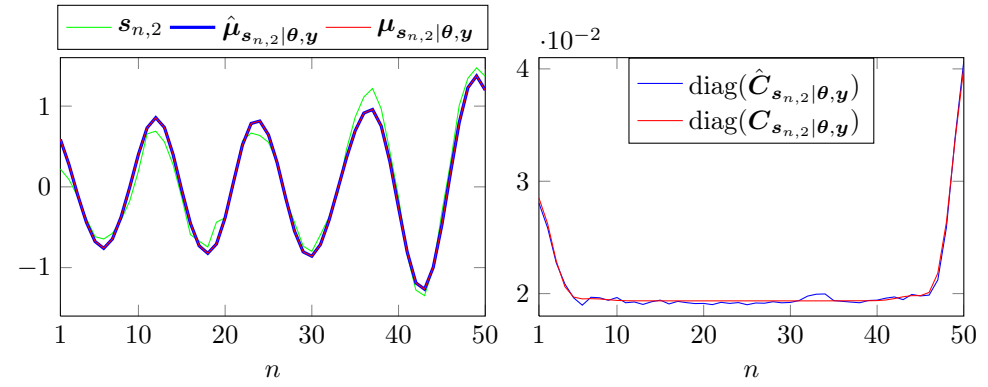

Figure 7.2: Comparison of the true, analytical and numerically estimated mean (left plot) for the second element of the 2-dimensional state vector. The right plot shows the diagonal elements of the analytical and numerically estimated covariance matrices.

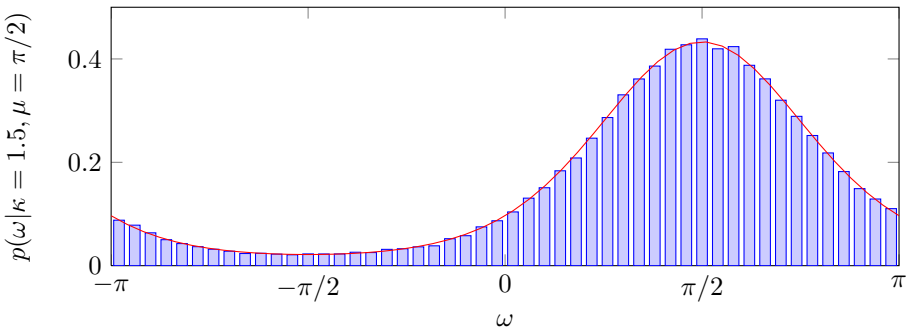

Figure 7.3: True von Mises distribution (red curve) for parameters $\kappa = 1.5$ and $\mu = \pi/2$ and 50-bin normalised histogram for the 100,000 samples generated by the Best-Fisher algorithm.

7.1.2 Simulating from the von Mises Distribution

Samples for the frequency parameters are obtained by simulating random variates from the von Mises distribution as we showed in section 6.2.4. The simulation can be efficiently implemented by using the Best-Fisher algorithm outlined in algorithm 6.2 which is a rejection sampling scheme guaranteeing a minimum expected acceptance ratio of $\sqrt{e/2\pi} \approx 0.66$. For the verification of the implementation of the Best-Fisher algorithm, we drew $T = 100,000$ samples by using this algorithm and compared the histogram of the obtained samples against the von Mises probability density function. Figure 7.3 shows the results of the simulation with the parameters of the distribution selected as $\kappa = 1.5$ and $\mu = \pi/2$. The histogram consists of 50 bins and the magnitude of each bin has been normalised such that the total area covered by the histogram equals one. The acceptance ratio of the simulation was $\eta_A = 0.757$. From the figure, we see that the magnitude of the bins of the histogram approximately equals the true probability. We therefore believe that our implementation of the Best-Fisher algorithm generates random variates from the von Mises distribution.

7.1.3 MH-based Sampling of the Log-Damping Coefficients

In section 6.2.5, we proposed an approximate and an exact way for generating samples from the conditional distribution for the log-damping coefficients. In this section, we evaluate both of these proposed sampling schemes by comparing analytical results against histograms for the obtained samples. We also demonstrate that using a MH-sampling step within the Gibbs sampler leads to the desired results. In the simulations, we generated $N = 50$ observations from a dynamic sinusoidal model with a single sinusoid with the parameters $\sigma_w^2 = 0.1$, $\omega = 0.5$, $\boldsymbol{\mu} = \sqrt{2}\begin{bmatrix}1 & 1\end{bmatrix}^T$ and $\boldsymbol{P} = 0.1\boldsymbol{I}_2$ and known states $\boldsymbol{s}_{1:N}$. The state noise variance was assumed to be isotropic and unknown,

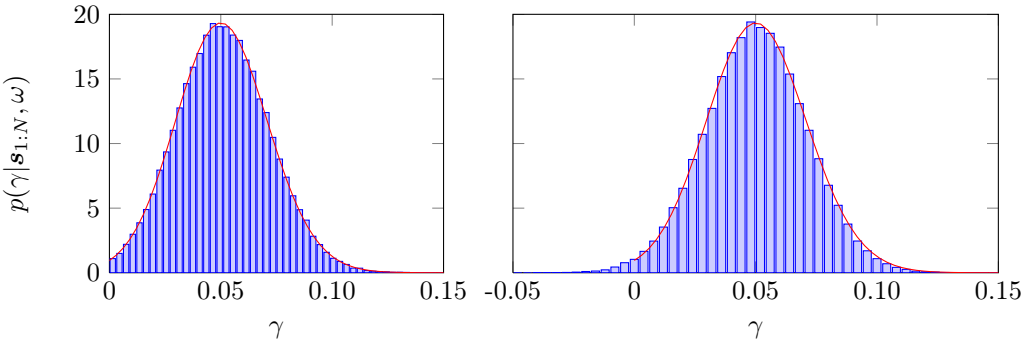

Figure 7.4: True marginal posterior distribution (red curve) for the log damping coefficient marginalised w.r.t. the isotropic state noise variance. The plots also show the 50-bin normalised histograms obtained by simulating 100,000 random variates from the true distribution by using the MH-sampling scheme (left plot) and the Laplace approximation (right plot).

i.e., $\boldsymbol{Q} = q\boldsymbol{I}_2$. Thus, we have two unknown parameters γ and q with the latter being a nuisance parameter. The true marginal posterior distribution for the log-damping coefficient can be found as

$$p(\gamma|\boldsymbol{s}_{1:N},\sigma_w^2,\omega,y_{1:N}) = p(\gamma|\boldsymbol{s}_{1:N},\omega) = \int p(\gamma,q|\boldsymbol{s}_{1:N},\omega)dq \tag{7.2}$$

$$\propto \int p(\gamma,q,\boldsymbol{s}_{1:N}|\omega)dq = \int p(\boldsymbol{s}_{1:N}|q,\omega,\gamma)p(\gamma,q|\omega)dq \tag{7.3}$$

$$= p(\gamma)\int p(\boldsymbol{s}_{1:N}|q,\omega,\gamma)p(q)dq \tag{7.4}$$

$$\propto p(\gamma)\int \left[\prod_{n=1}^{N-1} p(\boldsymbol{s}_{n+1}|\boldsymbol{s}_n,q,\omega,\gamma)\right] p(q)dq\ . \tag{7.5}$$

The distribution $p(\boldsymbol{s}_{n+1}|\boldsymbol{s}_n,q,\omega,\gamma)$ is the distribution governing the state equation given by Eq. (6.2) and is thus a Gaussian distribution. Since the prior distribution $p(q)$ is an inverse Gamma distribution given by Eq. (6.34), the true marginal posterior distribution for the log-damping coefficient can be found to be proportional to

$$p(\gamma|\boldsymbol{s}_{1:N},\omega) \propto p(\gamma)\left[b_v + \frac{1}{2}\sum_{n=1}^{N-1}(\boldsymbol{s}_{n+1}-\boldsymbol{A}\boldsymbol{s}_n)^T(\boldsymbol{s}_{n+1}-\boldsymbol{A}\boldsymbol{s}_n)\right]^{-(a_v+N-1)} \tag{7.6}$$

by using result B.3.

Since γ and q are the only unknown parameters, we can skip all sampling steps of the Gibbs sampler in algorithm 6.4 but step b) and step e). We have implemented step

e) using the exact way as given by the MH-based sampler in algorithm 6.3 as well as for using the Laplace approximation given by Eq. (6.73). For the prior distribution $p(\gamma)$ for the log-damping coefficients, we have used the improper uniform distribution. Figure 7.4 shows the results of the two simulations compared against the true distribution given by Eq. (7.6). The left plot shows the 50-bin normalised histogram obtained by drawing 100,000 samples using the MH-based sampler whereas the right plot shows 50-bin normalised histogram obtained by drawing 100,000 samples from the Laplace approximated marginal distribution. From the plots, we see that we are able to draw samples from $p(\gamma|\boldsymbol{s}_{1:N}, \omega)$ by using the MH-based sampling scheme. We also see that the Laplace approximation is a very good approximation although it generates negative log-damping coefficients.

7.2 Case 1: Inference for a Single Static Sinusoid

In the first simulation, we considered how our inference scheme performed on a single sinusoid generated by the static model in Eq. (1.1) in the case of missing observations. As compared against the simulations performed in chapter 4, the model in our simulation also included a log-damping coefficient. The true but unknown model parameters in the simulation were $\alpha = 1$, $\varphi = 0$, $\omega = 0.2$, $\gamma = 3 \cdot 10^{-3}$ and $\sigma_w^2 = 0.01$ for the amplitude, phase, frequency, log-damping coefficient and observation noise variance, respectively. The initial state statistics were assumed known and equal to $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{P} = 10\boldsymbol{I}_2$. For the prior distributions $p(q|a_v, b_v)$ and $p(\sigma_w^2|a_w, b_w)$ on the state noise variance and observation noise variance, respectively, we used non-informative priors to ensure that they played a minimal role by selecting $a_v = b_v = a_w = b_w = 10^{-5}$. For the prior $p(\omega|\kappa_0, \mu_0)$ on the frequency parameter, we used a prior favouring positive frequencies over negative frequencies by selecting $\kappa_0 = 5$ and $\mu_0 = \pi/2$. For the log-damping frequency, the prior $p(\gamma|\alpha_0, \beta_0)$ was selected such that values greater than 0.1 had very low probability by choosing $\alpha_{0,l} = 500$ and $\beta_{0,l} = 1000$. The initial values for $\boldsymbol{\omega}^{[0]}$ were found by using the subspace-based ESPRIT-estimator [Stoica and Moses, 2005, pp. 174-175], and the initial value for $\sigma_w^{2\,[0]}$ was found as the average noise-subspace eigenvalue computed as a bi-product by the ESPRIT estimator. In a rather heuristic way, the initial value for the state noise variance was selected as $q^{[0]} = \sigma_w^{2\,[0]}/10$ and the missing samples were all set to equal zero, i.e., $\boldsymbol{z}^{[\tau]} = \mathbf{0}$.

The samples from index 150 to index 250 of $N = 512$ observations were missing as shown in the top plot of figure 7.5. The remaining of the plots in the figure show the traces of generated samples as well as the histograms computed from these traces. Each trace consists of $T = 10,000$ samples generated by running algorithm 6.4. After a short burn-in time in which the underlying Markov chain converges to the correct posterior distribution, the traces are stationary and are constituted by samples from the marginal distributions defined in Eq. (6.5). The histograms in the margin are based on the

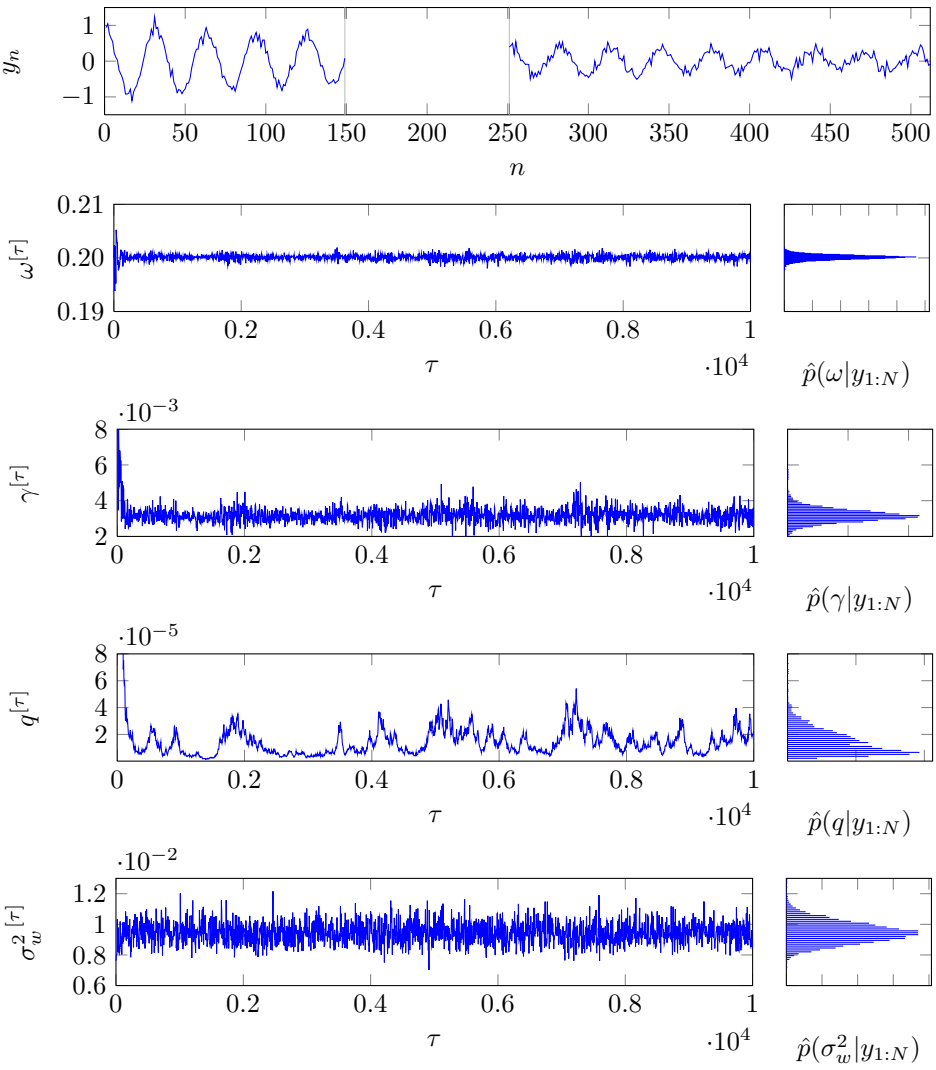

Figure 7.5: The observations with missing samples (top plot) and the traces of the 10,000 samples from the Gibbs sampler for the frequency, log-damping coefficient, state noise variance and observation noise variance. The plots in the margin show the histograms for the traces with the first 100 samples removed as burn-in samples.

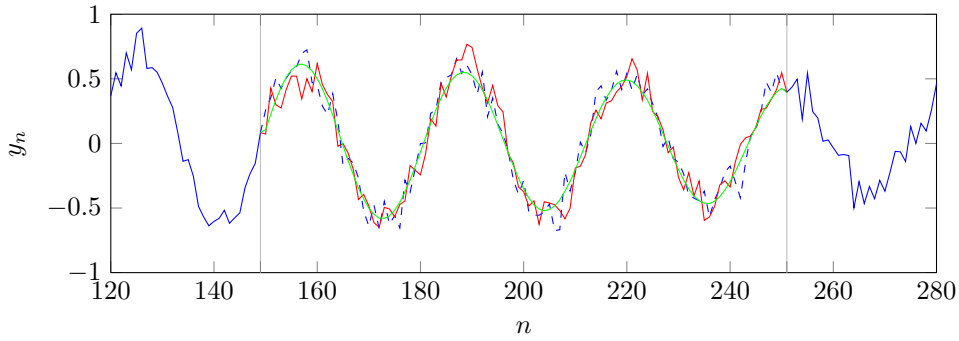

Figure 7.6: Missing samples (dashed line) and the reconstructed samples (red line) in the interpolation section marked by the vertical lines. The green line indicates the posterior mean value of the missing samples.

generated samples with the exception of the first 100 samples which were discarded as burn-in samples. In the limit of an infinite number of samples, these histograms converge to the true marginal distributions in Eq. (6.5). The histograms based on the $9,900$ samples, however, are a good approximation to these true distributions from which, e.g., means and confidence intervals can be derived. For example, the means of the histograms are in the limit of infinitely many samples equal to the MMSE estimate which we discussed in section 2.3. Computing those means yields the estimates $\hat{\mu}_{\omega|y_{1:N}} = 0.2001$, $\hat{\mu}_{\gamma|y_{1:N}} = 3.2 \cdot 10^{-3}$, $\hat{\mu}_{q|y_{1:N}} = 1.41 \cdot 10^{-5}$ and $\hat{\mu}_{\sigma_w^2|y_{1:N}} = 9.5 \cdot 10^{-3}$ which are very close the true values.

Figure 7.6 shows the result of the interpolation. The dashed line indicates the true missing samples and the red line indicates the reconstructed samples. As we discussed in section 6.3, the reconstructed samples are a sample from the posterior distribution for the missing samples. This ensures that the noise is also modelled. The green line indicates a recontruction based on the mean of the joint posterior distribution for the missing samples. Clearly, the curve is smooth and the noise is thus not modelled when using this approach for the reconstruction of the missing samples. This yields an interpolant which is atypical for the underlying signal [**?**].

7.3 Case 2: Inference in a Simplified Dynamic Model

In the second simulation, we increased the complexity of the simulations by considering the dynamic model with several unknown parameters. In this simulation, however, we did not consider the full model but a simplified version of it in which we assumed isotropic state noise covariance and known log-damping coefficients equal to zero. We did also assume that we did not have any missing or corrupted observations. The

states as well as the rest of the parameters were assumed unknown. Although we refer to these model assumptions as a simplified model, the complexity of the model is so high that it prohibits analytical evaluation. In the simulation, we observed $N = 512$ samples from the dynamic signal model consisting of two sinusoids, i.e., $L = 2$, with the unknown model parameters $\boldsymbol{Q} = q\boldsymbol{I}_4$ with $q = 0.01$, $\sigma_w^2 = 0.5$, $\boldsymbol{\omega} = \begin{bmatrix}\omega_1 & \omega_2\end{bmatrix}^T = \begin{bmatrix}0.3 & 0.35\end{bmatrix}^T$. The amplitude and the phase of the sinusoids were unknown and given by $\boldsymbol{\alpha} = \begin{bmatrix}\alpha_1 & \alpha_2\end{bmatrix}^T = \begin{bmatrix}0.8 & 1\end{bmatrix}^T$ and $\boldsymbol{\varphi} = \begin{bmatrix}\varphi_1 & \varphi_2\end{bmatrix}^T = \begin{bmatrix}0 & \pi/2\end{bmatrix}^T$. We selected the same hyperparameters as in the simulations for a single static sinusoid.

Inference in the dynamic model based on these assumption were performed by running $T = 10,000$ iterations of the Gibbs sampler in algorithm 6.4. Since the log-damping coefficients were known and no samples were missing, step e) and f) of the algorithm were skipped. Figure 7.7 shows the results of the simulations. The top plot shows the $N = 512$ observation whose amplitude seems to follow a non-stationary envelope. One of the strength of the dynamic model as compared with the static model is that this does not necessarily violate the model assumptions as we previously pointed out in section 5.2. The remaining three plots in the figure show the traces of samples generated by the inference scheme.

It should be mentioned that the computational complexity of algorithm 6.4 is rather high and renders the algorithm unsuitable for real-time applications. For this simulation, the generation of the traces took approximately 10 minutes on a powerful 2.66 GHz Core i7 desktop computer. This should be compared against the $N = 512$ observations which corresponds to approximately 12 ms of audio sampled at a standard rate of 44,100 kHz or 64 ms of speech sampled at a standard rate of 8000 kHz. Although the implementation of the sampling scheme probably can be heavily optimised for increased speed and the number of generated samples can be lowered significantly, the algorithm would still only be feasible for off-line applications.

