

# Signal Separation of Musical Instruments

Simulation-based methods for musical signal decomposition and transcription

---

*A thesis submitted to the University of Cambridge  
for the degree of Doctor of Philosophy*

PAUL JOSEPH WALMSLEY

PEMBROKE COLLEGE

September 2000

---



SIGNAL PROCESSING GROUP  
DEPARTMENT OF ENGINEERING  
UNIVERSITY OF CAMBRIDGE

*To my family*

# *Declaration*

---

---

The research described in this dissertation was carried out by the author between October 1996 and September 2000. This dissertation is the result of my own work and includes nothing that is the outcome of work done in collaboration. No part of this dissertation has been submitted to any other university. This dissertation contains not more than 50000 words.

Paul J. Walmsley



# *Abstract*

---

This thesis presents techniques for the modelling of musical signals, with particular regard to monophonic and polyphonic pitch estimation. Musical signals are modelled as a set of notes, each comprising of a set of harmonically-related sinusoids. An hierarchical model is presented that is very general and applicable to any signal that can be decomposed as the sum of basis functions. Parameter estimation is posed within a Bayesian framework, allowing for the incorporation of prior information about model parameters. The resulting posterior distribution is of variable dimension and so reversible jump MCMC simulation techniques are employed for the parameter estimation task. The extension of the model to time-varying signals with high posterior correlations between model parameters is described. The parameters and hyperparameters of several frames of data are estimated jointly to achieve a more robust detection. A general model for the description of time-varying homogeneous and heterogeneous multiple component signals is developed, and then applied to the analysis of musical signals.

The importance of high level musical and perceptual psychological knowledge in the formulation of the model is highlighted, and attention is drawn to the limitation of pure signal processing techniques for dealing with musical signals. Gestalt psychological grouping principles motivate the hierarchical signal model, and component identifiability is considered in terms of perceptual streaming where each component establishes its own context.

A major emphasis of this thesis is the practical application of MCMC techniques, which are generally deemed to be too slow for many applications. Through the design of efficient transition kernels highly optimised for harmonic models, and by careful choice of assumptions and approximations, implementations approaching the order of real-time are viable.



# *Acknowledgements*

---

My sincerest thanks are due to all those that have helped me to get this far. Thanks to my supervisors, Simon Godsill and Peter Rayner, for their inspiration, ideas and support, not to mention for coping with the mountains of paperwork I must have caused. Thanks to Bill Fitzgerald and Malcolm Macleod for the stimulating discussions arising from the marriage of signal processing and music. I'm deeply grateful to my hardened and voraciously enthusiastic team of proof-readers: Nick for making sense of it all, Patrick for the expurgation of redundant verbosity, Simon for the extra chapter, James for the extra commas and the royal 'we', and Steve for the elimination of foot-in-mouth disease. Thanks to everyone else in the lab who have made it a fun place to live for the last four years, particularly for the sellotape warfare. I am indebted to EPSRC for funding my research escapades.

Those to thank for my sanity: *Arco Iris* and *Galo já Cantou*, the excitable bunch of *sambistas* and *capoeristas*, for making loud noises and giving me something to hit. *Obrigado camarat!* The Grange Road gang for late nights and klaxxons. Allan and Parkin for the four steps of contemporary modern dance. Sibelius for giving me something to do in the real world.

My deepest thanks go to my family, for their love and support over the time its taken me to get this done.

And of course, Dominic's teeth, without whom...



# Notation

---

## Symbols

$\mathbf{b}$	Amplitudes of GLM basis functions
$\hat{\mathbf{b}}$	An estimate (generally least squares) for $\mathbf{b}$
$\Delta_\phi, \sigma_\phi^2$	Variational hyperparameters for parameter $\phi$
$\Gamma^q$	Boolean indicator variable for component $q$
$\mathbf{G}$	GLM basis matrix
$\mathbf{G}^c$	Composite GLM basis matrix
$\mathbf{g}$	GLM basis vector
$\theta$	Parameters of one component of a GLM, $\theta = \{\phi, M, \mathbf{b}\}$ .
$\tilde{\theta}$	Parameters of one component with linear amplitudes marginalised
$\{\theta^q\}_{\mathbb{Q}}$	Set of parameters for currently included components, $\{\theta^q : q \in \mathbb{Q}\}$
$\{\theta_i^q\}_{\mathbb{Q}, N_f}$	Ditto over all frames, $\{\theta_i^q : q \in \mathbb{Q}, i = 1 : N_f\}$
$\{\theta^q\}_{-\{q\}}$	All components in $\mathbb{Q}$ except for the $q^{\text{th}}$
$\mathbb{I}_{[x_1, x_2]}(x)$	Indicator function – unity for $x_1 \leq x \leq x_2$ , zero otherwise
$\mathcal{I}$	All implicit prior knowledge
$\mathbf{I}_N$	The $N \times N$ identity matrix
$M$	GLM model order (number of basis functions)
$\mathcal{M}_{\mathbb{Q}}$	Model composition for multiple component model
$\phi$	Parameter of a GLM basis function
$q(\theta^* ; \theta^k)$	Proposal distribution for $\theta^*$ conditional upon the previous state $\theta^k$
$\mathbb{Q}$	Set of indices of included components, $\{q : \Gamma^q = 1\}$
$Q$	Maximum number of components
$x \sim p(x)$	A sample $\mathbf{x}$ is drawn from the probability distribution $p(x)$
$p(x   \mathbf{d}, \bullet)$	Full conditional distribution of $x$ ( $\bullet$ denotes all other parameters in the model)

**Probability distributions****Poisson** for discrete  $H$ 

$$\text{Poisson}(H = h; \mathbb{H}) = \frac{(\mathbb{H})^h}{h!} e^{-\mathbb{H}}$$

**Gamma** which has mode  $(\alpha - 1)/\beta$ , mean  $\alpha/\beta$ , and variance  $\alpha/\beta^2$ 

$$\text{Ga}(z; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} z^{\alpha-1} e^{-\beta z} \mathbb{I}_{[0,+\infty)}(z) \quad (1)$$

**Inverse Gamma** which has mode  $\beta/(\alpha + 1)$ , mean  $\beta/(\alpha - 1)$  (defined for  $\alpha > 1$ ) and variance  $\beta^2/(\alpha - 1)^2(\alpha - 2)$  (defined for  $\alpha > 2$ ). To sample from the IG distribution, sample  $y \sim \text{Ga}(\alpha, \beta)$  then transform  $z = y^{-1}$ .

$$\text{IG}(z; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} z^{-\alpha-1} e^{-\frac{\beta}{z}} \mathbb{I}_{[0,+\infty)}(z) \quad (2)$$

**Lognormal** which is equivalent to  $\log(\omega)$  being normally distributed,  $\text{N}(\log(\nu), w^2)$ .

$$\text{LN}(\omega; \nu, w^2) = \frac{\mathbb{I}_{[0,+\infty)}(\omega)}{(2\pi w^2)^{\frac{1}{2}} \omega} \exp \left[ -\frac{(\log \omega - \log \nu)^2}{2w^2} \right] \quad (3)$$

**Multivariate Normal**

$$\text{N}(\mathbf{x}; \mu, \Sigma) = \frac{1}{|2\pi\Sigma|^{\frac{1}{2}}} \exp \left[ -\frac{(\mathbf{x} - \mu)^t \Sigma^{-1} (\mathbf{x} - \mu)}{2} \right] \quad (4)$$

**Abbreviations**

AIC	Akaike's Information Criterion
GLM	General Linear Model
MAP	Maximum <i>A Posteriori</i>
MCMC	Markov chain Monte Carlo
MDL	Minimum Description Length
MH	Metropolis Hastings
PR	Posterior probability Ratio
SNR	Signal to Noise Ratio
TR	Transition probability Ratio

# Contents

---

<b>1</b>	<b>Introduction</b>	<b>17</b>
<b>2</b>	<b>Approaches to Transcription, Modelling and Separation</b>	<b>23</b>
2.1	Introduction . . . . .	23
2.2	Auditory modelling . . . . .	24
2.2.1	Low level auditory modelling . . . . .	25
2.2.2	High level perceptual models . . . . .	29
2.3	Auditory and perceptual-based models . . . . .	33
2.4	Musical models . . . . .	36
2.4.1	Pitch models . . . . .	37
2.4.2	Harmony . . . . .	39
2.4.3	Statistical studies . . . . .	40
2.4.4	Streaming, grouping and segmentation . . . . .	41
2.5	Signal processing models . . . . .	45
2.5.1	Additive models . . . . .	45
2.5.2	Source-filter models . . . . .	55
2.5.3	Instrument models . . . . .	57
2.6	Conclusions . . . . .	67

<b>3</b>	<b>Bayesian Signal Processing</b>	<b>69</b>
3.1	Introduction . . . . .	69
3.2	Probabilistic signal modelling . . . . .	70
3.3	Bayes' theorem . . . . .	73
3.3.1	Prior probabilities . . . . .	74
3.3.2	Representation of prior knowledge . . . . .	75
3.3.3	Marginalisation . . . . .	76
3.3.4	Model selection . . . . .	76
3.4	Parameter estimation through MCMC . . . . .	78
3.4.1	Monte-Carlo Integration . . . . .	79
3.4.2	Markov chain overview . . . . .	80
3.4.3	The Metropolis-Hastings algorithm . . . . .	81
3.4.4	The Gibbs sampler . . . . .	83
3.5	Designing efficient transition kernels . . . . .	85
3.5.1	Types of transition kernel . . . . .	86
3.6	MCMC for model selection . . . . .	88
3.7	Conclusions . . . . .	90
<b>4</b>	<b>Detection and Estimation of Single Component Models</b>	<b>91</b>
4.1	Introduction . . . . .	91
4.2	The General Linear Model . . . . .	92
4.2.1	Formulation . . . . .	93
4.2.2	Conditional distributions for amplitudes and error variance . . . . .	96
4.2.3	The effect of parameter priors . . . . .	96
4.3	MCMC parameter estimation . . . . .	97
4.3.1	Fixed-scale shape function . . . . .	100
4.3.2	Multiple basis functions . . . . .	102

---

4.4	Variable model order . . . . .	102
4.4.1	Sensitivity of model order to amplitude prior . . . . .	104
4.5	Modifications to prior structure . . . . .	106
4.5.1	Marginalisation and joint proposals . . . . .	108
4.6	Multiple frame methods . . . . .	109
4.6.1	Graphical models . . . . .	110
4.6.2	Independent frames . . . . .	113
4.6.3	Markovian dependence . . . . .	113
4.6.4	Hierarchical multiple frame models . . . . .	115
4.7	Transition kernels for multiple frame methods . . . . .	121
4.7.1	Multiple frame simulation . . . . .	124
4.8	Conclusions . . . . .	127
<b>5</b>	<b>Detection and Estimation of Multiple Component Models</b>	<b>129</b>
5.1	Introduction . . . . .	129
5.2	The multiple component general linear model . . . . .	130
5.2.1	Model formulation . . . . .	130
5.2.2	Bayesian formulation . . . . .	132
5.3	Simulation and inference . . . . .	134
5.3.1	Residual methods for mixtures of GLMs . . . . .	137
5.3.2	Residual-dependent and signal-dependent kernels . . . . .	139
5.3.3	Sampling for indicator variables . . . . .	141
5.3.4	Non-marginalised amplitudes . . . . .	143
5.4	Multiple component example . . . . .	143
5.4.1	Model formulation and priors . . . . .	144
5.4.2	Proposal distributions . . . . .	145
5.4.3	Results . . . . .	147

5.5	Multiple component, multiple frame models . . . . .	149
5.5.1	Component birth/death move . . . . .	155
5.5.2	Joint block update move . . . . .	156
5.5.3	Perturbation update . . . . .	158
5.6	Conclusions . . . . .	159
<b>6</b>	<b>Application to Monophonic Musical Signals</b>	<b>161</b>
6.1	Introduction . . . . .	161
6.2	Harmonic modelling . . . . .	162
6.2.1	Formulation . . . . .	163
6.2.2	Periodogram estimator . . . . .	163
6.2.3	Posterior distribution . . . . .	164
6.2.4	Harmonic transform . . . . .	166
6.3	Monophonic pitch detection . . . . .	168
6.3.1	MCMC techniques . . . . .	169
6.3.2	Monophonic analysis example . . . . .	171
6.4	Multiple frame model . . . . .	175
6.4.1	Transition kernels . . . . .	176
6.4.2	Monophonic multiple frame example . . . . .	178
6.4.3	Harmonic evolution . . . . .	179
6.5	Conclusions . . . . .	181
<b>7</b>	<b>Application to Polyphonic Pitch Detection</b>	<b>185</b>
7.1	Introduction . . . . .	185
7.2	Single frame polyphonic model . . . . .	186
7.2.1	Motivation . . . . .	186
7.2.2	Model composition . . . . .	187

---

7.2.3	Transition kernels . . . . .	188
7.3	Multiple frame polyphonic model . . . . .	192
7.3.1	Simulation scheme . . . . .	194
7.4	Simulation results . . . . .	195
7.4.1	Synthetic harmonic data . . . . .	195
7.4.2	Duophonic example . . . . .	197
7.4.3	Polyphonic piano examples . . . . .	198
7.5	Appraisal of the harmonic model . . . . .	200
7.5.1	Benefits . . . . .	201
7.5.2	Limitations . . . . .	203
7.5.3	Suitability of MCMC methods . . . . .	205
7.6	Conclusions . . . . .	206
<b>8</b>	<b>Conclusions and Future Research</b>	<b>207</b>
8.1	Conclusions . . . . .	207
8.2	Discussion . . . . .	208
8.2.1	Importance of high level modelling . . . . .	208
8.2.2	Completing the loop . . . . .	210
8.2.3	The need for hierarchical modelling and feedback . . . . .	210
8.3	Future research . . . . .	212
8.3.1	Timbral context . . . . .	212
8.3.2	High level modelling . . . . .	213
8.3.3	Beat inference . . . . .	214
8.3.4	Software for multiple component analysis . . . . .	214
	<b>Bibliography</b>	<b>215</b>

<b>A</b>	<b>Block matrix inversion</b>	<b>229</b>
A.1	Inverse by partitioning . . . . .	229
A.2	Use with basis matrices . . . . .	230
<b>B</b>	<b>Marginalisation of amplitude parameters and error variance</b>	<b>231</b>
B.1	Amplitude parameters . . . . .	231
B.2	Error variance . . . . .	232
B.3	Approximation of conditional residual to joint marginal posterior . . . .	233
<b>C</b>	<b>Accompanying CD</b>	<b>237</b>

---

Musical signals have been of enigmatic interest to many scholars since the time of Pythagoras [14]. Considerable amounts of research have been devoted to their analysis, yet we do not appear appreciably closer to understanding those properties of musical signals which are capable of provoking cognitive and emotional responses in the listener. It is this inherent complexity which draws so much attention to the analysis of musical signals from such diverse backgrounds as engineering, physics, artificial intelligence, auditory psychology, music psychology and music theory.

Other aspects of musical technology have however progressed to the point that today comparatively few households are without digital audio media in the form of compact discs, minidisks or MP3 players. Many of today's commercial recordings are committed directly to the digital domain without a reel of magnetic tape in sight. The ubiquity of digital audio is further evidenced by the multimedia capabilities of modern personal computers and its profusion over the internet. The downloading of streamed audio in real time from radio stations and concert venues is now commonplace, and perceptually-motivated coding techniques are employed to obtain high compression rates whilst maintaining high audio quality.

One of the main attractions of digital audio is the ability to transfer and reproduce it in the digital domain without degradation. Many hardware and software tools exist to replace the array of traditional recording studio hardware, performing duties such as adding effects<sup>1</sup>, reducing noise and compensating for other undesired signal characteristics, all without introducing losses from the signal paths between the processing components.

Processing stages intended to remedy artefacts introduced in the signal path and from the media characteristics, *e.g.*, the removal of gramophone thumps, hiss, clicks and

---

<sup>1</sup>Including, perversely, simulating the characteristics of magnetic tape and other analogue equipment.

wow [54, 155, 164] and nonlinear distortions [152], operate upon the entire signal, as all signal components are affected.

Other operations may however be incompatible with this global approach: reducing or increasing the prominence of a particular instrument within the mix, correcting a wrong note played by one member of an ensemble or removing an errant mobile phone from a valuable live recording. In many instances the original multi-track recording is not available, and a one or two channel mix of a number of instruments must suffice. Often, particularly with live orchestral recordings, only a handful of microphones are used, each receiving a different proportion of the signal from each of the constituent instruments.

To perform these operations, the entire signal has first to be broken down into its constituent components, and the relevant processing can then be applied. The individual components are unlikely to be disjoint in time and frequency and so simple filtering and transformation-based methods will generally be inadequate. A parametric model of musical signals that can account for the contribution of several different sources to the overall sound is a necessary first step towards any form of separation or inference about the individual components.

Musical signals are characteristically very structured: at the lowest level, sinusoids are grouped together to form *notes* of particular *itches*. Notes are grouped in the pitch domain to form *chords* or *harmonies*, and grouped in the time domain to form *melodies*. Yet higher levels of structure may establish *themes* through repetition and simple transformation of smaller elements. This successive abstraction to higher levels is termed *musical context integration*. The perception of musical signals is consistent with these Gestalt grouping principles.

From a signal processing point of view, the existence of musical structure can in fact be a hindrance. Common assumptions of independence and orthogonality (which usually permit considerable simplifications to the problem at hand) do not necessarily apply. In most forms of (Western tonal) music the notes will not be independent events — they are likely to be highly correlated in the time and pitch domains. The nature of harmony is such that notes with common harmonics tend to sound concordant, therefore such notes will not be orthogonal in conventional representations, by virtue of the sharing of fundamental components. Consequently, blind application of raw signal processing

techniques to musical signals is unlikely to meet with much success.

This thesis addresses how musical signals may be modelled to reflect this hierarchical framework of grouping. Once the model has been formulated, it is necessary to estimate values of the parameters which characterise the data. Given these parameter values, any or all of the components can be resynthesised individually for subsequent processing (as for the applications described above), or the variation of the parameters (usually pitch) of each instrument over time can be logged. This latter operation relates to the task of polyphonic music transcription, which is of great interest to composers and music publishers for transforming audio input into a score.

The high degree of structure in the data makes a parametric model ideal. The model can be posed within a Bayesian framework which allows enormous potential for the representation of salient prior information. This information may affect parameter values, for instance by imposing physical constraints such as the parameter range, or convey some conviction about possible values given the current musical context, or alternatively we may wish to ensure that we are being sufficiently non-informative so as not to bias results when we know little *a priori*. Information can also be incorporated at a higher level to not only represent likely parameter values, but to also describe their behaviour over time using *hyperparameters*. The structure can be represented using Bayesian *graphical models* which reflect the statistical dependence between the parameters and hyperparameters. They can also result in more efficient estimation algorithms by identifying correlated parameters, such that these parameters can be estimated jointly.

Bayesian inference is employed for the parameter estimation task — typically we obtain the mode or the mean of the posterior distribution. The complexity of the model makes direct calculation of parameter estimates infeasible and necessitates the use of numerical techniques. In recent years, Markov chain Monte Carlo (MCMC) methods have become popular for such problems, as a consequence of the rapid growth in available computing power. MCMC methods generate a sequence of dependent samples from (*i.e.*, simulate a Markov chain from) the desired posterior distribution, such that inferences upon the distribution can be made from Monte Carlo integrals of the samples.

MCMC methods are generally held to be slow for practical applications, but this thesis shows how efficient algorithms can be produced by exploiting the structure of the model at hand. In many instances there are a number of simplifying assumptions and

approximations that may be made; these may permit a substantial reduction in the computational cost of the estimation scheme. Together with careful implementation, these techniques bring the prospect of real-time MCMC-based algorithms significantly closer. The layout of the rest of this thesis is as follows:

### **Chapter 2. Approaches to transcription and separation**

This chapter describes some of the very varied methods for the analysis of musical signals from the perspective of a number of different disciplines: auditory psychology, music psychology and signal processing. Key elements of different approaches are drawn on in the work presented in later chapters.

### **Chapter 3. Bayesian signal processing**

This chapter introduces a Bayesian methodology for the modelling of signals, and subsequent methods for performing the parameter estimation task, in particular Markov Chain Monte Carlo (MCMC) techniques, together with strategies for the implementation of efficient MCMC algorithms. The issues of model selection over variable sized dimensions are addressed with recourse to reversible jump techniques.

### **Chapter 4. Detection and estimation of single component models**

A flexible model formulation based upon the General Linear Model is introduced in this chapter that is suited to the analysis of many types of signal. The model is posed in a Bayesian framework and MCMC techniques for producing parameter estimates are described. The extension of the basic model to deal with time-varying signals is detailed that exploits the slow variation of signal parameters to produce an efficient algorithm. The parameters of several observations are estimated jointly, which increases robustness to transient disturbances; this can potentially achieve speed increases over independent analyses of each frame.

### **Chapter 5. Detection and estimation of multiple component models**

This chapter generalises the results of the previous chapter to multiple components, when the number and type of components is unknown. Careful formulation of the model is required to enable parameter estimation over a variable sized parameter space, and reversible jump techniques are applied to allow this. Residual methods are developed to reduce the computational load of the simulation.

Efficient parameter estimation techniques for time-varying multiple component signals are also presented.

### **Chapter 6 Application to monophonic musical signals**

The models of the previous chapters are applied to the analysis of monophonic musical signals. A general linear model composed of harmonically related basis functions is employed, and the motivation for the harmonic model is outlined. An efficient MCMC simulation scheme is described that exploits the structure of the harmonic model to produce some novel transition kernels. Several assumptions and approximations are outlined that can reduce the computational cost by an order of magnitude. Some analyses of monophonic data are shown that highlight the ability of the model to perform pitch estimation, or give detailed information about the harmonic character of the musical instrument.

### **Chapter 7 Application to polyphonic musical signals**

The monophonic model of the previous chapter is extended to polyphonic signals. A hierarchical model is used that incorporates Gestalt perceptual grouping principles to model the perception of musical signals in terms of streams of notes over time. An efficient MCMC simulation scheme is described that employs residual methods combined with methods optimised for harmonic signals to produce a method that is able to explore the posterior distribution very rapidly. Some examples for real and synthetic polyphonic datasets are shown. The chapter also gives an assessment of the model for musical signal modelling, describing its advantages and limitations.

### **Chapter 8. Conclusions**

This chapter presents a summary of the material in this thesis and outlines potential areas for future research. Some of the predominant issues arising from the analysis of audio signals and the rôle of Bayesian signal modelling in the area of musical signal analysis are discussed.

Throughout this thesis there are several important concepts which arise repeatedly. Their importance is reflected in the breadth of subject areas covered in the literature review chapters.

#### **Model structure**

- The exploitation of model structure in musical signals, which are highly structured at a number of levels. This can lead to efficient analysis methods and can permit simplifying assumptions.
- The importance of Gestalt grouping principles as an abstraction mechanism and the use of Bayesian hierarchical models for representing this structure.

### **Multi-disciplinary approaches**

- The need for expertise from many different subject areas for effective modelling.
- Signal processing principles alone are inadequate for modelling complex musical signals, whilst high level models do not consider the signal level.
- The need to unite high and low levels of modelling is very important: low level signals are modelled in order to make inferences in the high level problem domain.

### **Importance of context**

- The analysis of musical signals can lead to ambiguous results as many properties (*e.g.*, pitch, timbre) are ill-defined.
- Establishing a surrounding *context* in one dimension (*e.g.*, time, frequency) may provide a means to resolve the ambiguities by horizontal or vertical grouping.
- Joint estimation of signal parameters over longer timescales provides increased robustness through the exploitation of model structure.

# *Approaches to Transcription, Modelling and Separation*

---

# 2

## *2.1 Introduction*

The analysis of music and musical signals is an area which has attracted attention from a number of different disciplines, in particular physics, psychoacoustics, mathematics, music and engineering. Each has played a major part in the current state of understanding of the characteristics of musical signals. Pythagoras investigated the relative lengths of vibrating strings producing concordant sounds, thus establishing the twelve tone Pythagorean scale in terms of the ratios of integers. Mersenne and Galileo independently related musical pitch to the rate of oscillation. Fourier's decomposition of periodic oscillations into frequency components proved to be inspirational for many subsequent investigators, and physicists such as Helmholtz, Ohm and Seebeck sought to explain the perception of a single pitch given a harmonic series [108].

Physics has contributed much to musical research in terms of the understanding of the sound generating mechanisms of musical instruments. This can be of great importance in the design of new or improved instruments. Psychoacoustics deals not with the physical properties of sound itself but of its perceptual properties; *i.e.*, rather than dealing with sound generation it is concerned with sound reception. Intensity, pitch, timbre and roughness are examples of such perceptual properties, which are important to the perception of musical sounds, but which may not necessarily map directly onto signal characteristics. It is closely related to the field of cognitive psychology which deals with the high level organisation of perceived sound, and the area of auditory scene analysis.

Naturally, music theory and music psychology also have much to contribute to this field

of research. Rather than be concerned with low-level concerns of sound generation and auditory detection, they focus on high-level modelling of musical structure. Musical constructs such as key, harmony and metre are significant in these methods, and the emphasis of many techniques is of an *understanding* of musical signals such that high-level inferences and expectations may be generated.

More recently, researchers in several areas of engineering (particularly within the signal processing community) find themselves drawn towards the challenges of analysing and modelling musical sound. Many analysis techniques exist in signal processing for the representation, detection and transformation of signals. Much research is application-driven and closely correlated to developments in hardware, so efficient implementations are often readily achievable. At a more fundamental level, signal processing provides various calculi for the representation and transformation of signals.

Clearly each of the above areas has much to offer the realm of musical signal analysis; none, though, can offer (alone) a complete solution. A musical signal is a representation of a series of physical vibrations, perhaps varying over time, which produce sensations of perceptual characteristics (*e.g.*, pitch, timbre, loudness) over short time scales and gives rise to perceptions of musical structure (*e.g.*, melody, rhythm, harmony) over longer time scales. The remainder of this chapter will review salient contributions to the field of musical signal analysis in each of the above areas. Subsequent chapters will draw upon various aspects of these areas to improve the performance of the signal detection and estimation techniques.

## 2.2 *Auditory modelling*

Performance on a par with the human auditory apparatus is undeniably a goal of most pitch estimation and signal decomposition techniques. What the auditory system may lack in terms of sensitivity and its susceptibility to masking effects is more than compensated for by its speed, robustness and ability to perform high level inference. The ear is able to isolate a single speaker in a crowded room (the ‘cocktail-party effect’) or follow a single instrument in a polyphonic recording, even without the help of spatial

information. Researchers in psychoacoustics and cognitive psychology are concerned with understanding the physical, neural and psychological processes which enable us to perceive sound, and to build models of the human auditory system. There are several such areas which are directly relevant to this thesis. The following sections will review some of the salient contributions of (low-level) auditory modelling and (high-level) perceptual modelling research to the understanding of the auditory system. Some practical models drawing on auditory and perceptual principles will also be briefly reviewed.

### 2.2.1 Low level auditory modelling

In the mammalian auditory system, acoustic vibrations are channelled by the outer ear into the middle ear, via three bones: the *malleus*, *incus* and *stapes*. The middle ear effects impedance matching between the pressure upon the *tympanic membrane* (eardrum) and the *cochlea* (inner ear). The cochlea is filled with two different fluids which transmit oscillations to the *basilar membrane*. The basilar membrane (BM) has variable compliance along its length, and von Békésy [157] showed that travelling waves (rather than standing waves, as was previously thought) are induced in the BM. The consequence of this is that each point along the BM has maximum response at a different frequency, and hence it acts as a frequency-to-place convertor. There is also a direct correspondence between each place and the neurones sensitive to that particular characteristic frequency (CF), referred to as *tonotopic organisation*. This gives rise to *place* models of pitch perception.

The response of the BM to excitation at different frequencies can be quantified in terms of the magnitudes of the resonances of each place on the BM. One may talk about the ‘sharpness’ of the tuning in terms of the Q of the local resonance, which is in inverse proportion to the bandwidth [97]. The psychoacoustic response is often expressed in terms of *auditory filters*. Fletcher [41] suggests this approach, where the peripheral auditory system is modelled as a bank of overlapping bandpass filters. He measures the sensitivity to sinusoidal signals in noise by masking the sinusoid with band-limited noise (of constant power density) centred upon the sinusoidal frequency. The point at which an increase of the noise bandwidth does not increase the amount of noise in the auditory filter is the *critical bandwidth* (CB). This expression of the auditory filters is extensively used; Zwicker and his collaborators adopt measures based upon critical

bandwidth [168, 167]. The audible frequency range is split into around 24 abutting bands spaced at 100Hz intervals up to 500Hz and then spaced at approximately  $0.2f$ Hz above 500Hz.

The critical-band rate (*i.e.*, the cumulative number of bands as a function of increasing frequency) thus obtained is known as the Bark scale.<sup>1</sup> The Bark scale is also proportional to the position of maximum excitation along the length of the BM as a function of frequency. It is also proportional to another less common pitch scale, the *mel* scale, also known as *ratio pitch*. Ratio pitch is measured by presenting a pure tone  $f_1$  to the subject who must then adjust a second frequency  $f_{1/2}$  such that it appears to be half the pitch of  $f_1$ . The scale of  $f_{1/2}$  against  $f_1$  is linear up to around  $f_1 = 1\text{kHz}$  but then exhibits a marked departure from linearity such that 1.3kHz is perceived to be the half pitch of an 8kHz stimulus. Since the scale purportedly relates to the sensation of melody, the scale is named the *mel* and is defined relative to a pure tone of 125Hz having a ratio pitch of 125mel [143]. Its relation to the Bark scale is 1 Bark = 100 mel. Its departure from the familiar (linear) musical scale of pitch in terms of octaves accounts for its lack of popularity, despite being an empirical measure of pitch perception [63].

Critical-band based auditory filterbanks, on the other hand, are popular as a preprocessing stage in musical applications. Moore [97], however, notes that critical bands are rectangular — as originally proposed by Fletcher as a simplifying assumption in the analysis — whereas the shape of the auditory filters are not. A scale based upon Equivalent Rectangular Bandwidths (ERBs) is proposed [98] which differs significantly from the CB scale at low frequencies as more accurate measuring techniques have indicated that the auditory bandwidths continue to decrease below 500Hz. Experiments show that the human auditory filter has a rounded exponential shape (*roex*) which can be well approximated by a *gammatone* filter whose impulse response has the envelope of a gamma filter multiplied by a sinusoid at the centre frequency [107]. The gammatone filterbank can also be efficiently constructed, a feature which has made it popular in many applications.

---

<sup>1</sup>After Barkhausen, who studied the perception of loudness.

### *Place models*

The simplest place models of pitch are unable to explain many pitch phenomena, in particular the perception of complex tones. Those complex tones that evoke the greatest sensation of pitch are comprised of pure tones (*i.e.*, sinusoids) at harmonically related frequencies.<sup>2</sup> This gives rise to many peaks in the oscillation of the BM, and a place model based upon attending to the cochlear channel with the maximum response will fail to explain the percept of a single pitch. Ohm's acoustical law postulates that the individual frequency components in a complex tone can be 'heard out', but this is now known to be true only for the first five to seven harmonics, and experimental conditions are required to isolate individual components [109, 117].<sup>3</sup> Complex tones with harmonically related components usually generate a single sensation of pitch, as they 'fuse' perceptually. This effect will be discussed in greater detail in the next section. One limitation of the place model of pitch, as suggested by some authors (including Scheirer [127]) is that it may assume more frequency resolution than the ear possesses and hence isn't an accurate model of the auditory system.

Another effect which should be addressed by a pitch model is that of the missing fundamental. The first few harmonics of a complex tone can be omitted and yet the tone still evokes a sensation of pitch at the (now absent) fundamental frequency. This percept has become known by several names: *virtual pitch*, *low pitch* or *residue pitch*. Certain musical instruments, *e.g.*, the bassoon, have very low or non-existent fundamental frequencies yet they still appear to be pitched at that fundamental frequency. The virtual pitch theory of Terhardt *et al.* [147, 146] bases a perception of pitch on a combination of analytic listening and holistic perception. Analytic listening, *i.e.*, the ability to hear out individual partials in a complex tone yields spectral pitch. The holistic perception makes inferences from tones where the lower harmonics are not present. This is also an important mode of pitch perception for instruments such as chimes and bells which are inharmonic and have no modes of vibration near to the fundamental — the pitch is instead determined by the 4th to 6th harmonics [42].

---

<sup>2</sup>There are nevertheless other stimuli which can invoke a sensation of pitch, for instance narrow-band noise.

<sup>3</sup>For instance probe-tone techniques can be used where the subject has to decide whether a frequency is present in a complex tone. The single frequency stimulus is non-deterministically chosen to be a true component of the harmonic series or a value halfway between two harmonics.

Two broad classes of psychoacoustic pitch models are predominant in the literature: pattern recognition models and temporal models [97, chapter 5]. Pattern recognition approaches assume two stages of analysis: peripheral processing (spectral analysis) followed by central processing (pattern recognition). The most widely referenced model is Goldstein's optimum processor [55] which resolves individual spectral peaks from each ear and feeds them to a central processor which assumes that all frequencies are noisy representations of harmonic frequencies and produces a maximum likelihood estimate of the fundamental frequency. The model is able to explain phenomena such as missing fundamentals and dichotic pitch (where each ear receives different harmonics). The Schröder histogram [129] is a commonly referenced pitch model that tallies all integer sub-multiples of each active frequency, and the one with the most tallies is taken as the fundamental; Parsons [106] uses this technique to determine the pitch of speech signals.

### *Temporal models*

Temporal models operate rather differently by assuming that temporal information is also important in the perception of pitch. The auditory nerve firings are phase-locked to the excitation waveform such that they fire at the same phase on the waveform, though not necessarily on every cycle. This behaviour is exhibited for stimuli below around 5kHz [97]. Each neurone therefore transfers periodicity information in its temporal firing pattern, in addition to that suggested by its tonotopic organisation. Temporal models apply an initial cochlea filtering stage and pass each filterbank output into a periodicity detector. Hybrid place/temporal schemes have become popular, in particular the models of Patterson and Holdsworth [107] and Meddis and Hewitt [93].

Meddis and Hewitt's model applies an initial filter stage to account for the response of the outer and middle ear, followed by a filterbank of 128 overlapping critical-band filters to simulate the response of the basilar membrane. Each filter output is passed through a hair-cell simulation model and the resulting neural spikes undergo an auto-correlation analysis. This is then averaged across all channels to produce a summary autocorrelogram where the height of each pitch candidate corresponds to the strength or salience of that pitch.

Patterson and Holdsworth's model is intended as a functional rather than physiologically accurate model. The cochlea filtering stage employs a gammatone filterbank with

bandwidths based upon Moore and Glasberg's ERB formulae [98] and their neural detection stage includes logarithmic compression and an adaptation stage. This latter stage simulates the auditory mechanisms of adaptation and suppression, such that it adapts rapidly to changes in intensity and that areas of low activity are suppressed when other intense activity regions exist. The following temporal stage is based around the concept of an *auditory image model*. For periodic sounds, the auditory image produced should be stable, but the Neural Activity Pattern (NAP) produced from the adaptation stage does not present a steady image as each cochlear channel will consist of responses at multiples of the excitation pitch period. The temporal integration is therefore made adaptive, triggering at a particular point on the waveform and effectively simulating the phase-locking of neurons.

### 2.2.2 High level perceptual models

Our responses to everyday acoustic stimuli are of course much more complex than the psychoacoustic responses that might be expected from steady-state complex tones in noise. We are able to detect much more in sounds than simply a collection of instantaneous pitches. From musical signals many different kinds of information can be inferred:

- **Musical information:**  
Melody, rhythm, metre, harmony and modality
- **Environmental information:**  
Spatial location of the sound source, room characteristics, identification of background noises (*ecology*)
- **Cognition:**  
Identification of the type of instrument, recognition of the melody, recognition of a particular composer's style, recognition of musical genre
- **Emotional information:**  
Identifying the emotional intents of the composer and musician

All of the above form percepts which are of a much higher level than pitch sensations, and which cannot be explained at a physiological level. The area of study concerned

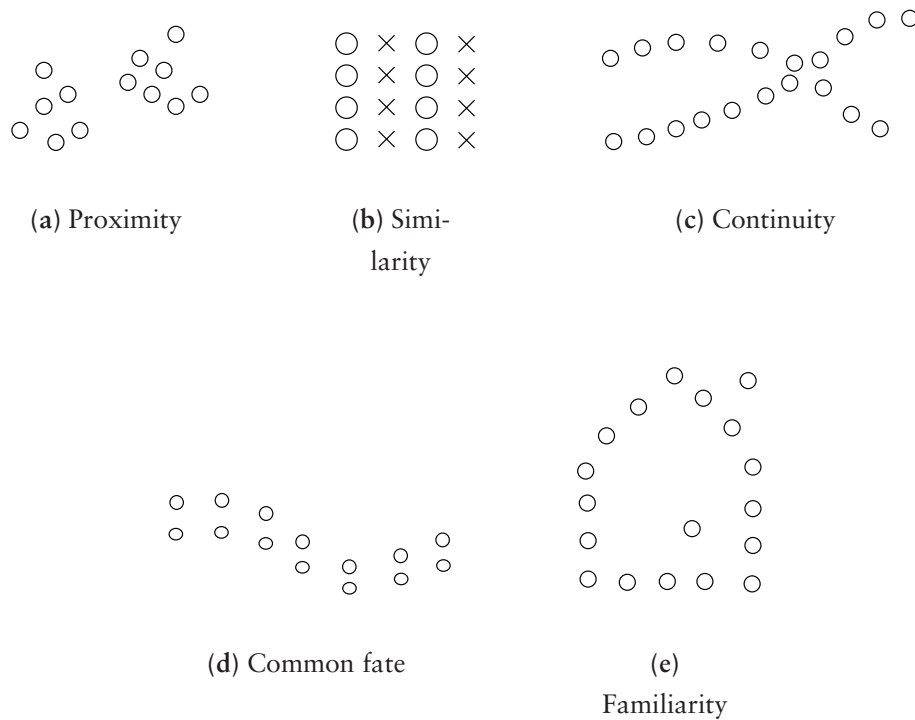


Figure 2.1: Illustration of some Gestalt grouping principles (see text for explanation).

with models of audition has become known as *Auditory Scene Analysis* after Bregman's text [11]. The name is taken from the auditory analog of visual scene analysis — the process of drawing inferences from the given environment. Bregman's approach is based upon the concepts of *grouping* and *streaming*. Low level components of sound are grouped using a variety of rules into *sources* which are associated with streams.

The models of grouping used almost universally in the field are those proposed in the early twentieth century by the Gestalt psychologists. A simple set of rules is employed to group fundamental elements together. Many principles of auditory organisation have visual analogs — the most common grouping *cues* are illustrated in figure 2.1. Proximity rules group elements which are close together. Similarity rules group elements with common characteristics — the figure comprises of a  $4 \times 4$  grid of elements, but rather than suggesting the formation of a square, the perception is of two sets of parallel lines by account of the similarity of the elements in each column. The principle of good continuity groups those elements which follow particular directions — the illustration

is perceived as two crossing curves rather than two curves which touch and then diverge. The principle of common fate groups elements which are subject to similar variations — in the figure we see two identical curves. Finally the principle of familiarity causes us to group elements into configurations which are familiar to us, and so the final figure forms itself into the image of a house.

These principles are of great importance when analysing musical signals, as argued by Deutsch [27]. Sounds which are similar suggest an origination from the same source, and conversely, sounds which are different suggest a number of sources. A sequence of sounds with smoothly changing characteristics (*e.g.*, frequency) suggest a stream from a single source.

Two of the most important cues in the low-level analysis of musical signals are common harmonicity and common onset. Harmonicity is a common fate cue which explains the perception of pitch in terms of the grouping of sinusoids with common periodicity. Psychoacoustic evidence suggests this is an integral part of auditory processing (see §2.2.1). Most pitch models assume the existence of sinusoidal partials at frequencies harmonically related to a fundamental (which may or may not be present).

The ear does not require perfect harmonicity for grouping. Moore *et al.* [99, 100] adjusted the frequency of a single partial in a harmonic series and found that for deviations of up to 3% the partial is still perceived as ‘belonging’ to the harmonic series, a concept known as the ‘harmonic sieve’. Further increases generate a sensation of ‘roughness’ and a shift of over 8% separates the partial entirely from the complex. This fits in with Goldstein’s optimal processor theory [55] described previously. Furthermore, shifts in the frequency of one harmonic affect the perceived pitch of the tonal complex, in a manner which is subject-dependent. The amount which each harmonic is capable of altering the pitch perception of the complex provides a measure of the dominance of that harmonic. The dominant harmonic is usually one of the first six [99].

The results also concur with the observations of inharmonicity in piano tones. Piano strings can be inharmonic by up to 3% which accounts for the ambiguous or ‘muddy’ pitch sensation for low notes. The inharmonicity increases with the string width and the effect is less marked if longer strings are used [42]. Pianos are often tuned to a ‘stretched’ scale beyond a 2:1 frequency ratio per octave in the high and low registers, partially due to the inharmonicity and also due to the non-uniform perception of pitch

at high registers, as suggested by the mel scale. Rasch [118] observes that harmonics of a single note fuse together, where *spectral fusion* plays a major part in the segregation of multiple sources. This effect is evident in organ stops where individual notes often fuse together as harmonics of a single note.

Common onset (also known as common amplitude onset) is a major auditory cue. It is the event marking the transition between silence and the simultaneous sounding of a number of frequencies. If a number of frequencies start at exactly the same time, it suggests that they might originate from the same instrument. Similarly, if a number of instruments are playing together it is quite plausible that a number of them may play on the same beat. Rasch [119] notes that there is typically a delay of between 30 and 50ms between the onsets of different instruments. The rise time of partials of most instruments tends to be within around 40ms in a small ensemble. These delays are usually long enough to allow the auditory system to detect each note in the ensemble.

Common onset is particularly important when considering the perception of non-tuned or inharmonic instruments such as percussion and bells. A struck cymbal generates a signal which has energy distributed over a wide range of frequencies, and evokes no definite pitch sensation<sup>4</sup> so common periodicity is not a viable cue. The spectral components are instead grouped according to their common moment of attack. Similarly, bells are inharmonic (but tuned), and the common onset of all frequency components allows them to be perceptually grouped. If two bells are struck at precisely the same time, the sensation of pitch is blurred and the common onset suggests the presence of a single bell. This phenomenon can also be observed in the Gamelan [76] whose bars generate inharmonically spaced partials<sup>5</sup> and are played in pairs.

Complementary to common onset is the cue of common offset, which describes the common termination of a number of frequency components as a cue for grouping. It is rarely discussed in the literature since it is not often encountered in musical or ecological sound stimuli. Few musical instruments have the characteristic of cutting off all frequency components simultaneously and furthermore the acoustic environment is often sufficiently reverberant to ensure that all frequencies decay at different rates.

---

<sup>4</sup>However, a second cymbal of a different size may sound higher or lower, although still unpitched, by virtue of the relative locations of the spectral centroids.

<sup>5</sup>This inharmonicity is essential to the sound of the Gamelan and is carefully controlled at the time the Gamelan is made.

Common frequency variation is a demonstrably important cue for spectral fusion. When the ear is presented with a harmonically rich mix of steady-state sounds, harmonicity alone may not be sufficient to effect perceptual grouping. If several components have frequencies which vary in an identical manner then they will be perceived to originate from the same source. This is demonstrated by McAdams [89] who applies frequency modulation to a synthetic vowel sound which was embedded within a mixture of other vowel sounds. The modulated frequency components fused perceptually and stood out from the other components. Common frequency variation is of particular interest in musical signals for two reasons: firstly, a change in the frequency of excitation of a musical instrument will change the frequency of each harmonic by an amount proportional to the harmonic number, and secondly, common frequency modulation in the form of *vibrato* is often added by a performer. This is encountered frequently in the parts played by the leading instruments of an ensemble, perhaps not only as an additional means of expression but also as a means of increasing the perceptual salience of the soloist over a spectrally rich backdrop. Rasch [118] finds that if a tone was played against a louder masking tone, the quieter tone could be reduced by 18dB and still be heard if vibrato was added.

Common amplitude variation doesn't appear to be as useful as common frequency variation since the harmonics of an instrument will decay at different rates. Mellody and Wakefield demonstrate an interesting example however where the sound of a violin played with vibrato was resynthesised with only amplitude or frequency modulation respectively [95]. The AM-only reconstruction sounded perceptually identical to the original, whereas the FM-only reconstruction sounded comparatively flat and lifeless. Amplitude variation can therefore be an important psychological cue, but may not be simple to model analytically.

## 2.3 Auditory and perceptual-based models

The psychoacoustics and perceptual psychology literature have proved to be extremely influential in problems of signal separation and musical transcription. In particular, the

key concepts of the cochlea as a filterbank and the perceptual organisation of sound through Gestalt grouping principles are central to many models, some of which will be reviewed in this section.

The sensitivity of the ear as a function of frequency is studied extensively by psychoacousticians. The representation of the cochlea as an auditory filterbank, based upon Equivalent Rectangular Bandwidths or Critical Bandwidths (see section 2.2.1) has led many authors to adopt this kind of frequency decomposition rather than an approach based upon the Fourier transform. This front-end processing may then be followed directly by a pattern processing stage, corresponding to a place model of pitch, *e.g.*, as in Goldstein's optimal processor [55] or each cochlear channel may be followed by a periodicity detection stage (temporal model of pitch perception) as in Meddis and Hewitt's model [93]. Hybrid place/temporal schemes are also common.

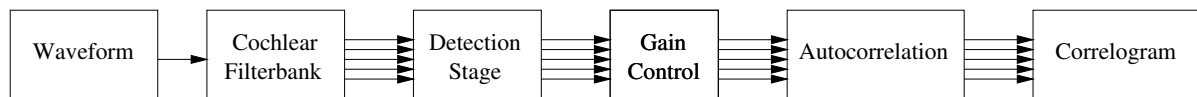


Figure 2.2: The processing steps of a typical correlogram-based auditory model.

The temporal integration of the cochlear channel outputs can be performed in different ways and a typical method is shown in figure 2.2. The cochlear filterbank outputs pass through a detection stage which is intended to simulate the firings of the auditory nerve. A gain control compresses the dynamic range of the detector outputs and the autocorrelation stage detects the periodicities in each channel, simulating the phase-locking behaviour of neurones which fire at the same relative phase on the excitation waveform. The outputs of the autocorrelation in each channel are then often combined into a 2-dimensional time-frequency or 3-dimensional time-lag-frequency representation called a *correlogram*.

Meddis and Hewitt propose a more complex model which also accounts for the transmission characteristics of the outer ear and has a more elaborate simulation of nerve fibre discharge [93]. Karjalainen and Tolonen employ a similar model for multiple pitch estimation, using half-wave rectification and low pass filtering in the detection stage [70].

Ellis' work [34, 35, 36] is built upon the 3d log-lag correlogram which also uses half-wave rectification and low pass filtering in the detection stage. A series of delay lines at logarithmically-spaced lags are then used to perform the autocorrelation with a constant number of bins per octave. The image of frequency (cochlear channel) against log-lag constitutes a correlogram 'slice' which are then concatenated over time to form a 3-d structure. Various statistics can be obtained from the correlogram, *e.g.*, summation over the cochlear channels yields the summary autocorrelation (a form of periodogram) similar to that of Meddis and Hewitt, and the zero-lag face of the correlogram shows a time-frequency representation of signal intensity.

A significant aspect of Ellis' model concerns the decomposition of signals into fundamental objects other than sinusoids. His basic building blocks are noise clouds (noise with a finite spectral spread), transient clicks, and *weft* (wideband periodic energy with a smooth time-varying envelope).<sup>6</sup> The subsequent processing in his model is based on a blackboard system (see also §2.4.4) which develops several inference hypotheses in parallel. The inference strategy is prediction-driven, where the source models generate expectations which are then confirmed or denied by a bottom-up signal analysis. His goal is computational auditory analysis, with an emphasis on ecological sound mixtures such as street ambience.

Mellinger [94] employs a cochleagram representation from which he makes inferences directly using techniques from the image processing literature. He develops a set of time-frequency filtering kernels<sup>7</sup> for extracting features from the correlogram. Mellinger's motivation is for a physiologically compatible model of auditory processing, and he claims the filtering kernels are plausible since they can be implemented as a delayed weighted summation which is within the functional capability of a neuron. The kernels which he produces enhance features such as note onsets and offsets and frequency variation, since the logarithmic frequency scale means that the frequency variation of harmonically related partials undergo coherent motion at a fixed spacing. He also deals with a higher level of modelling concerned with event detection by grouping partials using Gestalt grouping cues. Individual partials become part of an *affinity group* which corresponds to a note event and is comprised of partials that are closely related (*e.g.*, common onset, common periodicity).

---

<sup>6</sup>The name *weft* comes from an Anglo-Saxon term for the horizontal threads in a woven fabric.

<sup>7</sup>Or more precisely, time-height, where *height* is the logarithm of frequency.

Slaney, Naar and Lyon [135] describe techniques for inverting the auditory model directly from the correlogram. Components of the correlogram with common periodicity are separated, and are used to form short-time power spectra, and hence, estimates of the cochleagram. The component sounds are then resynthesised using an overlap-add technique. Weintraub [163] also computes the autocorrelation of the cochlear filterbank outputs to calculate the periodicity information in each frequency region, from which a spectral estimate of each sound is generated. An iterative algorithm locally maximises the probability of the spectral estimate, given the local periodicity information and the spectral continuity constraints. This approach is intended for speaker separation, and claims to separate unvoiced (*i.e.*, inharmonic) sounds. However, it must be trained with instances of all transitions, as it uses a Markov model to detect the voicing state for each speaker.

A neural comb filter is proposed by de Cheveigné [25] to isolate harmonic sounds in a cancellation model of auditory processing. Implementation of the filter requires knowledge of the fundamental frequencies, which are estimated from the average magnitude difference function (AMDF). This is searched over the lag domain ( $\tau$ ) for the minimum. For two simultaneous sounds, a comb filter with two lags is used, and the double difference function (DDF) must be searched over two dimensions of lag. The method performs well for samples where the fundamental frequencies didn't cross, but requires clean voiced sections which eliminated 25% of the data samples.

## 2.4 Musical models

Much of the work outlined in the preceding section is concerned with the auditory processes involved in the perception of sound. The high level auditory grouping processes described in the previous sections are not specific to music but rather applicable to the perception of complex sound mixtures of speech, music or our environment. This section will describe some of the considerations which are specific to musical signals. They have a high degree of structure at many different levels, and so algorithms or models must be designed specifically for them. The application of raw signal processing

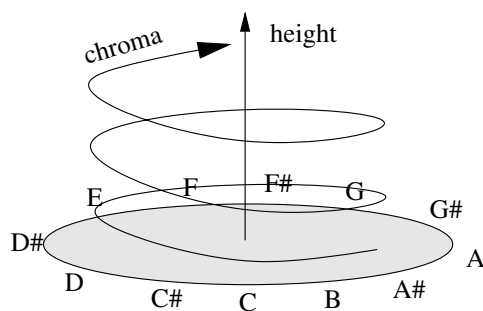
techniques is unlikely to meet with much success as many of the common engineering assumptions (for instance, the independence of multiple sources) are quite explicitly violated.

If the goal is a musical application, for instance musical transcription (as opposed to a signal processing application such as signal separation) then the output should be of a form suitable for human consumption which means that it should be expressed in terms of a musical structure. Frequencies are replaced by pitches, and the relative time-line is expressed in terms of musical *metre*. Hence a level of understanding of the signal is required in order to place it within a musical *context*. This process is often termed *musical context integration* and many of the methods described in this section are concerned with establishing various forms of musical context.

### 2.4.1 Pitch models

The perceptual models of pitch thus far encountered have psychoacoustic rather than musical motivations. Musical pitch, however, is rather different in that it has two dimensions — *chroma* and *height*. Pitch chroma is the label given to a musical note at a particular point in the scale and in Western music takes one of 12 labels – ‘A’ to ‘G’ plus five accidentals (sharps and flats). Pitch height increases with frequency, assuming the tone is harmonic with a perceived pitch at the same frequency as a sinusoid at the fundamental frequency. Pitch height tends to increase monotonically with pitch chroma, but after ascending by an octave, the chroma returns to the same value. Two pitches close in frequency sound similar as do two pitches whose separation is close to an whole number of octaves. A commonly used representation of this structure is the *pitch helix* (figure 2.3).

Shepard [134] presents a striking example of separating the dimensions of height and chroma through synthesis of a musical scale which appears to ascend indefinitely and yet which passes through the same point after each octave. The tones were constructed from a set of sinusoids whose amplitudes are determined by a spectral envelope which was a Gaussian with a fixed central frequency. The pitch chroma is increased in steps but the pitch height is approximately constant. The illusion is similar to that of the barber’s pole where a potentially infinite spiral is observed through a window of fixed



**Figure 2.3:** Helical representation of pitch. Two pitches are similar if they are close along the path of the helix or at the same chroma at different heights.