7.4 Case 3: Inference in a Full Dynamic Model

In the third simulation, we increased the complexity of the simulation in the previous section by introducing a third sinusoid with amplitude, phase and frequency and log-damping coefficient given by $\alpha_3 = 1.2$, $\varphi_3 = -0.1$, $\omega_3 = 0.6$ and $\gamma_3 = 0.02$, respectively. We also assumed the other log-damping coefficients as unknown with true values given by $\gamma_1 = \gamma_2 = 0$, and we no longer assumed the state noise to be isotropic but diagonal with $\boldsymbol{q} = \begin{bmatrix}q_1 & q_2 & q_3\end{bmatrix}^T = \begin{bmatrix}0.01 & 0.01 & 0.025\end{bmatrix}^T$. The prior distributions were selected in the same way as in the simulation for the single static sinusoid. The rest of the parameters were the same as in the previous simulation for the simplified dynamic model. We again observed $N = 512$ samples but this time with the samples from index 200 to index 250 missing as shown in figure 7.8.

As in the previous simulation, we performed the simulations by running $T = 10,000$

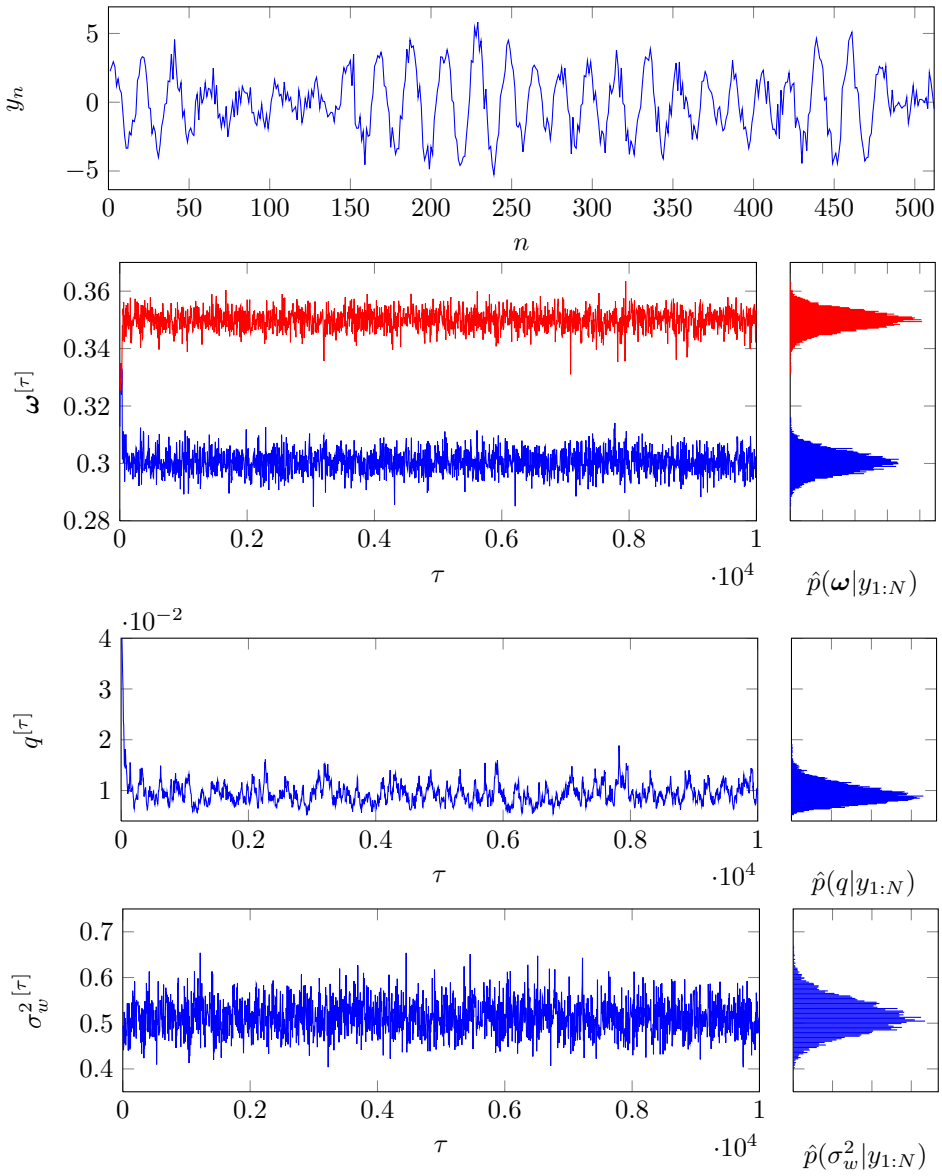

Figure 7.7: Observed sequence (top plot) and the traces of the 10,000 samples from the Gibbs sampler for the frequencies, state noise variance and observation noise variance. The margin shows the histograms for the traces with the first 100 samples removed as burn-in samples.

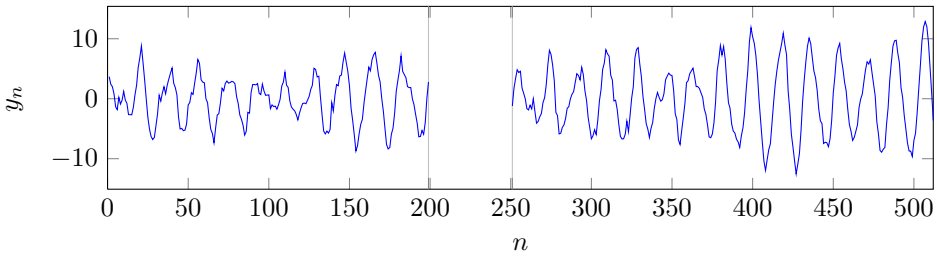

Figure 7.8: The $N = 512$ observations with missing samples from index 200 to index 250.

iterations of the Gibbs sampler in algorithm 6.4. However, step e) and f) were not skipped in this simulation since the log-damping coefficients were assumed unknown and observations were missing. Figure 7.9 shows the traces of generated samples and the histograms calculated from these samples. From the figure, we see that the sampling scheme quickly converged. It should be pointed out that this is not true in general for the algorithm whose convergence time critically depends on the ESPRIT-based initialisation.

As an example of performing inference based on the generated samples, we have computed the means of the marginal posterior distribution for the parameters shown in figure 7.9. As we discussed in section 2.3, these means are the MMSE estimators of the unknown parameters. For the frequencies, log-damping coefficients and state noise variances, the MMSE estimates were calculated as

$$\begin{bmatrix} \hat{\mu}_{\omega_1|y_{1:N}} & \hat{\mu}_{\gamma_1|y_{1:N}} & \hat{\mu}_{q_1|y_{1:N}} \\ \hat{\mu}_{\omega_2|y_{1:N}} & \hat{\mu}_{\gamma_2|y_{1:N}} & \hat{\mu}_{q_2|y_{1:N}} \\ \hat{\mu}_{\omega_3|y_{1:N}} & \hat{\mu}_{\gamma_3|y_{1:N}} & \hat{\mu}_{q_3|y_{1:N}} \end{bmatrix} = \begin{bmatrix} 0.2960 & 0.0048 & 0.0081 \\ 0.3496 & 0.0006 & 0.0195 \\ 0.6038 & 0.0154 & 0.0243 \end{bmatrix} \tag{7.7}$$

and the MMSE estimate of the observation noise variance was $\hat{\mu}_{\sigma_w^2|y_{1:N}} = 0.4082$. Comparing these values against the true values, we see that the frequency estimates are very close to the true values. For the other parameters, some of the estimates are close whereas others deviates more from the true value. The posterior distributions, from which the deviating estimators are computed, however, are rather broad. This indicates that a significant uncertainty is associated with these point estimates.

Figure 7.10 shows the results of the interpolation. In the figure, the interpolation section is marked by vertical lines and it contains the true missing samples indicated by the dashed curve along with the reconstruction indicated by the solid red curve.

7.5 Case 4: Inference for a Real Audio Signal

In the fourth and final simulation, we have considered restoration of a real audio signal with missing samples. The audio signal was a frame of length $N = 1024$ from a very

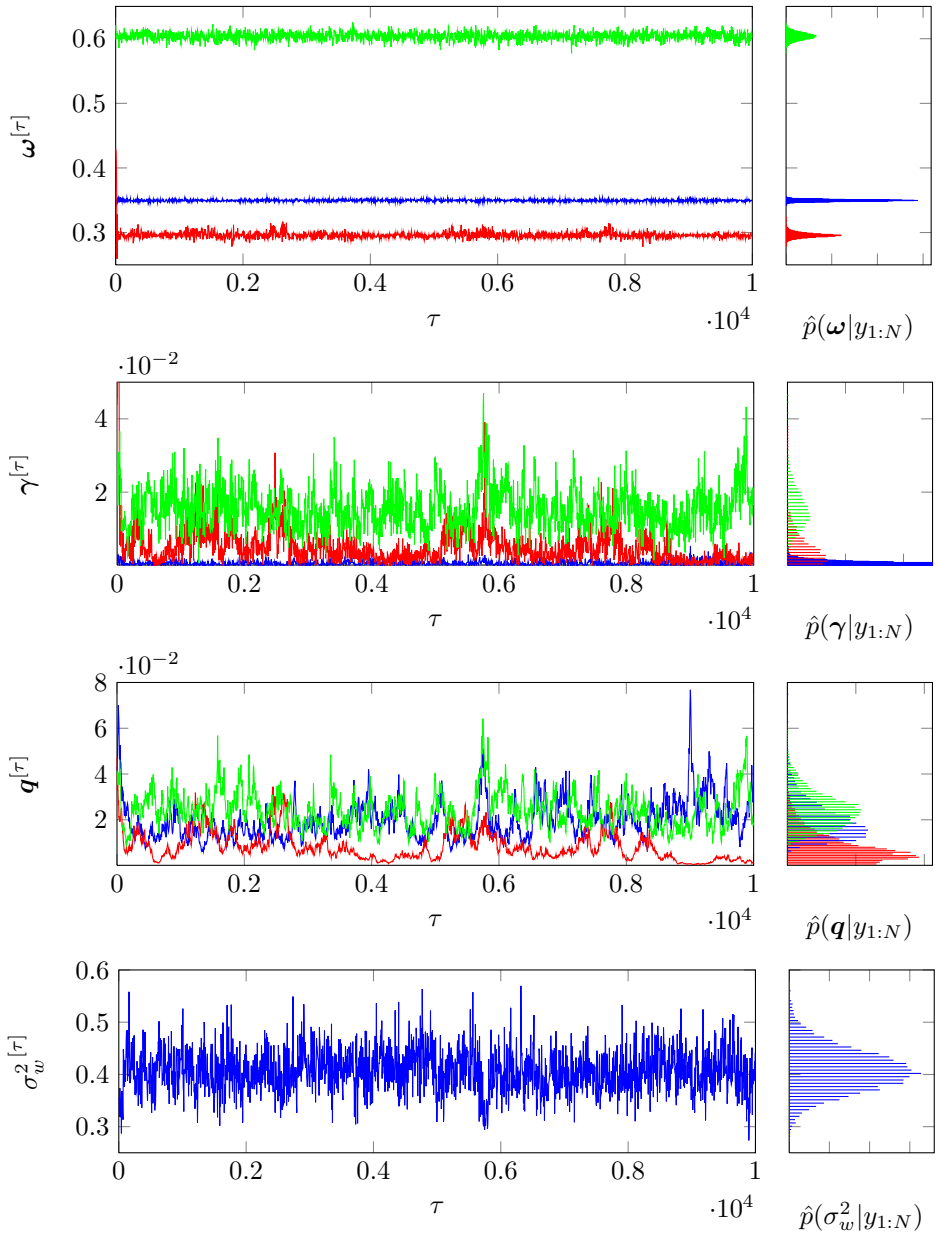

Figure 7.9: The traces of the 10,000 samples from the Gibbs sampler for the frequencies, log-damping coefficients, state noise variances and observation noise variance. The margin shows the histograms for the traces with the first 100 samples removed as burn-in samples.

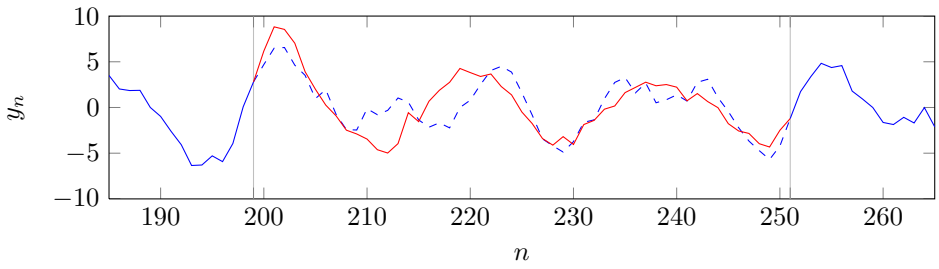

Figure 7.10: Missing samples (dashed line) and the reconstructed samples (red lines) in the interpolation section marked by the vertical lines.

simple piano signal downsampled by a factor of four to a sample rate of 11,025. In the top plot of figure 7.11, the time series of the piano signal is shown with a centred section of 220 missing observations. For a sample rate of of 11,025, the gap corresponds to 20 ms and it causes an audible click. The bottom plot of the figure shows the periodogram of the 1024 observations and it reveals that the piano signal is a harmonic signal with an (angular) fundamental frequency of 0.15. It should be noticed that the periodogram was computed from all 1024 samples with the true missing samples inserted in the gap. Although this is not possible in a real world application, we did this in the simulation in order to validate the results of our inference scheme.

In the simulations, we selected the model order as $L = 4$ and used a diagonal state noise covariance matrix $\boldsymbol{Q}$. For the prior distributions for the model parameters, we used the same values as in the previous simulations with one noticeable exception. The state as well as observation variances of the piano signal were very small which resulted in a very slow convergence time. In order to avoid this, we increased the hyperparameter b_w of the prior distribution for the observation noise variance σ_w^2 from 10^{-5} to 10^{-3}. By doing so, we prevented the observation noise variance from collapsing to zero [Lindley and Smith, 1972; Rajan et al., 1997] and increased the convergence speed significantly.

Figure 7.12 shows the traces of samples obtained by running algorithm 6.4 for $T = 10,000$ iterations. The traces of samples for the fundamental frequency and the first two harmonics quickly converged to the true vales and the corresponding log-damping coefficients and state noise variances decreased towards zero. The latter behaviour implies that the audio signal is stationary which, based on the time series plot in figure 7.11, seems to be reasonable. The samples on the fourth trace, however, does not appear to originate from neither the third nor the fourth harmonic but from a mixture of them and the variance of the samples is very high. However, since the corresponding log-damping coefficient was very high, the fourth frequency had negligible influence on the interpolation which is shown in figure 7.13. From the figure, it is seen that the missing samples were reconstructed in a satisfactory way close to the true values of the missing samples. Inserting these interpolated samples into the audio signal thus removes the

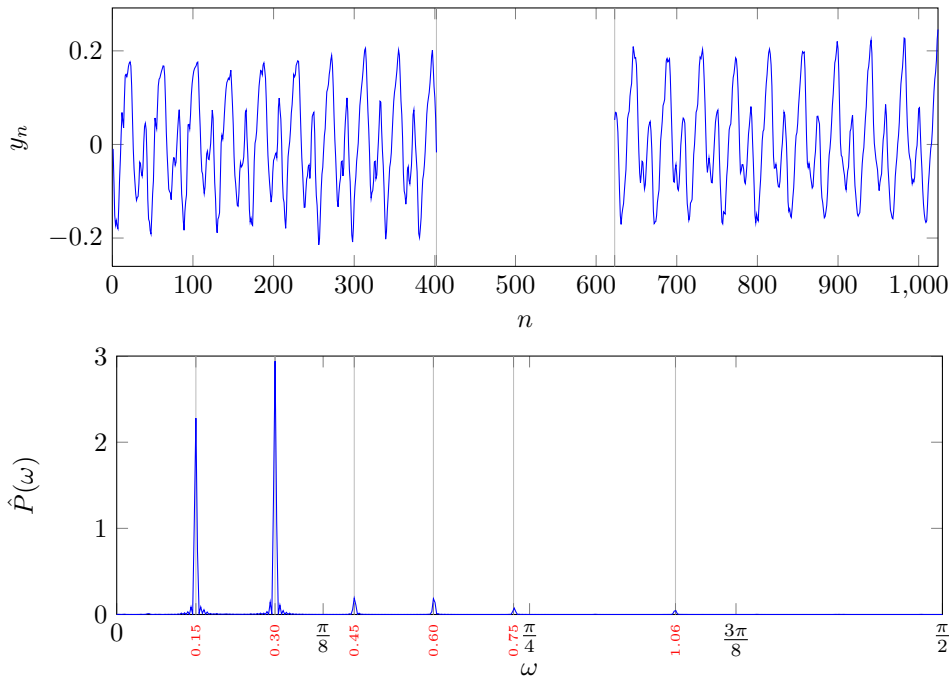

Figure 7.11: The $N = 1024$ observations (top plot) from a piano signal with a centred gap of 220 missing samples and the periodogram (bottom plot) of the observations with no missing samples.

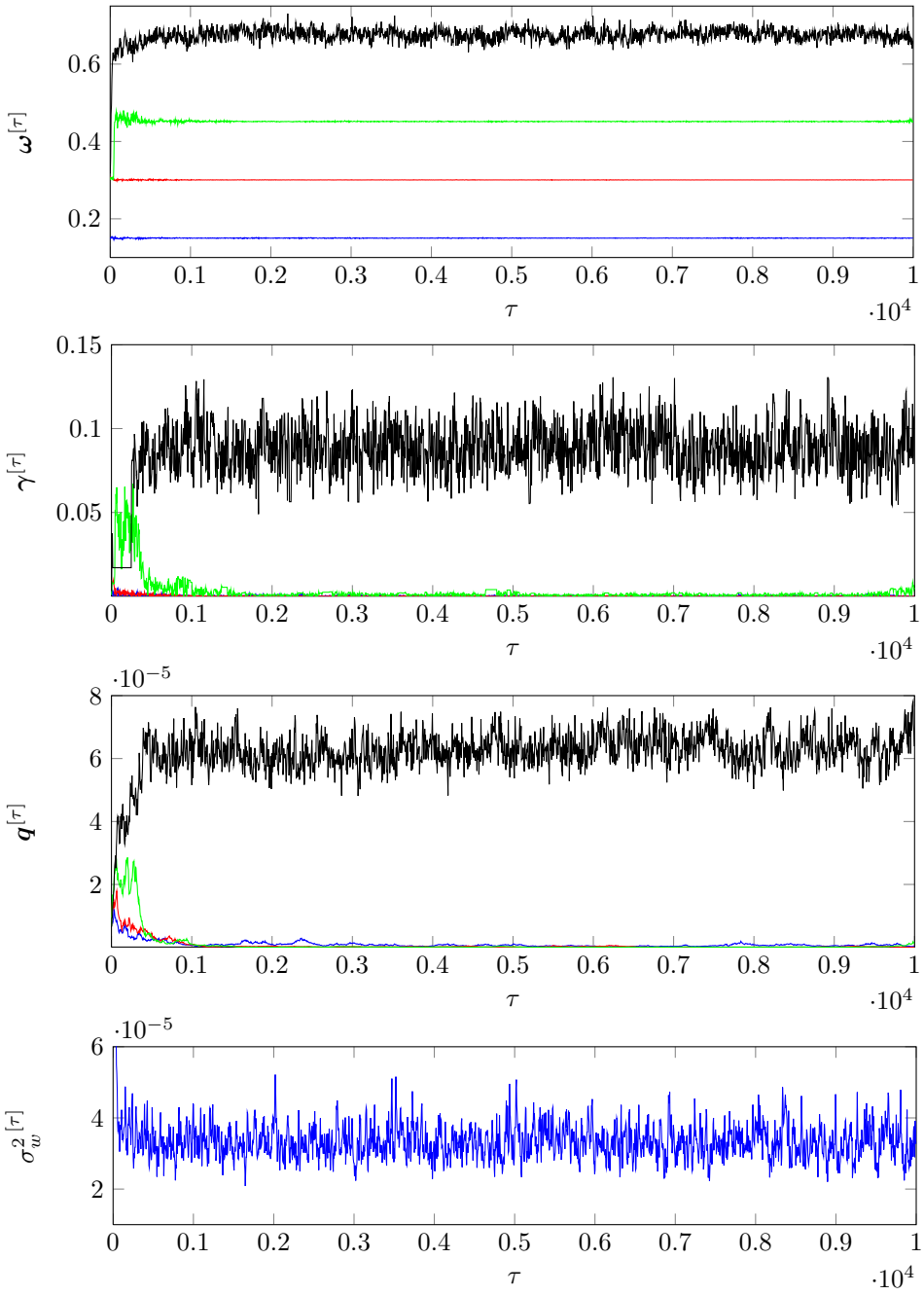

Figure 7.12: The traces of the 10,000 samples from the Gibbs sampler for the frequencies, log-damping coefficients, state noise variances and observation noise variance.

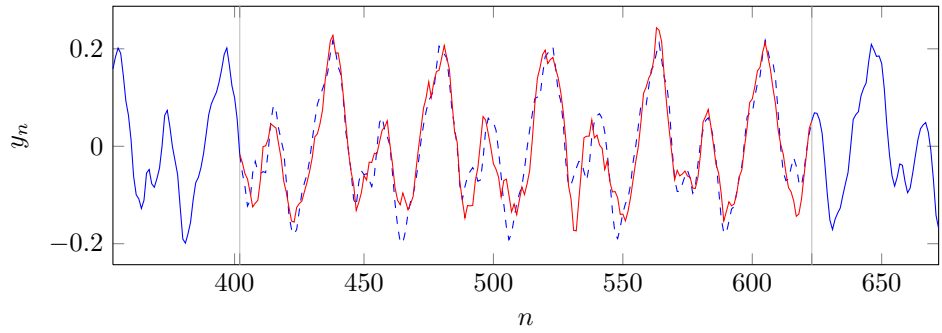

Figure 7.13: The true missing audio samples (dashed line) and the reconstructed samples (red lines) in the interpolation section marked by the vertical lines.

audible click.

7.6 Summary

In this chapter, we have demonstrated the applicability of the inference scheme derived in chapter 6. This was accomplished through several simulations on various synthetic signals as well as on a real audio signal. The inference scheme was demonstrated to be able to draw samples from the marginal posterior distributions of the unknown model parameters, and it was also shown to be able to handle signals with missing observations which were successfully reconstructed through interpolation. The computational complexity of the inference scheme is very high and this renders the algorithm unusable for real-time applications. This is a general problem for numerical Bayesian inference schemes.

Chapter 8

Conclusion

In this thesis, we have taken the Bayesian approach to statistical inference and applied it to the problem of performing inference for the parameters of the sinusoidal model. The Bayesian approach offers some attractive advantages over the classical approach to statistical inference as we outlined in the introduction and demonstrated using several small-scale examples throughout part I of this thesis. One of the major advantages is that the Bayesian approach always offers the complete and optimal solution in terms of the posterior distribution on which all probabilistic statements about the inference problem are based. However, this attractive advantage of Bayesian statistics is often dwarfed by the difficulties associated with deriving analytical solutions, which may not exist, or evaluating high-dimensional integrals using numerical techniques. Although computational algorithms such as the Metropolis-Hastings algorithm and the Gibbs sampler to some extent have remedied for this, the computational complexity associated with performing Bayesian inference may still render it infeasible for many applications.

Initially we considered the static sinusoidal model since it is the most popular sinusoidal model. In that connection, we reviewed two of the Bayesian inference schemes for the frequency parameter of the static sinusoidal model. In part II, however, we introduced the dynamic sinusoidal model as an extension to the static model, and it was shown to be able to model non-stationary signals. This is achieved by allowing the amplitudes and phases of the sinusoids to change as a function of time. By using this model, we are therefore able to model, e.g., audio signals much more accurately, and we do not have to restrict ourselves to analysing only short time-frames in order not to violate the local-stationarity assumption implicit in the static model.

The main contribution of this thesis consists in the proposed and developed inference scheme based on the Gibbs sampler for the sinusoidal parameters of the dynamic sinusoidal model. Whereas previous Bayesian inference schemes for this model are based on an assumed discrete frequency parameter and deterministic inference methods, the

proposed inference scheme in this thesis considers the frequency parameters as continuous random variables and is based on stochastic inference methods. This leads to a more general inference scheme which gives the exact results in the limit of an infinitely large sample size. Another contribution of this thesis is the way samples for the frequency parameters are generated. To our knowledge, it has not been established before that the frequency parameters, conditioned on the other model parameters and states, have a von Mises distribution in the case of a diagonal state noise covariance matrix. This discovery enables simple, exact and user-parameter free sampling of the frequency parameters. The applicability of the proposed inference scheme was demonstrated on several synthetic signals as well as on a real audio signal. These simulations showed that the inference scheme can be successfully applied to applications involving parameter estimation and, in particular, restoration problems.

Although applicable to analysis and synthesis of real audio signals, the proposed inference scheme is still subject to unsolved problems and open to further research. For example, the proposed method for sampling the log-damping coefficient is based on the Laplace approximation or the Metropolis-Hastings algorithm. Although the latter is demonstrated to work in the simulations, the computational complexity of the Metropolis-Hastings algorithm is very high for which reason it would be desirable to derive a more efficient sampling scheme. Including the log-damping coefficients in the sinusoidal model should also be coupled with some kind of adaptive segmentation of the signal so that decays of the signal envelope are observed in the beginning of a time-frame. For many audio signals such as the piano signal considered in the simulation, the frequencies are related to one or more fundamental frequencies. These signals are known as single and multi pitch signals, respectively, and it would be desirable to extend the proposed inference scheme to handle these kind of harmonic structures. Another obvious limitation of the proposed inference scheme is that it assumes the model order as being known. In real world applications, this is almost never the case so the inference should be extended to handle this as well. As we have shown in this thesis, the Bayesian approach offers some sound methods for achieving this. Finally, a significant amount of effort should be put into developing faster and more efficient implementations of the individual sampling steps.

Bibliography

Andrieu, C. and Doucet, A. (1999). Joint Bayesian model selection and estimation of noisy sinusoids via reversible jump MCMC. *IEEE Trans. Signal Process.*, 47(10):2667–2676.

Bernardo, J. M. and Smith, A. (1994). *Bayesian Theory*. John Wiley and Sons Ltd, 1. edition.

Best, D. J. and Fisher, N. I. (1979). Efficient simulation of the von Mises distribution. *J. Appl. Statist.*, 28(2):152–157.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Bolstad, W. M. (2007). *Introduction to Bayesian Statistics*. Wiley-Interscience, 2. edition.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley, 1. edition.

Bretthorst, G. L. (1988). *Bayesian Spectrum Analysis and Parameter Estimation*. Springer-Verlag, Berlin Heidelberg.

Carter, C. K. and Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika*, 81(3):541–553.

Cemgil, A. T. (2004). *Bayesian Music Transcription*. PhD thesis, Radboud University of Nijmegen.

Cemgil, A. T. and Godsill, S. J. (2005). Efficient variational inference for the dynamic harmonic model. In *Proc. IEEE Workshop on Appl. of Signal Process. to Aud. and Acoust.*, pages 271–274.

Cemgil, A. T., Kappen, H. J., and Barber, D. (2006). A generative model for music transcription. *IEEE Trans. Audio, Speech, Lang. Process.*, 14(2):679–694.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *J. Amer. Statistical Assoc.*, 90(432):1313–1321.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4):327–335.

Davy, M., Godsill, S. J., and Idier, J. (2006). Bayesian analysis of polyphonic western tonal music. *J. Acoust. Soc. Am.*, 119(4):2498–2517.

De Jong, P. and Shephard, N. (1995). The simulation smoother for time series models. *Biometrika*, 82(2):339–350.

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. Springer-Verlag, New York.

Dou, L. and Hodgson, R. J. W. (1995). Bayesian inference and Gibbs sampling in spectral analysis and parameter estimation I. *Inverse Problems*, 11(5):1069–1085.

Duda, R. O., Hart, P. E., and Stork, D. G. (2000). *Pattern Classification*. Wiley-Interscience, 2. edition.

Durbin, J. and Koopman, S. (2001). *Time series analysis by state space methods*. Oxford University Press.

Durbin, J. and Koopman, S. J. (2002). A simple and efficient simulation smoother for state space time series analysis. *Biometrika*, 89(3):603–615.

Evans, M., Hastings, N., and Peacock, B. (2000). *Statistical Distributions*. Wiley-Interscience, New York, 3. edition.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*. Chapman & Hall/CRC, 2. edition.

Godsill, S. J. and Rayner, P. J. W. (1998). *Digital Audio Restoration*. Springer-Verlag, London.

Grachev, I. A. (1977). General problems of metrology and measurement techniques. *Measurement Techniques*, 20(6):pp. 784–787.

Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732.

Guttorp, P. and Lockhart, R. (1988). Finding the location of a signal: A Bayesian analysis. *J. Amer. Statistical Assoc.*, 83:pp. 322–330.

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press.

Jaynes, E. T. (2003). *Probability Theory : The Logic of Science*. Cambridge University Press.

Kay, S. (2005). *Intuitive Probability and Random Processes using MATLAB*. Springer-Verlag New York, Inc.

Kay, S. M. (1993). *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*. Prentice Hall PTR.

Koopman, S. J. (1993). Disturbance smoother for state space models. *Biometrika*, 80(1):117–126.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society.*, 34(1):1–41.

Liu, J. S. (2002). *Monte Carlo Strategies in Scientific Computing*. Springer.

MacKay, D. J. C. (2002). *Information Theory, Inference & Learning Algorithms*. Cambridge University Press.

Mazet, V., Brie, D., and Idier, J. (2005). Simulation of positive normal variables using several proposal distributions. In *Proc. IEEE Workshop on Stat. Signal Process.*, pages 37–42.

Petersen, K. B. and Pedersen, M. S. (2008). The matrix cookbook. Online. Version 20081110.

Rajan, J., Rayner, P., and Godsill, S. (1997). Bayesian approach to parameter estimation and interpolation of time-varying autoregressive processes using the Gibbs sampler. *IEE Proc. Vis. Image Signal Process.*, 144(4):249–256.

Stark, H. and Woods, J. W. (2001). *Probability and Random Processes with Application to Signal Processing*. Prentice Hall, 3. edition.

Stoica, P. and Moses, R. L. (2005). *Spectral Analysis of Signals*. Prentice Hall.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques*. Elsevier.

Appendices

Appendix A

Probability Distributions

In this appendix, we sum up the various probability distributions encountered in this thesis. The summary comprises a short description as well as a list of important properties such as probability density function, mean, mode and variance. Special emphasis is put on the Gaussian distribution for which we give and proof an important result.

A.1 Probability Distributions

The following list of probability distributions and their properties is not exhaustive in any way since it only comprises the few distributions encountered in this thesis. A much more complete list of probability distributions can be found in, e.g., [Bishop, 2006, ap. B], [Gelman et al., 2003, ap. A] and [Evans et al., 2000]. In the presentation, we denote the random variable as x in the univariate case, as the D-dimensional vector $\boldsymbol{x}$ in the multivariate case, and as the $D \times D$-dimensional matrix $\boldsymbol{X}$ in the matrix-variate case.

A.1.1 Inverse Gamma Distribution

The inverse gamma distribution is a univariate distribution with two parameters and denoted as Inv-$\mathcal{G}(x; \alpha, \beta)$. It is the inverse of the Gamma distribution and is encountered as the conjugate prior for the variance of a univariate Gaussian distribution or a multivariate Gaussian distribution with an isotropic covariance matrix and known mean. Figure A.1 shows a few plots of the inverse gamma density function and some of

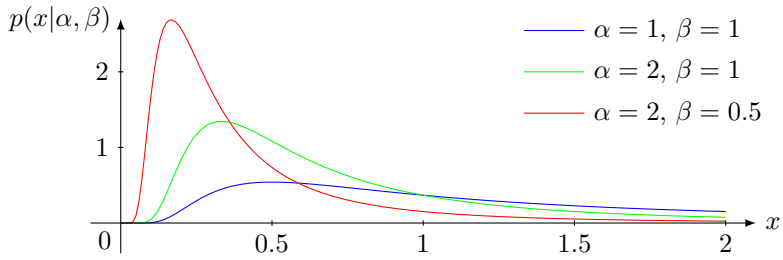

Figure A.1: Three examples of the inverse gamma probability density function.

the main characteristics of the distribution are summarised below.

$$\text{Density function} : p(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} \exp\left\{-\frac{\beta}{x}\right\} \tag{A.1}$$

$$\text{Support} : x \in (0,\infty)$$

$$\text{Parameters} : \alpha > 0 \quad \text{(Shape)}$$

$$\beta > 0 \quad \text{(Scale)}$$

$$\text{Mean (for } \alpha > 1) : E(x) = \frac{\beta}{\alpha - 1} \tag{A.2}$$

$$\text{Mode} : \text{Mode}(x) = \frac{\beta}{\alpha + 1} \tag{A.3}$$

$$\text{Variance (for } \alpha > 2) : \text{Var}(x) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)} \tag{A.4}$$

For $\alpha = \nu/2$ and $\beta = 1/2$ the inverse Gamma distribution is identical to inverse chi-square distribution Inv-$\chi^2(x;\nu)$. The matrix-variate generalisation of the inverse gamma is the inverse Wishart distribution.

A.1.2 Inverse Wishart Distribution

The inverse Wishart distribution is a matrix-variate distribution with two parameters and denoted as Inv-$\mathcal{W}(\boldsymbol{X};\nu,\boldsymbol{\Psi})$. It is the conjugate prior for the covariance matrix of a multivariate Gaussian distribution with known mean vector. Its main characteristics

are summarised below.

$$\text{Density function}: \quad p(\boldsymbol{X}|\nu,\boldsymbol{\Psi}) = B(\nu,\boldsymbol{\Psi})|\boldsymbol{X}|^{-(\nu+D+1)/2}\exp\left\{-\frac{1}{2}\text{tr}(\boldsymbol{\Psi}\boldsymbol{X}^{-1})\right\} \tag{A.5}$$

$$B(\nu,\boldsymbol{\Psi}) = \frac{|\boldsymbol{\Psi}|^{\nu/2}}{2^{\nu D/2}\pi^{D(D-1)/4}\prod_{i=1}^{D}\Gamma(\frac{\nu+1-i}{2})} \tag{A.6}$$

$$\text{Support} : \quad \boldsymbol{X} \text{ is p.d.}$$

$$\text{Parameters} : \quad \nu > D-1 \quad \text{(Degrees of freedom)}$$

$$\boldsymbol{\Psi} \text{ is sym. and p.d.} \quad \text{(Scale matrix)}$$

$$\text{Mean} : \quad E(\boldsymbol{X}) = \frac{\boldsymbol{\Psi}}{\nu-D-1} \tag{A.7}$$

$$\text{Mode} : \quad \text{Mode}(\boldsymbol{X}) = \frac{\boldsymbol{\Psi}}{\nu+D+1} \tag{A.8}$$

In the univariate case, i.e., $D = 1$, the inverse Wishart distribution is identical to the inverse gamma distribution Inv-$\mathcal{G}(x;\alpha,\beta)$ with parameters $\alpha = \nu/2$ and $\beta = \Psi/2$.

A.1.3 Gaussian Distribution

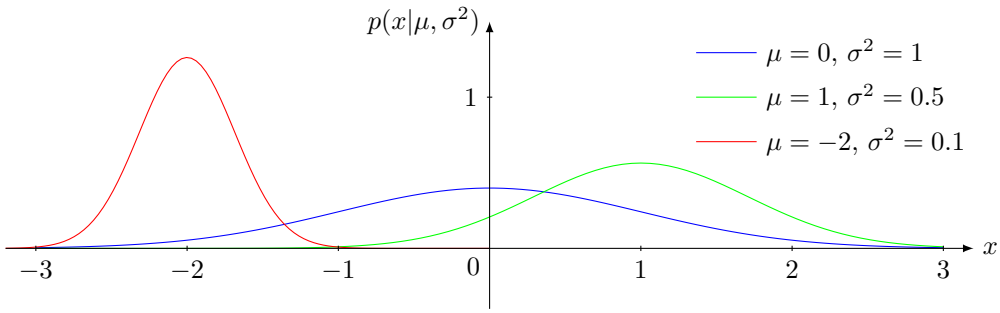

Figure A.2: Three examples of the univariate Gaussian probability density function.

The Gaussian distribution, which is also known as the normal distribution, has two parameters and denoted as $\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$. The Gaussian distribution is frequently encountered and it is the conjugate prior of the mean of the Gaussian distribution with known covariance matrix. Figure A.2 shows a few plots of the univariate Gaussian distribution. The main characteristics for the multivariate Gaussian distribution are

summarised below.

$$\text{Density function} : \quad p(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \tag{A.9}$$

$$\text{Support} : \quad \boldsymbol{x} \in \mathbb{R}^D$$

$$\text{Parameters} : \quad \boldsymbol{\mu} \in \mathbb{R}^D \qquad \text{(Mean)}$$

$$\boldsymbol{\Sigma} \text{ is sym. and p.d.} \qquad \text{(Covariance)}$$

$$\text{Mean} : \quad E(\boldsymbol{x}) = \boldsymbol{\mu} \tag{A.10}$$

$$\text{Mode} : \quad \text{Mode}(\boldsymbol{x}) = \boldsymbol{\mu} \tag{A.11}$$

$$\text{Covariance} : \quad \text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma} \tag{A.12}$$

The covariance matrix is sometimes constrained to a particular structure such as the diagonal or isotropic form. In the diagonal case, the non-diagonal elements are zero which yields a covariance matrix of the form $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \cdots, \sigma_D^2)$. In the isotropic case, the diagonal elements are further constrained to be equal, i.e., $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}_D$.

The Gaussian distribution has many attractive analytical properties. We state one of the most important ones below and several others in appendix B.

Result A.1 If the subvectors $\boldsymbol{x} = \begin{bmatrix}\boldsymbol{x}_1^T & \boldsymbol{x}_2^T\end{bmatrix}^T$ are jointly Gaussian distributed $\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$ with $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ partitioned as

$$\boldsymbol{\mu} = \begin{bmatrix}\boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2\end{bmatrix} \qquad \text{and} \qquad \boldsymbol{\Sigma} = \begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix}, \tag{A.13}$$

then the conditional distribution $p(\boldsymbol{x}_1|\boldsymbol{x}_2)$ is also Gaussian distributed and given by

$$p(\boldsymbol{x}_1|\boldsymbol{x}_2) = \mathcal{N}(\boldsymbol{x}_1;\boldsymbol{\mu}_{1|2},\boldsymbol{\Sigma}_{1|2}) \tag{A.14}$$

$$\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2) \tag{A.15}$$

$$\boldsymbol{\Sigma}_{1|2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\ . \tag{A.16}$$

Further, the marginal distribution $p(\boldsymbol{x}_1)$ is also Gaussian distributed and given by

$$p(\boldsymbol{x}_1) = \mathcal{N}(\boldsymbol{x}_1;\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_{11})\ . \tag{A.17}$$

■

Even tough the proof is rather involved mathematically, we restate it here since it highlights some important points regarding manipulation of the Gaussian. One of these manipulation techniques is called *completing the squares* and it is used frequently throughout this thesis.