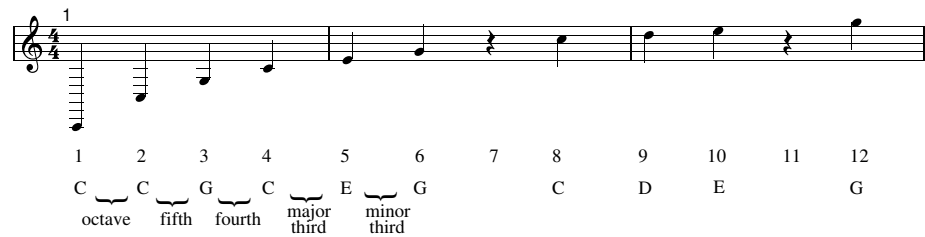
size, and the spiral appears to ascend continuously.<sup>8</sup>

Musicians perceive pitch differently than non-musicians. Musicians are likely to be aware of the importance of chroma and the intervals between notes, whereas non-musicians are less likely to appreciate the importance of octaves, although they should recognise the similarity of notes an octave apart.

The pitch helix is not restricted to the Western 12-tone scale. Most musical cultures exhibit a notion of *octave equivalence* regardless of the number of notes in their scale. Many also show preferences for scales which include consonant intervals of fourths and fifths (*i.e.*, with relative fundamental frequencies of 4:3 and 3:2). It is debatable whether our preference for octaves has a physiological or cognitive basis (*e.g.*, because of phase-locking in neurons or from familiarity, respectively) since it is almost universal across musical cultures, but differs for infants and older children [14].

Pythagoras constructed a 12-note scale based upon the concordant intervals of 3:2 and 4:3 which he determined by experimentation with strings of varying lengths [108]. The note intervals are derived from the ratio of powers of 2 and 3. An alternative to the Pythagorean scale is just intonation where the intervals are defined in terms of integer ratios, but tuning has to be performed relative to a particular key which makes the scale impractical for fixed-tuning instruments. The predominant scale in use today is the equal temperament scale which divides each octave into 12 logarithmically equal steps. This enables transposition into any key but fourths and fifths are no longer exact integer

<sup>8</sup>A demonstration of Shepard's tones may be found at <http://asa.aip.org/sound.html>



**Figure 2.4:** The relation between the harmonics of C and other notes. Common intervals are labelled.

ratios — they now take the values 1.335 and 1.498 rather than  $4/3$  and  $3/2$ . Despite the apparent simplicity of octave equivalence in terms of the concordance of vibrations in a 2:1 frequency ratio, exact octaves are actually perceived to be ‘too short’: the pitch separation must be increased by 10% of a semitone (0.6% frequency increase) to be perceived as a ‘true’ octave [42]. Pianos are tuned according to a stretched octave, due to the inharmonicity of the strings [42]. Burns [14] also suggests musicians’ preference for sharp rather than flat intervals as another explanation.

### 2.4.2 Harmony

The basis of harmony in Western music arose from the concordance of particular pitch combinations. It was found that various combinations of notes in the scale can be played together as a *chord* to create a harmonious effect, or other combinations could produce a sensation of tension and instability. One of the most harmonious and common chords is a major chord whose pitches are in the ratio 4:5:6. If each note is played by a tone which is a harmonic series with a fundamental frequency equal to the pitch then the frequency of every third harmonic of the first note will be the same as the frequency of every second harmonic of the third note, *etc.*. It is this coincidence of harmonics which makes the chord sound pleasant.

The chord also suggests the presence of a note further down the scale at a pitch of one quarter (*i.e.*, two octaves) below the first. This is the *harmonic root* of the chord. Figure 2.4 illustrates the harmonic relationships of the note ‘C’. Ten of the first twelve harmonics happen to coincide with the other notes in the scale. Several commonly

encountered musical intervals are shown in the figure, and the chord of C major appears in the 4th, 5th and 6th harmonics, as do its inversions (CEG, EGC, GCE).

If several notes are combined in this way then the perception can sometimes be of a single tone at the harmonic root. Pipe organs are capable of producing the sensation of a low pitch by combining several higher tones [118]. Debussy and his contemporaries explored the possibility of combining the sounds of several instruments playing simultaneously to create complex timbres rather than the impression of chords [27].

The identification of chords can be an ambiguous problem. The idea that a particular stimulus can be perceived in different ways by different subjects complicates both the construction of algorithms to detect chords and also the measurement of its success. It also provides an indication of the inadequacy of making inferences based upon a single frame of data. If the individual tones do not start simultaneously but have a small onset delay between them then this may yield enough information to resolve the ambiguity — the first note is heard for long enough to establish a context for itself. The onset of a second note now appears as a new note. If the context of the first note is not established before analysing the two tone complex then the detection may instead yield the harmonic root of the chord.

Techniques for resolving the inherent ambiguities of chordal detection are not common in the literature. Most pitch estimation techniques in fact ignore the problem entirely. The prevalence of chords in Western tonal music suggests that this high degree of structure should be capitalised upon, by modelling it explicitly and incorporating it into the musical context.

### 2.4.3 *Statistical studies*

A great deal of useful information about musical signals can be obtained from statistical studies of various musical phenomena. In a probabilistic modelling framework this information may provide suitable (frequentist) *a priori* probabilities for model parameters.

Mathematical analysis of musical scores between 1500 and 1960 [43] has highlighted variations in the frequency distribution of pitch over successive historical periods. Music before 1900 tends to have a roughly normal frequency distribution, whereas mid-

twentieth century music has a flatter distribution, showing a tendency to make more use of the extremes of the spectrum. Also the size of the transitions from one note to the next are much larger in twentieth century music. Studies on the duration of notes suggest that lengths less than 150ms are rare, and most notes which form part of melodic themes last between 150 and 900ms. Baroque music was found to have a median tone rate of 6.3 tones/s [11, p462]. Mellody and Wakefield have found the mean rate of vibrato for violin sounds to be 5.9Hz with a mean excursion of  $\pm 15.2$  cents [95].

Tempo analyses by van Noorden and Moelants [154] show a strong tendency for music randomly selected from the radio to have a tempo of around 130 bpm (beats per minute). Analysis of different musical styles show that jazz and baroque both have a much wider variation, with a more diffuse mode between 100–170 bpm, whilst modern dance music also has a mean of around 130 bpm but with a much smaller variance, and music from the Flemish hit parade shares similar characteristics. A potential explanation put forward for this ‘resonance’ around 130 bpm (or equivalently, the mean beat duration of 450ms) is that it corresponds roughly to the speed of human locomotion.

Rasch [119] conducted experiments to determine the degree of synchronisation between players in an ensemble, finding that the average delay between two instruments was between 30 and 50ms. Longer delays of between 100 and 200ms occurred occasionally due to mistakes, tempo changes or the first note after rests. Experiments also showed that melody instruments tend to lead, followed by bass, and middle registers are last. Longer onset times also appear to occur for string instruments than for wind instruments.

#### 2.4.4 *Streaming, grouping and segmentation*

In addition to being very structured in terms of the mathematical relations between scales and intervals, musical signals have many more levels of structure than can be observed at a microscopic level. The smallest musically significant unit is the note *event* — a pitch sounding at a particular instant in time. Many signal processing algorithms are concerned with producing a list of note events as their output. Conversely, many musical algorithms take a list of note events as their input and produce higher level inferences. Given a sequence of note events it is often desired to determine how these

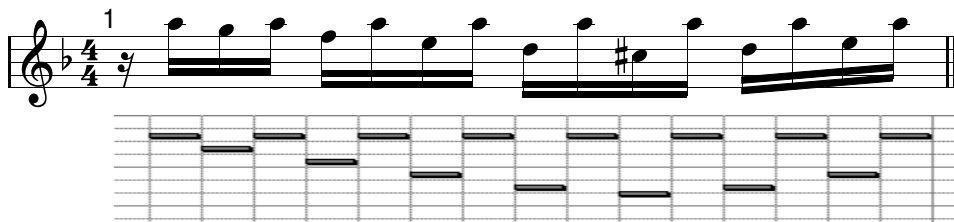


Figure 2.5: Fugue from Bach's Toccata and Fugue in Dm — an example of virtual polyphony.

are formed into perceptual streams such that a signal comprised of several instruments may be parsed into its constituent parts.

Generally, the grouping of notes into perceptual streams appears to take place according to the Gestalt principles of grouping outlined previously. There are two dimensions of grouping to be concerned with: horizontal and vertical.

### *Horizontal grouping*

Horizontal grouping or *sequential integration* refers to the connection of events at different points in time, *e.g.*, for the formation of melody. It is the mechanism by which individual melodic lines can be heard out from a dense background, or similarly how an individual conversation can be followed in a room full of interfering conversations (the *cocktail party effect*). Gestalt principles of continuity and similarity are important for horizontal grouping.

When we are presented with a monophonic melody, *i.e.*, with only one note sounding at a time, then we tend to perceive a pitch moving in time, rather than a set of separate pitches [24]. This percept can be defeated by the composer, however, as illustrated in figure 2.5. The musical excerpt is of a Bach organ fugue and it shows the use of *virtual polyphony*. A single melodic line alternates between a high and a low note, and as their separation increases the percept is of two separate melodic streams, which is also apparent from the graphical representation in the lower figure.

The analysis of the melodic 'surface' is an active area of research in musical psychology

and musicology.<sup>9</sup> Lerdahl and Jackendoff [77] propose a generative theory of tonal music which accounts for the perception of a musical extract by applying a hierarchical grouping. They propose several different types of structure which operate on low-level features such as pitch and note durations, in addition to metrical information (a higher level of time-variation) and more abstract musical concepts such as symmetry, inversion and variation of previously encountered passages. Each grouping is based upon a set of ‘well-formedness rules’ based upon musical intuition. The grouping boundaries are dictated by changes in the local structure. Cambouropoulos [15] extends Lerdahl and Jackendoff’s grouping rules to improve the detection of local boundaries in the melodic surface by adding an identity-change rule to prevent the segmentation between two identical elements. In a later paper [16] he proposes a clustering algorithm to detect high level structure, *e.g.*, the *rondo* form ABAC.

A key element of the horizontal organisation of music lies in its rhythm. The rhythmic structure of the music often ensures that the timing of musical events will be regular; once the tempo is known, it is possible to predict the likely points in time that the next event will occur. Chafe and Jaffe [18] used a metrical grid to investigate event timing and spot weak or missing events and developed a form of *rhythmic expectation*. They also propose the concept of a *cognitive flywheel* to process high level structures once the low-level musical context has been established, for instance to recognise features such as ‘thematic repeat’.

An interesting line of development of horizontal organisation draws on Artificial Intelligence (AI) principles. Streaming paradigms, as employed by Nakatani *et al.* [101], simulate Bregman-style streaming using a number of *agents*, each of which is dynamically allocated to a sound stream and is either a *watcher* or a *tracer*. Watchers use the DFT to find new streams and generate tracers which follow the sound stream until it ends. The output of each tracer is fed back to the input for signal cancellation.

### *Vertical grouping*

Vertical grouping or *simultaneous integration* refers to the grouping of concurrent events to produce a single sound percept, *e.g.*, a pitch or a chord [11, p30]. Common

---

<sup>9</sup>The term ‘surface’ is employed because melody is a visible feature of a musical extract. In contrast, the key of a melody is not directly observable and must be inferred.

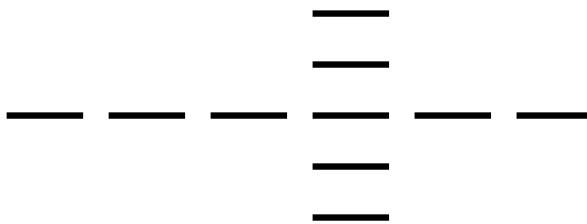


Figure 2.6: Ambiguity between horizontal and vertical grouping.

---

onset and harmonic cues are the dominant cues for the first level of grouping fundamental sound elements such as sinusoids, transients and noise into percepts of individual musical notes. A second level of grouping collects sets of notes together into chords. Much classical and modern music is chordal, *i.e.*, there is an implication of a particular chord, suggested by the choice of bass note and the combination of notes played by the other instruments.

Some composers, for instance Debussy, have blended together different instruments to produce chords which sound more like rich timbres, defeating the perceptual grouping mechanism by combining several different notes to create the percept of only one. Similar effects are evident in pipe organ stops where several pitches are combined to give the effect of a single low pitch (see §2.4.2). Horizontal grouping also comes into effect here to determine whether a note should be grouped into a chord or into a melodic stream. This is illustrated in figure 2.6: a note establishes a melody through horizontal grouping, but vertical grouping attempts to integrate that note into a chord.

Artificial Intelligence methods based upon blackboard systems are effectively employed by Ellis [34] and Martin [85]. This is an AI metaphor of a number of experts stood around a blackboard each adding their own expertise. These experts have ‘knowledge’ about physical sound production, auditory physiology and musical practice to effect a hierarchical grouping of successively FFT, tracks, partials, notes, intervals, chords and tonality. Good results are reported by Martin on four-voice Bach chorales, although subject to several restrictions. Nonetheless, the effectiveness of Martin’s method is proof of the necessity of incorporating both low level (signal) and high level (musical structure) information for the processing of musical signals.

## 2.5 Signal processing models

Signal processing provides various calculae for the description and manipulation of signals, and many of the techniques described in previous sections incorporate some element of signal processing. In particular, many signal processing techniques are concerned with the analysis of time and frequency variation of signals. This section will describe some of the broad domains of signal processing and their contributions to the analysis of musical signals. In many cases a distinction will be made between *parametric* and *non-parametric* models. The division between the two classes is blurred but in general the spirit of a parametric model is that an explicit signal model is constructed which is parametrised in terms of a set of features which are often physically meaningful. The subsequent process of inverting the model to find the parameters is termed *parameter estimation*. By contrast, non-parametric methods are often based upon signal transformations (e.g., between time and frequency domains), and do not necessarily have an explicit underlying signal model. Inferences are then made from this transformed representation. Additive models are described, encompassing sinusoidal and Fourier-based models which are highly relevant to musical signals by virtue of their formulation in terms of frequency components. Alternative methods of describing frequency characteristics are described, including time-frequency representations and source-filter models. Explicit models of musical instruments drawing upon physical modelling techniques are also described towards the end of the section.

### 2.5.1 Additive models

A large class of models can be described as *additive* signal models, in which the data is represented as the sum of a number of elemental components. Discrete data is expressed as the linear summation of a set of basis functions  $\{g_j[n]\}$  with amplitude coefficients  $\{b_j\}$ ,

$$d[n] = \sum_{j=1}^J g_j[n] b_j \quad (2.1)$$

or in vector notation

$$\mathbf{d} = \sum_{j=1}^J \mathbf{g}_j b_j \quad (2.2)$$

the basis functions can be reorganised into a *basis matrix*  $\mathbf{G}$

$$\mathbf{G} = [\mathbf{g}_1 \ \mathbf{g}_2 \ \dots \ \mathbf{g}_J] \quad (2.3)$$

$$\mathbf{b} = [b_1 \ b_2 \ \dots \ b_J]^t \quad (2.4)$$

$$\mathbf{d} = \mathbf{G}\mathbf{b}. \quad (2.5)$$

Estimates of the linear coefficients can be obtained from the pseudoinverse of  $\mathbf{G}$ ,

$$\hat{\mathbf{b}} = (\mathbf{G}^t \mathbf{G})^{-1} \mathbf{G}^t \mathbf{d}. \quad (2.6)$$

Of particular interest are sets of orthogonal basis functions, *i.e.*,  $\langle \mathbf{g}_j, \mathbf{g}_k \rangle = 0$ ,  $j \neq k$ , (where  $\langle \cdot, \cdot \rangle$  denotes the vector inner product) as in this case  $(\mathbf{G}^t \mathbf{G})$  is diagonal,

$$\mathbf{G}^t \mathbf{G} = \begin{bmatrix} \langle \mathbf{g}_1, \mathbf{g}_1 \rangle & & 0 \\ & \ddots & \\ 0 & & \langle \mathbf{g}_J, \mathbf{g}_J \rangle \end{bmatrix}, \quad (2.7)$$

and so

$$b_j = \frac{\langle \mathbf{g}_j, \mathbf{d} \rangle}{\langle \mathbf{g}_j, \mathbf{g}_j \rangle}. \quad (2.8)$$

This permits considerable simplifications as the amplitudes may be evaluated independently from the projection of the data onto each basis function, obviating the need for a matrix inversion.

Parametric additive models aim to find a minimal spanning set  $\{\mathbf{g}_j\}$  to economically represent the given data. Each model typically specifies a particular form of basis function whose parameters are to be determined. This type of model is described as *adaptive* since the parameters of the model are chosen specifically for each data vector [57]. Model overfitting should be avoided as it reduces the efficiency of the representation and its ability to generalise (the comparison of different models is discussed in chapter 3). Non-parametric methods may in be useful in determining the basis functions to use, *e.g.*, by peak-picking from a non-parametric transformation.

Non-parametric basis expansions generally take the form of a reversible transformation whose basis functions form a spanning set over the entire signal space. Since the basis functions are fixed, it is only the amplitude coefficients which have to be calculated and this calculation can often be highly optimised (*e.g.*, the Fast Fourier Transform).

Musical signals have strong features in the time and frequency domains and so basis function expansions which encapsulate time-frequency variation are popular. The most common class of basis functions used in audio applications are complex exponentials, which naturally arise as the eigenfunctions of linear systems.

### *Sinusoidal models*

Sinusoidal basis functions are closely related to the Fourier transform. The (continuous) Fourier transform is defined for a signal  $x(t)$  as

$$\mathcal{F}\{x(t)\} : X(\omega) = \int_{-\infty}^{\infty} x(t) e^{-j\omega t} dt \quad (2.9)$$

and its inverse is given by

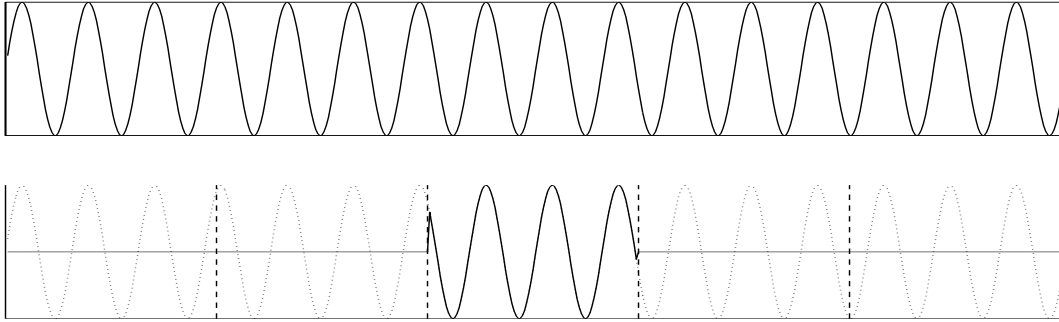
$$\mathcal{F}^{-1}\{X(\omega)\} : x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\omega) e^{j\omega t} d\omega. \quad (2.10)$$

For a sampled signal  $x[n]$  the Discrete Fourier Transform (DFT) is calculated for  $0 \leq k < N$

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-\frac{2\pi knj}{N}} \quad (2.11)$$

$$x[n] = \frac{1}{N} \sum_{k=0}^{N-1} X[k] e^{\frac{2\pi knj}{N}}. \quad (2.12)$$

The inverse DFT is also therefore a basis expansion of complex exponentials with amplitudes given by the DFT coefficients  $X[k]$ . The frequency of the  $k^{\text{th}}$  component is  $\omega_k = 2\pi k/N\Delta t$  where  $\Delta t$  is the sampling interval and  $N$  is the number of data points used in the analysis. The frequency resolution therefore increases as the duration  $N\Delta t$  of the data increases. Orthogonality of the complex exponentials can be seen from the inner



**Figure 2.7:** Windowing applied to a sinusoid which extends over all time to perform a time-localised Fourier transform.

product of  $g_1(t) = e^{j\omega_1 t}$  and  $g_2(t) = e^{j\omega_2 t}$

$$\langle g_1(t), g_2(t) \rangle = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} e^{j(\omega_1 - \omega_2)t} dt \quad (2.13)$$

$$= \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} \cos(\omega_1 - \omega_2)t + j \sin(\omega_1 - \omega_2)t dt \quad (2.14)$$

For  $\omega_1 \neq \omega_2$  the periodic nature of the real and imaginary components means that the integral will tend to zero as  $T$  tends to infinity. If  $\omega_1 = \omega_2$  then the integral will be finite and will tend to unity in this instance, hence for a sufficiently large analysis interval the complex exponentials form an orthogonal basis set. For the discrete form this requires  $N\Delta t \gg 2\pi/\omega_{\min}$ , where  $\omega_{\min}$  is the lowest frequency in the data.

Musical signals are characterised by significant variations in their amplitudes and frequencies over time. The Fourier transform is defined for two-sided signals of infinite duration, which is inappropriate for musical signals. A variation on the DFT is the Short-Time Fourier Transform (STFT) which applies a window  $w[n]$  before calculating the transform,

$$X[k, m] = \sum_{n=-\infty}^{\infty} w[n] x[n + m] e^{-\frac{2\pi knj}{N}} \quad (2.15)$$

The window is non-zero only within the interval  $[-N/2, N/2]$ . The subscript  $m$  is now used to indicate a time-localised transformation — only the part of the signal in the interval  $[m - N/2, m + N/2]$  is analysed, as shown in figure 2.7.

It is very inefficient to calculate the STFT for all time steps  $m$  and where the time-variation of the parameters is slow compared to the window length it is usually effective to subsample the STFT at intervals of length  $L$  (the *hop*) by only evaluating it at these points. The STFT can be reformulated to make this explicit:

$$X[k, i] = \sum_{n=-\infty}^{\infty} w[n] x[n + iL] e^{-\frac{2\pi knj}{N}}. \quad (2.16)$$

One of the most useful properties of the Fourier transform is the transformation of convolution in the time-domain to multiplication in the frequency domain,

$$\mathcal{F}\{f(t) * g(t)\} = F(\omega) G(\omega). \quad (2.17)$$

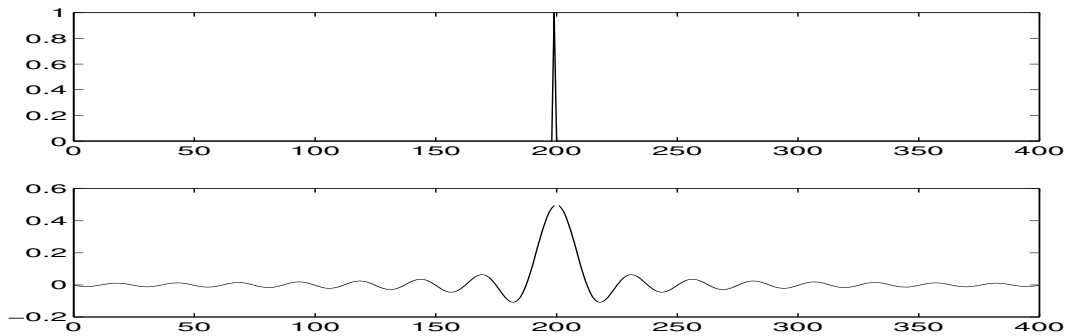
The effect of applying a symmetrical window before the transformation becomes equivalent to multiplying the FT of the window with the FT of the infinite time signal. The FT of a sinusoid of frequency  $\omega_0$  is an impulse at  $\omega = \omega_0$ . However, when a window is applied the FT will suffer a loss of sharpness. In the simplest case of a rectangular window  $w(t)$  the FT is a sinc function (figure 2.8)

$$w(t) = \begin{cases} 1, & -T/2 < t \leq T/2 \\ 0, & \text{otherwise} \end{cases} \quad (2.18)$$

$$\therefore W(\omega) = \int_{-T/2}^{T/2} e^{-j\omega t} dt \quad (2.19)$$

$$= T \frac{\sin(\omega T/2)}{(\omega T/2)}. \quad (2.20)$$

The convolution with the windowing function introduces artifacts into the short-time signal spectrum which are undesirable. Sidelobes are introduced which complicate the extraction of spectral peaks and the main lobe may spread over several frequency bins which makes it more difficult to resolve closely-spaced sinusoids. A variety of windowing functions are commonly used to compensate for these drawbacks which trade-off between main lobe width and sidelobe amplitudes. The Hamming window for instance has better sidelobe characteristics than the rectangular window but has a broadened main lobe. Depalle and Hélie [26] propose a window without sidelobes which consists of a Gaussian multiplied by a triangular window raised to a power. The triangular part



**Figure 2.8:** The Fourier transform of a sinusoid of infinite length is an impulse (upper plot). When convolved with a window the Fourier transform is multiplied by the transform of the window itself which for a rectangular window is a sinc function (shown in lower plot).

controls the main lobe width whilst the Gaussian part controls the asymptotic frequency behaviour away from the main lobe.

In the subsampled STFT formulation, artifacts may be introduced into the resynthesised signal when the frequencies or amplitudes in the data signal are time-varying. This can result in discontinuities at the frame boundaries which sound harsh when played back (in the case that the hop size is the same as the window size,  $L = N$ ). These artifacts can be reduced by making the hop size less than the frame size, typically  $N/2 \leq L \leq N$ . A tapered window such as a triangular window is applied in the STFT. The overlapping tapered windows, when reconstructed and added together yield a smoother parameter variation. This method is termed overlap-add synthesis and provides improved audio quality at the expense of extra computation.

### *Resolution and interpolation*

The resolution of the DFT is related to the number of points in the transform, since the number of points used in the transform equates to the number of basis functions used in the reconstruction. The time-varying behaviour of the signal precludes us from choosing an arbitrarily large frame size as this will introduce smearing of the spectral peaks. The number of data points is therefore governed by the amount of time over which the frequencies and amplitudes are almost constant. For musical signals this interval is usually chosen to be around 20–30ms (*e.g.*, [54]). Using the STFT, the resolution can

be increased further by zero padding, *i.e.*, appending zeros to the data to increase the number of basis functions in the transform whilst effectively keeping the number of data points constant. Increasing the amount of zero padding does not change the shape or size of the main or sidelobes in the spectrum since these are a function of the window.

Although the resolution of the STFT and DFT is finite, it is possible to improve upon the resolution of frequency estimates generated from them. For instance, if the discrete signal  $f[n]$  is a complex exponential with unit amplitude and a frequency which lies between bins  $m$  and  $m + 1$ , such that  $0 \leq \Delta \leq 1$

$$f[n] = \exp \left[ j \frac{2\pi n}{N} (m + \Delta) \right] \quad (2.21)$$

then the DFT is evaluated at bins  $m$  and  $m + 1$  to yield

$$F[m] \approx \frac{N}{\pi\Delta} \sin(\pi\Delta) \quad (2.22)$$

$$F[m + 1] \approx \frac{N}{\pi(\Delta - 1)} \sin(\pi(\Delta - 1)). \quad (2.23)$$

These can be combined to give an estimate of  $\Delta$

$$\hat{\Delta} = \frac{|F[m + 1]|}{|F[m]| + |F[m + 1]|} \quad (2.24)$$

An interpolation technique which utilises the change of phase around the DFT peak is described in [64, 105] (in the above example  $F[m]$  and  $F[m + 1]$  have near-zero imaginary values, but with opposite signs). Depalle and Hélie [26] perform a Taylor expansion of the Fourier transform of the window function in order to produce a least-squares estimate of the spectral peak. An interpolator employing a parabolic approximation to the spectral peak is suggested by McIntyre and Dermott [92]. Macleod [81, 82] interpolates the discrete Fourier Spectrum (DFS) between the spectral peak and its two neighbours using amplitude and phase information, since at least 85% of the energy (sum of squares) of the DFS of a cisoid is contained within these three sample points. This is also extended to a five point interpolator and an iterative method for multiple tone estimation is proposed, based upon successive parameter estimation, reconstruction and subtraction.