Proof. According to Bayes' Theorem, we can factor the joint distribution as

$$p(\boldsymbol{x}_1,\boldsymbol{x}_2) = p(\boldsymbol{x}_1|\boldsymbol{x}_2)p(\boldsymbol{x}_2)\ . \tag{A.18}$$

Initially, we wish to determine $p(\boldsymbol{x}_1|\boldsymbol{x}_2)$ from $p(\boldsymbol{x}_1, \boldsymbol{x}_2)$. The distribution $p(\boldsymbol{x}_2)$ acts as a pure normalisation factor for a given $\boldsymbol{x}_2$ w.r.t. $p(\boldsymbol{x}_1|\boldsymbol{x}_2)$ for which reason we only consider

$$p(\boldsymbol{x}_1|\boldsymbol{x}_2) \propto p(\boldsymbol{x}_1, \boldsymbol{x}_2) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} . \tag{A.19}$$

This formulation greatly simplifies the problem as it allows us to focus our attention to the terms dependent on $\boldsymbol{x}_1$ without having to compute the part which is independent of $\boldsymbol{x}_1$, i.e., the normalisation.

The inverse of the covariance matrix can be computed using the analytic inversion formula [Bishop, 2006, p. 87]

$$\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \boldsymbol{M} & -\boldsymbol{M}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ -\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{M} & \boldsymbol{\Sigma}_{22}^{-1} + \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{M}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \end{bmatrix} \triangleq \begin{bmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} \end{bmatrix} \tag{A.20}$$

where $\boldsymbol{M} = (\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})^{-1}$ and we have introduced the precision matrices $\boldsymbol{\Lambda}_{ij}$ for notational convenience. Now, expanding the quadratic term in the exponent of the exponential in Eq. (A.19) and using the fact that $\boldsymbol{\Lambda}_{12} = \boldsymbol{\Lambda}_{21}^T$ yields

$$\begin{aligned} -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = & -\frac{1}{2}(\boldsymbol{x}_1-\boldsymbol{\mu}_1)^T\boldsymbol{\Lambda}_{11}(\boldsymbol{x}_1-\boldsymbol{\mu}_1) - \frac{1}{2}(\boldsymbol{x}_1-\boldsymbol{\mu}_1)^T\boldsymbol{\Lambda}_{12}(\boldsymbol{x}_2-\boldsymbol{\mu}_2) \\ & -\frac{1}{2}(\boldsymbol{x}_2-\boldsymbol{\mu}_2)^T\boldsymbol{\Lambda}_{21}(\boldsymbol{x}_1-\boldsymbol{\mu}_1) - \frac{1}{2}(\boldsymbol{x}_2-\boldsymbol{\mu}_2)^T\boldsymbol{\Lambda}_{22}(\boldsymbol{x}_2-\boldsymbol{\mu}_2) \\ = & -\frac{1}{2}\boldsymbol{x}_1^T\boldsymbol{\Lambda}_{11}\boldsymbol{x}_1 + \boldsymbol{x}_1^T[\boldsymbol{\Lambda}_{11}\boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{12}(\boldsymbol{x}_2-\boldsymbol{\mu}_2)] + c_2 \end{aligned} \tag{A.21}$$

where

$$c_2 = -\frac{1}{2}\boldsymbol{x}_2^T\boldsymbol{\Lambda}_{22}\boldsymbol{x}_2 + \boldsymbol{x}_2^T(\boldsymbol{\Lambda}_{22}\boldsymbol{\mu}_2 + \boldsymbol{\Lambda}_{21}\boldsymbol{\mu}_1) - \frac{1}{2}\boldsymbol{\mu}_1^T\boldsymbol{\Lambda}_{11}\boldsymbol{\mu}_1 - \frac{1}{2}\boldsymbol{\mu}_2^T\boldsymbol{\Lambda}_{22}\boldsymbol{\mu}_2 \tag{A.22}$$

is a constant representing all terms independent of $\boldsymbol{x}_1$. We see that the function of $\boldsymbol{x}_1$ is also quadratic which means that $p(\boldsymbol{x}_1|\boldsymbol{x}_2)$ is Gaussian distributed completely specified by its mean $\boldsymbol{\mu}_{1|2}$ and covariance $\boldsymbol{\Sigma}_{1|2}$. In order to determine these parameters, we expand the exponent of the exponential for $p(\boldsymbol{x}_1|\boldsymbol{x}_2)$ and compare it to the expansion in Eq. (A.21). This procedure is referred to as *completing the squares*. Expanding yields

$$-\frac{1}{2}(\boldsymbol{x}_1-\boldsymbol{\mu}_{1|2})^T\boldsymbol{\Sigma}_{1|2}^{-1}(\boldsymbol{x}_1-\boldsymbol{\mu}_{1|2}) = -\frac{1}{2}\boldsymbol{x}_1^T\boldsymbol{\Sigma}_{1|2}^{-1}\boldsymbol{x}_1 + \boldsymbol{x}_1^T\boldsymbol{\Sigma}_{1|2}^{-1}\boldsymbol{\mu}_{1|2} - \frac{1}{2}\boldsymbol{\mu}_{1|2}^T\boldsymbol{\Sigma}_{1|2}^{-1}\boldsymbol{\mu}_{1|2} . \tag{A.23}$$

By comparing the second order terms of Eq. (A.21) and Eq. (A.23) we immediately see that

$$\boldsymbol{\Sigma}_{1|2} = \boldsymbol{\Lambda}_{11}^{-1} = \boldsymbol{M}^{-1} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} . \tag{A.24}$$

By a similar comparison for the first order terms, we obtain the equation

$$\boldsymbol{x}_1^T\boldsymbol{\Sigma}_{1|2}^{-1}\boldsymbol{\mu}_{1|2} = \boldsymbol{x}_1^T[\boldsymbol{\Lambda}_{11}\boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{12}(\boldsymbol{x}_2-\boldsymbol{\mu}_2)] \tag{A.25}$$

from which we readily derive

$$\boldsymbol{\mu}_{1|2} = \boldsymbol{\Sigma}_{1|2}[\boldsymbol{\Lambda}_{11}\boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{12}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)] = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2) \ . \tag{A.26}$$

This concludes the proof for the conditional distribution $p(\boldsymbol{x}_1|\boldsymbol{x}_2)$.

The marginal distribution is derived using similar procedures. First, we write

$$p(\boldsymbol{x}_1) = \int p(\boldsymbol{x}_1, \boldsymbol{x}_2) d\boldsymbol{x}_2 \propto \int \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right\} d\boldsymbol{x}_2 \ . \tag{A.27}$$

By expanding and collecting terms in the same way as before, but w.r.t. $\boldsymbol{x}_2$ instead of $\boldsymbol{x}_1$, we obtain

$$p(\boldsymbol{x}_1) \propto \exp\left\{c_1 + \boldsymbol{\mu}_{2|1}^T \boldsymbol{\Sigma}_{2|1}^{-1} \boldsymbol{\mu}_{2|1}\right\} \int \exp\left\{-\frac{1}{2}(\boldsymbol{x}_2 - \boldsymbol{\mu}_{2|1})^T \boldsymbol{\Sigma}_{2|1}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_{2|1})\right\} d\boldsymbol{x}_2 \tag{A.28}$$

where c_1 and $\boldsymbol{\mu}_{2|1}^T \boldsymbol{\Sigma}_{2|1}^{-1} \boldsymbol{\mu}_{2|1}$ are independent of $\boldsymbol{x}_2$. Expressions for them are given by following the same procedure as in the first part of this proof, i.e., with swapped indices c_1, $\boldsymbol{\Sigma}_{2|1}$ and $\boldsymbol{\mu}_{2|1}$ are given by Eq. (A.22), Eq. (A.24) and Eq. (A.26), respectively. The integral is recognised as the integral over an unnormalised Gaussian proportional to $p(\boldsymbol{x}_2|\boldsymbol{x}_1)$ and is thus easy to evaluate; it simply gives a constant independent of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. By expanding and collecting terms of the exponent of the remaining exponential, we obtain after some algebra

$$c_1 + \boldsymbol{\mu}_{2|1}^T \boldsymbol{\Sigma}_{2|1}^{-1} \boldsymbol{\mu}_{2|1} = -\frac{1}{2}\boldsymbol{x}_1^T \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{x}_1 + \boldsymbol{x}_1^T \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\mu}_1 + \text{const} \tag{A.29}$$

where 'const' denotes the terms independent of $\boldsymbol{x}_1$. The marginal distribution $p(\boldsymbol{x}_1)$ is thus Gaussian since the exponent is a quadratic function of $\boldsymbol{x}_1$. Completing the squares yields easily the parameters and we have that

$$p(\boldsymbol{x}_1) = \mathcal{N}(\boldsymbol{x}_1; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) \ . \tag{A.30}$$

□

A.1.4 Student's t-Distribution

The student's t-distribution has three parameters and is denoted as $\text{St}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$. It is, for example, encountered when marginalising the covariance matrix of a multivariate Gaussian pdf with respect to a conjugate inverse Wishart prior. Figure A.3 shows a few plots of the density function of the univariate student's t-distribution. The main

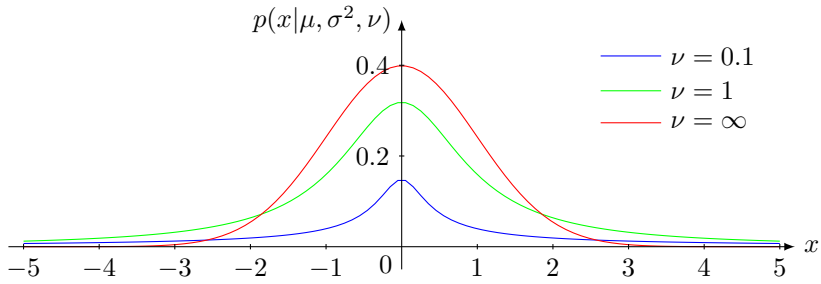

Figure A.3: Three examples of the univariate Student's t-probability density function. In all examples, the mean and variance are fixed to 0 and 1, respectively.

characteristics of the multivariate student's t-distribution are summarised below.

$$\text{Density function} : p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \frac{\Gamma(D/2 + \nu/2)}{\Gamma(\nu/2)\sqrt{(\pi\nu)^D|\boldsymbol{\Sigma}|}} \left[1 + \frac{\Delta^2}{\nu}\right]^{\frac{-D-\nu}{2}} \tag{A.31}$$

$$\Delta^2 = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \tag{A.32}$$

$$\text{Support} : \boldsymbol{x} \in \mathbb{R}^D$$

$$\text{Parameters} : \boldsymbol{\mu} \in \mathbb{R}^D \quad \text{(Mean)}$$

$$\boldsymbol{\Sigma} \text{ is sym. and p.d.} \quad \text{(Covariance)}$$

$$\nu > 0 \quad \text{(Degrees of freedom)}$$

$$\text{Mean(for } \nu > 1) : E(\boldsymbol{x}) = \boldsymbol{\mu} \tag{A.33}$$

$$\text{Mode} : \text{Mode}(\boldsymbol{x}) = \boldsymbol{\mu} \tag{A.34}$$

$$\text{Covariance(for } \nu > 2) : \text{Cov}(\boldsymbol{x}) = \frac{\nu}{\nu - 2}\boldsymbol{\Sigma} \tag{A.35}$$

The student's t-distribution can be interpreted as an infinite mixture of Gaussian distributions with equal mean values but with different variances, and it converges to the Gaussian distribution given by $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ for $\nu \to \infty$.

A.1.5 Uniform Distribution

The uniform distribution is a simple distribution with two parameters and denoted as $\mathcal{U}(x; a, b)$. It is important in a wide range of applications and is often used as a building block for random variate generations from other distributions. Figure A.4 shows a few plots of the uniform density function and some of the main characteristics of the

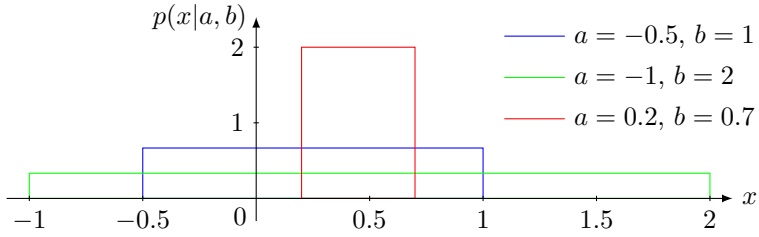

Figure A.4: Three examples of the uniform probability density function.

distribution are summarised below.

$$\text{Density function}: \quad p(x|a,b) = \begin{cases} (b-a)^{-1} & \text{for } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases} \tag{A.36}$$

$$\text{Support} : \quad x \in (a,b)$$

$$\text{Parameters} : \quad a \in (-\infty, b) \quad \text{(lower boundary)}$$

$$b \in (a, +\infty) \quad \text{(upper boundary)}$$

$$\text{Mean} : \quad E(x) = \frac{a+b}{2} \tag{A.37}$$

$$\text{Mode} : \quad \text{any point in } (a,b) \tag{A.38}$$

$$\text{Variance} : \quad \text{Var}(x) = \frac{(b-a)^2}{12} \tag{A.39}$$

For Bayesian statistical inference, a uniform distribution with a sufficiently large support can be selected as a prior distribution in cases where little or no prior information is available.

A.1.6 Von Mises Distribution

The von Mises distribution is also known as the circular Gaussian distribution since it is the Gaussian distribution for *periodic variables*. An example of a periodic variable is an angle. The distribution has two parameters and is denoted as $\mathcal{VM}(x; \kappa, \mu)$. Figure A.5 shows a few plots of the von Mises density function and some of the main characteristics

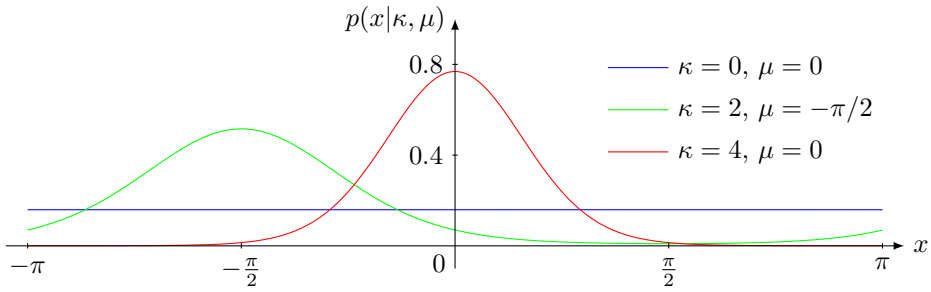

Figure A.5: Three examples of the von Mises probability density function.

of the distribution are summarised below.

$$\text{Density function} : \quad p(x|\kappa,\mu) = \frac{1}{2\pi I_0(\kappa)} \exp\{\kappa \cos(x-\mu)\} \tag{A.40}$$

$$\text{Support} : \quad \text{any interval } I \text{ of length } 2\pi$$

$$\text{Parameters} : \quad \kappa > 0 \quad \text{(Concentration)}$$

$$\mu \in I \quad \text{(Location)}$$

$$\text{Mean} : \quad E(x) = \mu \tag{A.41}$$

$$\text{Mode} : \quad \text{Mode}(x) = \mu \tag{A.42}$$

$$\text{Circular variance} : \quad \text{Var}(x) = 1 - \frac{I_1(\kappa)}{I_0(\kappa)} \tag{A.43}$$

The function $I_k(\kappa)$ is the modified Bessel function of the first kind of order k.

Appendix B

Bayesian Inference for the Gaussian

The single most important building block in statistical inference is the Gaussian distribution. This stems from the fact that it is an accurate model for many real world random phenomena, that it is the maximum entropy distribution if nothing but the mean and the variance is specified, and that it has convenient analytical properties. In Bayesian statistics, the Gaussian distribution is also very popular and in this section we state some of the important results associated with Bayesian inference for the Gaussian distribution. The results are useful in the situation in which we observe M D-dimensional i.i.d. random vectors

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_1^T & \boldsymbol{x}_2^T & \cdots & \boldsymbol{x}_M^T \end{bmatrix}^T \tag{B.1}$$

from a Gaussian distribution whose mean $\boldsymbol{\mu}$ and/or covariance $\boldsymbol{\Sigma}$ are unknown. The $N = MD$-dimensional vector $\boldsymbol{x}$ has also a Gaussian distribution with mean vector

$$\boldsymbol{m} = \begin{bmatrix} \boldsymbol{\mu}^T & \cdots & \boldsymbol{\mu}^T \end{bmatrix}^T \tag{B.2}$$

and covariance matrix

$$\boldsymbol{C} = \text{diag}(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}, \cdots, \boldsymbol{\Sigma}) \tag{B.3}$$

where the block diagonal structure of the covariance matrix follows from the fact that the $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_M$ are independent random vectors. Using Bayes' Theorem, we wish to derive the posterior distributions for the mean $\boldsymbol{\mu}$ and/or the covariance $\boldsymbol{\Sigma}$ as well as for the model evidence. In some of the results, we assume the mean vector to be an affine function of an L-dimensional parameter vector $\boldsymbol{\theta}$, i.e., $\boldsymbol{\mu} = \boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{b}$ with $\boldsymbol{A}$ and $\boldsymbol{b}$ being

known, and an isotropic covariance matrix $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_D$. For these assumptions, the mean and covariance for $\boldsymbol{x}$ are $\boldsymbol{m} = \boldsymbol{H}\boldsymbol{\theta} + \boldsymbol{r}$ with

$$\boldsymbol{H} \triangleq \begin{bmatrix} \boldsymbol{A}^T & \boldsymbol{A}^T & \cdots & \boldsymbol{A}^T \end{bmatrix}^T \tag{B.4}$$

$$\boldsymbol{r} \triangleq \begin{bmatrix} \boldsymbol{b}^T & \boldsymbol{b}^T & \cdots & \boldsymbol{b}^T \end{bmatrix}^T \tag{B.5}$$

and $\boldsymbol{C} = \sigma^2 \boldsymbol{I}_N$, respectively. We constrain ourselves to these structures of the mean vector and covariance matrix in some of the results below since these are encountered frequently in this thesis.

B.1 Inference for an Unknown Mean

We observe M D-dimensional i.i.d. random vectors $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_M$ from a Gaussian distribution $\mathcal{N}(\boldsymbol{x}_m; \boldsymbol{A}\boldsymbol{\theta}+\boldsymbol{b}, \boldsymbol{\Sigma})$ with unknown mean parameter vector $\boldsymbol{\theta}$ and known covariance $\boldsymbol{\Sigma}$. Assuming a conjugate Gaussian prior $\mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$ over the mean parameter vector $\boldsymbol{\theta}$ with known hyperparameters $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$, and using Bayes' theorem

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{x})} \ , \tag{B.6}$$

we wish to find the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{x})$ as well as the evidence $p(\boldsymbol{x})$.