Another problem which may arise from the finite resolution and smearing of peaks is that a frequency component may be obscured by another which is close in frequency

and of greater amplitude. Analysis-by-synthesis methods (such as that proposed in [81, 82]) are able to resolve close peaks by subtracting the reconstruction of the dominant frequency component and analysing the residual waveform. Parsons [106] uses knowledge of the shape of the smearing function (*i.e.*, the window function) to resolve overlapping peaks. The peak shapes are calculated from the initial frequency estimates and subtracted from the spectrum. The previously masked peaks are then discernible in the residual spectrum.

### *Parametric and non-parametric interpretations*

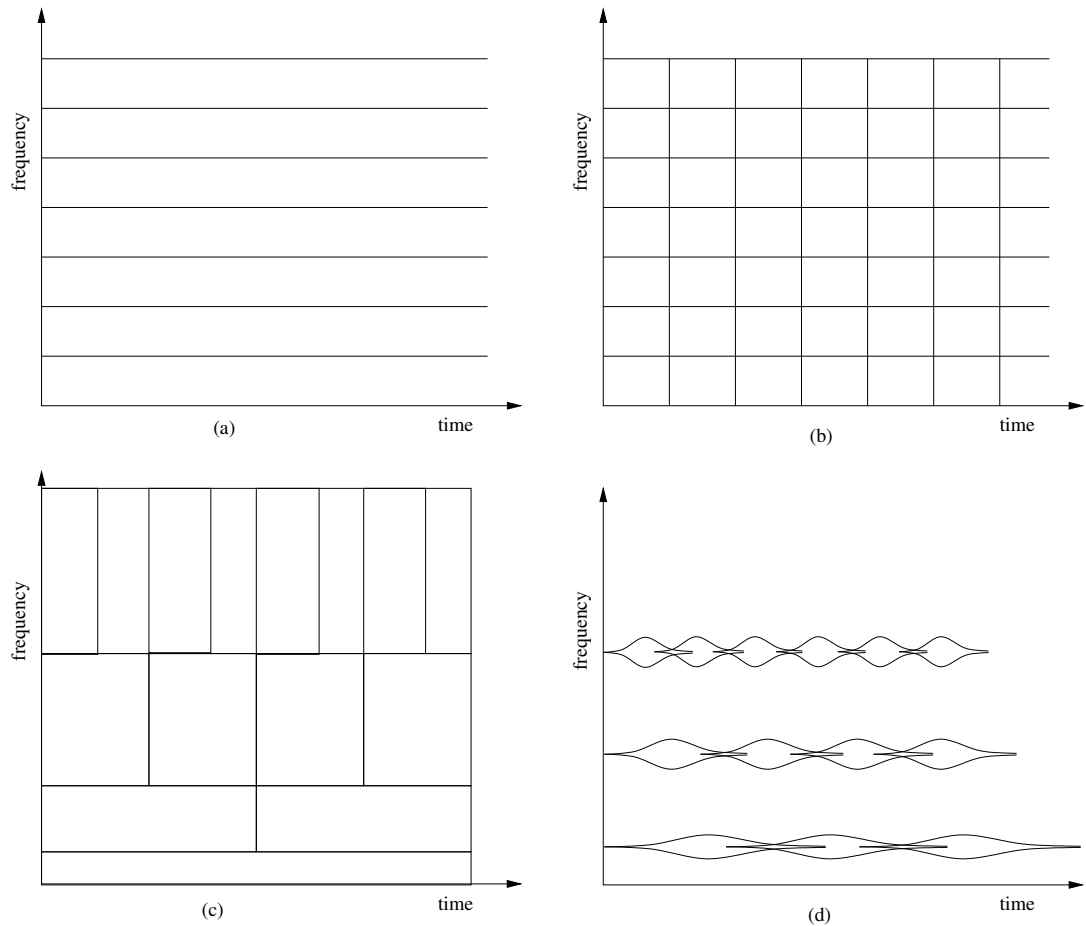
Sinusoidal and Fourier-based models can be either parametric or nonparametric. In some applications the Fourier transform is regarded as being a bank of bandpass filters. This interpretation is particularly common in audio coding applications. The phase vocoder [40] is one such example, which is based upon the STFT and produces a subband decomposition. Transform coders, for instance as used by the MPEG-1 standard [2, 10] rely on a Fourier transform as an initial processing step.

Terhardt [145] makes the Fourier transformation into a parametric representation by letting the complex amplitudes and phases of each component be variable over time, allowing for frequency modulation and birth and death of frequency tracks. However there are still an infinite set of parameters. This spectrum is known as the *Fourier Amplitude Spectrum*.

Sinusoidal models can be written in a parametric form where the data is represented using a parsimonious set of sinusoids, and the STFT is employed to produce estimates of the most salient frequency components by peak-picking, *e.g.*, [90] (see also §2.5.3). It is an important distinction that the Fourier transform can be viewed as a transformation or as a parameter estimator — this notion will be revisited in chapter 6.

### *Time-frequency representations*

Other time-frequency basis functions are commonly used, corresponding to different tilings of the time-frequency plane, as illustrated in figure 2.9. Each has particular characteristics in terms of the resolution in the time and frequency domains. For instance the STFT can have good frequency resolution at the expense of time resolution. This



**Figure 2.9:** Different tilings of the time-frequency plane. The Fourier transform is shown in (a) which requires a signal of infinite extent in the time domain. The short-time Fourier transform is shown in (b) which uses an analysis window of a fixed length to obtain a degree of time-localisation. A multiresolution tiling is shown in (c) where each frequency region is analysed using a different window length to yield better time-localisation (*e.g.*, wavelets). Figure (d) shows schematically a decomposition using Gabor kernels, where each sinusoidal basis function is windowed with a Gaussian with parameters corresponding to scale and translation (the frequency is fixed however).

may be significant if it is required that sharp instrument attacks are to be represented. These cannot be modelled well if the moment of attack does not fall upon the frame boundary.<sup>10</sup> Much better time-localisation can be achieved through the use of wavelet basis functions [83]. The time-localisation is obtained at the expense of frequency localisation. For the tiling shown in figure 2.9(c) resolution in time is much improved, but each tile spans an octave which is of limited use for characterising complex musical signals. Such tilings have found application to the modelling of wideband frequency content and residual waveforms [57].

Evangelista [37] acknowledges the limitations of conventional wavelet bases in respect of musical signal analysis, namely the octave frequency resolution of dyadic wavelets and the fact that signal pseudo-periodicity is not exploited. He extends the definition of the wavelet transform to other bases which are better able to exploit periodicity and time-varying behaviour using frequency warping techniques. Frequency warping is used to re-map the frequency plane such that, for instance, inharmonic sounds can be transformed to a harmonic series. Newland [104, 103] generates new wavelet basis functions (namely harmonic and musical wavelets) which are more suited to musical signals. Wavelet methods are *multiresolution* analysis techniques where each frequency region is analysed over a different time scale. This can be applied to sinusoidal models to circumvent the need for analysis windows long enough to capture the lowest frequencies in the signal. Instead, higher frequencies are modelled with progressively shorter windows so that the model has increased time-resolution at high frequencies. This principle is employed for coding to produce an efficient representation of wideband audio with improved time-localisation for transients [78].

An alternative representation in the time-frequency plane is the Gabor atom. A Gabor atom is defined as a sinusoid at a frequency  $\omega_0$  with a Gaussian amplitude envelope of a particular *scale* at a particular *epoch* in time

$$g(t, t_0, \sigma^2, \omega) = \exp \left[ -\frac{(t - t_0)^2}{2\sigma^2} + j\omega t \right] \quad (2.25)$$

The Fourier transform can in fact be considered as a special case of the Gabor expansion where  $\sigma \rightarrow \infty$ . Dörfler and Feichtinger [31] discuss how Gabor kernels can be adapted

---

<sup>10</sup>Sinusoidal models are in any case not ideal for the representation of percussive attacks which tend to have a damped oscillatory characteristic.

to particular musical instruments. Gabor atoms have also been adopted for *granular* sound synthesis as a flexible means of synthesising musical sounds [136].

There is a family of bilinear (*i.e.*, quadratic) time-frequency representations built upon the Wigner-Ville distribution (WVD) which is the Fourier transform of the signal's local autocorrelation function,

$$W(t, \omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} s^*(t - \frac{\tau}{2}) s(t + \frac{\tau}{2}) e^{-j\omega\tau} d\tau. \quad (2.26)$$

For data comprised of sinusoidal components the WVD has components at the frequencies of these components, but it also includes cross-terms at the midpoints of each pair of frequencies. A transformation kernel specific to the type of signal under consideration is used to minimise these effects. An adaptive time-frequency kernel optimised for musical signals is proposed by Sterian and Wakefield [142]. Their modal time-frequency distribution belongs to Cohen's class of bilinear time-frequency distributions [21] and applies frequency-dependent smoothing to achieve a multi-resolution analysis whilst also suppressing the cross-terms. The cross-term at the frequency  $(\omega_j + \omega_k)/2$  oscillates at a frequency of  $\omega_j - \omega_k$  and so a low-pass filter is used with a breakpoint just below the closest expected partial separation to minimise these effects without affecting the more slowly varying partials. Mellody and Wakefield [95] show how the modal distribution can produce high resolution estimates of amplitude and frequency variation for vibrato in a violin signal.

### 2.5.2 Source-filter models

A physically-motivated approach to signal modelling represents the signal as the excitation of a resonant cavity. Such methods are commonly referred to as source-filter methods, and are popular for the coding of speech signals, where the speech production method is modelled as a glottal excitation passing through a vocal tract filter. The excitation takes the form of a pseudo-periodic stream of either pulses or noise and the vocal tract has a response with several peaks (*formants*) which vary over time to produce different vowel sounds. Fant's vocal production model [38] switches the excitation between pulse and noise waveforms to reproduce both voiced and unvoiced speech; this technique has formed the basis of early speech synthesisers. Linear prediction (LP) of

speech is built upon Fant's model [84]. The pulse or noise excitation is passed through (slowly) time varying linear filters, and hence the output at time instant  $n$  can be predicted by a linear combination of the preceding values (via the  $z$ -transform of the filter function). Linear prediction is extensively employed for speech coding (linear predictive coding, or LPC) as the essential elements of the signal (voicing state, pitch and vocal tract response) may be represented much more efficiently than the raw waveform data.

Closely related to LPC is the autoregressive (AR) signal model. Autoregressive models seek to model a random process whilst LP methods seek to estimate it [148]. Several techniques for the restoration of degraded audio signals employ AR models as an underlying signal representation [54, 53, 126, 152].

An AR model can be expressed as a time series

$$s[n] = \sum_{i=1}^P s[n-i]\alpha_i + e[n] \quad (2.27)$$

as the weighted sum of the  $P$  previous inputs and an excitation term (generally Gaussian). The AR model is also often referred to as the all-pole model; the poles may appear close to the unit circle (for near-periodic signals) or close to the origin (for noise-like signals). Autoregressive moving average (ARMA) models include zeros as well as poles. This can result in a much more compact representation for complex systems, but the estimation of ARMA parameters is more difficult [50, 148]. However, an AR model extended to a higher order can be used to approximate a given ARMA model. Typical musical signals may require an AR model of order 30-100 [53].

In the taxonomy of models, AR models may be described as semi-parametric. One motivation for their use is their ability to represent time series. The poles describe the frequency domain behaviour of the signal but not in a way that makes it simple to extract pitch information, particularly for polyphonic sources. The AR model itself is not an ideal representation for periodic data, particularly where the excitation is periodic (rather than a noise source, *e.g.*, as with voiced speech). Vermaak *et al.* describe an extension to the AR model which explicitly models periodic variations [156].

Another method based upon the source-filter model is *homomorphic deconvolution* or *cepstral analysis*<sup>11</sup> [8]. This method relies on the convolution property of the Fourier

---

<sup>11</sup>Or more correctly, cepstral *analysis*.

transform whereby the process of convolving the excitation with the impulse response of the resonant cavity can be effected by multiplying the respective Fourier transforms in the frequency domain.

The logarithm of the Fourier spectrum is equal to the sum of the logarithm of the excitation spectrum and the logarithm of the response spectrum.

$$x(t) = e(t) * f(t) \quad (2.28)$$

$$X(\omega) = E(\omega) F(\omega) \quad (2.29)$$

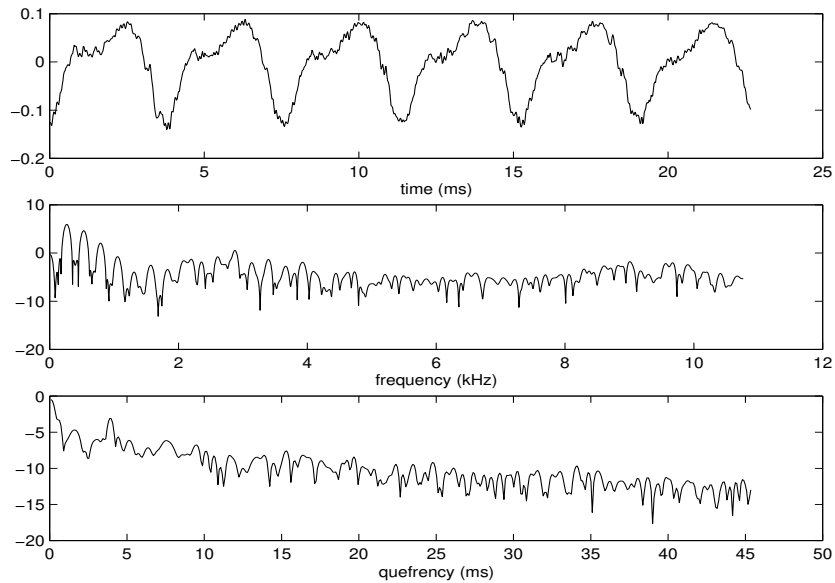
$$\therefore \log[X(\omega)] = \log[E(\omega)] + \log[F(\omega)] \quad (2.30)$$

For voiced speech the excitation spectrum will be a series of impulses whilst the vocal tract response spectrum will be a relatively smooth function with several spectrum peaks at the formant frequencies. By taking the Fourier transform of the log spectrum, these effects can be separated. The top plots in figure 2.10 shows the waveform of a voiced speech signal, the logarithm of the spectrum and the Fourier transform of the log spectrum (the *cepstrum*). The impulses in the excitation give rise to a cepstral peak at around 4ms (in the *quefrequency* domain) corresponding to the pitch, whilst the slowly varying characteristic of the vocal tract are represented by the low quefrequency values. The lower plots show the cepstrum of an unvoiced speech signal which has no peak at the pitch period.

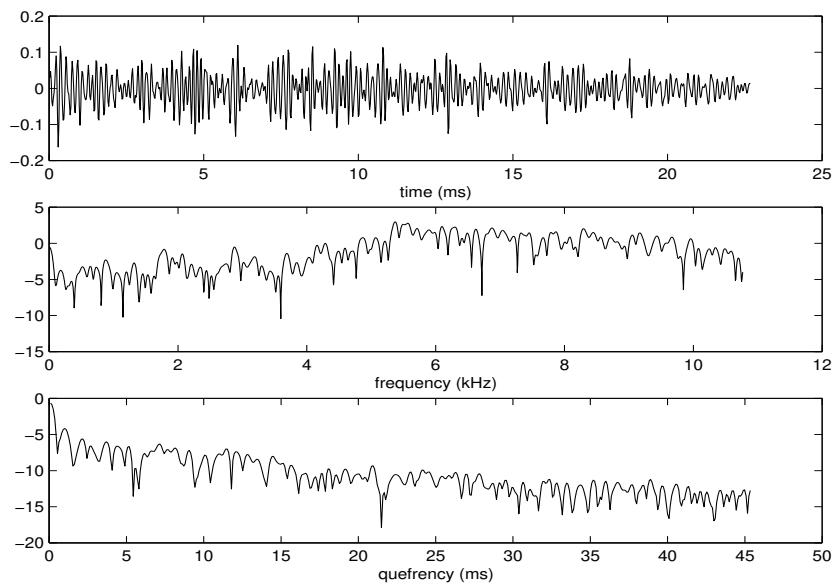
Cepstral methods are popular for front-end processing in speech recognition applications in order to isolate speech formants. Cepstral methods are not in general robust to additive noise, as the logarithm operation used to separate the cepstra produces noise correlated with the signal [4].

### 2.5.3 Instrument models

One approach to the analysis of musical signals is to explicitly model musical instruments. In contrast to psychoacoustic models, it is the generating rather than receiving mechanism which is modelled. There are three main motivations for this type of modelling. One is to produce invertible models which are able to extract information about the instrument and the notes it is playing, the second to synthesise realistic sounding instruments, and the third to gain a better understanding of the physics of instruments in order to improve their design.



(a) Voiced



(b) Unvoiced

**Figure 2.10:** Cepstral analysis of voiced and unvoiced speech signals. In each figure, the top plot is the waveform in the time domain, the middle plot is the logarithm of the Fourier spectrum and the lower plot is the cepstrum. See the text for further discussion.

### *Sinusoidal instrument models*

Many of the techniques for analysing musical signals can be described as ‘analysis by synthesis’. A parametric or semiparametric signal model is used to represent the data, and a parameter estimation stage calculates parameter values which are then used to synthesise a signal. The parameters are chosen to minimise the error, usually in a least-squares sense. The most common class of analysis models are those based upon a sinusoidal representation, building on the framework of McAulay and Quatieri [90] (hereafter referred to as M+Q models). Their model was originally intended for speech but has found applicability in many fields. Speech production is modelled using a source/filter paradigm — a glottal excitation waveform is passed through a time-varying vocal tract filter. The model is intended for perfectly voiced speech, and so the signal is expressed as the sum of sinusoids whose parameters are constant over an analysis frame.

The frame length is chosen to be a multiple of the pitch period as this minimises the spectrum sidelobes. The method is *pitch synchronous* as it requires a rough estimate of the pitch to construct the analysis window. The estimates of the amplitudes and phases of the constituent sinusoids are taken from the peaks of the short-time Fourier transform (STFT). The subsequent synthesis stage uses an overlap-add technique to achieve parameter interpolation between frames by overlapping each analysis frame. The estimation of phases used a cubic polynomial to achieve phase unwrapping, where the maximally smooth phase trajectory is chosen.

There are many refinements and extensions to the M+Q model in the literature, including scope for signal transformations [113], pitch estimation [91], use of quadratic phase variation and improved residual coding [29], and representation in terms of instantaneous amplitudes in order to obviate the need for phase unwrapping [79].

Pitch-synchronous methods assume slow time variation of the frequency components in order to use a rough estimate of the pitch at that point in time. The advantage of a pitch synchronous approach is that if the window length is an integer multiple of the pitch period then this minimises the artifacts introduced by the window function. A major limitation of pitch-synchrony, however, is that it is suited only to the analysis of monophonic sounds, *i.e.*, those comprised of a single set of related harmonics. For sounds composed of unrelated notes from two different instruments, there can be no

concept of pitch-synchrony since each has its own pitch. Attempts to find a single pitch may yield that of the strongest of the two sources or the *harmonic root* of the two harmonic series.

### *Sinusoidal plus residual models*

Sinusoidal models are able to represent voiced speech and tuned musical instruments quite well. They are able to capture the steady state frequency components of the signal and when the signal is reconstructed, the sounds are usually very intelligible and recognition of the speaker or type of instrument is possible. The reproduction will not reproduce all perceptual characteristics of the original signal, however. For percussive instruments, the transients which accompany the attack of each note are of great perceptual importance, but these cannot be accurately represented in a sinusoidal model since the sinusoidal components are not localised in time and amplitude variations are constrained to occur at time boundaries. Similarly, in unvoiced speech or instruments with a noise component such as bowed strings or woodwind, the noise component is essential, or at least important, to the characteristic and intelligibility of the sounds. There are several popular models that are used to model the *residual* waveform which remains following extraction of the sinusoidal components.

Serra (amongst others) develops techniques for extending the sinusoidal model to include a stochastic component [132, 133]. The residual signal is assumed to be stochastic (the sinusoidal components are assumed deterministic) and is modelled as filtered white noise,

$$\hat{e}(t) = \int_0^t h(t, t - \tau)u(\tau)d\tau \quad (2.31)$$

with a slowly time-varying filter  $h(t, \cdot)$  and white noise process  $u(t)$ . The estimated sinusoidal components are resynthesised and the STFT is taken and subtracted from the original STFT to yield the residual spectrum. This is simplified by calculating a piecewise-linear, continuous approximation to the spectrum. The stochastic part is then resynthesised by taking the inverse Fourier transform of the residual spectrum multiplied by a random phase term, since only the amplitude characteristics of the noise spectrum are considered perceptually important in the reconstruction. Goodwin [56] proposes a noise model based upon perceptual properties in which he uses equivalent

rectangular bandwidths (ERBs) to obtain a more perceptually salient noise representation.

Rodet [124] develops a different sinusoidal plus residual model. The sinusoidal estimation stage has several refinements over the M+Q method. Most notably the problem of partial tracking, where trajectory information is included, is formulated in terms of transition probabilities of a hidden Markov model, and is solved using the Viterbi algorithm. Rodet suggests approximating the spectral envelope of the residual with an autoregressive time series (an all-pole filter) or using cepstral techniques to separate the effects of the excitation (which is a noise signal) and the (vocal tract) filter response.

Sinusoidal models assume a quasi-periodic excitation where the excitation frequency variation is piecewise constant and slowly varying. This assumption can be violated by certain signals, for instance in speech where *vocal fry* can occur at low frequencies. The effect is due to the ‘chaotic’ nature of the glottal excitation, and means that the spacing of the glottal impulses will not be constant. Rodet notes that sinusoidal plus stochastic residual methods are not suited to representing this type of signal.

The stochastic residual model is good for noise-like characteristics, and when the spectral envelope is parametrised, as in the models of Rodet and Serra, it can be transformed in the time or frequency domains along with the sinusoidal components to achieve time-stretching or pitch-shifting. It is not able to represent transient events which are important for percussive instruments, and so Rodet employs a dictionary of elementary waveforms to decompose the residual into waveforms which are well-localised in time. Time-localised characteristics are preserved in the high-resolution matching pursuit algorithm that he uses to perform the decomposition.

The sinusoidal plus residual models are suitable for analysis and synthesis. The representation is also becoming popular for audio coding, and is incorporated into the MPEG-4 audio coder [112, 68]. These models are intended to capture the important perceptual characteristics of musical signals. The residual components, however, are not necessarily physically meaningful — the residual spectral envelope carries no intuitively useful information about the nature of the instrument (in contrast to the more meaningful parameter of pitch).

### *Tracking*

Much of the musical signal processing literature focuses upon the problems of modelling audio signals over short durations (of the order of 20ms). It is common to break the signal down into frames of this length and process each frame separately, perhaps with a small overlap to reduce the artifacts on playback (see §2.5.1). However, it is also important to consider the longer term variation of musical signals. Many authors consider musical signals to be slowly varying sets of sinusoids with more significant changes in the data represented by the birth and death of sinusoids. McAulay and Quatieri [90, 113] advocate this method and implement a scheme to track changes in frequency between adjacent frames whilst allowing for the birth and death of sinusoids. This yields a set of frequency tracks which may then be used for coding, pitch/time transformations or pitch detection via a least squares fit of the set of partials [91]. They note, however, that the identification of two concurrently sounding pitches is inherently ill-conditioned due to the possibility of closely-spaced harmonics between the two notes. Godsill employs a similar birth/death sinusoid tracking approach for the detection of wow in gramophone recordings, with an explicit dependence on the cyclic frequency variation parameter [54, 49].

Doval and Rodet [33, 124] adopt a more explicit tracking model formulated in terms of a probabilistic harmonic association problem. The set of frequency observations from the STFT peaks are modelled as a set of harmonically related partials plus some ‘noise’ partials. A hidden Markov model is constructed to emit a frequency observation at each time step with a transition probability that is low for small frequency changes, and the optimal state sequence is found via the Viterbi algorithm. Sterian *et al.* [140] formulate a similar association problem, using cues such as common harmonicity, common onset and statistics of the partials, for grouping. They employ a Kalman filter to model the power and frequency variation of each partial.

Fernández-Cid and Casajus-Quirós [39] propose to build up a history for a note before it is accepted. This increases robustness against signal transients which cause spurious detections and acknowledges the importance of context. However, the method does not explicitly track note frequency changes and would be unsuitable for rapidly varying pitches. An *ad hoc* rule-based system is used for validating note candidates on the basis of their fundamental frequency and harmonic amplitudes; this helps to suppress

the detection of harmonic roots by ensuring that the lower harmonics have appreciable amplitudes. The pitch tracking is a post-processing stage which requires the note candidates to be present for at least 90ms of the preceding 150ms.

Signal processing approaches to frequency tracking all rely upon the slowly varying characteristics of musical signals. Such behaviour is evident in simple cases, but for faster note variations or polyphonic signals the added complexity invalidates the slow variation assumption over longer time scales. An alternative approach is to explicitly use musical knowledge to track the variation. Discontinuities in frequency tracks are introduced when a note is moved between discrete pitches. These changes generally correspond to the discrete pitches of the musical scale. Kashino and Murase [71] create a Bayesian belief network to represent the probabilities of different note transitions, obtained from the analysis of a database of musical scores. They report a much lower error rate with the introduction of note transition information and further decreases with musical rôle consistency measures (*e.g.*, an instrument playing the bass line will not cross into the pitch regions of the melodic part). Unfortunately, their work does not address the problem of pitch estimation in each frame, as they assume that the multiple fundamental frequency estimation problem is solved. Their method does serve, however, to highlight the importance of musical context and structure within an application designed to deal with musical signals.

### *Physical instrument models*

Physical modelling of musical instruments is an active area of research whose aim is to capture the important characteristics of acoustic instruments and control them in real-time. Conventional techniques employed in musical synthesisers for synthesising realistic instruments are based on *sample and synthesis* or *wavetable* methods. The instrument is sampled and the sustained part of the waveform is looped. The sound is played back at different pitches merely by playing the waveform back at different rates and each instrument must be sampled at several points over its natural pitch range, called *multisampling*. Only simple modifications to the sound are possible with sample and synthesis (*e.g.*, envelope manipulation and filtering), and so the synthetic instrument lacks expressiveness [88].

Physical modelling overcomes this limitation by constructing a model to simulate the

actual physical characteristics of the instrument. A popular class of physical models, summarised by Smith [137] are based upon digital waveguides. These simulate distributed media such as vibrating strings and bores (in contrast to ‘lumped’ elements such as point masses connected by springs and dampers) which act as transmission lines for forward and backward travelling wave components. Losses can be introduced into the waveguides to simulate effects such as friction and drag, and scattering junctions join together sections with different impedances, introducing reflections. Smith shows how these simple building blocks, may produce realistic sound synthesis for instruments such as clarinets and bowed strings.

Cook [22] describes how to harness physical models for the automatic production of sound effects. Despite the huge advances made in computer animation and the advent of feature-length computer-generated visuals, the sound backdrop is usually created manually by the Foley artist.<sup>12</sup> Techniques which have been around for around 50 years are still in use — wooden chairs to simulate creaking floorboards, coconut shells for horse steps, and baseball bats striking frozen chickens for kung fu films. Cook presents physical models based upon a taxonomy of elemental sound types such as blowing, striking, rubbing, scratching, *etc.*, which may be combined and scaled to simulate other sounds. For instance a gourd model (of grains within a gourd) can be used to model tambourines, shakers and feet on gravel. The technique, called physically informed stochastic event modelling (PHISEM), models the collision of individual ‘grains’ on the gourd as an exponentially decaying noise signal and the output is fed to a filter which represents the resonances of the gourd, *e.g.*, hollow cavities will have a single resonance peak, whereas sleighbell jingles can be modelled with a more complex resonance structure. The model was motivated by stochastic simulations of collisions obtained using Newtonian mechanics.<sup>13</sup>

Such physical instrument models are very powerful for synthesis but due to their complexity and non-linearities they are generally not invertible, making them of limited use

---

<sup>12</sup>Named after Jack Foley, the first practitioner of the technique, who calculated that he had walked over 5000 miles in a studio re-recording the sound of footsteps.

<sup>13</sup>The potential for sound effect automation is demonstrated in a short animation, “Music for unprepared piano”, where balls of varying mass and hardness are projected onto the strings of a piano. The sound and animation are generated from object models given the raw composition data of the speed and mass of the balls.

for analysis. Even if it were possible to invert such a model, there is no guarantee that the resulting parameters would be perceptually salient.

Physical models can also be of great use in the understanding of the sound generation mechanisms of musical instruments. The physics behind the behaviour of vibrating strings, air columns, plates and membranes is well understood (see [42] for a review). This knowledge can be applied to improve the design of instruments; Kausel [72] describes a technique that assists in the design of trumpets, shortening the design cycle required to obtain a desired set of trumpet characteristics.

### *Timbre*

Another aspect of instrument modelling is concerned with the representation of timbre. Timbre is rather nebulously defined by the American National Standards Institute as,

“...that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar.” [3]

For several reasons it is difficult to construct a precise mathematical definition of timbre. Timbre is a multidimensional property and certain aspects of timbral perception may be subjective and qualitative. Ohm’s acoustical law states that it is the relative amplitudes of harmonic partials rather than their relative phases which determine timbre. It is generally claimed that the ear is insensitive to phase<sup>14</sup> and examples by McAulay and Quatieri [90] and Risset and Wessel [122] suggest this is largely true for speech intelligibility and timbral perception respectively. Phase changes also occur in listening environments from surface reflections, and these do not appear to have any bearing on perceived timbre. Hence timbre cannot be dependent solely on the waveform shape.

Whilst spectral amplitudes are clearly an important aspect of timbre, it also depends greatly on time-domain variations of sound. Percussive sounds are characterised by the fast rise times of their harmonic partials, and changing the envelope amplitude significantly alters the apparent timbre of the instrument. Similarly, if the sound of a piano is

---

<sup>14</sup>This is often stated quite ambiguously — what is usually implied is insensitivity to relative phase differences between steady-state harmonic partials [122].

played backwards then it no longer sounds like a piano, despite having the same time-averaged spectral content. A possible set of quantitative features to describe timbre include the relative amplitudes of harmonics and their time-varying envelopes, resonance characteristics (*e.g.*, formants), inharmonicity and spectral centroid (which represents ‘brightness’, although this too is a subjective measure). Martin [86] illustrates some such characteristics for flute, violin and trumpet tones and shows the marked differences between them. He later extends the feature set to a 31-dimensional vector for the purposes of instrument recognition [87]. He identifies around ten characteristics, in addition to signal statistics such as mean and variance both in the steady state and the initial attack phase; the characteristics include spectral centroid location and modulation, vibrato frequency, onset rate and ratio of odd to even harmonics. (These features are obtained using Ellis’ log-lag correlogram [34].) Using these features he proposes a taxonomic hierarchy of instruments organised by properties such as pizzicato or sustained. Most instrument families he found could be organised into disjoint regions of the timbre space with the exception of woodwind which has to be split into flute/piccolo or reed sub-families.

Brown [13] describes a method for discriminating the timbres of the oboe and the saxophone in which cepstral coefficients are calculated and used as features in a subsequent pattern analysis. The classifier is trained using a  $k$ -means algorithm. The cepstral coefficients are averaged over the duration of the training samples (solo saxophone and oboe extracts of around 1 minute duration). No time domain information is used in the classification, but for the given two class problem, the spectral method, which decouples the excitation and response (see §2.5.2), performs well. However, it is not clear that this method could extend to classification of a much larger number of instruments since spectral information, whilst important, is not sufficient for the recognition of many instruments. Time domain variation (in particular the nature of the attack and decay phases) and the presence of noise (*e.g.*, breath noise for a flute) are also major cues to recognition. Furthermore, Risset and Wessel [122] observe that, “a saxophone remains a saxophone whether it is heard over a distortion-ridden pocket-sized transistor radio or directly in a concert hall.” Hence, spectral information may be even less important than one might intuitively think.

## 2.6 *Conclusions*

This chapter has presented a wide variety of techniques from very different disciplines, each of which are highly relevant to the analysis of musical signals. The rôle of the ear as the detector of musical sound is investigated in section 2.2. The frequency decomposition properties of the cochlea and its interpretation as a filter-bank are of particular interest, along with its implications for pitch detection. Higher level auditory processes are also discussed which are responsible for the integration of the low level frequency stimuli into meaningful percepts (such as musical notes) by grouping and streaming principles. Some practical models built upon these principles are described in section 2.3. Some of the considerations specific to musical signals are detailed in section 2.4, which describes some of the forms of structure found in musical signals at medium and high levels of modelling. Some statistical properties of musical structure are mentioned which may form a useful basis for probabilistic modelling of high level musical features (which will be described in chapter 3). Finally, section 2.5 describes some of the techniques for analysis at a signal level, drawn from several areas of signal processing. Parametric and nonparametric methods are described which either construct an explicit signal model and then perform a parameter estimation step, or extract salient features from the data by means of signal transformation and filtering operations. Techniques used to model musical instruments for synthesis are also described.

Currently, many of the aforementioned fields of research occupy distinct territories sharing little overlap. For an application such as musical transcription one must draw upon applicable techniques from different fields. This challenging problem has many ambiguities which can only be resolved in a meaningful manner by the application of specific knowledge about the expected musical structure, characteristics of the sound generating mechanisms and low level properties of the signal. Some of these important features will be employed in the models developed later in this thesis.



## *3.1 Introduction*

The previous chapter has shown how a model based approach is an important tool for the description, transformation, and analysis of musical signals, and how the motivations for a particular model may come from diverse sources. Once an explicit signal model has been specified, it is necessary to assign values to each of the parameters in order to make the model representative of the given data. This process is known as *parameter estimation*. In this chapter, probabilistic modelling techniques are described which provide a flexible framework for parameter estimation and also the comparison of different candidate models. Model selection is an important part of data modelling as it allows us to assess the ability of several different models to describe the data whilst favouring economy above overfitting. Signal models can be posed in a probabilistic form by reasoning about the statistics of the error. Any salient prior information about the model parameters can be represented explicitly and incorporated into the modelling framework. This can be invaluable for avoiding unrealistic model configurations, applying physical constraints, or resolving model ambiguities.

Bayes' theorem provides a calculus for addressing these requirements, and the comparatively recent field of Bayesian signal processing has been able to attack a range of complex signal processing problems through a process of signal model formulation and parameter estimation. Recent applications include artifact removal, missing sample interpolation, and correction of signal clipping and quantisation in audio signals [54, 126, 152], the restoration of degraded video sequences [75], and spectral analysis [5].

Section 3.2 describes how a signal model can be posed in a probabilistic form, and how the likelihood function is inadequate for many problems. Bayes' theorem is then introduced in section 3.3 to form a probability distribution over the space of the model parameters which allows for the incorporation of prior information and selection between competing models. The need for a method to perform the complex parameter estimation task necessitates the introduction of Markov chain Monte Carlo (MCMC) methods in section 3.4. Some of the issues of implementing an efficient MCMC algorithm are discussed in section 3.5 and the use of MCMC for model selection is described in section 3.6.

### 3.2 Probabilistic signal modelling

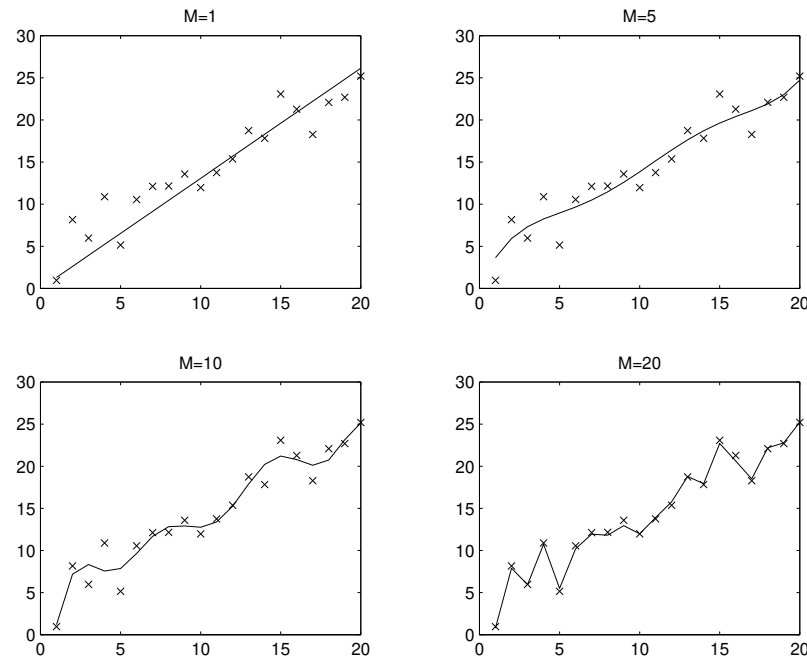
Parametric signal models construct an explicit model of the data with a set of parameters which may have a direct physical interpretation (*e.g.*, physical parameters of musical instruments), or they may be more abstract (*e.g.*, as with coefficients of an AR process). In either case, it is of interest to find the values of the parameters which produce a close approximation to the signal, since knowledge of these parameters enables us to recreate the signal, make further inferences about the signal characteristics, or perhaps transform the signal by manipulation of the parameters, for instance, to effect time-scaling.

For a data sample  $d_i$  at time instant  $i$ , the model produces  $x_i$  which is a function of the parameters  $\theta$ ,  $x_i = g_i(\theta)$ . The error in the reconstruction is  $e_i$ , and so

$$d_i = x_i + e_i. \quad (3.1)$$

Naturally, it is desired to reduce the error in the model's approximation by finding the set of values which minimises this error, or, more specifically, the sum of the squares of the error terms over the length of the signal. Introducing a vector notation to represent the sequence of values over the analysis interval of length  $N$  it is required to minimise the expression

$$\|\mathbf{e}\|^2 = \|\mathbf{d} - \mathbf{g}(\theta)\|^2 \quad (3.2)$$



**Figure 3.1:** The problem of overfitting. The observed data is linear with some additive Gaussian noise. The figures show the effect of increasing the order of the polynomial used to model the data. As the order,  $M$ , of the polynomial increases, the error between the data points and the polynomial decreases, since the measurement error is modelled and thus the model loses its ability to generalise.

with respect to all possible values of the model parameters  $\theta$ ,

$$\hat{\theta}_{\text{LS}} = \underset{\theta}{\operatorname{argmin}} \|\mathbf{e}\|^2. \quad (3.3)$$

The *least squares* parameter estimate produced by this criterion is unsatisfactory in many situations. Success is measured solely by the model's proximity to the observation. The model is therefore susceptible to overfitting by allowing more parameters to be added to the model to reduce the error term. In the most extreme case, when the number of parameters is equal to  $N$ , the data can be represented with zero error. Figure 3.1 illustrates the tendency to overfit with increasing model order.

Overfitting is undesirable for several reasons:

- Errors in the data are modelled (*e.g.*, measurement error) and so sensitivity to

noise is increased.

- The model loses the ability to generalise since minor variations in the data are modelled.
- The model produced is much larger (and therefore more expensive to represent and calculate) than the ‘true’ underlying model.

To obtain better parameter estimation criteria, the task is posed in a probabilistic framework. The error,  $\mathbf{e}$ , is assumed to be a zero-mean Gaussian, independent, identically distributed (iid) process with variance  $\sigma^2$ . This is generally chosen since the Gaussian distribution is the least informative probability distribution consistent with a given second moment [12]. This allows the probability of the error sequence to be expressed as

$$\begin{aligned} p_{\mathbf{e}}(\mathbf{e} | \sigma^2, \mathcal{I}) &= \prod_{i=1}^N p_{e_i}(e_i | \sigma^2, \mathcal{I}) \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{\|\mathbf{e}\|^2}{2\sigma^2}\right] \end{aligned} \quad (3.4)$$

where  $\mathcal{I}$  denotes all other prior information known about the problem. The rôle of  $\mathcal{I}$  is subtle and can be very important, and will be discussed in the next section. A transformation of variables  $\mathbf{e} \mapsto \mathbf{d}$  is performed so as to produce a probability expression which is dependent on the model parameters  $\theta$ , (temporarily omitting the dependence on  $\sigma^2$  and  $\mathcal{I}$  for clarity)

$$p_{\mathbf{d}|\theta}(\mathbf{d} | \theta) = p_{\mathbf{e}}(\mathbf{e}) \left| \frac{\partial \mathbf{e}}{\partial \mathbf{d}} \right| \quad (3.5)$$

The Jacobian of the transformation is unity for causal systems.<sup>1</sup> The *likelihood* is defined as:

$$p(\mathbf{d} | \theta, \sigma^2, \mathcal{I}) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{\|\mathbf{e}\|^2}{2\sigma^2}\right]. \quad (3.6)$$

The maximum likelihood parameter estimate is that which maximises the likelihood expression

$$\hat{\theta}_{\text{ML}} = \underset{\theta}{\operatorname{argmax}} p(\mathbf{d} | \theta, \sigma^2, \mathcal{I}) \quad (3.7)$$

---

<sup>1</sup>As the matrix will be lower triangular with ones on the leading diagonal [114].

The likelihood expression is sometimes written as  $L(\theta; \mathbf{d})$  to emphasise that it is a function of the parameters. Strictly speaking, the interpretation of (3.6) is that it is a function of the data.

Modifications of the likelihood expression have been proposed that penalise model complexity by adding terms, which are dependent on the number of parameters in the model, to the logarithm of the likelihood. Akaike's Information Criterion (AIC) [1], Rissanen's Minimum Description Length (MDL) [121], and a similar measure proposed by Schwartz [131], are of this form, and are motivated by information-theoretic considerations. These measures when considered in a probabilistic setting, however, appear to make restrictive assumptions about the nature of the parameters. A more flexible approach is permitted by Bayesian methods, as discussed in the next section.

### 3.3 Bayes' theorem

A major limitation of the maximum likelihood estimator and other measures such as AIC and MDL is their inability to exploit prior information about the model. This prior information may take many forms, but one of the most common cases is to represent knowledge about the likely values a parameter may take before the data is observed. This may reflect the expected variation ascertained from a number of previous observations, or the enforcement of physical constraints upon the possible parameter values. Bayes' theorem takes account of prior information by transforming the likelihood via the *a priori* probability of the model parameters.

Bayes' theorem is written as:

$$p(\theta | \mathbf{d}, \mathcal{I}) = \frac{p(\mathbf{d} | \theta, \mathcal{I}) p(\theta | \mathcal{I})}{p(\mathbf{d} | \mathcal{I})}, \quad (3.8)$$

where  $\mathcal{I}$  represents all prior information and assumptions about the model,  $p(\mathbf{d} | \mathcal{I})$  is the *evidence*, which generally may be regarded as a normalising factor,  $p(\theta | \mathcal{I})$  is the *prior probability* density of the parameters before the data is observed, and  $p(\theta | \mathbf{d}, \mathcal{I})$  is the *posterior* or *a posteriori* probability density.

The interpretation of the posterior is rather different to the likelihood. The likelihood is the probability the data  $\mathbf{d}$  could be observed if the parameters are  $\theta$ . The posterior on the other hand is the probability that  $\theta$  are the model parameters, given that the data  $\mathbf{d}$  was observed. The difference is very important as a model with a high likelihood may have a low posterior due to a low prior probability.<sup>2</sup>

The maximum *a posteriori* (MAP) parameter estimate is

$$\hat{\theta}_{\text{MAP}} = \underset{\theta}{\operatorname{argmax}} p(\theta | \mathbf{d}, \sigma^2, \mathcal{I}) \quad (3.9)$$

This is superior to the ML estimate as it can balance the prior expectations against the data observed. In contrast, the likelihood is dependent solely on the data and takes no account of prior information. In the absence of any prior information about  $\theta$ , where all values are equally likely, the MAP and the ML estimates are equivalent.

### 3.3.1 Prior probabilities

Among the criticisms of Bayesian techniques, which have impeded their general acceptance until relatively recently, is that the results of experiments using Bayesian techniques could be unduly influenced by the choice of the priors, rather than being solely determined by the data itself. This can certainly be true if the priors are not specified carefully, but this also embodies much of the power of Bayesian methods. Further, the concept of probability as an expression of belief, rather than purely as a frequentist interpretation, is often a cause for concern, and has generated much polemical debate. As a vehement defender of the Bayesian paradigm, Jaynes [65] expressed his *Desideratum of Consistency* that, “in two problems where we have the same prior information, we should assign the same prior probability”, such that two experimenters, given the same initial information would assign the same priors. ‘Consistency’ here means ‘consistent with the state of prior information’.

However, in some applications there may be sufficient conviction about the nature of the parameters to impose more ‘informative’ priors. Ardent frequentists are opposed

---

<sup>2</sup> The distinction can sometimes be subtle, but an example should highlight the important difference. The probability  $p(\text{spots}|\text{measles})$  is different to  $p(\text{measles}|\text{spots})$  since, although spots may be an inevitable consequence of measles ( $p(\text{spots}|\text{measles}) = 1$ ), the existence of spots may be due to a number of other diseases. Also  $\sum_d p(d|\text{spots}) = 1$ , and so  $p(\text{measles}|\text{spots}) < 1$ .

to the use of prior probabilities to represent subjective beliefs, but here the possibility of incorporating salient prior information into the model is left open. Where little is known of the likely values of a parameter it is desirable to assign a prior probability which is non-informative or *vague*. The prior should be diffuse compared to the likelihood [9], and carry no more information than is available. Maximum entropy priors are often used for this purpose as they have the maximum entropy consistent with the prior knowledge [65]. In addition to merely assigning numerical prior probability distributions to parameters, there is the flexibility of imposing prior structures for model parameters which are dependent on further *hyperparameters*.

There is another aspect to the use of prior information. Strictly, when writing probability expressions  $p(a|b, \mathcal{I})$  should be written to mean “the probability of  $a$  given  $b$  and any other prior information  $\mathcal{I}$  we may have”. Often, for notational simplicity  $\mathcal{I}$  is omitted, but its importance shouldn't be underestimated. It embodies the tacit assumptions of the model under consideration. All posterior inferences upon model parameters are conditional upon  $\mathcal{I}$  and must be recognised as such. In particular, one of the prime assumptions is that the model can form a valid representation of the input data. If the observation is a different type of signal to the one that was expected (*e.g.*, a sinusoid is expected but a Gaussian is observed) then the resulting parameter estimates will be meaningless. Thus it is assumed that the data we encounter will be adequately represented by the model, or by one of the set of candidate models, which Bernardo and Smith [7] describe as an  $M$ -complete scenario. Jaynes [66] warns of the pitfalls of inadequately specifying the states of prior information which can lead to apparent paradoxes if  $\mathcal{I}$  is either ill-specified or not specified at all. Embodied within  $\mathcal{I}$  is also knowledge, or expectation, of the physical behaviour of the system being modelled, which leads to the particular choices of the form of the model and the prior distributions of the parameters.

### 3.3.2 Representation of prior knowledge

Many authors find the inclusion of subjective prior information in a model unpalatable. It is argued that subjectivity has no place in algorithms since it lacks repeatability, as described above. Non-Bayesian algorithms, however, may have a more subtle dependence upon the state of prior knowledge. For instance, a non-parametric spectral estimation

scheme may produce frequency estimates from a spectrogram or other time-frequency representation. A method must be specified to extract the desired information from the intermediate representation, for instance by picking the highest  $n$  peaks or only those peaks above a certain threshold. The question arises as to how the number  $n$  or the threshold is to be determined. Generally, values which tend to perform well for ‘typical’ data might be chosen, but this may still be a subjective decision. For instance, the application of a detection threshold 30dB below the signal power is merely an expression of the subjective prior knowledge that all signal components of interest will be within 30dB of the peak power; another experimenter may have chosen a different threshold. The difference with Bayesian models is that the subjective values can be made explicitly part of the model formulation, rather than being embedded in the algorithm. The same subjective views may underpin both algorithms but the Bayesian approach makes the dependence explicit.

### 3.3.3 Marginalisation

A useful technique available to Bayesian modellers is the ability to perform *marginalisation* on model parameters. Typically it would be employed for removing *nuisance parameters* from consideration, *i.e.*, those parameters which unavoidably form part of the model, but about whose values are of little interest. If the parameter space is split  $\theta = \{\theta_1, \theta_2\}$  and it is desired to eliminate  $\theta_2$  then it can be marginalised through integration of the joint posterior of  $\theta$ ,

$$p(\theta_1 | \mathbf{d}, \mathcal{I}) = \int_{\Theta_2} p(\theta_1, \theta_2 | \mathbf{d}, \mathcal{I}) d\theta_2. \quad (3.10)$$

Unfortunately, marginalisation integrals are often analytically intractable, and numerical methods must be used instead (see section 3.4). The conceptual simplifications which arise from not having to be concerned about nuisance variables are however generally worthwhile.

### 3.3.4 Model selection

Another very powerful aspect to Bayesian analysis is the process of *model selection*, the importance of which has been touched upon briefly up to now. The need for model selection techniques arises from the need to compare several candidate models and assess

their suitability in a manner consistent with *Ockham's razor*: where model parsimony is traded off against the goodness-of-fit. In this thesis, model *order* selection will be used to describe the selection between models of the same type which differ in the number of components, or in the order of the functional form of the model. The term *model selection* is usually intended to refer to the comparison of models of different types. Bayes' theorem provides a unified framework for comparing different types of models probabilistically. Bayesian inference has a very important rôle to play in statistical data analysis, as an iterative process of inference and criticism. A model is chosen so that, "in the light of the then available knowledge, it best takes account of relevant phenomena in the simplest way possible" [9]. The model is evaluated for its ability to describe the data, *e.g.*, by analysing the residuals (which should have the same statistics as the assumed error model). Following this criticism the model is modified and the procedure repeats.

A total of  $K$  models are proposed, denoted by  $\mathcal{M}_k$ ,  $k = \{1, \dots, K\}$ , of which one is assumed to be the correct model for the data. The parameters of model  $\mathcal{M}_k$  are  $\theta_k$ . The posterior expression must now also take the model type into consideration

$$p(\theta_k, \mathcal{M}_k | \mathbf{d}, \mathcal{I}) = \frac{p(\mathbf{d} | \theta_k, \mathcal{M}_k, \mathcal{I}) p(\theta_k | \mathcal{M}_k, \mathcal{I}) p(\mathcal{M}_k | \mathcal{I})}{p(\mathbf{d} | \mathcal{I})}. \quad (3.11)$$

over which inferences will be made. If only information about the most likely model type is required, then the marginal posterior for  $\mathcal{M}_k$  is needed,<sup>3</sup>

$$p(\mathcal{M}_k | \mathbf{d}, \mathcal{I}) = \int_{\Theta_k} p(\theta_k, \mathcal{M}_k | \mathbf{d}, \mathcal{I}) d\theta_k \quad (3.12)$$

and hence the maximum marginal posterior model is

$$\hat{\mathcal{M}} = \underset{\mathcal{M}_k}{\operatorname{argmax}} p(\mathcal{M}_k | \mathbf{d}, \mathcal{I}). \quad (3.13)$$

This criterion for model selection can be informally compared to non-Bayesian model selection criteria, namely AIC and MDL. AIC is a function of the likelihood with a penalisation term for the number of parameters, where for the moment, it is assumed that model  $\mathcal{M}_k$  has  $k$  parameters,

$$\text{AIC} = -2 \log(\text{maximum of likelihood}) + 2k. \quad (3.14)$$

---

<sup>3</sup>This integration, however, is analytically intractable in most cases.

Similarly for MDL,

$$\text{MDL} = -\log(\text{maximum of likelihood}) + \frac{k}{2} \log(N). \quad (3.15)$$

The posterior can be written in a similar form, assuming equal priors for all model types and noting that the evidence term is a constant,

$$-\log(p(\theta_k, \mathcal{M}_k | \mathbf{d}, \mathcal{I})) = -\log(p(\mathbf{d} | \theta_k, \mathcal{M}_k, \mathcal{I})) - \log(p(\theta_k | \mathcal{M}_k, \mathcal{I})) + c$$

It is apparent that AIC and MDL are approximately equivalent to a criterion based upon the posterior but with restrictive assumptions of priors.<sup>4</sup> AIC effectively corresponds to the assumption  $p(\theta_k | \mathcal{M}_k, \mathcal{I}) \propto e^{-k}$ . Further, making the simplifying assumption that all parameters are independent and are uniformly distributed over the same interval  $[0, Z]$ , AIC corresponds to  $Z = e$ . For MDL, the effective prior is  $p(\theta_k | \mathcal{M}_k, \mathcal{I}) \propto N^{-k/2}$  and under similar assumptions this corresponds to a value of  $Z = N^{1/2}$ . The tendency for AIC to overfit is clear when the true range of each parameter is much greater than  $e$ . MDL tends to select smaller model sizes as it effectively assumes that the range of each parameter is larger than for AIC (when  $N > e^2$ ) and so the prior will be lower. A more detailed discussion of model comparison measures is presented by Wu [165].