Result B.1 Let the prior distribution and the likelihood be given by

$$p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}) \tag{B.7}$$

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{H}\boldsymbol{\theta} + \boldsymbol{r}, \boldsymbol{C}) \ , \tag{B.8}$$

respectively. If the covariance matrix $\boldsymbol{C}$ and the hyperparameters $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ are known, the posterior distribution for the parameter vector is a Gaussian distribution given by

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}_{\boldsymbol{\theta}|\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}|\boldsymbol{x}}) \tag{B.9}$$

$$\boldsymbol{\mu}_{\boldsymbol{\theta}|\boldsymbol{x}} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}|\boldsymbol{x}}(\boldsymbol{H}^T\boldsymbol{C}^{-1}(\boldsymbol{x} - \boldsymbol{r}) + \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\theta}}) \tag{B.10}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}|\boldsymbol{x}} = (\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} + \boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{H})^{-1} \ . \tag{B.11}$$

The marginal distribution over $\boldsymbol{\theta}$ is also a Gaussian distribution and given by

$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{H}\boldsymbol{\mu}_{\boldsymbol{\theta}} + \boldsymbol{r}, \boldsymbol{C} + \boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\boldsymbol{H}^T) \ . \tag{B.12}$$

■

Proof. From result A.1, we know how to compute the conditional distribution $p(\boldsymbol{\theta}|\boldsymbol{x})$ and the marginal distribution $p(\boldsymbol{x})$ from their joint distribution $p(\boldsymbol{\theta}, \boldsymbol{x})$. Thus, the proof

focuses on deriving an expression for this joint distribution. For notational convenience, we initially define the Gaussian random variable $\boldsymbol{y} = \begin{bmatrix}\boldsymbol{\theta}^T & \boldsymbol{x}^T\end{bmatrix}^T$ with mean $\boldsymbol{\mu}_{\boldsymbol{y}}$ and covariance $\boldsymbol{\Sigma}_{\boldsymbol{y}}$. We now have

$$p(\boldsymbol{y}) = p(\boldsymbol{\theta}, \boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{H\theta} + \boldsymbol{r}, \boldsymbol{C})\mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}) \tag{B.13}$$

$$\propto \exp\left\{\frac{-1}{2}(\boldsymbol{x} - \boldsymbol{H\theta} - \boldsymbol{r})^T\boldsymbol{C}^{-1}(\boldsymbol{x} - \boldsymbol{H\theta} - \boldsymbol{r})\right\}\exp\left\{\frac{-1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta}})^T\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta}})\right\}$$

$$\propto \exp\Bigg\{\frac{-1}{2}\big[\boldsymbol{\theta}^T(\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{H} + \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1})\boldsymbol{\theta} + \boldsymbol{x}^T\boldsymbol{C}^{-1}\boldsymbol{x} - \boldsymbol{\theta}^T\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{x} - \boldsymbol{x}^T\boldsymbol{C}^{-1}\boldsymbol{H\theta}$$

$$- 2\boldsymbol{\theta}^T(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{r}) - 2\boldsymbol{x}^T\boldsymbol{C}^{-1}\boldsymbol{r}\big]\Bigg\}\,. \tag{B.14}$$

By comparing the latter expression with the exponent of the exponential of $p(\boldsymbol{y})$ given by

$$-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu}_{\boldsymbol{y}})^T\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_{\boldsymbol{y}}) = -\frac{1}{2}\left(\boldsymbol{y}^T\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\boldsymbol{y} - 2\boldsymbol{y}^T\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\boldsymbol{\mu}_{\boldsymbol{y}} + \boldsymbol{\mu}_{\boldsymbol{y}}^T\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\boldsymbol{\mu}_{\boldsymbol{y}}\right)\,, \tag{B.15}$$

we obtain by completing the squares that

$$\boldsymbol{y}^T\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\boldsymbol{y} = \begin{bmatrix}\boldsymbol{\theta}^T & \boldsymbol{x}^T\end{bmatrix}\begin{bmatrix}\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{H} + \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} & -\boldsymbol{H}^T\boldsymbol{C}^{-1} \\ -\boldsymbol{C}^{-1}\boldsymbol{H} & \boldsymbol{C}^{-1}\end{bmatrix}\begin{bmatrix}\boldsymbol{\theta} \\ \boldsymbol{x}\end{bmatrix} \tag{B.16}$$

and

$$-2\boldsymbol{y}^T\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\boldsymbol{\mu}_{\boldsymbol{y}} = -2\begin{bmatrix}\boldsymbol{\theta}^T & \boldsymbol{x}^T\end{bmatrix}\begin{bmatrix}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{r} \\ \boldsymbol{C}^{-1}\boldsymbol{r}\end{bmatrix}\,. \tag{B.17}$$

From these two equations, we readily obtain that the mean and covariance for the joint distribution $p(\boldsymbol{y}) = p(\boldsymbol{\theta}, \boldsymbol{x})$ are given by

$$\boldsymbol{\mu}_{\boldsymbol{y}} = \boldsymbol{\Sigma}_{\boldsymbol{y}}\begin{bmatrix}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{r} \\ \boldsymbol{C}^{-1}\boldsymbol{r}\end{bmatrix} \tag{B.18}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{y}} = \begin{bmatrix}\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{H} + \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} & -\boldsymbol{H}^T\boldsymbol{C}^{-1} \\ -\boldsymbol{C}^{-1}\boldsymbol{H} & \boldsymbol{C}^{-1}\end{bmatrix}^{-1} = \begin{bmatrix}\boldsymbol{\Sigma}_{\boldsymbol{\theta}} & \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\boldsymbol{H}^T \\ \boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{\theta}} & \boldsymbol{C} + \boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\boldsymbol{H}^T\end{bmatrix} \tag{B.19}$$

where the last equality follows from Eq. (A.20). Using result A.1 and some straightforward algebra, we therefore obtain for the conditional distribution that

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}_{\boldsymbol{\theta}|\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}|\boldsymbol{x}}) \tag{B.20}$$

$$\boldsymbol{\mu}_{\boldsymbol{\theta}|\boldsymbol{x}} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}|\boldsymbol{x}}(\boldsymbol{H}^T\boldsymbol{C}^{-1}(\boldsymbol{x} - \boldsymbol{r}) + \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\theta}}) \tag{B.21}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}|\boldsymbol{x}} = (\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} + \boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{H})^{-1} \tag{B.22}$$

and for the marginal distribution that

$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{H}\boldsymbol{\mu}_{\boldsymbol{\theta}} + \boldsymbol{r}, \boldsymbol{C} + \boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\boldsymbol{H}^T) \ . \tag{B.23}$$

This concludes the proof. □

B.2 Inference for an Unknown Covariance

We observe M D-dimensional i.i.d. random vectors $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_M$ from a Gaussian distribution $\mathcal{N}(\boldsymbol{x}_m; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with known mean vector $\boldsymbol{\mu}$ and unknown covariance matrix $\boldsymbol{\Sigma}$. Assuming a conjugate inverse Wishart prior Inv-$\mathcal{W}(\boldsymbol{\Sigma}; \nu, \boldsymbol{\Psi})$ over the covariance matrix $\boldsymbol{\Sigma}$ with known hyperparameters ν and $\boldsymbol{\Psi}$, and using Bayes' theorem

$$p(\boldsymbol{\Sigma}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{\Sigma})p(\boldsymbol{\Sigma})}{p(\boldsymbol{x})} \ , \tag{B.24}$$

we wish to find the posterior distribution $p(\boldsymbol{\Sigma}|\boldsymbol{x})$.

Result B.2 Let the prior distribution and the likelihood be given by

$$p(\boldsymbol{\Sigma}) = \text{Inv-}\mathcal{W}(\boldsymbol{\Sigma}; \nu, \boldsymbol{\Psi}) \tag{B.25}$$

$$\begin{aligned} p(\boldsymbol{x}|\boldsymbol{\Sigma}) &= \mathcal{N}(\boldsymbol{x}; \boldsymbol{m}, \boldsymbol{C}) \\ &= \prod_{m=1}^{M} p(\boldsymbol{x}_m|\boldsymbol{\Sigma}) = \prod_{m=1}^{M} \mathcal{N}(\boldsymbol{x}_m; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \ , \end{aligned} \tag{B.26}$$

respectively. If the mean vector $\boldsymbol{\mu}$ and the hyperparameters ν and $\boldsymbol{\Psi}$ are known, the posterior distribution for the noise covariance is an inverse Wishart distribution given by

$$p(\boldsymbol{\Sigma}|\boldsymbol{x}) = \text{Inv-}\mathcal{W}(\boldsymbol{\Sigma}; \nu_{\boldsymbol{\Sigma}|\boldsymbol{x}}, \boldsymbol{\Psi}_{\boldsymbol{\Sigma}|\boldsymbol{x}}) \tag{B.27}$$

$$\nu_{\boldsymbol{\Sigma}|\boldsymbol{x}} = \nu + M \tag{B.28}$$

$$\boldsymbol{\Psi}_{\boldsymbol{\Sigma}|\boldsymbol{x}} = \boldsymbol{\Psi} + \sum_{m=1}^{M} (\boldsymbol{x}_m - \boldsymbol{\mu})(\boldsymbol{x}_m - \boldsymbol{\mu})^T \ . \tag{B.29}$$

■

Proof. From Bayes' theorem, we have that

$$p(\boldsymbol{\Sigma}|\boldsymbol{x}) \propto p(\boldsymbol{x}|\boldsymbol{\Sigma})p(\boldsymbol{\Sigma}) = \left[\prod_{m=1}^{M} \mathcal{N}(x_m; \boldsymbol{\mu}, \boldsymbol{\Sigma})\right] \text{Inv-}\mathcal{W}(\boldsymbol{\Sigma}; \nu, \boldsymbol{\Psi}) \tag{B.30}$$

$$\propto \left[\prod_{m=1}^{M} |\boldsymbol{\Sigma}|^{-1/2} \exp\left\{\frac{-1}{2}(\boldsymbol{x}_m - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_m - \boldsymbol{\mu})\right\}\right] \times |\boldsymbol{\Sigma}|^{-(\nu+D+1)/2} \exp\left\{\frac{-1}{2}\text{tr}(\boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1})\right\} \tag{B.31}$$

$$= |\boldsymbol{\Sigma}|^{-M/2}|\boldsymbol{\Sigma}|^{-(\nu+D+1)/2} \times \exp\left\{\frac{-1}{2}\text{tr}\left(\sum_{m=1}^{M}(\boldsymbol{x}_m - \boldsymbol{\mu})(\boldsymbol{x}_m - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\right) - \frac{1}{2}\text{tr}(\boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1})\right\} \tag{B.32}$$

$$= |\boldsymbol{\Sigma}|^{-(\nu+M+D+1)/2} \times \exp\left\{\frac{-1}{2}\text{tr}\left(\left[\sum_{m=1}^{M}(\boldsymbol{x}_m - \boldsymbol{\mu})(\boldsymbol{x}_m - \boldsymbol{\mu})^T + \boldsymbol{\Psi}\right]\boldsymbol{\Sigma}^{-1}\right)\right\} \tag{B.33}$$

$$\propto \text{Inv-}\mathcal{W}\left(\boldsymbol{\Sigma}; \nu + M, \boldsymbol{\Psi} + \sum_{m=1}^{M}(\boldsymbol{x}_m - \boldsymbol{\mu})(\boldsymbol{x}_m - \boldsymbol{\mu})^T\right) \tag{B.34}$$

where the last proportional sign follows from comparing the second last factorisation against the expression for the inverse Wishart distribution given by Eq. (A.5). □

If we restrict the covariance matrix to be isotropic, i.e., $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}_D$, the conjugate prior is the inverse gamma distribution Inv-$\mathcal{G}(\sigma^2; a, b)$. Assuming the hyperparameters to be known and using Bayes' theorem

$$p(\sigma^2|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\sigma^2)p(\sigma^2)}{p(\boldsymbol{x})} \ , \tag{B.35}$$

we wish to find the posterior distribution $p(\sigma^2|\boldsymbol{x})$ as well as the evidence $p(\boldsymbol{x})$.

Result B.3 Let the prior distribution and the likelihood be given by

$$p(\sigma^2) = \text{Inv-}\mathcal{G}(\sigma^2; a, b) \tag{B.36}$$

$$p(\boldsymbol{x}|\sigma^2) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{m}, \sigma^2\boldsymbol{I}_N) \ , \tag{B.37}$$

respectively. If the mean vector $\boldsymbol{m}$ and the hyperparameters a and b are known, the

posterior distribution for the noise variance is an inverse gamma distribution given by

$$p(\sigma^2|\boldsymbol{x}) = \text{Inv-}\mathcal{G}(\sigma^2; a_{\sigma^2|\boldsymbol{x}}, b_{\sigma^2|\boldsymbol{x}}) \tag{B.38}$$

$$a_{\sigma^2|\boldsymbol{x}} = a + N/2 \tag{B.39}$$

$$b_{\sigma^2|\boldsymbol{x}} = b + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^T(\boldsymbol{x}-\boldsymbol{m}) \ . \tag{B.40}$$

The marginal distribution over σ^2 is a multivariate Student's t-distribution given by

$$p(\boldsymbol{x}) = \text{St}(\boldsymbol{x}; \boldsymbol{m}, \frac{b}{a}\boldsymbol{I}_N, 2a) \ . \tag{B.41}$$

■

Proof. From Bayes' theorem, we have that

$$p(\sigma^2|\boldsymbol{x}) \propto p(\boldsymbol{x}|\sigma^2)p(\sigma^2) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{m}, \sigma^2\boldsymbol{I}_N)\text{Inv-}\mathcal{G}(\sigma^2; a, b) \tag{B.42}$$

$$\propto |\sigma^2\boldsymbol{I}_N|^{-1/2} \exp\left\{\frac{-1}{2\sigma^2}(\boldsymbol{x}-\boldsymbol{m})^T(\boldsymbol{x}-\boldsymbol{m})\right\}(\sigma^2)^{-(a+1)}\exp\left\{\frac{-b}{\sigma^2}\right\} \tag{B.43}$$

$$= (\sigma^2)^{-N/2}(\sigma^2)^{-(a+1)}\exp\left\{\frac{-1}{2\sigma^2}(\boldsymbol{x}-\boldsymbol{m})^T(\boldsymbol{x}-\boldsymbol{m}) - \frac{b}{\sigma^2}\right\} \tag{B.44}$$

$$= (\sigma^2)^{-(a+N/2+1)}\exp\left\{-\frac{b+\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^T(\boldsymbol{x}-\boldsymbol{m})}{\sigma^2}\right\} \tag{B.45}$$

$$\propto \text{Inv-}\mathcal{G}\left(\sigma^2; a+N/2, b+\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^T(\boldsymbol{x}-\boldsymbol{m})\right) \tag{B.46}$$

where the last proportional sign follows from comparing the second last factorisation against the expression for the inverse gamma distribution given by Eq. (A.1).

For the marginal distribution, we have that

$$p(\boldsymbol{x}) = \int p(\sigma^2, \boldsymbol{x})d\sigma^2 = \int p(\boldsymbol{x}|\sigma^2)p(\sigma^2)d\sigma^2 \tag{B.47}$$

$$\propto \int (\sigma^2)^{-(a+N/2+1)}\exp\left\{-\frac{b+\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^T(\boldsymbol{x}-\boldsymbol{m})}{\sigma^2}\right\}d\sigma^2 \tag{B.48}$$

where the proportional sign follows from the derivation for $p(\sigma^2|\boldsymbol{x})$ given above. The expression can be extended as

$$p(\boldsymbol{x}) \propto \left[b+\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^T(\boldsymbol{x}-\boldsymbol{m})\right]^{-(a+N/2)} \int \left[b+\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^T(\boldsymbol{x}-\boldsymbol{m})\right]^{(a+N/2)}$$
$$\times (\sigma^2)^{-(a+N/2+1)}\exp\left\{-\frac{b+\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^T(\boldsymbol{x}-\boldsymbol{m})}{\sigma^2}\right\}d\sigma^2 \tag{B.49}$$

so the integral is taken over the inverse gamma distribution scaled by the factor $\Gamma(a + N/2)$ and is therefore equal to $\Gamma(a + N/2)$. Thus, we have that

$$p(\boldsymbol{x}) \propto \left[b + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{m})^T(\boldsymbol{x} - \boldsymbol{m})\right]^{-(a+N/2)} \tag{B.50}$$

$$\propto \left[1 + \frac{(\boldsymbol{x} - \boldsymbol{m})^T \frac{a}{b}(\boldsymbol{x} - \boldsymbol{m})}{2a}\right]^{\frac{-N-2a}{2}} \tag{B.51}$$

$$\propto \mathrm{St}(\boldsymbol{x}; \boldsymbol{m}, \frac{b}{a}\boldsymbol{I}_N, 2a) \tag{B.52}$$

where the last proportional sign follows from comparing the second last factorisation against the expression for the student's t-distribution given by Eq. (A.31). □

B.3 Inference for an Unknown Mean and an Unknown Isotropic Covariance

We observe M D-dimensional i.i.d. random vectors $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_M$ from a Gaussian distribution $\mathcal{N}(\boldsymbol{x}_m; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with unknown L-dimensional mean parameter vector $\boldsymbol{\theta}$ and unknown isotopic covariance $\sigma^2\boldsymbol{I}_D$. The conjugate priors for $\boldsymbol{\theta}$ and σ^2 are the Gaussian distribution $\mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}_{\boldsymbol{\theta}}, \sigma^2\boldsymbol{C}_{\boldsymbol{\theta}})$ and inverse Gamma distribution Inv-$\mathcal{G}(\sigma^2; a, b)$, respectively. Notice that the covariance of the Gaussian depends on the noise variance which is required for conjugacy, i.e., $\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \sigma^2\boldsymbol{C}_{\boldsymbol{\theta}}$. The joint distribution $p(\boldsymbol{\theta}, \sigma^2) = p(\boldsymbol{\theta}|\sigma^2)p(\sigma^2)$ constituted by these priors is often referred to as the *Gaussian-inverted gamma* distribution. Using Bayes' theorem

$$p(\boldsymbol{\theta}, \sigma^2|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta}, \sigma^2)p(\boldsymbol{\theta}|\sigma^2)p(\sigma^2)}{p(\boldsymbol{x})} \ , \tag{B.53}$$

we wish to find the marginal posterior distributions $p(\boldsymbol{\theta}|\boldsymbol{x})$ and $p(\sigma^2|\boldsymbol{x})$ as well as the evidence $p(\boldsymbol{x})$.