A Bayesian approach to model selection is not limited by the effective assumptions embodied in AIC or MDL, principally that all model order coefficients carry equal weight and that the cost of a model is function only of the number of terms. The cost of a model is also determined by how deterministic the parameters are, in the sense that a model with two parameters which have very narrow priors is likely to have a higher posterior probability than a model with a single, vague, parameter.

### 3.4 Parameter estimation through MCMC

There are two major aspects to signal modelling — the design of a model to capture the structure of the data, and the techniques used to produce parameter estimates for that model. Box and Tiao [9] describe the iterative approach to modelling: inference, model

<sup>4</sup>This also assumes that the maximum likelihood value  $\hat{\theta}$  falls within the range of the prior.

appraisal and subsequent reformulation using Bayesian inference as an exploratory tool. Here, it is assumed that the functional form of the model is known, and concentrate on the task of parameter estimation.

Bayesian inference can be applied to the joint posterior to produce the desired statistics of the distribution (MAP estimates, moments, marginal posterior estimates, confidence intervals, etc.). The form of the joint posterior distribution is often too complex to permit the direct calculation of such statistics. To overcome this, Markov chain Monte Carlo (MCMC) methods are employed to generate a stream of samples which are drawn from the joint posterior distribution. The method exploits the duality between probability densities and the samples produced from them. Given a probability density, samples can be generated from it. Conversely, given a large number of samples, an estimate of their probability distribution — or statistics thereof — can be obtained from Monte Carlo integrations. A Markov chain is set up which has the desired posterior distribution as its invariant distribution and whose state space is (in general) the parameter space of the model.

A set of *transition kernels* are specified which define the transition probabilities between states at each iteration. The form of these transitions is dependent upon the type of algorithm used. Once a sufficiently long Markov chain has been produced, Monte Carlo integrations over the chain allow Bayesian inference of any desired statistic of the posterior.

### 3.4.1 Monte-Carlo Integration

Monte Carlo integration can be employed to produce expectations of functions of a parameter,

$$E[f(x)] = \sum_{x \in X} f(x) p(x) \quad (3.16)$$

but, if samples  $\{\bar{x}_1, \dots, \bar{x}_N\}$  drawn from  $p(x)$  are available, the Monte Carlo estimator of  $E[f(x)]$  is

$$\hat{f} = \frac{1}{N} \sum_{i=1}^N f(\bar{x}_i). \quad (3.17)$$

Thus statistics of the distribution  $p(x)$  can be inferred when samples  $\bar{x}_i \sim p(x)$  are available. The estimator is unbiased, so

$$\lim_{N \rightarrow \infty} |E[f(x)] - \hat{f}| = 0 \quad (3.18)$$

but the error standard deviation decreases in proportion to  $\sqrt{N}$ , so Monte Carlo integration can be inefficient as many samples may be required for accurate results.

### 3.4.2 Markov chain overview

This section presents a brief overview of some of the characteristics of Markov chains. This discussion will be limited to discrete variables; a more detailed exposition is given by Roberts [123]. The extension of these concepts to continuous state spaces is presented by Tierney [151].

The Markov chain represents a series of random variables,  $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ , each of which may vary over the *state space*  $\mathbf{X}$  of the Markov chain,<sup>5</sup> and have the property that the distribution of each variable depends solely on the previous one,

$$p(\mathbf{x}_{n+1} | \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) = p(\mathbf{x}_{n+1} | \mathbf{x}_n). \quad (3.19)$$

The procession between subsequent states is determined by a transition kernel  $T_n(\mathbf{x}; \hat{\mathbf{x}})$  which defines the probability of moving from state  $\hat{\mathbf{x}}$  to  $\mathbf{x}$  at the  $n^{\text{th}}$  point in the chain. Hence, the probability of being in state  $\mathbf{x}$  at step  $n$  is calculated from all the paths to it from the previous step

$$p_{n+1}(\mathbf{x}) = \sum_{\hat{\mathbf{x}} \in \mathbf{X}} p_n(\hat{\mathbf{x}}) T_n(\mathbf{x}; \hat{\mathbf{x}}) \quad (3.20)$$

The desired behaviour of the Markov chain is that the series converges to a *stationary distribution*  $\pi(\mathbf{x})$ , such that further transitions do not affect the probability distribution — a property known as *positive recurrence*:

$$\pi(\mathbf{x}) = \sum_{\hat{\mathbf{x}}} \pi(\hat{\mathbf{x}}) T_n(\mathbf{x}; \hat{\mathbf{x}}). \quad (3.21)$$

---

<sup>5</sup>The state space is also the parameter space of the model in most cases. Hybrid techniques exist which also employ dynamical information in addition to current parameter values, e.g., see [165].

For the chain to converge to the stationary distribution, several conditions have to be met. In addition to positive recurrence, the chain must be irreducible, which ensures that there is a non-zero probability of reaching all states eventually. The chain must also be *aperiodic* which prevents it from periodically oscillating between sets of states.

We are interested in Markov chains which are *time reversible* which requires that the condition for *detailed balance* be satisfied,

$$\pi(\mathbf{x}) T(\hat{\mathbf{x}}; \mathbf{x}) = \pi(\hat{\mathbf{x}}) T(\mathbf{x}; \hat{\mathbf{x}}). \quad (3.22)$$

If the chain satisfies the above conditions then it is *ergodic*, such that averages over the Markov chain converge to the ensemble average, and that the state probabilities converge to the stationary distribution,

$$\lim_{n \rightarrow \infty} p_n(\mathbf{x}) = \pi(\mathbf{x}). \quad (3.23)$$

The initial portion of the Markov chain is termed the *burn-in* period, during which the chain converges to the stationary distribution. The burn-in is typically discarded.

### 3.4.3 The Metropolis-Hastings algorithm

The Metropolis-Hastings (MH) algorithm [59, 96] is a popular method for constructing the Markov chain. It is a versatile algorithm which is often useful when the posterior distribution is of a complex non-standard form. A transition kernel  $q(\theta^*; \theta^k)$  proposes a state  $\theta^*$  from the current state  $\theta^k$  using a proposal density which is generally dependent on the current state. This state is accepted with a probability  $Q(\theta^k, \theta^*)$  determined by the Metropolis-Hastings acceptance function

$$Q(\theta^k, \theta^*) = \min(1, \text{PR}(\theta) \text{TR}(\theta)) \quad (3.24)$$

$$\text{PR}(\theta) = \frac{p(\theta^* | \mathbf{d}, \bullet)}{p(\theta^k | \mathbf{d}, \bullet)} \quad \text{TR}(\theta) = \frac{q(\theta^k; \theta^*)}{q(\theta^*; \theta^k)}$$

where  $q(\theta^k; \theta^*)$  is the probability of proposing the reverse transition from state  $\theta^*$  to state  $\theta^k$ . This acceptance function ensures that detailed balance is achieved. This form of the MH algorithm is due to a generalisation by Hastings [59] of the original method of Metropolis *et al.* [96] to the case where the proposal distributions are not symmetrical. This is considerably more powerful than the original method, and allows efficient

transition kernels to be designed by tailoring them to the posterior density under consideration. The acceptance probability is a function of the ratio of the posteriors of each state (PR) and the transition ratios between these states (TR). The Markovian property of the MH algorithm arises as each new state is dependent upon the previous state.

It is also possible to propose *independence sampling steps* [150] where the proposal is independent of the current state, and so  $q(\theta^*; \theta^k) = q(\theta^*)$ . These proposals, if well designed, are beneficial for improving the mixing of the Markov chain as the parameter space can be explored very quickly and correlations in the Markov chain are reduced. It is important to note that different types of transition kernel can be combined in the MH algorithm, to exploit the advantages of each type.

The simplest form of transition kernel is the *random walk* where the proposal state is constructed from a random perturbation to the current state,  $q(\theta^*; \theta^k) = q(\theta^* - \theta^k)$ . This is simple to implement, but tends to be very inefficient for global exploration of the parameter space, particularly for multimodal distributions. Its success is usually critically dependent on the complexity of the posterior distribution and the size of the proposal distribution: small steps have a high acceptance rate but do not explore much of the posterior distribution, large steps are capable of exploring the parameter space but have lower acceptance rates. Random walk kernels can however be useful for local exploration of the posterior distribution. Neal [102] suggests a rule of thumb of proposing a step of the order of the standard deviation of the width of the posterior mode in the most confined dimension. Gilks *et al.* [47] suggest adjusting the size of the standard deviation to obtain an acceptance rate of 20–40%.

A *global* MH algorithm proposes a change for all parameters. For all but very simple models this can be highly inefficient due to the low acceptance probabilities which may result from proposing a point in a high dimensional space. A more useful implementation for larger models performs a proposal for a small set of parameters in turn, updating the state if the new parameter values are accepted, or keeping the same values if it is rejected. This is described as a *local* MH algorithm, but it is often known as *single component Metropolis-Hastings* or *Metropolis-Hastings-within-Gibbs*.<sup>6</sup> If a move is proposed for the parameter set  $\gamma \subset \theta$  whilst keeping the remaining parameters at their current values, such that  $\theta^* = \{\gamma^*, \theta_{-\{\gamma\}}^k\}$ , and  $\theta^k = \{\gamma^k, \theta_{-\{\gamma\}}^k\}$ , then the ratio

<sup>6</sup>The Gibbs sampler is introduced in the next section

of posteriors reduces to the ratio of full conditionals<sup>7</sup> and the transition kernels affect only  $\gamma$

$$\text{PR}(\theta) = \frac{p(\gamma^* | \theta_{-\{\gamma\}}^k, \mathbf{d}, \bullet)}{p(\gamma^k | \theta_{-\{\gamma\}}^k, \mathbf{d}, \bullet)} \quad \text{TR}(\theta) = \frac{q(\gamma^k; \gamma^*, \theta_{-\{\gamma\}}^k)}{q(\gamma^*; \gamma^k, \theta_{-\{\gamma\}}^k)}. \quad (3.25)$$

The full conditionals can be derived from the joint posterior [46]

$$p(\gamma | \mathbf{d}, \theta_{-\{\gamma\}}) = \frac{p(\theta | \mathbf{d})}{\int p(\theta | \mathbf{d}) d\gamma} \quad (3.26)$$

and, since the term on the denominator is not a function of  $\gamma$ , the full conditional is proportional to the joint posterior. The set of parameters which are proposed together are generally chosen to be parameters that are highly correlated. Local moves for each parameter would have a low probability of acceptance since the other parameters would not be changed, so a proposal distribution is formed that can create joint proposals for high probability regions of the posterior. The local form of the MH algorithm will be employed extensively in the following chapters. It is summarised in algorithm 3.1. For generality, the full form of the MH acceptance function is shown, but for single component moves it can be simplified as described above.

#### 3.4.4 The Gibbs sampler

The Gibbs sampler [44] is another method of constructing a Markov chain with the posterior as its invariant distribution. For a parameter space  $\theta = \{\theta_1, \dots, \theta_M\}$  the  $n^{\text{th}}$  state of the Markov chain is, for each parameter, sampled from the full conditional using the most recent values of the other parameters,

$$\theta_1^n \sim p(\theta_1 | \mathbf{d}, \theta_2^{n-1}, \dots, \theta_M^{n-1}) \quad (3.27)$$

$$\vdots$$

$$\theta_j^n \sim p(\theta_j | \mathbf{d}, \theta_1^n, \dots, \theta_{j-1}^n, \theta_{j+1}^{n-1}, \dots, \theta_M^{n-1}) \quad (3.28)$$

$$\vdots$$


---

<sup>7</sup>The full conditional distribution of a parameter  $\theta_i$  is the posterior for  $\theta_i$  conditional upon all other model parameters,  $p(\theta_i | \mathbf{d}, \theta_{-(i)})$ .

---

**Algorithm 3.1:** The local (or single component) Metropolis-Hastings algorithm

```

initialise  $\theta^0 \in \Theta$ 
for iteration  $k = 1 \dots N_{\text{iter}}$  do
  for parameter  $i = 1 \dots N_\theta$  do
    create proposal  $\theta_i^* \sim q(\theta_i^*; \theta^k)$ 
    MH-accept(  $\theta^*, \theta^k$  )
  end for
end for

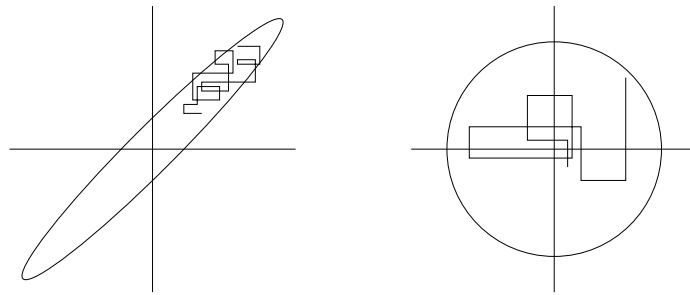
function: MH-accept(  $\theta^*, \theta^k$  )
  evaluate  $Q(\theta^*, \theta^k) = \min \left( 1, \frac{p(\theta^* | \mathbf{d}) q(\theta^k; \theta^*)}{p(\theta^k | \mathbf{d}) q(\theta^*; \theta^k)} \right)$ 
  draw  $v \sim \mathbb{U}_{[0,1]}$ 
  if  $v < Q(\theta^*, \theta^k)$  then
    accept proposal  $\theta^k = \theta^*$ 
  else
    keep old state  $\theta^k$ 
  end if

```

---

This is, in fact, a special case of the single component Metropolis-Hastings algorithm using full conditionals as proposal distributions,  $q(\theta^*; \theta^k) = p(\theta_j | \mathbf{d}, \theta_{-j})$ . The ratio of the transition probabilities is the reciprocal of the ratio of the full conditionals and so the move is unconditionally accepted. The Gibbs sampler is popular in circumstances where the full conditionals are easily sampled from. However, Wu [165] suggests that, even in these cases, the Metropolis-Hastings algorithm may be more efficient due to its greater mobility around the state-space. Nevertheless there are many statistical problems, where the probability distributions are of standard forms and can be sampled from directly, for which a Gibbs framework is well suited. In these cases a Gibbs scheme may be deployed mechanically — the BUGS software (Bayesian inference Using Gibbs Sampling) [138, 149] allows for automation of this process, expressing models in graphical form.

Each transition in the Gibbs sampler changes only one parameter at a time, and the



**Figure 3.2:** Convergence of correlated (left) and uncorrelated (right) parameters with the Gibbs Sampler. The progress of the Markov chain is shown. The uncorrelated case is able to explore the parameter space much more quickly than the correlated case.

net effect of the algorithm is a *random walk* along each parameter in turn. Correlation between the parameters reduces the convergence rate markedly [102], as illustrated in figure 3.2. A variant on the Gibbs sampler is the *Hit & Run* algorithm which picks a random direction for the co-ordinate axes [128]. It has the advantage that detailed balance holds for any distribution of directions, but sampling along all directions must be possible. Chen [19] generalises the algorithm to give more flexibility over the choice of direction distribution. Gilks and Roberts [48] describe how parameter transformations can be employed to improve the behaviour of the Markov chain. They also note that Metropolis-Hastings proposals can be made along the elliptical major axis which would produce a rapidly converging Markov chain.

### 3.5 Designing efficient transition kernels

It is a commonly held belief, amongst non-practitioners, that MCMC methods are far too slow to be of use in practical applications. However, significant speed increases can be obtained by the design of kernels which are both computationally efficient and succeed in exploring the high probability regions of the posterior surface. Combined with a good choice of initial values — for instance from the results of an analysis on data close temporally or spatially to the current observation, or estimates obtained from a fast non-parametric method — respectable performance can be obtained. Although,

theoretically, the Markov chain will converge to the posterior distribution regardless of the choice of starting value, the length of the burn-in period can be greatly reduced if the chain is started in a region of high posterior probability.

In this section the careful choice of Metropolis-Hastings transition kernels is advocated, in order to exploit the posterior characteristics of the model under consideration.

### 3.5.1 *Types of transition kernel*

For a problem of any appreciable complexity, any MCMC sampling scheme must be designed specifically for the particular type of model if an efficient algorithm is desired. The inadequacy of random walk samplers for complex multimodal distributions, and the likely appearance of non-standard forms for the posterior densities, necessitates careful choice of the transition kernels. Here, several generic types of transition kernel are identified which achieve different types of movement around the parameter space. The kernels may be combined in mixtures or cycles [150], *e.g.*, using a stochastic or deterministic selection of a particular kernel in each iteration.

#### *Global exploration kernels*

Perhaps the most important of the kernels, from an efficiency point of view, are those which attempt to propose transitions to high probability regions elsewhere in the parameter space. These are particularly useful to have during the initial burn-in period of the Markov chain.

Since a sample from the full conditional will result in an unconditional acceptance (corresponding to a Gibbs sampler move), drawing a sample from the full conditional is often a goal which is unachievable due to the complexity of the conditional distribution. Consequently, it is desirable to construct a transition kernel which concentrates its support in the same regions as the full conditional.

One method of achieving this is to construct an independence sampling step [150] whose proposal distribution approximates the full conditional. The resulting state, if the proposal is accepted, will be independent of the current state. This is beneficial for the mixing of the Markov chain as correlations are reduced. It also serves to speed the convergence since it is possible to traverse the state space very quickly. It is important

that a sample can be drawn from the proposal distribution with a low computational burden, otherwise little would be gained over a method which samples from the full conditional but at great expense (*e.g.*, rejection sampling [158]).

### *Local exploration kernels*

In addition to locating interesting regions of the posterior, it is advantageous to perform local explorations. The transition kernels which achieve this are often random perturbations of the current state, such as random walks. These transitions form Markov chains which have high correlations between states and they tend to explore the parameter space very slowly. Small perturbations are likely to yield a high acceptance rate but slow exploration whilst large perturbations can explore more of the local region at the expense of a lower acceptance rate. It is also possible to use heavy tailed distributions for the random walk proposals which are more likely to produce large excursions occasionally, for instance using a Cauchy distribution rather than a Gaussian distribution. Heavy tailed moves may also be achieved by using a Gaussian distribution with a variance which may vary over time [52].

### *Related-mode transitions*

Many models are likely to result in multimodal posterior distributions, and, in such cases, care must be taken to avoid becoming trapped in local modes. A global exploration kernel may be useful here, as described previously, but, if it is created as an approximation to the full conditional distribution, it may not be able to propose transitions into all modes. Many types of model will generate a posterior distribution whose modes which are related, such that, given the location of one mode, a number of others may be reached by appealing to the structure of the model. As an example, consider a model to detect a periodic pulse train parametrised on the shape of the pulse, the location of the first pulse (epoch) and the period. Given the shape and the period, the conditional distribution for the epoch will be multimodal with modes separated by the period. The conditional distribution of the period will have modes at multiples of the true period. Hence, proposals which move the epoch by an amount equal to the period,

or multiply the period by a ratio of integers, can be used to explore the related modes of the distribution.

If there is redundancy in the model, such that a particular signal can be modelled identically in more than one way, then a related-mode transition may be useful for exploring the alternative representations, of which it is hoped the most parsimonious model or the one most consistent with the prior information is favoured. A specific instance of this type of transition kernel, for harmonic signals, is described in §6.3.

### 3.6 MCMC for model selection

Until recently, inference on the posterior probabilities of different models has relied upon a separate analysis for each model type as MCMC methods were only applicable to models of fixed dimensionality. Reversible jump techniques now exist for variable-order models where a single sample-based approach is adequate to cover all model types. Reversible jump, developed by Green [58], constructs a Markov chain which is capable of jumping between parameter subspaces of differing dimensionality. It is a generalisation of the Metropolis-Hastings algorithm where the transition kernels can create proposals in different parameter subspaces.

The model is defined by the model type  $\mathcal{M}$ , taken from the set of all models  $\mathbb{M}$ ; the parameters  $\theta$  are specific to each model type. The joint posterior of this model is

$$p(\theta, \mathcal{M} | \mathbf{d}) \propto p(\mathbf{d} | \mathcal{M}, \theta) p(\theta | \mathcal{M}) p(\mathcal{M}) \quad (3.29)$$

In reversible jump, transitions may be of an *update* type, where the model type is unchanged and the proposal is within the space of  $\theta$ , or they may be of a *subspace transition* type, where the model type is changed along with the parameters. For an update move a proposal  $\theta^*$  is sampled,

$$\theta^* \sim q(\theta^* ; \theta^k, \mathcal{M}^k) \quad (3.30)$$

and then accepted according to the M-H acceptance function. The proposal distribution for  $\theta^*$  could be dependent upon the previous state, or an independent distribution may

be used (see §3.5.1). For a subspace transition move, a joint proposal for  $\{\theta^*, \mathcal{M}^*\}$  is obtained

$$\{\theta^*, \mathcal{M}^*\} \sim q(\theta^*, \mathcal{M}^*; \theta^k, \mathcal{M}^k). \quad (3.31)$$

This move may be performed in two steps, firstly, a new model type is proposed and, secondly, a new set of parameter values for that model is proposed,

$$\mathcal{M}^* \sim q(\mathcal{M}^*; \theta^k, \mathcal{M}^k) \quad (3.32)$$

$$\theta^* \sim q(\theta^*; \mathcal{M}^*, \theta^k, \mathcal{M}^k). \quad (3.33)$$

Care must be taken when specifying subspace transition kernels to ensure reversibility — typically moves of this sort are birth/death or split/combine moves which increase or reduce the number of components by one. This allows for heuristic proposal densities to generate parameter values for a new component which are likely to be accepted (*e.g.*, from an independence sampling distribution). These methods are becoming popular for practical problems: Richardson and Green [120] apply reversible jump techniques to mixture distributions, Andrieu and Doucet [5] use reversible jump transitions for the detection of noisy sinusoids and Troughton and Godsill [153] apply these methods to autoregressive time series.

Other methods have been proposed for variable dimension problems. Carlin and Chib's model [17] defines the global parameter space as the product of the parameter spaces of all model parameters. This has the disadvantage that *linkage densities* or *pseudo-priors* are required between unused model parameters, and their choice is critical to the success of the algorithm. George and McCulloch's stochastic search variable selection (SSVS) [45] includes all parameters in the model, but assigns low values for parameters which are not deemed important to the model, hence the model size is fixed. Godsill [51] unites several of these different methods using a composite model space which can be used for common model choice problems such as nested models and variable selection.

### 3.7 *Conclusions*

In this chapter a probabilistic approach to signal modelling has been described. The likelihood function for the model parameters can be obtained using an assumption of Gaussian error statistics, but reliance on the likelihood can lead to overfitting and the inability to generalise. The application of Bayes' theorem transforms a state of prior knowledge upon the observation of data and produces a probability distribution over the parameter space of the model (§3.3). Prior probability distributions can be employed to represent available prior knowledge about the model parameters and provide a basis for probabilistic model selection. Bayesian model selection operates consistently with Ockham's razor such that overfitting is avoided and the most parsimonious model is selected.

For non-trivial models the parameter estimation task requires sophisticated techniques, such as Markov chain Monte Carlo methods, which simulate a Markov chain whose stationary distribution is asymptotically equivalent to the posterior distribution of the model parameters (§3.4). The Metropolis-Hastings algorithm and related techniques are generally regarded to be too slow for practical applications, but this is often as a result of poor choices for the transition kernels. Several types of transition kernel have been suggested in section 3.5 which exploit the structure of the posterior distribution to perform different types of move around the parameter space. The extension of the Metropolis-Hastings algorithm to model selection has been described, in particular the reversible jump sampler which performs jumps between different subspaces of the parameter space.

# *Detection and Estimation of Single Component Models*

---

# 4

## *4.1 Introduction*

In this chapter a Bayesian signal analysis is developed for the class of single component models. This class encompasses signals comprised of a single fundamental element (for instance a Gaussian, rectangular pulse or sinusoid) and also those comprised of more than one such element, but which can be logically grouped together due to the sharing of a common parameter, or due to more abstract Gestalt grouping principles. For instance, signals composed of a periodic pulse train or harmonically related sinusoids would be classed as single component models.

General linear models are employed for modelling signals as the composition of a number of basis functions. Section 4.2 poses them in a Bayesian setting and shows how nuisance variables can be marginalised to obtain a posterior expression for the unknown basis function parameters. MCMC techniques are employed in section 4.3 to draw samples from the posterior distribution for the purposes of Bayesian inference. In section 4.4 it is shown how the model can be extended to allow a variable number of basis functions, and the modifications required to the Metropolis-Hastings algorithm to produce an efficient simulation are described. In section 4.5, changes to the prior structure are detailed which alleviate some of the problems associated with model order selection. The extension of the general linear model for time-varying signals is discussed in section 4.6, and an analysis framework is introduced which expresses a multiple frame model as a Bayesian graphical model. Efficient transition kernels for the multiple frame model which exploit correlations between neighbouring frames are described in section 4.7.

One of the predominant themes in this chapter is the choice of transition kernels in the Metropolis-Hastings algorithm, exploiting different aspects of the structure of the posterior distribution. Analysis methods based upon MCMC techniques are generally held to be far too slow for practical applications. In the literature, little attention appears to have been directed towards producing efficient implementations. This chapter seeks to explore several avenues for efficiency increases by exploiting prior knowledge in the formulation of the model and in the function of the algorithm. Where justifiable, approximations may be introduced which can yield an order of magnitude increase in speed. In many applications, this speed gain may be of far higher importance than the resulting slight loss of accuracy. Aspects of the material in this chapter are presented in [159].

## 4.2 *The General Linear Model*

The *general linear model* (GLM) is a form of model which is extensively employed in this thesis, and which has had a brief exposition in section 2.5.1. This is a powerful representation for the modelling of ‘structured’ data, in the sense that there is an *a priori* expectation that the data will conform to a known mathematical form. Knowledge of the underlying physical processes generating the data is an important step towards obtaining a representative model. On the other hand, the generating mechanism may be unclear, as with many problems in statistical inference, but exploratory analyses using non-parametric methods, such as from time-frequency representations, may reveal apparent structure in the data which could subsequently be captured in a parametric model.

In cases where the data can be well represented by a known parametric form (or as one of a number of competing parametric forms), and is such that it can be expressed as the linear combination of a number of *basis vectors* (which need not be orthogonal), then a GLM is often an ideal candidate. GLMs, as described here, are well suited to joint model selection and parameter estimation — see [126] for a general exposition of their rôle in Bayesian signal processing. GLMs can be used to represent an autoregressive

(AR) process [115, 126], and are also popular for regression problems in statistics [62].

#### 4.2.1 Formulation

The signal model is composed of a linear combination of a number of basis functions with parameters  $\{\phi\}$  which can be concisely written in vector form:

$$\mathbf{d} = \mathbf{G}\mathbf{b} + \mathbf{e} \quad (4.1)$$

where  $\mathbf{G}(\phi)$  is the *basis matrix* whose  $M$  columns are the basis vectors (each of length  $N$ ),  $\mathbf{b}$  are the amplitudes, and  $\mathbf{e}$  is the error component, usually assumed zero-mean Gaussian iid with variance  $\sigma_e^2$ .

The likelihood can be formulated from the error term

$$p(\mathbf{d} | \phi, \mathbf{b}, \sigma_e^2, \mathcal{I}) = \frac{1}{(2\pi\sigma_e^2)^{\frac{N}{2}}} \exp \left[ -\frac{\|\mathbf{d} - \mathbf{G}\mathbf{b}\|^2}{2\sigma_e^2} \right] \quad (4.2)$$

where for the moment the number of basis functions and the type of model is assumed to be known. The posterior is

$$p(\phi, \mathbf{b}, \sigma_e^2 | \mathbf{d}, \mathcal{I}) = \frac{p(\mathbf{d} | \phi, \mathbf{b}, \sigma_e^2, \mathcal{I}) p(\phi, \mathbf{b}, \sigma_e^2 | \mathcal{I})}{p(\mathbf{d} | \mathcal{I})} \quad (4.3)$$

and the prior structure  $p(\phi, \mathbf{b}, \sigma_e^2 | \mathcal{I})$  must be suitably specified. It is generally convenient to assume prior independence for the basis function parameters, the amplitudes, and the error variance.<sup>1</sup> Hence the prior can be written

$$p(\phi, \mathbf{b}, \sigma_e^2 | \mathcal{I}) = p(\phi | \mathcal{I}) p(\mathbf{b} | \mathcal{I}) p(\sigma_e^2 | \mathcal{I}). \quad (4.4)$$

Notwithstanding the importance of the state of prior information  $\mathcal{I}$  (see section 3.3), for simplicity of notation it shall be omitted for the remainder of this chapter. The prior  $p(\phi)$  will be highly specific to each type of model under consideration. Several cases will be detailed in the next section. Its specification depends upon the strength of prior knowledge about the model parameters and there is much potential for the use of subjective or empirical knowledge in its formulation. If little is known *a priori* then it should be sufficiently uninformative to prevent biasing the posterior. Different forms of the amplitude prior  $p(\mathbf{b})$  will be discussed in this chapter.

<sup>1</sup>Or, more strictly, conditional independence given  $\mathcal{I}$ .

The choice of the error variance prior can be motivated from a few different considerations. One is that it should be uninformative. The Jeffreys prior [67] is a popular choice for scale parameters, such as variance, as it is scale-invariant and is also a maximum entropy prior. The Jeffreys prior is

$$p(\sigma_e^2) \propto \frac{1}{\sigma_e^2}. \quad (4.5)$$

This prior, however, is improper since its integral is not finite, so limits must be put on the prior for it to be made proper. Improper priors are problematic for model comparison since the model choice will be dependent upon the ratio of two unnormalised distributions. Djurić *et al.* describe a method for circumventing this problem by splitting the data into estimation and validation sets to determine the unknown proportionality constants [30].

The simplest prior distribution for  $\mathbf{b}$  is a uniform distribution

$$p(\mathbf{b}) = \frac{1}{B^M} \prod_{m=1}^M \mathbb{I}_{[-\frac{B}{2}, \frac{B}{2}]}(b_m) \quad (4.6)$$

where  $B$  is the maximum allowable range for each element  $b_m$  and  $M$  is the number of basis functions. If  $B$  is sufficiently large to cover all likely values of  $b_i$  then the value which maximises the conditional posterior  $p(\mathbf{b} | \mathbf{d}, \phi, \sigma_e^2, \mathcal{I})$  is also the value of  $\mathbf{b}$  which minimises the squared error  $\|\mathbf{e}\|^2$ ,

$$\left. \frac{\partial \|\mathbf{e}\|^2}{\partial \mathbf{b}} \right|_{\mathbf{b}=\hat{\mathbf{b}}} = 0 \quad (4.7)$$

and so

$$\begin{aligned} \frac{\partial \|\mathbf{e}\|^2}{\partial \mathbf{b}} &= -2\mathbf{G}^t(\mathbf{d} - \mathbf{G}\mathbf{b}) \\ \hat{\mathbf{b}} &= (\mathbf{G}^t\mathbf{G})^{-1}\mathbf{G}^t\mathbf{d} \end{aligned} \quad (4.8)$$

which is the familiar least-squares estimate. It is further possible to marginalise the amplitudes from the posterior [126]. Writing the error term as a quadratic in  $\mathbf{b}$ ,

$$\begin{aligned} \|\mathbf{d} - \mathbf{G}\mathbf{b}\|^2 &= (\mathbf{b} - \hat{\mathbf{b}})^t (\mathbf{G}^t\mathbf{G})(\mathbf{b} - \hat{\mathbf{b}}) + (\mathbf{d}^t\mathbf{d} - \mathbf{f}^t\mathbf{f}) \\ \mathbf{f} &= \mathbf{G}\hat{\mathbf{b}} \end{aligned} \quad (4.9)$$

the integration of the exponential term of the likelihood is comprised of the multiplication of a multivariate Gaussian  $N_M(\mathbf{b}; \hat{\mathbf{b}}, \sigma_e^2(\mathbf{G}^t \mathbf{G})^{-1})$  and a term which is independent of  $\mathbf{b}$ . The marginalised posterior expression becomes

$$\begin{aligned} p(\phi, \sigma_e^2 | \mathbf{d}) &= \int p(\phi, \mathbf{b}, \sigma_e^2 | \mathbf{d}) d\mathbf{b} \\ &= B^{-M} |\mathbf{G}^t \mathbf{G}|^{-\frac{1}{2}} (2\pi\sigma_e^2)^{\frac{M-N}{2}} \exp \left[ -\frac{\|\mathbf{d}\|^2 - \|\mathbf{f}\|^2}{2\sigma_e^2} \right]. \end{aligned} \quad (4.10)$$

The effect of the exponential term is similar to its effect in the likelihood. The term  $\|\mathbf{d}\|^2 - \|\mathbf{f}\|^2$  is still a measure of the error;  $\mathbf{f}$  is the least-squares projection of the model formed by the parameters  $\{\phi\}$  onto the data  $\mathbf{d}$ . The error term, therefore, is the difference between the energy of the signal and the energy of the least-squares projection, as a function of  $\{\phi\}$ . If the basis is composed of orthogonal basis vectors, then  $\|\mathbf{d}\|^2 - \|\mathbf{f}\|^2 = \|\mathbf{d} - \mathbf{f}\|^2$ . If  $\mathbf{G}$  formed a complete orthogonal basis set for the data then this error would be zero. This posterior expression is sensitive to ill-conditioned basis matrices. If the columns of  $\mathbf{G}$  are close to being linearly dependent then the determinant will be small and the posterior may become numerically unstable.

The adoption of a Jeffreys prior for the error variance allows for its marginalisation using the identity<sup>2</sup>

$$\int_0^\infty \frac{(2\pi\sigma_e^2)^{-\varepsilon}}{\sigma_e^2} \exp \left[ -\frac{Q}{2\sigma_e^2} \right] d\sigma_e^2 = \frac{\Gamma(\varepsilon)}{\pi^\varepsilon Q^\varepsilon} \quad (4.11)$$

to yield

$$\begin{aligned} p(\phi | \mathbf{d}) &\propto \frac{\Gamma(\varepsilon) p(\phi)}{B^M |\mathbf{G}^t \mathbf{G}|^{\frac{1}{2}} \pi^\varepsilon [\|\mathbf{d}\|^2 - \|\mathbf{f}\|^2]^\varepsilon} \\ \varepsilon &= \frac{N-M}{2}. \end{aligned} \quad (4.12)$$

The posterior is now solely a function of the data and the model parameters  $\{\phi\}$ . In the instances where the basis functions are orthogonal to each other this inevitably leads to considerable computational savings,

$$p(\phi | \mathbf{d}) \propto \frac{\Gamma(\varepsilon) p(\phi)}{B^M (\prod_m \|\mathbf{g}_m\|^2)^{\frac{1}{2}} \pi^\varepsilon [\|\mathbf{d}\|^2 - \sum_m \|\mathbf{g}_m^t \mathbf{d}\|^2 / \|\mathbf{g}_m\|^2]^\varepsilon}. \quad (4.13)$$

<sup>2</sup>Adapted from [126].

### 4.2.2 Conditional distributions for amplitudes and error variance

Estimates for the amplitudes and error variance, if required, can be produced from their conditional distributions. The conditional distribution for the amplitudes (with the error variance marginalised) is

$$p(\mathbf{b} | \mathbf{d}, \phi) \propto \frac{p(\mathbf{b})}{[\|\mathbf{d} - \mathbf{G}\mathbf{b}\|^2]^\varepsilon}. \quad (4.14)$$

When the prior for  $\mathbf{b}$  is uniform and the least squares value  $\hat{\mathbf{b}}$  lies within the non-zero range of the prior then the least-squares value also maximises the conditional posterior.

The conditional posterior for the error distribution, following marginalisation of the amplitudes is

$$p(\sigma_e^2 | \phi, \mathbf{d}) \propto \frac{1}{|\mathbf{G}^t \mathbf{G}|^{\frac{1}{2}} (\sigma_e^2)^{\frac{N-M}{2}+1}} \exp \left[ -\frac{\|\mathbf{d}\|^2 - \|\mathbf{f}\|^2}{2\sigma_e^2} \right] \quad (4.15)$$

which is an inverse gamma (IG) distribution

$$p(\sigma_e^2 | \mathbf{d}, \phi) = \text{IG} \left( \sigma_e^2; \frac{N-M}{2}, \frac{\|\mathbf{d}\|^2 - \|\mathbf{f}\|^2}{2} \right) \quad (4.16)$$

$$\text{IG}(z; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} z^{-(\alpha+1)} e^{-\frac{\beta}{z}} \quad (4.17)$$

The inverse gamma distribution has a mode at  $z = \beta/(\alpha + 1)$  and a mean  $z = \beta/(\alpha - 1)$  (for  $\alpha > 1$ ), so an estimate of the error variance is obtained as

$$\hat{\sigma}_e^2 = \frac{\|\mathbf{d}\|^2 - \|\mathbf{f}\|^2}{N - M - 2} \quad (4.18)$$

### 4.2.3 The effect of parameter priors

To illustrate the effect on the posterior of the parameter priors a GLM is considered that models a signal composed of two Gaussians. Both have the same variance  $\sigma_g^2$ , and are a fixed distance  $\tau_g$  apart, but have different amplitudes. Both  $\sigma_g^2$  and  $\tau_g$  are known. The basis matrix is

$$\mathbf{G} = [\mathbf{g}_1 \ \mathbf{g}_2] \quad (4.19)$$

$$\mathbf{g}_1(n) = \text{N}(n; \mu_g, \sigma_g^2) \quad \mathbf{g}_2(n) = \text{N}(n; \mu_g + \tau_g, \sigma_g^2)$$

The unknown parameter is  $\mu_g$ , the location of the first peak. The prior for  $\mu_g$  is assigned a Gaussian distribution,  $p(\mu_g) = N(\mu_g; \hat{\mu}_g, \sigma_\mu^2)$  which can be made diffuse if little is known *a priori*, or narrow if compelling prior information is available.

It is frequently convenient to deal with the log of the posterior since the computer implementation of probabilistic schemes can be prone to arithmetic underflow or overflow due to the exponential functions involved. The log posterior is, ignoring additive constants,

$$\log p(\mu_g | \mathbf{d}) \approx -\frac{(\mu_g - \hat{\mu}_g)^2}{2\sigma_\mu^2} - \varepsilon \log[\|\mathbf{d}\|^2 - \|\mathbf{f}\|^2] \quad (4.20)$$

Figure 4.2 shows the sensitivity of the shape of the posterior density to several parameters. The observed data is shown in figure 4.1. The prior carries a prior expectation that the value of  $\mu_g$  is closer to the second peak. The general observation is that compelling data (*i.e.*, data which can be modelled with a low error or where many points of data are available) outweighs prior knowledge. Where the data is unable to discriminate well between different model parameters, prior knowledge compensates.

### 4.3 MCMC parameter estimation

In this section it is shown how MCMC techniques can be employed for parameter estimation of a single component signal model. The basis matrix is a function of the model parameters  $\phi$  such that  $\mathbf{G} = \mathbf{G}(\phi)$  where  $\phi = \{\phi_1, \dots, \phi_{N_\phi}\}$ . The number of basis functions  $M$  is known *a priori* and the amplitudes  $\mathbf{b}$  and error variance  $\sigma_e^2$  are marginalised.

A family of proposal distributions  $q()$  must be found to suit the model under consideration. As summarised in §3.5, there are several motivations for choosing proposal distributions according to the desired type of movement around the parameter space. For global exploration, the most useful type of move is an independence sampling step [150] with a proposal distribution which concentrates its support in similar regions to the full conditional. In the case that the proposal is equal to the full conditional then this becomes a Gibbs sampler move. Strictly speaking this type of transition might be labelled

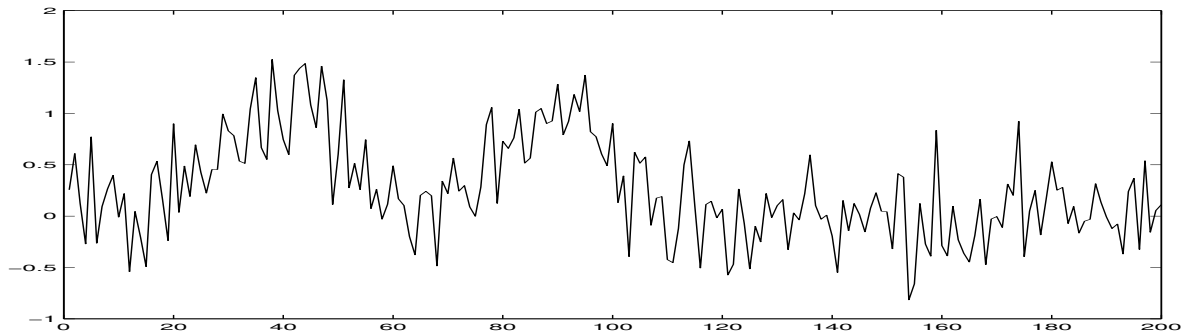


Figure 4.1: Two peak Gaussian data in additive Gaussian noise.

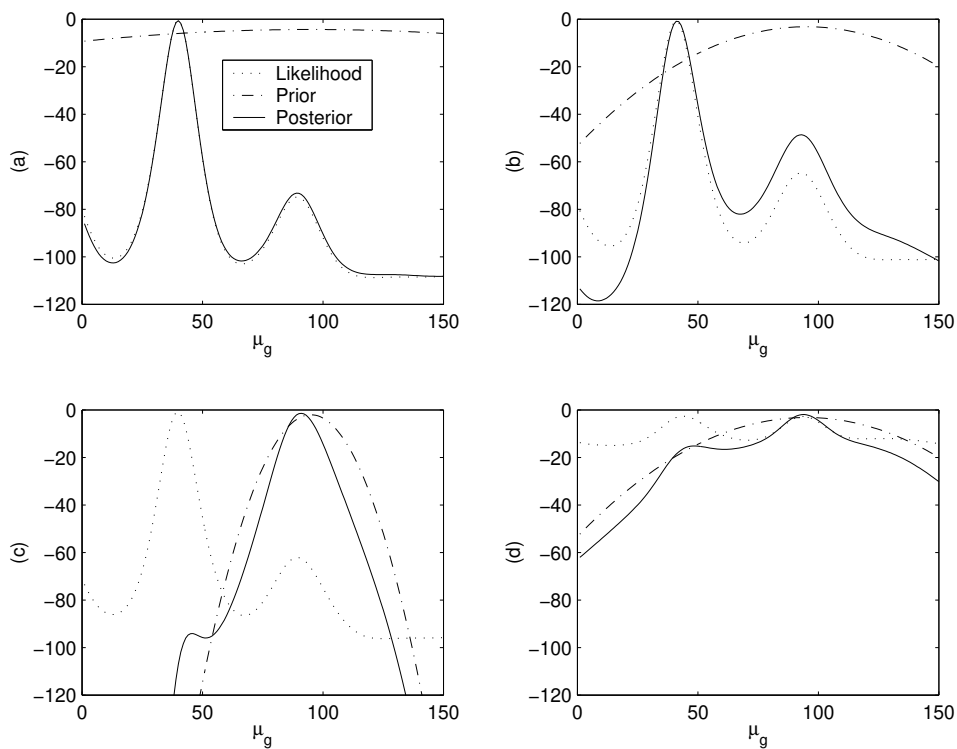


Figure 4.2: The effect of parameter priors. Each figure shows the log of the likelihood, prior and conditional posterior  $p(\mu_g | \mathbf{d}, \bullet)$ . The data observed is shown in figure 4.1. The prior is centred at  $\hat{\mu}_g = 95$ . In figure (a) the prior is vague ( $\sigma_\mu^2$  is large) and has little effect. The prior standard deviation is reduced by a factor of 10 in (b) and 100 in (c) and in this plot the prior dominates the likelihood. Figure (d) shows the effect of increasing the noise in the data; prior knowledge compensates for poor data.

a conditional-independence sampling step since it is conditionally dependent on the current values of the other model parameters, and is actually a hybrid between a true independence sampling step and a single component transition kernel.

---

**Algorithm 4.1:** Multiple transition kernels for single-component Metropolis-Hastings algorithm

```

function: update  $\phi_i$ 
  draw  $u \sim \mathbb{U}_{[0,1]}$ 
  if  $u < \lambda_{\text{cond}}$  then
    draw  $\phi_i^* \sim q_{\text{cond}}(\phi_i^*; \phi^k)$ 

  else if  $u < \lambda_{\text{cond}} + \lambda_{\text{per}}$  then
    draw  $\Delta_i^* \sim q_{\text{per}}(\Delta_i^*)$ 
     $\phi_i^* = \phi_i^k + \Delta_i^*$ 

  else
    draw  $\phi_i^* \sim q_{\text{ind}}(\phi_i^*)$ 
  end if
  MH-accept(  $\theta^*, \theta^k$  )

function: MH-accept(  $\theta^*, \theta^k$  )
  evaluate  $Q(\theta^*, \theta^k) = \min \left( 1, \frac{p(\theta^* | \mathbf{d}) q(\theta^k; \theta^*)}{p(\theta^k | \mathbf{d}) q(\theta^*; \theta^k)} \right)$ 
  draw  $v \sim \mathbb{U}_{[0,1]}$ 
  if  $v < Q(\theta^*, \theta^k)$  then
    accept proposal  $\theta^k = \theta^*$ 
  else
    keep old state  $\theta^k$ 
  end if

```

---

The single component Metropolis-Hastings algorithm (algorithm 3.1, §3.4.3) can be modified to permit the non-deterministic selection between different types of transition kernel. Three types are used in this algorithm, a conditional independence step, a perturbation (random walk) step and an independence step, with probabilities  $\lambda_{\text{cond}}$ ,  $\lambda_{\text{per}}$

and  $\lambda_{\text{ind}}$  respectively, where  $\lambda_{\text{cond}} + \lambda_{\text{per}} + \lambda_{\text{ind}} = 1$ . The conditional independence step uses  $q_{\text{cond}}(\phi_i^*; \phi^k)$  as its proposal distribution. The perturbation step samples a perturbation from a distribution  $q_{\text{per}}(\Delta_i^*)$  where  $\Delta_i^*$  has the same dimensionality as  $\phi_i$ . A scheme which appears to work well in practice is to non-deterministically select a proposal variance from a range of several values (e.g., geometrically spaced by a factor of 10); this allows local exploration with the opportunity to make larger jumps, and is beneficial for the mixing of the Markov chain. The independence sampling step draws its proposal from  $q_{\text{ind}}(\phi_i^*)$ . These steps are summarised in algorithm 4.1.

The posterior ratio for the marginalised  $\{\mathbf{b}, \sigma_e^2\}$  model is (from (4.12))

$$\frac{p(\phi^* | \mathbf{d})}{p(\phi^k | \mathbf{d})} = \frac{p(\phi^*)}{p(\phi^k)} \frac{|\mathbf{G}^{k^t} \mathbf{G}^k|^{\frac{1}{2}} [ \|\mathbf{d}\|^2 - \|\mathbf{f}^k\|^2 ]^\varepsilon}{|\mathbf{G}^{*t} \mathbf{G}^*|^{\frac{1}{2}} [ \|\mathbf{d}\|^2 - \|\mathbf{f}^*\|^2 ]^\varepsilon} \quad (4.21)$$

To achieve a reasonable acceptance rate it is desirable to obtain a transition probability similar to the posterior ratio. It is often useful to optimise the evaluation of the projection energy  $\|\mathbf{f}(\phi_{n_\phi})\|^2$  for a range of values of  $\phi_{n_\phi}$ . Specific examples will be presented in later chapters. If the conditional distribution for  $\phi_{n_\phi}$  is slowly varying then it may suffice to calculate the projection energy for a few well-spaced values of  $\phi_{n_\phi}^*$  and form a piecewise linear continuous proposal distribution between these points.<sup>3</sup> If the form of the conditional is more complex then it may be necessary to evaluate many more points of the distribution to resolve the peaks of the distribution before forming a proposal distribution. It is now considered how efficient independence sampling distributions can be constructed for the common case that the parameters of the basis functions can be split into two subsets which define shape and location respectively.

### 4.3.1 Fixed-scale shape function

In the Metropolis-Hastings scheme described above, independence sampling steps for one parameter are conditioned upon the current values of the others. It is therefore possible to consider the proposals for shape and location independently.

The vector  $\mathbf{g}$  is a basis function with a fixed shape, for instance a Gaussian with a known variance. Its unknown parameter is the location parameter  $\mu_g$  whose conditional

<sup>3</sup> To ensure irreducibility, care must be taken at the endpoints for variables which have infinite support.

posterior is

$$\begin{aligned}
 p(\mu_g | \mathbf{d}) &\propto p(\mu_g) |\mathbf{g}^t \mathbf{g}|^{-\frac{1}{2}} [ \|\mathbf{d}\|^2 - \|\mathbf{f}\|^2 ]^{-\varepsilon} \\
 \mathbf{f} &= \mathbf{g} \hat{\mathbf{b}} = \mathbf{g} (\mathbf{g}^t \mathbf{g})^{-1} \mathbf{g}^t \mathbf{d} \\
 \therefore \|\mathbf{f}\|^2 &= \mathbf{d}^t \mathbf{g} (\mathbf{g}^t \mathbf{g})^{-1} \mathbf{g}^t \mathbf{d}
 \end{aligned} \tag{4.22}$$

Noting that  $\mathbf{g}^t \mathbf{g}$  is a scalar and making the dependence on  $\mu_g$  explicit,

$$\|\mathbf{f}(\mu_g)\|^2 = \frac{\|\mathbf{g}(\mu_g)^t \mathbf{d}\|^2}{\|\mathbf{g}(\mu_g)\|^2} \tag{4.23}$$

and, since  $\mu_g$  is a location parameter, the summation can be written

$$\begin{aligned}
 \mathbf{g}(\mu_g)^t \mathbf{d} &= \sum_{i=1}^N g(i; \mu_g) d(i) \\
 &= \sum_{i=1}^N g(i - \mu_g; 0) d(i).
 \end{aligned} \tag{4.24}$$

If  $\mathbf{g}$  is time-reversed to produce the function  $\overleftarrow{\mathbf{g}}$  then a convolution can be used to perform the projection for  $\mu_g = \{1, \dots, N\}$

$$\mathbf{D} = \mathbf{d} * \overleftarrow{\mathbf{g}} \tag{4.25}$$

This is a matched filter which is an intuitive result for finding the location in the observation of a fixed shape. The projection energy becomes

$$\|\mathbf{f}(\mu_g)\|^2 = \frac{\|\mathbf{D}(\mu_g)\|^2}{\|\mathbf{g}(\mu_g)\|^2}. \tag{4.26}$$

If the energy of the basis function is invariant to changes in location then the denominator of this expression is constant, as is the determinant in the posterior.

### *Gaussian example*

The basis function is parametrised on location  $\mu_g$  and scale (variance)  $\sigma_g^2$

$$g(i; \mu_g, \sigma_g^2) = \frac{1}{(2\pi\sigma_g^2)^{\frac{1}{2}}} \exp \left[ -\frac{(i - \mu_g)^2}{2\sigma_g^2} \right]. \tag{4.27}$$

Over 99% of the mass of  $\mathbf{g}$  is located within  $\pm 3\sigma_g$  of the centre and so, for  $3\sigma_g < \mu_g < N - 3\sigma_g$ , the energy of the basis function can be considered approximately constant for fixed  $\sigma_g$ . An expression can be obtained from the observation that the square of a Gaussian is also Gaussian with variance of  $\sigma_g^2/2$ ,

$$\|\mathbf{g}(\mu_g)\|^2 \approx (4\pi\sigma_g^2)^{-\frac{1}{2}}. \quad (4.28)$$

At the endpoints,  $\mu_g \leq 3\sigma_g$  and  $\mu_g \geq N - 3\sigma_g$ , the energy  $\|\mathbf{g}(\mu_g)\|^2$  should be calculated explicitly. This reduces end effects which would otherwise cause the approximation to erroneously rise at each end of the distribution.

### 4.3.2 Multiple basis functions

For a basis matrix composed of more than basis function,

$$\begin{aligned} \mathbf{G} &= [\mathbf{g}_1 \ \mathbf{g}_2 \ \dots \ \mathbf{g}_M] \\ \|\mathbf{f}\|^2 &= \mathbf{d}^t \mathbf{G} (\mathbf{G}^t \mathbf{G})^{-1} \mathbf{G}^t \mathbf{d} \\ &= \mathbf{D}(\mu_g)^t (\mathbf{G}(\mu_g)^t \mathbf{G}(\mu_g))^{-1} \mathbf{D}(\mu_g) \end{aligned} \quad (4.29)$$

where

$$\begin{aligned} \mathbf{D}(\mu_g) &= [\mathbf{D}_1(\mu_g)^t \ \dots \ \mathbf{D}_M(\mu_g)^t]^t \\ \mathbf{D}_m(\mu_g) &= \mathbf{d} * \overleftarrow{\mathbf{g}}_m \end{aligned} \quad (4.30)$$

If the basis functions are orthogonal with respect to each other, then  $\mathbf{g}_i^t \mathbf{g}_j = 0$ , and  $\mathbf{G}(\mu_g)^t \mathbf{G}(\mu_g)$  becomes diagonal such that

$$\|\mathbf{f}(\mu_g)\|^2 = \sum_{m=1}^M \frac{\|\mathbf{D}_m(\mu_g)\|^2}{\|\mathbf{g}_m(\mu_g)\|^2} \quad (4.31)$$

This shows how the output from several matched filters may be combined to calculate the projection from a higher order model.

## 4.4 Variable model order

For many applications, the model presented in the previous section is not sufficiently flexible. It is often required that not only the parameters of the basis functions, but also

the number of basis functions (which is termed the *model order*) should be determined. If the model describes a periodic pulse train and each basis function represents the pulse shape at a different location then it is generally desirable to determine the number of pulses in the data.