Result B.4 Let the prior distributions and the likelihood be given by

$$p(\sigma^2) = \text{Inv-}\mathcal{G}(\sigma^2; a, b) \tag{B.54}$$

$$p(\boldsymbol{\theta}|\sigma^2) = \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}_{\boldsymbol{\theta}}, \sigma^2\boldsymbol{C}_{\boldsymbol{\theta}}) \tag{B.55}$$

$$p(\boldsymbol{x}|\boldsymbol{\theta}, \sigma^2) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{H}\boldsymbol{\theta} + \boldsymbol{r}, \sigma^2\boldsymbol{I}_N) \ , \tag{B.56}$$

respectively. If the hyperparameters a, b, $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ and $\boldsymbol{C}_{\boldsymbol{\theta}}$ are known, the marginal posterior

distribution for the variance is an inverse gamma distribution given by

$$p(\sigma^2|\boldsymbol{x}) = \text{Inv-}\mathcal{G}(\sigma^2; a_{\sigma^2|\boldsymbol{x}}, b_{\sigma^2|\boldsymbol{x}}) \tag{B.57}$$
$$a_{\sigma^2|\boldsymbol{x}} = a + N/2 \tag{B.58}$$
$$b_{\sigma^2|\boldsymbol{x}} = b + \frac{1}{2}\Delta_{\boldsymbol{x}}^2 \tag{B.59}$$
$$\Delta_{\boldsymbol{x}}^2 = (\boldsymbol{x} - \boldsymbol{H}\boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{r})^T(\boldsymbol{I}_N + \boldsymbol{H}\boldsymbol{C}_{\boldsymbol{\theta}}\boldsymbol{H}^T)^{-1}(\boldsymbol{x} - \boldsymbol{H}\boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{r}) \tag{B.60}$$
$$= (\boldsymbol{x} - \boldsymbol{H}\boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{r})^T(\boldsymbol{I}_N - \boldsymbol{H}\boldsymbol{C}_{\boldsymbol{\theta}|\boldsymbol{x}}\boldsymbol{H}^T)(\boldsymbol{x} - \boldsymbol{H}\boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{r}) \tag{B.61}$$
$$= (\boldsymbol{x} - \boldsymbol{H}\boldsymbol{\mu}_{\boldsymbol{\theta}|\boldsymbol{x}} - \boldsymbol{r})^T(\boldsymbol{x} - \boldsymbol{r}) + (\boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{\mu}_{\boldsymbol{\theta}|\boldsymbol{x}})^T\boldsymbol{C}_{\boldsymbol{\theta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\theta}} \ . \tag{B.62}$$

The the second expression for $\Delta_{\boldsymbol{x}}^2$ follows from the matrix inversion lemma, and the latter expression is partly due to [Bernardo and Smith, 1994, p. 442] with $\boldsymbol{\mu}_{\boldsymbol{\theta}|\boldsymbol{x}}$ and $\boldsymbol{C}_{\boldsymbol{\theta}|\boldsymbol{x}}$ given below. The marginal posterior distribution for the mean parameter vector is a multivariate Student's t-distribution given by

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \text{St}\left(\boldsymbol{\theta}; \boldsymbol{\mu}_{\boldsymbol{\theta}|\boldsymbol{x}}, \frac{b_{\sigma^2|\boldsymbol{x}}}{a_{\sigma^2|\boldsymbol{x}}}\boldsymbol{C}_{\boldsymbol{\theta}|\boldsymbol{x}}, 2a_{\sigma^2|\boldsymbol{x}}\right) \tag{B.63}$$
$$\boldsymbol{\mu}_{\boldsymbol{\theta}|\boldsymbol{x}} = \boldsymbol{C}_{\boldsymbol{\theta}|\boldsymbol{x}}(\boldsymbol{H}^T(\boldsymbol{x} - \boldsymbol{r}) + \boldsymbol{C}_{\boldsymbol{\theta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\theta}}) \tag{B.64}$$
$$\boldsymbol{C}_{\boldsymbol{\theta}|\boldsymbol{x}} = (\boldsymbol{H}^T\boldsymbol{H} + \boldsymbol{C}_{\boldsymbol{\theta}}^{-1})^{-1} \ . \tag{B.65}$$

The marginal distribution over $\boldsymbol{\theta}$ and σ^2 is also a multivariate Student's t-distribution and given by

$$p(\boldsymbol{x}) = \text{St}\left(\boldsymbol{x}; \boldsymbol{H}\boldsymbol{\mu}_\theta + \boldsymbol{r}, \frac{b}{a}(\boldsymbol{I}_N + \boldsymbol{H}\boldsymbol{C}_{\boldsymbol{\theta}}\boldsymbol{H}^T), 2a\right) \ . \tag{B.66}$$

■

Proof. From Bayes' theorem, we have that

$$p(\sigma^2|\boldsymbol{x}) \propto p(\boldsymbol{x}|\sigma^2)p(\sigma^2) \tag{B.67}$$

for the marginal posterior distribution for the noise variance. The conditional distribution $p(\boldsymbol{x}|\sigma^2)$ can be expressed as

$$p(\boldsymbol{x}|\sigma^2) = \int p(\boldsymbol{x}, \boldsymbol{\theta}|\sigma^2)d\boldsymbol{\theta} = \int p(\boldsymbol{x}|\boldsymbol{\theta}, \sigma^2)p(\boldsymbol{\theta}|\sigma^2)d\boldsymbol{\theta} \tag{B.68}$$
$$= \mathcal{N}(\boldsymbol{x}; \boldsymbol{H}\boldsymbol{\mu}_{\boldsymbol{\theta}} + \boldsymbol{r}, \sigma^2(\boldsymbol{I}_N + \boldsymbol{H}\boldsymbol{C}_{\boldsymbol{\theta}}\boldsymbol{H}^T)) \tag{B.69}$$

where the last equality follows from result B.1. Inserting this into Eq. (B.67) and using result B.3 yield

$$p(\sigma^2|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{H}\boldsymbol{\mu}_{\boldsymbol{\theta}} + \boldsymbol{r}, \sigma^2(\boldsymbol{I}_N + \boldsymbol{H}\boldsymbol{C}_{\boldsymbol{\theta}}\boldsymbol{H}^T))\text{Inv-}\mathcal{G}(\sigma^2; a, b) \tag{B.70}$$
$$= \text{Inv-}\mathcal{G}(\sigma^2; a_{\sigma^2|\boldsymbol{x}}, b_{\sigma^2|\boldsymbol{x}}) \tag{B.71}$$

where

$$a_{\sigma^2|\boldsymbol{x}} = a + N/2 \tag{B.72}$$

$$b_{\sigma^2|\boldsymbol{x}} = b + \frac{1}{2}\Delta_{\boldsymbol{x}}^2 \tag{B.73}$$

$$\Delta_{\boldsymbol{x}}^2 = (\boldsymbol{x} - \boldsymbol{H}\boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{r})^T(\boldsymbol{I}_N + \boldsymbol{H}\boldsymbol{C}_{\boldsymbol{\theta}}\boldsymbol{H}^T)^{-1}(\boldsymbol{x} - \boldsymbol{H}\boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{r}) \ . \tag{B.74}$$

This concludes the proof for the expression of the marginal posterior distribution $p(\sigma^2|\boldsymbol{x})$.

For the marginal posterior distribution for the mean parameter vector, we have that

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \int p(\boldsymbol{\theta}, \sigma^2|\boldsymbol{x})d\sigma^2 = \int p(\boldsymbol{\theta}|\sigma^2, \boldsymbol{x})p(\sigma^2|\boldsymbol{x})d\sigma^2. \tag{B.75}$$

The expressions for the two distribution inside the integral are known from result B.1 and from the first part of this proof. Inserting the expressions yields

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \int \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}_{\boldsymbol{\theta}|\boldsymbol{x}}, \sigma^2\boldsymbol{C}_{\boldsymbol{\theta}|\boldsymbol{x}})\text{Inv-}\mathcal{G}(\sigma^2; a_{\sigma^2|\boldsymbol{x}}, b_{\sigma^2|\boldsymbol{x}})d\sigma^2 \tag{B.76}$$

$$= \text{St}\left(\boldsymbol{\theta}; \boldsymbol{\mu}_{\boldsymbol{\theta}|\boldsymbol{x}}, \frac{b_{\sigma^2|\boldsymbol{x}}}{a_{\sigma^2|\boldsymbol{x}}}\boldsymbol{C}_{\boldsymbol{\theta}|\boldsymbol{x}}, 2a_{\sigma^2|\boldsymbol{x}}\right) \tag{B.77}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\theta}|\boldsymbol{x}} = \sigma^2\boldsymbol{C}_{\boldsymbol{\theta}|\boldsymbol{x}}$ and the last equality follows from result B.3. This concludes the proof for the expression of the marginal posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{x})$.

The marginal distribution over $\boldsymbol{\theta}$ and σ^2 can be found from

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}, \sigma^2)d\sigma^2 = \int p(\boldsymbol{x}|\sigma^2)p(\sigma^2)d\sigma^2 \ . \tag{B.78}$$

From the first part of this proof, we know $p(\boldsymbol{x}|\sigma^2)$. Thus

$$p(\boldsymbol{x}) = \int \mathcal{N}(\boldsymbol{x}; \boldsymbol{H}\boldsymbol{\mu}_{\boldsymbol{\theta}} + \boldsymbol{r}, \sigma^2(\boldsymbol{I}_N + \boldsymbol{H}\boldsymbol{C}_{\boldsymbol{\theta}}\boldsymbol{H}^T))\text{Inv-}\mathcal{G}(\sigma^2; a, b)d\sigma^2 \tag{B.79}$$

$$= \text{St}\left(\boldsymbol{x}; \boldsymbol{H}\boldsymbol{\mu}_{\boldsymbol{\theta}} + \boldsymbol{r}, \frac{b}{a}(\boldsymbol{I}_N + \boldsymbol{H}\boldsymbol{C}_{\boldsymbol{\theta}}\boldsymbol{H}^T), 2a\right) \tag{B.80}$$

where the last equality follows from result B.3. This concludes the proof. □

Appendix C

The Kalman Filter and Smoother

In this appendix, we derive the Kalman filter and smoother for the linear time-invariant Gaussian state space model. In contrary to many other derivations (see, e.g., [Kay, 1993, ch. 13], [Durbin and Koopman, 2001] and [Harvey, 1989]), we derive the Kalman filter and smoother using the Bayesian inference methods for manipulating probability distributions as described in chapter 2, appendix A and appendix B.

The linear time-invariant Gaussian state space model is given by

$$\begin{aligned} \boldsymbol{y}_n &= \boldsymbol{B}\boldsymbol{s}_n + \boldsymbol{w}_n \qquad \text{(observation equation)} \\ \boldsymbol{s}_{n+1} &= \boldsymbol{A}\boldsymbol{s}_n + \boldsymbol{v}_n \qquad \text{(state equation)} \end{aligned} \tag{C.1}$$

for $n = 1, \cdots, N$ where $\boldsymbol{y}_n$, $\boldsymbol{s}_n$, $\boldsymbol{w}_n$, $\boldsymbol{v}_n$, $\boldsymbol{B}$ and $\boldsymbol{A}$ are the $N \times 1$ observation vector, $M \times 1$ state vector, $N \times 1$ observation noise vector, $M \times 1$ state noise vector, $N \times M$ output matrix and $M \times M$ observation matrix, respectively, at time index n. The observation and state noise are both white Gaussian vector processes with covariance matrices $\boldsymbol{\Sigma}_{\boldsymbol{w}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{v}}$, respectively. The prior distribution for the initial state $\boldsymbol{s}_1$ is also Gaussian distributed with mean vector $\boldsymbol{\mu}_1$ and covariance matrix $\boldsymbol{P}_1$. Thus, we can summarise the state space equation in Eq. (C.1) as the three Gaussian distributions given by

$$p(\boldsymbol{y}_n|\boldsymbol{s}_n) = \mathcal{N}(\boldsymbol{y}_n; \boldsymbol{B}\boldsymbol{s}_n, \boldsymbol{\Sigma}_{\boldsymbol{w}}) \tag{C.2}$$

$$p(\boldsymbol{s}_{n+1}|\boldsymbol{s}_n) = \mathcal{N}(\boldsymbol{s}_{n+1}; \boldsymbol{A}\boldsymbol{s}_n, \boldsymbol{\Sigma}_{\boldsymbol{v}}) \tag{C.3}$$

$$p(\boldsymbol{s}_1) = \mathcal{N}(\boldsymbol{s}_1; \boldsymbol{\mu}_1, \boldsymbol{P}_1) \tag{C.4}$$

If we assume the model parameters $\boldsymbol{B}$, $\boldsymbol{A}$, $\boldsymbol{\Sigma}_{\boldsymbol{w}}$, $\boldsymbol{\Sigma}_{\boldsymbol{v}}$, $\boldsymbol{\mu}_1$ and $\boldsymbol{P}_1$ to be known, the Kalman filter and smoother are used for solving the following two inference problems:

- In a real-time application, we have observed $\boldsymbol{y}_{1:n} = \{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n\}$ at time index n and we wish to compute the posterior distribution for the state at time index n

given these observations, i.e., $p(\boldsymbol{s}_n|\boldsymbol{y}_{1:n})$. This inference problem is referred to as *filtering*.

- In an off-line application, we have observed $\boldsymbol{y}_{1:N} = \{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_N\}$ and we wish to compute the posterior distribution for the state at time index n given these observations, i.e., $p(\boldsymbol{s}_n|\boldsymbol{y}_{1:N})$. This inference problem is referred to as *smoothing*.

C.1 The Kalman Filter

At time index n, the posterior distribution for the state $\boldsymbol{s}_n$ given the observations $\boldsymbol{y}_{1:n}$ is

$$p(\boldsymbol{s}_n|\boldsymbol{y}_{1:n}) = \frac{p(\boldsymbol{y}_n, \boldsymbol{s}_n|\boldsymbol{y}_{1:n-1})}{p(\boldsymbol{y}_n)} = \frac{p(\boldsymbol{y}_n|\boldsymbol{s}_n, \boldsymbol{y}_{1:n-1})p(\boldsymbol{s}_n|\boldsymbol{y}_{1:n-1})}{p(\boldsymbol{y}_n)} \tag{C.5}$$

$$= \frac{p(\boldsymbol{y}_n|\boldsymbol{s}_n)p(\boldsymbol{s}_n|\boldsymbol{y}_{1:n-1})}{p(\boldsymbol{y}_n)} \propto p(\boldsymbol{y}_n|\boldsymbol{s}_n)p(\boldsymbol{s}_n|\boldsymbol{y}_{1:n-1}) \tag{C.6}$$

where the one-step predictive posterior distribution is given by

$$p(\boldsymbol{s}_n|\boldsymbol{y}_{1:n-1}) = \int p(\boldsymbol{s}_n, \boldsymbol{s}_{n-1}|\boldsymbol{y}_{1:n-1})d\boldsymbol{s}_{n-1} \tag{C.7}$$

$$= \int p(\boldsymbol{s}_n|\boldsymbol{s}_{n-1}, \boldsymbol{y}_{1:n-1})p(\boldsymbol{s}_{n-1}|\boldsymbol{y}_{1:n-1})d\boldsymbol{s}_{n-1} \tag{C.8}$$

$$= \int p(\boldsymbol{s}_n|\boldsymbol{s}_{n-1})p(\boldsymbol{s}_{n-1}|\boldsymbol{y}_{1:n-1})d\boldsymbol{s}_{n-1} \ . \tag{C.9}$$

Thus, the posterior distribution $p(\boldsymbol{s}_n|\boldsymbol{y}_{1:n})$ can be calculated recursively by inserting Eq. (C.9) into Eq. (C.6). The recursion is initiated with $p(\boldsymbol{s}_1)$ given by Eq. (C.4). Since this distribution as well as $p(\boldsymbol{y}_n|\boldsymbol{s}_n)$ given by Eq. (C.2) and $p(\boldsymbol{s}_n|\boldsymbol{s}_{n-1})$ given by Eq. (C.3) are all Gaussian distributions, $p(\boldsymbol{s}_n|\boldsymbol{y}_{1:n})$ and $p(\boldsymbol{s}_n|\boldsymbol{y}_{1:n-1})$ are also Gaussian distributions and given by

$$p(\boldsymbol{s}_n|\boldsymbol{y}_{1:n}) = \mathcal{N}(\boldsymbol{s}_n; \boldsymbol{\mu}_{n|1:n}, \boldsymbol{P}_{n|1:n}) \tag{C.10}$$

$$p(\boldsymbol{s}_n|\boldsymbol{y}_{1:n-1}) = \mathcal{N}(\boldsymbol{s}_n; \boldsymbol{\mu}_{n|1:n-1}, \boldsymbol{P}_{n|1:n-1}) \ . \tag{C.11}$$

Now, suppose we have computed the means and covariances of Eq. (C.10) and Eq. (C.11) at time index $n-1$. At time index n, we observe $\boldsymbol{y}_n$ and we wish to find the posterior distribution for $\boldsymbol{s}_n$ given the new observation $\boldsymbol{y}_n$ as well as the old observations $\boldsymbol{y}_{1:n-1}$.

Using Eq. (C.9), we first compute the posterior predictive distribution, i.e.,

$$p(\boldsymbol{s}_n|\boldsymbol{y}_{1:n-1}) = \int p(\boldsymbol{s}_n|\boldsymbol{s}_{n-1})p(\boldsymbol{s}_{n-1}|\boldsymbol{y}_{1:n-1})d\boldsymbol{s}_{n-1} \tag{C.12}$$

$$= \int \mathcal{N}(\boldsymbol{s}_n|\boldsymbol{A}\boldsymbol{s}_{n-1}, \boldsymbol{\Sigma}_{\boldsymbol{v}})\mathcal{N}(\boldsymbol{s}_{n-1}; \boldsymbol{\mu}_{n-1|1:n-1}, \boldsymbol{P}_{n-1|1:n-1})d\boldsymbol{s}_{n-1} \tag{C.13}$$

$$= \mathcal{N}(\boldsymbol{s}_n; \boldsymbol{A}\boldsymbol{\mu}_{n-1|n-1}, \boldsymbol{\Sigma}_{\boldsymbol{v}} + \boldsymbol{A}\boldsymbol{P}_{n-1|1:n-1}\boldsymbol{A}^T) \tag{C.14}$$

$$= \mathcal{N}(\boldsymbol{s}_n; \boldsymbol{\mu}_{n|1:n-1}, \boldsymbol{P}_{n|1:n-1}) \tag{C.15}$$

where the second last equality follows from result B.1. Then, we compute the posterior distribution from Eq. (C.6) which yields

$$p(\boldsymbol{s}_n|\boldsymbol{y}_{1:n}) \propto p(\boldsymbol{y}_n|\boldsymbol{s}_n)p(\boldsymbol{s}_n|\boldsymbol{y}_{1:n-1}) \tag{C.16}$$

$$= \mathcal{N}(\boldsymbol{y}_n|\boldsymbol{B}\boldsymbol{s}_n, \boldsymbol{\Sigma}_{\boldsymbol{w}})\mathcal{N}(\boldsymbol{s}_n; \boldsymbol{\mu}_{n|1:n-1}, \boldsymbol{P}_{n|1:n-1}) \tag{C.17}$$

$$\propto \mathcal{N}\Big(\boldsymbol{s}_n; (\boldsymbol{P}^{-1}_{n|1:n-1} + \boldsymbol{B}^T\boldsymbol{\Sigma}^{-1}_{\boldsymbol{w}}\boldsymbol{B})^{-1}(\boldsymbol{B}^T\boldsymbol{\Sigma}^{-1}_{\boldsymbol{w}}\boldsymbol{y}_m + \boldsymbol{P}^{-1}_{n|1:n-1}\boldsymbol{\mu}_{n|1:n-1}),$$
$$(\boldsymbol{P}^{-1}_{n|1:n-1} + \boldsymbol{B}^T\boldsymbol{\Sigma}^{-1}_{\boldsymbol{w}}\boldsymbol{B})^{-1}\Big) \tag{C.18}$$

$$= \mathcal{N}(\boldsymbol{s}_n; \boldsymbol{\mu}_{n|1:n}, \boldsymbol{P}_{n|1:n}) \tag{C.19}$$

where the last proportional sign follows from result B.1. The update of the mean values and covariances of the posterior and predictive posterior distributions given by Eq. (C.19) and Eq. (C.15), respectively, constitute what is referred to as the Kalman filter equations. They are typically stated in two phases. The *prediction* phase consists of the two equations

$$\boldsymbol{\mu}_{n|1:n-1} = \boldsymbol{A}\boldsymbol{\mu}_{n-1|n-1} \tag{C.20a}$$

$$\boldsymbol{P}_{n|1:n-1} = \boldsymbol{\Sigma}_{\boldsymbol{v}} + \boldsymbol{A}\boldsymbol{P}_{n-1|1:n-1}\boldsymbol{A}^T\ , \tag{C.20b}$$

and the *update* phase consists of the two equations

$$\boldsymbol{P}_{n|1:n} = (\boldsymbol{P}^{-1}_{n|1:n-1} + \boldsymbol{B}^T\boldsymbol{\Sigma}^{-1}_{\boldsymbol{w}}\boldsymbol{B})^{-1} \tag{C.21}$$

$$\boldsymbol{\mu}_{n|1:n} = \boldsymbol{P}_{n|1:n}(\boldsymbol{B}^T\boldsymbol{\Sigma}^{-1}_{\boldsymbol{w}}\boldsymbol{y}_n + \boldsymbol{P}^{-1}_{n|1:n-1}\boldsymbol{\mu}_{n|1:n-1})\ . \tag{C.22}$$

In order to avoid the many matrix inversions in the update phase, it can be rewritten into another form given by [Harvey, 1989, pp. 105-106]

$$\boldsymbol{\mu}_{n|1:n} = \boldsymbol{\mu}_{n|1:n-1} + \boldsymbol{K}_n\boldsymbol{e}_n \tag{C.23a}$$

$$\boldsymbol{P}_{n|1:n} = (\boldsymbol{I}_M - \boldsymbol{K}_n\boldsymbol{B})\boldsymbol{P}_{n|1:n-1} \tag{C.23b}$$

where

$$\boldsymbol{e}_n = \boldsymbol{y}_n - \boldsymbol{B}\boldsymbol{\mu}_{n|1:n-1} \tag{C.23c}$$

$$\boldsymbol{F}_n = \boldsymbol{\Sigma}_{\boldsymbol{w}} + \boldsymbol{B}\boldsymbol{P}_{n|1:n-1}\boldsymbol{B}^T \tag{C.23d}$$

$$\boldsymbol{K}_n = \boldsymbol{P}_{n|1:n-1}\boldsymbol{B}^T\boldsymbol{F}_n^{-1} \tag{C.23e}$$

are referred to as the *innovation*, *innovation covariance* and *Kalman gain*, respectively. The innovation is also sometimes referred to as the prediction error or the residual. Some authors prefer joining the prediction and update phases yielding a set of equations involving only the moments of the posterior distribution or the moments of the one-step predictive posterior distribution. In the latter case, this set of equations is given by [Durbin and Koopman, 2001, p. 67]

$$\begin{aligned} \boldsymbol{e}_n &= \boldsymbol{y}_n - \boldsymbol{B}\boldsymbol{\mu}_{n|1:n-1} & \boldsymbol{F}_n &= \boldsymbol{\Sigma}_{\boldsymbol{w}} + \boldsymbol{B}\boldsymbol{P}_{n|1:n-1}\boldsymbol{B}^T \\ \boldsymbol{K}_n &= \boldsymbol{A}\boldsymbol{P}_{n|1:n-1}\boldsymbol{B}^T\boldsymbol{F}_n^{-1} & \boldsymbol{L}_n &= \boldsymbol{A} - \boldsymbol{K}_n\boldsymbol{B} \\ \boldsymbol{\mu}_{n+1|1:n} &= \boldsymbol{A}(\boldsymbol{\mu}_{n|1:n-1} + \boldsymbol{K}_n\boldsymbol{e}_n) & \boldsymbol{P}_{n+1|1:n} &= \boldsymbol{A}\boldsymbol{P}_{n|1:n-1}\boldsymbol{L}_n^T + \boldsymbol{\Sigma}_{\boldsymbol{v}} \ . \end{aligned} \tag{C.24}$$

C.2 The Kalman Smoother

The Kalman smoother is used for finding the moments of the posterior distribution for state $\boldsymbol{x}_n$ given all observations $\boldsymbol{y}_{1:N}$. This is in contrast to the Kalman filter which only use the current and past observations for this inference task. Of course, the posterior distributions computed by using the Kalman smoother has on average a smaller variance since future observations are used. The price paid for including these future observations is that the Kalman smoother is not feasible for use in real-time applications. In off-line applications, however, it is very useful.

At time index n, the posterior distribution for the state $\boldsymbol{s}_n$ given the observations $\boldsymbol{y}_{1:N}$ is

$$p(\boldsymbol{s}_n|\boldsymbol{y}_{1:N}) = \int p(\boldsymbol{s}_n, \boldsymbol{s}_{n+1}|\boldsymbol{y}_{1:N})d\boldsymbol{s}_{n+1} \tag{C.25}$$

$$= \int p(\boldsymbol{s}_n|\boldsymbol{s}_{n+1}, \boldsymbol{y}_{1:N})p(\boldsymbol{s}_{n+1}|\boldsymbol{y}_{1:N})d\boldsymbol{s}_{n+1} \tag{C.26}$$

$$= \int p(\boldsymbol{s}_n|\boldsymbol{s}_{n+1}, \boldsymbol{y}_{1:n})p(\boldsymbol{s}_{n+1}|\boldsymbol{y}_{1:N})d\boldsymbol{s}_{n+1} \tag{C.27}$$

where the last equality follows from the fact that $\boldsymbol{s}_n$ is independent of $\boldsymbol{y}_{n+1:N}$ given $\boldsymbol{s}_{n+1}$ and $\boldsymbol{y}_{1:n}$[1]. The distribution $p(\boldsymbol{s}_{n+1}|\boldsymbol{y}_{1:N})$ is the posterior distribution for the next

[1] It can sometimes be hard to determine whether two random variables are conditional independent. A systematic way of determining this is to use graphical models. For an introduction to graphical models see [Bishop, 2006, ch. 8].

smoothed state distribution. This suggest a recursive computation of $p(\boldsymbol{s}_n|\boldsymbol{y}_{1:N})$ in the reverse time direction. The distribution $p(\boldsymbol{s}_n|\boldsymbol{s}_{n+1}, \boldsymbol{y}_{1:n})$ can by using Bayes' theorem be written as

$$p(\boldsymbol{s}_n|\boldsymbol{s}_{n+1}, \boldsymbol{y}_{1:n}) = \frac{p(\boldsymbol{s}_{n+1}|\boldsymbol{s}_n, \boldsymbol{y}_{1:n})p(\boldsymbol{s}_n|\boldsymbol{y}_{1:n})}{p(\boldsymbol{s}_{n+1}|\boldsymbol{y}_{1:n})} = \frac{p(\boldsymbol{s}_{n+1}|\boldsymbol{s}_n)p(\boldsymbol{s}_n|\boldsymbol{y}_{1:n})}{p(\boldsymbol{s}_{n+1}|\boldsymbol{y}_{1:n})} \tag{C.28}$$

where $p(\boldsymbol{s}_{n+1}|\boldsymbol{s}_n)$ is given by Eq. (C.3), $p(\boldsymbol{s}_n|\boldsymbol{y}_{1:n})$ is the posterior distribution computed by the Kalman filter and given by Eq. (C.19), and $p(\boldsymbol{s}_{n+1}|\boldsymbol{y}_{1:n})$ is the predictive posterior distribution computed by the Kalman filter and given by Eq. (C.15). Since all of these distributions are Gaussian, $p(\boldsymbol{s}_n|\boldsymbol{s}_{n+1}, \boldsymbol{y}_{1:n})$ is also Gaussian and given by

$$p(\boldsymbol{s}_n|\boldsymbol{s}_{n+1}, \boldsymbol{y}_{1:n}) = \frac{\mathcal{N}(\boldsymbol{s}_{n+1}; \boldsymbol{A}\boldsymbol{s}_n, \boldsymbol{\Sigma}_{\boldsymbol{v}})\mathcal{N}(\boldsymbol{s}_n; \boldsymbol{\mu}_{n|1:n}, \boldsymbol{P}_{n|1:n})}{\mathcal{N}(\boldsymbol{s}_{n+1}; \boldsymbol{\mu}_{n+1|1:n}, \boldsymbol{P}_{n+1|1:n})} \tag{C.29}$$

$$= \mathcal{N}\Big(\boldsymbol{s}_n; (\boldsymbol{P}^{-1}_{n|1:n} + \boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}_{\boldsymbol{v}}\boldsymbol{A})^{-1}(\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}_{\boldsymbol{v}}\boldsymbol{s}_{n+1} + \boldsymbol{P}^{-1}_{n|1:n}\boldsymbol{\mu}_{n|1:n}), (\boldsymbol{P}^{-1}_{n|1:n} + \boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}_{\boldsymbol{v}}\boldsymbol{A})^{-1}\Big) \tag{C.30}$$

$$= \mathcal{N}\Big(\boldsymbol{s}_n; \boldsymbol{\mu}_{n|1:n} + \boldsymbol{\Gamma}_n(\boldsymbol{s}_{n+1} - \boldsymbol{\mu}_{n+1|1:n}), \boldsymbol{P}_{n|1:n} - \boldsymbol{\Gamma}_n\boldsymbol{P}_{n+1|1:n}\boldsymbol{\Gamma}^T_n\Big) \tag{C.31}$$

where the second equality follows from result B.1 and $\boldsymbol{\Gamma}_n \triangleq \boldsymbol{P}_{n|1:n}\boldsymbol{A}^T\boldsymbol{P}^{-1}_{n+1|1:n}$. Inserting this into Eq. (C.27) and performing the marginalisation using result B.1 yield the posterior distribution for the smoothed state as

$$p(\boldsymbol{s}_n|\boldsymbol{y}_{1:N}) = \mathcal{N}(\boldsymbol{s}_n; \boldsymbol{\mu}_{n|1:N}, \boldsymbol{P}_{n|1:N}) \tag{C.32}$$

where

$$\boldsymbol{\mu}_{n|1:N} = \boldsymbol{\mu}_{n|1:n} + \boldsymbol{\Gamma}_n(\boldsymbol{\mu}_{n+1|1:N} - \boldsymbol{\mu}_{n+1|1:n}) \tag{C.33}$$

$$\boldsymbol{P}_{n|1:N} = \boldsymbol{P}_{n|1:n} + \boldsymbol{\Gamma}_n(\boldsymbol{P}_{n+1|1:N} - \boldsymbol{P}_{n+1|1:n})\boldsymbol{\Gamma}^T_n \ . \tag{C.34}$$

In order to compute the moments of $p(\boldsymbol{s}_n|\boldsymbol{y}_{1:N})$, we must know the moments of $p(\boldsymbol{s}_n|\boldsymbol{y}_{1:n})$ and $p(\boldsymbol{s}_{n+1}|\boldsymbol{y}_{1:n})$, which can be computed using the Kalman filter, and the moments of $p(\boldsymbol{s}_{n+1}|\boldsymbol{y}_{1:N})$. This suggest that the moments of $p(\boldsymbol{s}_n|\boldsymbol{y}_{1:N})$ can be computed by, first, running the Kalman filter for $n' = 1, \cdots, N$ and, second, running the Kalman smoother for $n' = N, \cdots, n$.

For the case where only the moments of the one-step predictive posterior distribution are computed by using the Kalman filter given by Eq. (C.24), the smoothed moments are calculated by the recursion [Durbin and Koopman, 2001, p. 73]

$$\begin{aligned} \boldsymbol{r}_{n-1} &= \boldsymbol{B}^T\boldsymbol{F}^{-1}_n\boldsymbol{e}_n + \boldsymbol{L}^T_n\boldsymbol{r}_n & \boldsymbol{H}_{n-1} &= \boldsymbol{B}^T\boldsymbol{F}^{-1}_n\boldsymbol{B} + \boldsymbol{L}^T_n\boldsymbol{H}_n\boldsymbol{L}_n \\ \boldsymbol{\mu}_{n|1:N} &= \boldsymbol{\mu}_{n|1:n-1} + \boldsymbol{P}_{n|1:n-1}\boldsymbol{r}_{n-1} & \boldsymbol{P}_{n|1:N} &= \boldsymbol{P}_{n|1:n-1}(\boldsymbol{I}_M - \boldsymbol{H}_{n-1}\boldsymbol{P}_{n|1:n-1}) \end{aligned} \tag{C.35}$$

where $\boldsymbol{r}_N = \boldsymbol{0}$ and $\boldsymbol{H}_N = \boldsymbol{0}$.

Appendix D

WASPAA 2009 Paper

In this appendix, we have included a publication written in connection with this project. The included paper was submitted for the Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) 2009 on April 15, 2009. At the time of writing, the paper is still in the reviewing process so we have not been notified about whether it has been accepted or not.

BAYESIAN SINUSOIDAL PARAMETER ESTIMATION AND INTERPOLATION USING A STATE SPACE APPROACH

Jesper Kjær Nielsen†, Ali Taylan Cemgil††, Simon J. Godsill†††, Mads Græsbøll Christensen†, Søren Holdt Jensen†, and Torben Larsen†

†Aalborg University
Department of Electronic Systems
Niels Jernes Vej 12, DK-9220 Aalborg
{jkjaer,mgc,shj,tl}@es.aau.dk

††Boğaziçi University
Department of Computer Engineering
34342 Bebek Istanbul, TR
taylan.cemgil@boun.edu.tr

†††University of Cambridge
Department of Engineering
Trumpington St., Cambridge, CB2 1PZ, UK
sjg@eng.cam.ac.uk

ABSTRACT

In this paper, we consider Bayesian estimation and interpolation in the dynamic sinusoidal model. This model is more flexible than the static sinusoidal model since it enables the amplitude and phase of the sinusoids to vary as a function of time. Based on a Gibbs sampler, we derive a Bayesian inference scheme for the frequencies, the state and observation noise variances as well as for the case of missing observations. The problem of obtaining samples for the frequency parameters is given particular attention and we show that it can be solved in a simple and efficient way by sampling directly from the von Mises distribution. Finally, we demonstrate the application of the proposed method to analysis of synthetically generated signals consisting of multiple sinusoids.

Index Terms— Sinusoidal signal model, Bayesian signal processing, Gibbs sampler, state space modelling.

1. INTRODUCTION

The problem of estimating sinusoidal model parameters is an integral part of many audio signal processing applications such as compression, signal enhancement and restoration, music transcription and genre classification. In these applications, the problem is to make point estimates of the sinusoidal parameters or predict unobserved observations based on a set of observations. We consider initially the real static sinusoidal model given by

$$x_n = \sum_{l=1}^{L} \alpha_l e^{-\gamma_l n} \alpha_l \cos(\omega_l n + \varphi_l) + w_n \,, \tag{1}$$

for $n = 1, \cdots, N$, where $\alpha_l > 0$, $\varphi_l \in [-\pi, \pi]$, $\omega_l \in [0, \pi]$ and $\gamma_l > 0$ are the amplitude, phase, (angular) frequency and log-damping coefficient of the l'th sinusoid, respectively. The observed signal x_n at time index n is the sum of L such sinusoids and a white Gaussian noise term w_n with variance σ_w^2. The amplitudes and phases of the static model are assumed to be constant over the set of observations. We later relax this assumption by introducing the dynamic model. The problem of estimating the frequency parameters and log-damping coefficients is complicated by the fact that they enter the model in a non-linear fashion. Numerous approaches have been suggested in the scientific literature to overcome this difficulty (see e.g. [1]) and the problem is still an ongoing research area. Most of the suggested estimators are based on frequentist statistics whereas only a few are based on Bayesian statistics, despite the conceptual advantages of this approach. The primary reason for this is that Bayesian methods struggle with practical problems such as evaluation of high dimensional and intractable integrals which arise frequently in the Bayesian framework. In recent years, however, many computational algorithms such as Markov chain Monte Carlo (MCMC) sampling have been embraced and developed by the Bayesian community. Although these suffer from a high computational complexity, they have to a large extent overcome many of the practical problems and led to various developments in Bayesian frequency estimation (see e.g. [2], [3], [4] and references therein).

In this paper, we extend this work by proposing an inference scheme for the parameters of the sinusoidal model in which the amplitude and phase are allowed to vary stochastically as a function of time. This is obtained by firstly, rewriting the static sinusoidal model (1) as a dynamic model, the linear time-invariant state space model, as in [5, 6] and secondly, by performing Bayesian inference in this model using the Gibbs sampler. The detailed formulation is considered in Sec. 2 and the derivation of the inference scheme is given in Sec. 3. The inference scheme is evaluated in Sec. 4 and Sec. 5 concludes this paper.

2. SINUSOIDAL STATE SPACE MODEL

Using complex notation, we rewrite (1) as

$$x_n = \sum_{l=1}^{L} (c_l z_l^n + c_l^* z_l^{*n}) + w_n \tag{2}$$

where $(\cdot)^*$ denotes complex conjugation, $c_l = (\alpha_l/2)e^{j\varphi_l}$ and $z_l = e^{-\gamma_l} e^{j\omega_l}$. In matrix notation, we write this as

$$x_n = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix}^T \begin{bmatrix} z_1^n & 0 & \cdots & 0 & 0 \\ 0 & z_1^{*n} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & z_L^n & 0 \\ 0 & 0 & \cdots & 0 & z_L^{*n} \end{bmatrix} \begin{bmatrix} c_1 \\ c_1^* \\ \vdots \\ c_L \\ c_L^* \end{bmatrix} + w_n \tag{3}$$

$$\triangleq \tilde{\boldsymbol{b}}^T \tilde{\boldsymbol{A}}^n \tilde{\boldsymbol{c}} + w_n \tag{4}$$

The work of J.K. Nielsen was supported by the Oticon Foundation's Scholarship

The work of M.G. Christensen was supported by the Parametric Audio Processing project, Danish Research Council for Technology and Production Sciences grant no. 274–06–0521

The work of S.H. Jensen was partly supported by the Danish Technical Research Council, through the framework project Intelligent Sound, www.intelligentsound.org (STVF No. 26-04-0092)

where $(\cdot)^T$ denotes the transpose, $\tilde{\boldsymbol{b}}$ and $\tilde{\boldsymbol{c}}$ are $2L \times 1$ column vectors, and $\tilde{\boldsymbol{A}}$ is a $2L \times 2L$ diagonal matrix. This formulation constrains the log-damping coefficients and the frequency parameters to be fully determined by the time-invariant complex diagonal matrix $\tilde{\boldsymbol{A}}$. The matrix $\tilde{\boldsymbol{A}}$ and vector $\tilde{\boldsymbol{c}}$ are in general complex valued. In order to avoid the complex terms, we define the Hermitian complex block diagonal matrix $\boldsymbol{T} = \text{diag}(\boldsymbol{T}_1, \cdots, \boldsymbol{T}_l, \cdots \boldsymbol{T}_L)$ with

$$\boldsymbol{T}_l = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ j & -j \end{bmatrix} \tag{5}$$

where j is the imaginary unit. By use of $\boldsymbol{T}$, we now obtain

$$x_n = \tilde{\boldsymbol{b}}^T \left(\boldsymbol{T}^{-1} \boldsymbol{T} \tilde{\boldsymbol{A}} \boldsymbol{T}^{-1} \boldsymbol{T} \right)^n \tilde{\boldsymbol{c}} + w_n \triangleq \boldsymbol{b}^T \boldsymbol{A}^n \boldsymbol{c} + w_n \tag{6}$$

where $\boldsymbol{b}$, $\boldsymbol{A}$ and $\boldsymbol{c}$ are all real and given by

$$\boldsymbol{b} = \left(\tilde{\boldsymbol{b}}^T \boldsymbol{T}^{-1} \right)^T = \sqrt{2} \begin{bmatrix} 1 & 0 & \cdots & 1 & 0 \end{bmatrix}^T \tag{7}$$

$$\boldsymbol{A} = \boldsymbol{T} \tilde{\boldsymbol{A}} \boldsymbol{T}^{-1} = \text{diag}(\boldsymbol{A}_1, \cdots, \boldsymbol{A}_l, \cdots \boldsymbol{A}_L) \tag{8}$$

$$\boldsymbol{A}_l = e^{-\gamma_l} \begin{bmatrix} \cos\omega_l & \sin\omega_l \\ -\sin\omega_l & \cos\omega_l \end{bmatrix} \tag{9}$$

$$\boldsymbol{c} = \boldsymbol{T} \tilde{\boldsymbol{c}} = \sqrt{2} \left[\alpha_1 \begin{bmatrix} \cos\varphi_1 \\ \sin\varphi_1 \end{bmatrix}^T \cdots \quad \alpha_L \begin{bmatrix} \cos\varphi_L \\ \sin\varphi_L \end{bmatrix}^T \right]^T. \tag{10}$$

We are now able to write the sinusoidal model in (1) as a linear Gaussian time-invariant state space model. The state space formulation can be obtained from (6) as

$$\begin{aligned} y_n &= \boldsymbol{b}^T \boldsymbol{s}_n + w_n \qquad \text{(observation equation)} \\ \boldsymbol{s}_{n+1} &= \boldsymbol{A} \boldsymbol{s}_n + \boldsymbol{v}_n \qquad \text{(state equation)} \end{aligned} \tag{11}$$

for $n = 1, \cdots, N$, if we introduce white Gaussian state noise $\boldsymbol{v}_n$ with covariance matrix $\boldsymbol{Q}$ in the state equation and assume a Gaussian prior for the initial state vector $\boldsymbol{s}_1$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{P}$. The model in (11) is slightly different (hence the use of y_n instead of x_n for the observations) than the original model in (1), since we have introduced state noise with non-zero covariance matrix in the state equation. This allows the amplitude and phase to be a time-varying AR(1) process. In the case of initial state vector given by $\boldsymbol{s}_1 = \boldsymbol{A}\boldsymbol{c}$ and zero state-noise, the two models are identical, i.e. $y_n = x_n$.

3. BAYESIAN INFERENCE

In this section, we consider Bayesian inference in the state model model derived in the previous section and given by (11). We assume the log-damping coefficients γ_l to be 0 (see [7] for unknown γ_l), the model order L to be known and the state noise covariance matrix to be diagonal, i.e. $\boldsymbol{Q} = \text{diag}(q_1, q_1, \cdots, q_L, q_L)$. The latter assumption is made in order to keep the number of unknown parameters at a reasonable level. Thus, the unknown parameters of the model are the N state vectors $\boldsymbol{s}_{1:N}$, the L frequency parameters in $\boldsymbol{\omega}$ constituting the block diagonal system matrix $\boldsymbol{A}$, the L diagonal elements of the state noise covariance matrix $q_1, \cdots, q_L$ and the observation noise variance σ_w^2. In the Bayesian framework, all statistical inference is based on the joint posterior distribution over the unknown variables or a marginal posterior distribution over some of these. For the model in (11), the joint posterior distribution is $p(\boldsymbol{s}_{1:N}, \boldsymbol{\omega}, \boldsymbol{Q}, \sigma_w^2 | y_{1:N})$ from which point estimates cannot be computed in closed form. We therefore have to resort to numerical techniques in order to enable statistical inference based on this distribution. One of the simplest and most popular numerical techniques is the Gibbs sampler [8] which is an MCMC-based algorithm and suitable for this task. The Gibbs sampler draws samples from the joint posterior distribution over the unknown variables by breaking it into a number of conditional distributions of smaller dimensionality from which samples are obtained in an alternating pattern. After an initial burn-in time during which the sampling scheme converges, the samples obtained from sampling these lower dimensional conditional distributions can be regarded as samples from the joint posterior distribution. Applying the Gibbs sampler on the joint posterior distribution $p(\boldsymbol{s}_{1:N}, \boldsymbol{\omega}, \boldsymbol{Q}, \sigma_w^2 | y_{1:N})$ yields the set of conditional distributions given by

$$\begin{aligned} &\text{State:} && p(\boldsymbol{s}_{1:N} | \boldsymbol{\omega}, \boldsymbol{Q}, \sigma_w^2, y_{1:N}) && (12) \\ &\text{Frequency:} && p(\omega_l | \boldsymbol{s}_{1:N}, \boldsymbol{\omega}_{\setminus l}, \boldsymbol{Q}, \sigma_w^2, y_{1:N}) && (13) \\ &\text{State variance:} && p(q_l | \boldsymbol{s}_{1:N}, \boldsymbol{\omega}, \boldsymbol{Q}_{\setminus l}, \sigma_w^2, y_{1:N}) && (14) \\ &\text{Observation variance:} && p(\sigma_w^2 | \boldsymbol{s}_{1:N}, \boldsymbol{\omega}, \boldsymbol{Q}, y_{1:N}) && (15) \end{aligned}$$

where (13) and (14) are evaluated for $l = 1, \cdots, L$, and $(\cdot)_{\setminus l}$ denotes 'without element l'. In the following four sections, these distributions are derived from (11) and by introducing prior distributions over the unknown parameters.

3.1. Conditional Distribution for States

We can write the state space model in (11) as

$$p(\boldsymbol{s}_1) = \mathcal{N}(\boldsymbol{s}_1; \boldsymbol{\mu}, \boldsymbol{P}) \tag{16}$$

$$p(y_n | \boldsymbol{s}_n, \sigma_w^2) = \mathcal{N}(y_n; \boldsymbol{b}^T \boldsymbol{s}_n, \sigma_w^2) \tag{17}$$

$$p(\boldsymbol{s}_{n+1} | \boldsymbol{s}_n, \boldsymbol{\omega}, \boldsymbol{Q}) = \mathcal{N}(\boldsymbol{s}_{n+1}; \boldsymbol{A}\boldsymbol{s}_n, \boldsymbol{Q}) \tag{18}$$

from which the conditional state distribution in (12) can be shown to be a multivariate Gaussian distribution [7]. However, the dimensionality of this distribution is $2LN \times 1$ which would render direct sampling from it infeasible for most applications. Instead, we have used the the simulation smoother [9], which is an efficient sampling scheme using standard Kalman smoothing, for drawing samples from (12).

3.2. Conditional Distribution for Frequencies

By use of conditional independence and Bayes' Theorem, we factor (13) as

$$\begin{aligned} p(\omega_l | \boldsymbol{s}_{1:N}, \boldsymbol{\omega}_{\setminus l}, \boldsymbol{Q}, \sigma_w^2, y_{1:N}) &= p(\omega_l | \boldsymbol{s}_{1:N}, \boldsymbol{\omega}_{\setminus l}, \boldsymbol{Q}) && (19) \\ &\propto p(\boldsymbol{s}_{1:N} | \boldsymbol{\omega}, \boldsymbol{Q}) p(\omega_l) && (20) \end{aligned}$$

where $p(\omega_l)$ is the prior distribution for ω_l yet to be defined. Since $\boldsymbol{A}$ is 2×2 block diagonal and $\boldsymbol{Q}$ is diagonal, we can factor the likelihood of the state equation into L bivariate likelihoods with the l'th given by

$$\begin{aligned} &p(\boldsymbol{s}_{1:N,l} | \omega_l, q_l) = (2\pi q_l)^{-1} \\ &\times \exp\left\{ \frac{-1}{2q_l} \sum_{n=1}^{N-1} (\boldsymbol{s}_{n+1,l} - \boldsymbol{A}_l \boldsymbol{s}_{n,l})^T (\boldsymbol{s}_{n+1,l} - \boldsymbol{A}_l \boldsymbol{s}_{n,l}) \right\} \end{aligned} \tag{21}$$

where $\boldsymbol{A}_l$ is given by (9) and $\boldsymbol{s}_{(\cdot),l}$ is the l'th 2×1 subvector of $\boldsymbol{s}_{(\cdot)}$. Rearranging terms of this equation yields after some algebra

$$p(\boldsymbol{s}_{1:N,l} | \omega_l, q_l) = Z_l^{-1} \exp\{ d_{1,l} \cos\omega_l + d_{2,l} \sin\omega_l \} \tag{22}$$

where Z_l is a normalisation factor independent of ω_l and

$$\begin{bmatrix} d_{1,l} \\ d_{2,l} \end{bmatrix} = \frac{1}{q_l} \sum_{n=1}^{N-1} \begin{bmatrix} \boldsymbol{s}_{n,l} & , & \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \boldsymbol{s}_{n,l} \end{bmatrix} \boldsymbol{s}_{n+1,l} \,. \quad (23)$$

The exponent in (22) is a superposition of two sinusoids so the likelihood has a parametric form proportional to the von Mises distribution, i.e.

$$p(\boldsymbol{s}_{1:N,l}|\omega_l, q_l) = Z_l^{-1} \exp\{\kappa_l \cos(\theta_l - \omega_l)\} \quad (24)$$

where $\kappa_l = \sqrt{d_{1,l}^2 + d_{2,l}^2}$ and $\theta_l = \arctan(d_{2,l}/d_{1,l})$. The conjugate prior for this likelihood is also a von Mises distribution [10], $\mathcal{VM}(\omega_l; \kappa_0, \mu_0)$, which equals the uniform distribution on the interval $[-\pi, \pi]$ in the limit of $\kappa_0 \to 0$. Thus, (19) can be written as

$$p(\omega_l|\boldsymbol{s}_{1:N}, q_l) \propto \mathcal{VM}(\theta_l; \kappa_l, \omega_l)\mathcal{VM}(\omega_l; \kappa_0, \mu_0) \quad (25)$$

$$\propto \mathcal{VM}\left(\omega_l; \sqrt{\delta_{1,l}^2 + \delta_{2,l}^2}, \arctan(\delta_{2,l}/\delta_{1,l})\right) \quad (26)$$

where $\delta_{1,l} = d_{1,l} + \kappa_0 \cos\mu_0$ and $\delta_{2,l} = d_{2,l} + \kappa_0 \sin\mu_0$. Samples can be drawn efficiently from this distribution by use of the Best-Fisher Algorithm [11].

3.3. Conditional Distribution for State Variances

Since $\boldsymbol{Q}$ is diagonal, we can write (14) as

$$p(q_l|\boldsymbol{s}_{1:N}, \boldsymbol{\omega}, \boldsymbol{Q}_{\setminus l}, \sigma_w^2, y_{1:N}) = p(q_l|\boldsymbol{s}_{1:N,l}, \omega_l) \quad (27)$$

$$\propto p(\boldsymbol{s}_{1:N,l}|\omega_l, q_l)p(q_l) \quad (28)$$

where $p(\boldsymbol{s}_{1:N,l}|\omega_l, q_l)$ is given by (21) and $p(q_l)$ is the prior distribution for q_l. We select the conjugate inverse gamma prior, $\mathcal{G}^{-1}(q_l; \alpha_0, \beta_0)$. Thus, (14) is also an inverse gamma distribution, $\mathcal{G}^{-1}(q_l; \alpha_{q_l}, \beta_{q_l})$, with parameters

$$\alpha_{q_l} = \alpha_0 + N - 1 \quad (29)$$

$$\beta_{q_l} = \beta_0 + \frac{1}{2} \sum_{n=1}^{N-1} (\boldsymbol{s}_{n+1,l} - \boldsymbol{A}_l \boldsymbol{s}_{n,l})^T (\boldsymbol{s}_{n+1,l} - \boldsymbol{A}_l \boldsymbol{s}_{n,l}) \quad (30)$$

In the case of an isotropic state noise covariance, i.e. $\boldsymbol{Q} = q\boldsymbol{I}$, the posterior distribution in also an inverse gamma distribution, $\mathcal{G}^{-1}(q; \alpha_q, \beta_q)$, with parameters

$$\alpha_q = \alpha_0 + (N-1)L \quad (31)$$

$$\beta_q = \beta_0 + \frac{1}{2} \sum_{n=1}^{N-1} (\boldsymbol{s}_{n+1} - \boldsymbol{A}\boldsymbol{s}_n)^T (\boldsymbol{s}_{n+1} - \boldsymbol{A}\boldsymbol{s}_n) \,. \quad (32)$$

3.4. Conditional Distribution for Observation Variance

We can write (15) as

$$p(\sigma_w^2|\boldsymbol{s}_{1:N}, \boldsymbol{\omega}, \boldsymbol{Q}, y_{1:N}) = p(\sigma_w^2|\boldsymbol{s}_{1:N,l}, y_{1:N}) \quad (33)$$

$$\propto p(y_{1:N}|\boldsymbol{s}_{1:N,l}, \sigma_w^2)p(\sigma_w^2) \quad (34)$$

where $p(y_{1:N}|\boldsymbol{s}_{1:N,l}, \sigma_w^2)$ is the likelihood of the observation equation in (17) and $p(\sigma_w^2)$ is the prior distribution for σ_w^2. We select the conjugate inverse gamma prior, $\mathcal{G}^{-1}(\sigma_w^2; \alpha_0, \beta_0)$. Thus, (15) is also an inverse gamma prior, $\mathcal{G}^{-1}(\sigma_w^2; \alpha_{\sigma_w^2}, \beta_{\sigma_w^2})$, with parameters

$$\alpha_{\sigma_w^2} = \alpha_0 + N/2 \quad (35)$$

$$\beta_{\sigma_w^2} = \beta_0 + \frac{1}{2} \sum_{n=1}^{N} (\boldsymbol{y}_n - \boldsymbol{b}^T \boldsymbol{s}_n)^T (\boldsymbol{y}_n - \boldsymbol{b}^T \boldsymbol{s}_n) \,. \quad (36)$$

The Gibbs sampling scheme is summarised in algorithm 1. Compared to the inference scheme in [3], our algorithm does not require any tuning of user defined proposal distributions and user defined parameters.

Algorithm 1
Given $\boldsymbol{\omega}^{[0]}$, $\boldsymbol{Q}^{[0]}$ and $\sigma_w^{2\,[0]}$ do for $i = 1, \cdots, M$

1) Draw $\boldsymbol{s}_{1:N}^{[i]}$ from (16)-(18) by use of the simulation smoother.
2) For $l = 1, \cdots, L$ do
   a) $\omega_l^{[i]} \sim \mathcal{VM}\left(\sqrt{\delta_{1,l}^{2\,[i-1]} + \delta_{2,l}^{2\,[i-1]}}, \arctan\frac{\delta_{2,l}^{[i-1]}}{\delta_{1,l}^{[i-1]}}\right)$
   b) $q_l^{[i]} \sim \mathcal{G}^{-1}(\alpha_{q_l}^{[i-1]}, \beta_{q_l}^{[i-1]})$
3) $\sigma_w^{2\,[i]} \sim \mathcal{G}^{-1}(\alpha_{\sigma_w^2}^{[i-1]}, \beta_{\sigma_w^2}^{[i-1]})$

3.5. Interpolation of Missing Samples

In the case where some observations are missing, algorithm 1 can easily be extended to handle this interpolation task. First, partition the observations into $y_{1:N} = \{y_{1:K}, y_{K+1:K+R}, y_{K+R+1:N}\}$ where $\boldsymbol{z} = y_{K+1:K+R}$ are the R missing observations. Then, we simply add a fourth stage to algorithm 1 in which we draw a sample from

$$p(\boldsymbol{z}|\boldsymbol{s}_{1:N}, \boldsymbol{\omega}, \boldsymbol{Q}, y_{1:K}, y_{K+R+1:N}, \sigma_w^2) = p(\boldsymbol{z}|\boldsymbol{s}_{K+1:K+R}, \sigma_w^2)$$

which can be factored into R univariate Gaussian distributions given by (17). The first three steps of the algorithm remain the same.

4. SIMULATIONS

We demonstrate the applicability of the inference scheme on a small-scale example. In the simulations, we generated $N = 512$ observations from the model in (11) with a model order of $L = 3$. The amplitude, phase and (angular) frequency of the sinusoids were $\boldsymbol{\alpha} = [0.8, 0.5, 1.1]$, $\boldsymbol{\varphi} = [0, \pi/2, 0.2]$ and $\boldsymbol{\omega} = [0.3, 0.4, 0.5]$, respectively. The state noise covariance was selected to be isotropic with $q = 0.1$ and the observation noise variance was set to $\sigma_w^2 = 0.5$. The hyperparameters of the prior distribution for the unknown parameters were selected such that these distributions were diffuse. The initial values for the frequency and the observation noise variance were computed by use of the ESPRIT estimator. The initial value for the state noise covariance was somewhat heuristically set to $q = \sigma_w^{2\,[0]}/10$. The observations from time index 50 to 100 were removed and considered as missing. The Gibbs sampler in algorithm 1 was iterated $M = 10000$ times and the burn-in time was set to 100.

Figure 1 shows the result of the simulation. Plots (a)-(c) show the traces of samples obtained for the unknown parameters. It is easy to make meaningful inference about the unknown parameters based on the histograms in the right margin of the plots.

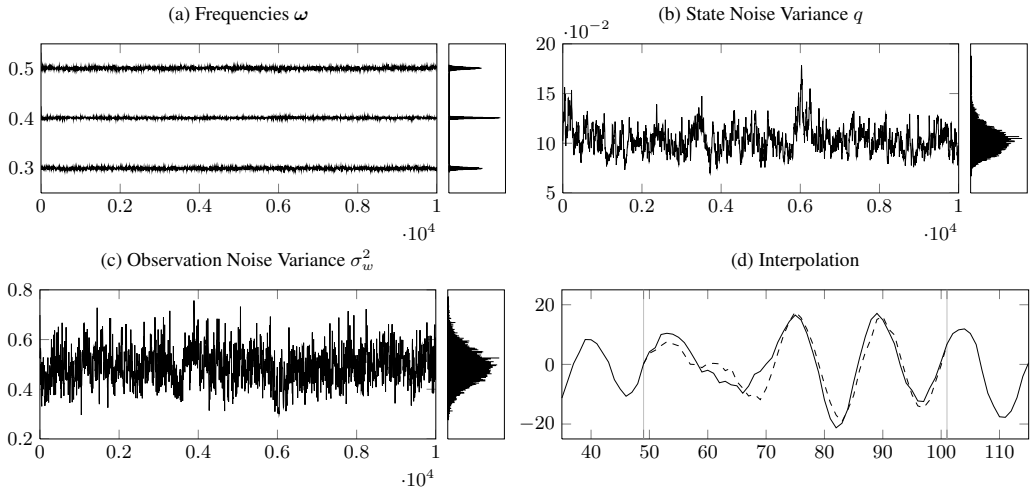

Figure 1: Traces of samples from the Gibbs sampler for the frequencies (a), state noise variance (b) and observation noise variance (c). These plots also show the histogram based on all $M = 10000$ samples with burn-in samples removed. Figure (d) shows the result of the interpolation (solid line) compared to the true signal (dotted line).

Plot (d) shows a typical sample for the missing observations compared to the true signal. The interpolation section is marked by vertical lines. Notice, that unlike maximum likelihood- and EM-restoration techniques, the noise is also modelled when performing the interpolation in the Bayesian framework. This yields a more typical interpolant [12].

5. CONCLUSION

Based on Gibbs sampler, we have presented a Bayesian inference scheme in which the frequency parameters of the dynamical sinusoidal model were sampled in a simple and efficient way from the von Mises distribution. By considering the more flexible an realistic dynamic sinusoidal model in (11), we have enabled the use of Gibbs sampler which, except for the initial values, does not require any user interaction or tuning. In the simulations, we demonstrated that the algorithm can be used for parameter estimation and interpolation. This is a vital part of many audio applications [12]. It is interesting to note that it is not possible to use the Gibbs sampler for the static sinusoidal model in (1) without using analytical approximations. For the static case, the Bayesian inference is usually based on sampling schemes involving user defined proposal distributions and parameters which require careful tuning [3].

6. REFERENCES

[1] P. Stoica and R. L. Moses, *Spectral Analysis of Signals*. Prentice Hall, 2005.

[2] G. L. Bretthorst, *Bayesian Spectrum Analysis and Parameter Estimation*. Springer-Verlag Berlin Heidelberg, 1988.

[3] C. Andrieu and A. Doucet, "Joint Bayesian model selection and estimation of noisy sinusoids via reversible jump MCMC," *IEEE Trans. Signal Process.*, vol. 47, no. 10, pp. 2667–2676, 1999.

[4] M. Davy, S. J. Godsill, and J. Idier, "Bayesian analysis of polyphonic western tonal music," *J. Acoust. Soc. Am.*, vol. 119, pp. 2498–2517, 2006.

[5] A. T. Cemgil and S. Godsill, "Efficient variational inference for the dynamic harmonic model," *Proc. IEEE Workshop on Appl. of Signal Process. to Aud. and Acoust.*, pp. 271–274, Oct. 2005.

[6] A. T. Cemgil, H. J. Kappen, and D. Barber, "A generative model for music transcription," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 2, pp. 679–694, 2006.

[7] J. K. Nielsen, "Sinusoidal parameter estimation - a Bayesian approach," Master Thesis, Aalborg University, June 2009.

[8] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 6, no. 6, pp. 721–741, November 1984.

[9] J. Durbin and S. Koopman, "A simple and efficient simulation smoother for state space time series analysis," *Biometrika*, vol. 89, no. 3, pp. 603–615, 2002.

[10] P. Guttorp and R. Lockhart, "Finding the location of a signal: A Bayesian analysis," *J. Amer. Statistical Assoc.*, vol. 83, pp. pp. 322–330, 1988.

[11] D. Best and N. Fisher, "Efficient simulation of the von Mises distribution," *J. Appl. Statist.*, vol. 28, pp. 152–157, 1979.

[12] S. J. Godsill and P. J. W. Rayner, *Digital Audio Restoration*. Springer-Verlag London, 1998.