The incorporation of another parameter into the parameter space of the general linear model necessitates a slight change to the prior structure, as the amplitude vector  $\mathbf{b}$  is now explicitly dependent on the number of basis functions  $M$ ,

$$p(\phi, \mathbf{b}, M, \sigma_e^2) = p(\phi | M) p(\mathbf{b} | M) p(M) p(\sigma_e^2) \quad (4.32)$$

For generality the prior for the model parameters  $\phi$  is written to be dependent on  $M$ . The prior for  $M$  may be chosen to reflect prior knowledge about the number of basis functions, *e.g.*, a truncated uniform or Poisson distribution. Marginalisation of  $\mathbf{b}$  and  $\sigma_e^2$  proceeds as before and the resulting conditional posterior is (*cf.* (4.12))

$$p(\phi, M | \mathbf{d}) \propto \frac{\Gamma(\varepsilon) p(\phi | M) p(M)}{B^M |\mathbf{G}^t \mathbf{G}|^{\frac{1}{2}} \pi^\varepsilon [ \|\mathbf{d}\|^2 - \|\mathbf{f}\|^2 ]^\varepsilon} \quad (4.33)$$

$$\varepsilon = \frac{N - M}{2}.$$

The complexity penalisation action of the Bayesian formulation now becomes evident, largely through the prior on  $\mathbf{b}$ . Attempts to increase the model order to produce a better fit are met with an increased cost dependent on the range of the amplitudes  $B$ .

A single component updating scheme (see algorithm 3.1, §3.4.3) might choose to update the model order much less often than the other model parameters if the range of expected values of  $M$  is small, as will often be the case. For the transition kernels of the  $M$  update move (algorithm 4.1), it can be useful to employ an independence step, *e.g.*, proposing a value from the prior,  $M^* \sim p(M)$  and a perturbation step, *e.g.*, randomly choosing between  $M^* = M^k - 1$  and  $M^* = M^k + 1$ . A conditional independence step could also be employed, but unless the range of  $M$  values is large, this may be unnecessarily time-consuming.

A major impediment to the speed of execution of a sampling scheme based upon this model is the amount of time required to compute  $\|\mathbf{f}\|^2$  upon each update move; this must be evaluated for the current state and the proposal state when the M-H acceptance probability is calculated, and involves a matrix inversion. Major savings can be made

for the computation when the basis functions are orthogonal. In the implementation, if extra values are stored in addition to the parameters of the current state, for instance,  $\{\mathbf{f}^k, \hat{\mathbf{b}}^k, \mathbf{G}^{k^t} \mathbf{G}^k\}$  and the posterior probability then this can reduce the computational overhead quite significantly. In the perturbation move for  $M$  described above, where the size of the basis matrix is increased, efficiency savings can be made by circumventing the need for a large matrix inversion. If the proposal is to increase the model order, then this adds new columns to the basis matrix,

$$\begin{aligned} M^* &= M^k + M^\Delta \\ \mathbf{G}^* &= [\mathbf{G}^k \ \mathbf{G}^\Delta] \end{aligned}$$

The determinant and inverse of  $\mathbf{G}^{*t} \mathbf{G}^*$  can subsequently be calculated efficiently, as described in appendix A.

#### 4.4.1 Sensitivity of model order to amplitude prior

Any attempt to increase the model order is accompanied by an increase in the size of the parameter space, since the dimension of the basis function amplitudes increases. The improvement in model fit must outweigh the extra cost associated with the increase in the size of the parameter space. It is instructive to consider the conditions for which the more complex model is favoured by the posterior.

Consider two model states; the first  $\{M^k, \phi\}$  is the true model, the second increases the model order  $\{M^*, \phi\}$  where  $M^* = M^k + 1$  but the basis function parameters are held constant. It is assumed that the joint posteriors of both states have their dominant modes centred at  $\phi$  with width  $d\phi$  such that the ratio of the marginal posteriors can be approximated by the ratio of the joint posteriors, *i.e.*,

$$\frac{p(M^* | \mathbf{d})}{p(M^k | \mathbf{d})} \approx \frac{p(M^*, \phi^* | \mathbf{d}) d\phi}{p(M^k, \phi^k | \mathbf{d}) d\phi}. \quad (4.34)$$

Further, it is assumed that the basis vectors are orthonormal, that the priors  $p(\phi, M)$  are equal for both states, and that  $N \gg M$  such that  $\varepsilon \approx N/2$ . At the threshold  $p(M^*, \phi^* | \mathbf{d}) = p(M^k, \phi^k | \mathbf{d})$ , the ratio of posteriors reduces to<sup>4</sup>

$$\left[ \frac{\|\mathbf{d}\|^2 - \|\mathbf{f}^k\|^2}{\|\mathbf{d}\|^2 - \|\mathbf{f}^*\|^2} \right]^\varepsilon \approx B \sqrt{\frac{\pi}{\varepsilon}}. \quad (4.35)$$

<sup>4</sup>The approximation  $\Gamma(\varepsilon^*)/\Gamma(\varepsilon^k) \approx \sqrt{\varepsilon}$  is obtained from Stirling's formula, *e.g.*, see [110].

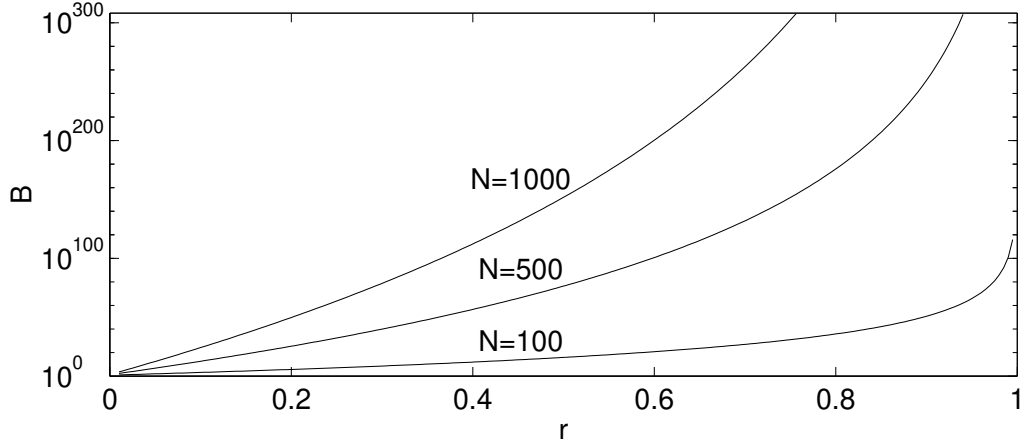


Figure 4.3: Threshold of detection of a variable order single component model using a uniform amplitude prior.

Owing to the orthogonality assumption, the energy of the projection  $\mathbf{f}^*$  can be split into the contributions from the  $k^{\text{th}}$  order model  $\mathbf{f}^k$  and from the additional basis functions  $\mathbf{f}^\Delta$

$$\|\mathbf{f}^*\|^2 = \|\mathbf{f}^k\|^2 + \|\mathbf{f}^\Delta\|^2 \quad (4.36)$$

Defining  $r$  to be the relative energy of the error absorbed by  $\mathbf{f}^\Delta$  to the residual  $\mathbf{d} - \mathbf{f}^k$

$$r = \frac{\|\mathbf{f}^\Delta\|^2}{\|\mathbf{d}\|^2 - \|\mathbf{f}^k\|^2} \quad (4.37)$$

the detection threshold expression can be rewritten as a function of  $r$

$$B = \sqrt{\frac{\varepsilon}{\pi}} [1 - r]^{-\varepsilon} \quad (4.38)$$

This expression shows how the acceptance of an increase in model order is dependent upon the amplitude range prior  $B$  and the fractional improvement of fit  $r$ . This variation is shown in figure 4.3 for several values of  $N$ . The initial choice of a uniform prior for  $\mathbf{b}$  was motivated by the assumption that it would be relatively uninformative as  $B$  was to be chosen such that it was sufficiently large to encompass all likely values of  $b_m$ ,  $\forall m \in \{1..M\}$ . However, from the above analysis it is apparent that the

choice of  $B$  has implications for the model order selection, and in choosing a value of  $B$  which is ‘sufficiently large’ the sensitivity of the detection is affected. In the next section improvements to the prior formulation which remove this dependence are considered.

In a typical real-world application the signal may be contaminated by crosstalk, echoes and other forms of ‘intelligible noise’ (*i.e.*, noise which shares similar characteristics with the desired signal). Owing to the complexity of incorporating these extra factors into a model explicitly, a Gaussian error model, such as the one above, is often used. A consequence of the preceding analysis is that, unless an incredibly high value is specified for  $B$ , the model will be perceived to overfit quite considerably since the extraneous signal components will also be detected. One may wish to be pragmatic about the remedy to this situation. The value of  $B$  could be raised to a sufficiently high level such that only the dominant components are detected. Alternatively the model can be left to overfit and a subsequent inference or post-processing step applies some form of thresholding; this, however, is computationally inefficient as detected components are then discarded. Both options may still be more attractive than a non-Gaussian error formulation. If the signal is composed of multiple components and it is desired to detect them all, then an explicit multiple component model is suitable, which is the subject of chapter 5.

### 4.5 *Modifications to prior structure*

The inadequacy of the choice of the uniform prior for the basis function amplitudes has been shown in the previous section. A more satisfying prior can be obtained by reasoning about the expected signal-to-noise ratio (SNR) of the observation. The SNR is expressed as  $\|\mathbf{G}\mathbf{b}\|^2 / \|\mathbf{e}\|^2$  and note that, since the error term is Gaussian,  $E[\|\mathbf{e}\|^2] = \sigma_e^2 N$ ;  $\delta^2$  is defined to be the expected SNR of the data, so

$$\delta^2 = E \left[ \frac{\|\mathbf{G}\mathbf{b}\|^2}{\|\mathbf{e}\|^2} \right] = \frac{E[\mathbf{b}^t (\mathbf{G}^t \mathbf{G}) \mathbf{b}]}{\sigma_e^2 N} \quad (4.39)$$

The Gaussian distribution makes the least assumptions about the specific form of the distribution given this second order moment [12], and so a multivariate Gaussian prior

is employed

$$\begin{aligned} p(\mathbf{b} | \phi, M, \sigma_e^2, \delta^2) &= \mathcal{N}(\mathbf{b}; \mathbf{0}, \sigma_e^2 \Sigma) \\ \Sigma(\phi, M, \delta^2) &= \delta^2 (\mathbf{G}^t \mathbf{G})^{-1} \end{aligned} \quad (4.40)$$

More rigorous motivations for this prior, known as the  $g$ -prior, from maximum entropy considerations can be found in Zellner [166] and Andrieu [6]; it is also invariant to changes in scale. The Gaussian form has the rather useful property that it can be marginalised easily since it is also a conjugate prior. The hyperparameter  $\delta^2$  may be specified *a priori* or treated as an unknown — it acts to control the sensitivity of the detection.

Unlike the uniform prior, the  $g$ -prior takes the energy of the basis functions into account and is invariant to changes in scale. Also the support of the prior is  $\mathbb{R}^M$  and so there is no need to ensure that each element of  $\mathbf{b}$  lies within  $\pm B/2$ .

A different prior is also specified for the error variance  $\sigma_e^2$ . The Jeffreys prior employed in the previous section has the advantages that it is uninformative, invariant to changes in scale and can be marginalised easily. On the other hand it is an improper prior such that its integral over the positive real axis is not finite. A popular choice for scale parameters is the inverse gamma distribution, which can be made relatively uninformative, and reduces to Jeffreys prior as a special case. A further motivation is that it is a conjugate prior, as the conditional distribution for  $\sigma_e^2$  is an inverse gamma distribution (4.16),

$$p(\sigma_e^2) = \text{IG}(\sigma_e^2; \alpha_e, \beta_e) \quad (4.41)$$

$$\text{IG}(z; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} z^{-\alpha-1} e^{-\frac{\beta}{z}} \mathbb{I}_{[0,+\infty)}(z). \quad (4.42)$$

Jeffreys prior is obtained when  $\alpha = 0, \beta = 0$ . The joint posterior now reflects the modified prior dependencies,

$$p(\phi, M, \mathbf{b}, \sigma_e^2 | \mathbf{d}, \delta^2) \propto p(\mathbf{d} | \phi, M, \mathbf{b}, \sigma_e^2) p(\phi) p(\mathbf{b} | \phi, M, \sigma_e^2, \delta^2) p(M) p(\sigma_e^2). \quad (4.43)$$

The marginalisation of  $\mathbf{b}$  and  $\sigma_e^2$  is performed as described in appendix B to obtain the

marginal posterior

$$\begin{aligned}
 p(\phi, M | \mathbf{d}, \delta^2) &= (1 + \delta^2)^{-\frac{M}{2}} [\mathbf{d}^t \mathbf{P} \mathbf{d} + 2\beta_e]^{-\varepsilon} p(\phi, M) \\
 \mathbf{P} &= \mathbf{I}_N - \frac{\delta^2}{1 + \delta^2} \mathbf{G} (\mathbf{G}^t \mathbf{G})^{-1} \mathbf{G}^t \\
 \varepsilon &= \frac{N}{2} + \alpha_e.
 \end{aligned} \tag{4.44}$$

The conditional distribution for  $\sigma_e^2$  is an inverse Gamma distribution with a mode close to  $\mathbf{d}^t \mathbf{P} \mathbf{d} / (N + 2)$

$$p(\sigma_e^2 | \mathbf{d}, \phi, M, \delta^2) = \text{IG}(\sigma_e^2; \frac{N}{2} + \alpha_e, \frac{\mathbf{d}^t \mathbf{P} \mathbf{d}}{2} + \beta_e) \tag{4.45}$$

The conditional distribution for  $\mathbf{b}$  is a multivariate Gaussian

$$\begin{aligned}
 p(\mathbf{b} | \mathbf{d}, \phi, M, \sigma_e^2, \delta^2) &= \text{N}(\mathbf{b}; \mathbf{m}, \sigma_e^2 \mathbf{M}) \\
 \mathbf{M} &= \frac{\delta^2}{1 + \delta^2} (\mathbf{G}^t \mathbf{G})^{-1} \\
 \mathbf{m} &= \mathbf{M} \mathbf{G}^t \mathbf{d} = \frac{\delta^2}{1 + \delta^2} \hat{\mathbf{b}}_{\text{ls}}
 \end{aligned} \tag{4.46}$$

#### 4.5.1 Marginalisation and joint proposals

If there are high posterior correlations between model parameters then this can have a profound negative effect upon the convergence rate of the Markov chain. In this model, the amplitude coefficients  $\mathbf{b}$  are, necessarily, highly correlated with the model parameters  $\{\phi, M\}$  since they respectively describe the amplitudes, parameters and model order of the basis functions. By marginalising  $\mathbf{b}$  some of the problems of slow convergence are circumvented, in the sense that updates for  $\mathbf{b}$  are handled ‘automatically’ and it is not required to explicitly construct proposals for it. The  $g$ -prior has also incorporated some dependence upon the model parameters in terms of the energy of the basis functions. It is generally convenient to marginalise ‘nuisance variables’ such as amplitude coefficients and error variances as they are usually of little interest, often have high correlations with other parameters, and they increase the dimensionality of the problem. In the case of  $\mathbf{b}$  and  $\sigma_e^2$ , estimates, if required, can be easily obtained from their full conditional distributions.

Suppose that the amplitude coefficients are no longer marginalised from the joint posterior. For a local MH algorithm, any attempt to propose a candidate parameter  $\phi^*$

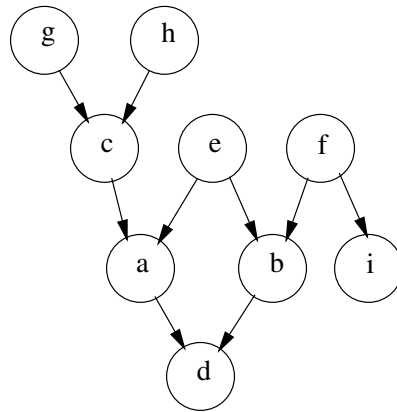
into a different region of the parameter space, whilst keeping  $\mathbf{b}$  constant, will have a low acceptance probability since it corresponds to a different set of basis functions and therefore a different  $\mathbf{b}$ . In contrast, perturbation moves for  $\phi^*$  generally will not change the shape of the basis functions much and so  $\mathbf{b}$  may not need to be updated. For all other moves, however, a change in the component parameters must be accompanied by an update for  $\mathbf{b}$ . This could be accomplished by sampling from the full conditional (4.46). The joint proposal  $\{\phi^*, \mathbf{b}^*\}$  will lead to a more efficient exploration of the parameter space than either parameter alone.

## 4.6 *Multiple frame methods*

Many signal processing applications deal with the analysis of signals over some interval of time or space. Frequently the signal will be of a time-varying nature and so a frame-based analysis method is often required, where the parameters are considered constant across each frame and the frames are analysed independently. In this section, it is shown how a prior expectation of the high correlations between the parameters of adjacent frames can be exploited in a Bayesian modelling framework.

There are several factors involved in selecting a suitable analysis interval in a conventional frame-based analysis method. As the frame length is increased, the variances of the parameter estimates are reduced, in the limit, towards the Cramer-Rao lower bound. This often carries an extra computational cost if the parameter estimation scheme is not  $\mathcal{O}(N)$ . If the signal behaviour changes significantly over time then this can reduce the accuracy of the parameter estimates since the original model is no longer strictly correct. The choice of  $N$  is chosen to trade off accuracy against the rate of time-variation of the parameters [61]. Additional constraints on frame size may appear due to the choice of basis functions, for instance to capture low frequency behaviour.

For a large class of signals, the parameters are expected to vary in a slow or well-defined manner over time. It is desired to capitalise upon this prior expectation in the model structure. One possibility is to explicitly model the entire signal and its time-varying parameters. This is likely to lead to a complex model whose parameters are difficult



**Figure 4.4:** A simple graphical model. Statistical dependencies are shown by directed arrows such that the parent node points towards the child.

to estimate. Additionally, the observation consists of a large number of data points which requires an  $\mathcal{O}(N)$  estimation scheme to be efficient. An alternative methodology is to assume that the parameters are sufficiently slowly varying such that they can be held constant across the duration of a frame. This constant parameter is subsequently estimated inside each frame.

There are several ways in which the dependencies between the parameters in different frames can be specified to express the prior expectations. In the next section the concept of graphical modelling is introduced as a representation for these dependencies.

#### 4.6.1 Graphical models

Graphical models are a convenient method for the representation of the dependencies between observations, model parameters and their hyperparameters. They are of particular benefit as the model increases in complexity, as it affords much simpler reasoning about the parameter dependencies.

Figure 4.4 shows an example of a simple graphical model — each node represents a stochastic variable which may be an observation or may be a parameter. The nodes are connected by directed arrows and form a *directed acyclic graph*<sup>5</sup>. Each arrow denotes

<sup>5</sup>More complex forms of graphical models are also possible which are capable of representing deterministic nodes and constants, and distinguish between unknown and observed variables.

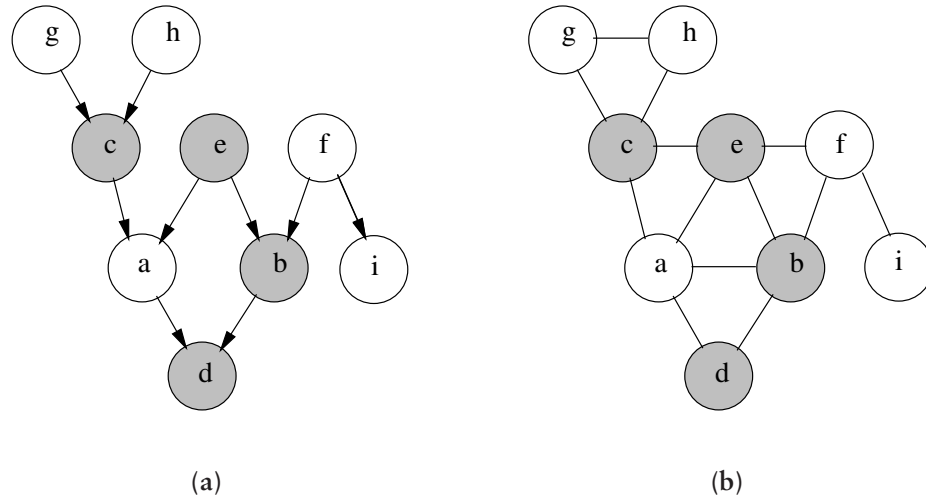


Figure 4.5: Figure (a) shows the Markov blanket of node  $a$  as a directed acyclic graph (DAG). The corresponding undirected conditional independence ('moral') graph is shown in (b).

a statistical dependence, *i.e.*, the probability distribution of that variable is conditional upon the *parents* of that node. In the example,  $a$  and  $b$  are parents of  $d$  ( $a$  is  $b$ 's coparent) which means that the conditional distribution for  $d$  is dependent only on  $a$  and  $b$ :

$$p(d | \cdot) = p(d | a, b) \quad (4.47)$$

If the set of all nodes is  $V$ , the joint distribution for  $V$  is:

$$p(V) = \prod_{v \in V} p(v | \text{parents}[v]). \quad (4.48)$$

In order to simulate a Markov chain (e.g. via the Metropolis-Hastings algorithm) for the purpose of Bayesian inference, it is required to sample from the full conditional distributions of each parameter. If the set of all nodes except  $v$  is  $V_{-\{v\}}$ , the full conditional distribution for  $v$  can be written,

$$\begin{aligned} p(v | V_{-\{v\}}) &\propto \text{all terms containing } v \text{ in joint distribution} \\ &\propto p(v | \text{parents}[v]) \prod_{w \in \text{children}[v]} p(w | \text{parents}[w]). \end{aligned} \quad (4.49)$$

The variables which appear in this expression are the *Markov blanket* of  $v$ . As an example, the Markov blanket for  $a$  is shown in figure 4.5(a) by the shaded nodes.

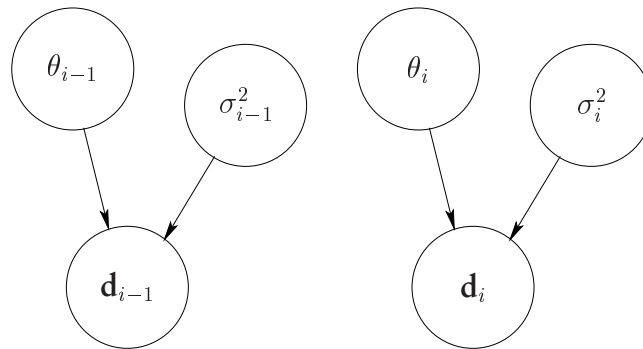


Figure 4.6: Independent frames.

The benefits of this graphical approach also allow for the possibility of an automatic tool for Bayesian inference — in fact, such software exists, BUGS [138, 149] being one such example. By specifying probability distributions which can be sampled from easily, a Gibbs sampler can be implemented automatically. The ability to obtain full conditional distributions in this manner is also beneficial for efficient implementation of the Metropolis-Hastings algorithm since the ratio of posteriors in the M-H acceptance function reduces to the ratio of the full conditionals for the updated parameters (§3.4.3).

It is useful to note that nodes with no ascendants usually have prior distributions which are constant, and so the constants describing the distributions are usually omitted in the interests of clarity. Spiegelhalter *et al.* provide an interesting case study into the use of graphical models [139].

The dependencies can be made more explicit by forming an undirected conditional independence graph, also known as a *moral graph*. It is constructed from the directed acyclic graph by dropping directions and marrying parents (‘moralising’). Conditional independence is observed through separation. The Markov blanket of node  $a$  is shown in the moral graph of figure 4.5(b). Node  $a$  is connected directly to each variable in its Markov blanket. Variables  $\{i, b\}$  are conditionally independent given (that is, separated by)  $f$ .

For multiple frames of estimation there are several configurations which might be employed, as illustrated in figures 4.6–4.8.

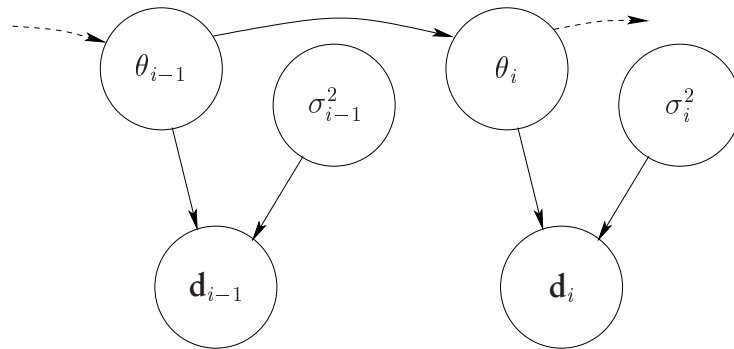


Figure 4.7: Markovian parameter dependence.

#### 4.6.2 Independent frames

In the first configuration, each frame of data is analysed independently of the others (figure 4.6). It is the simplest and most naïve approach since it makes the fewest assumptions about the variation of the model parameters  $\theta_i$  and the error variance  $\sigma_{e_i}^2$ . In a typical time-varying signal processing problem, the parameter inferences over time may be tracked or plotted, and further inferences about long-term behaviour made. This approach does have the disadvantage that inferences are made at two different levels and that knowledge assumed or inferred about the long-term behaviour cannot affect the low level estimation.

#### 4.6.3 Markovian dependence

One way of incorporating basic assumptions about the variation of parameters over time is to impose a Markovian dependence between parameters in successive frames (figure 4.7). The fundamental assumption of this model is that the parameters vary slowly between frames such that the value in one frame is close to that in the previous frame. The prior distribution for the parameters is centred upon the estimate obtained from the previous frame, with a spread parameter reflecting the expected rate of variation. This approach is still quite simple to implement as it requires only the specification of a different prior to the independent frames method. A further advantage is that a low-latency implementation is possible (determined by the frame length). The simplest first order Markovian dependence means that each estimate of the parameter embodies all

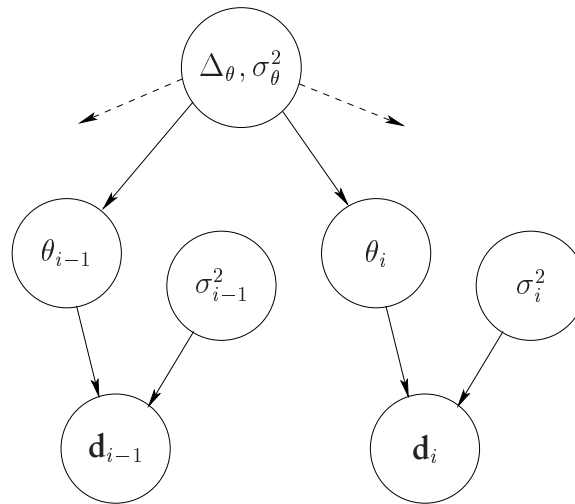


Figure 4.8: Hierarchical parameter dependence.

of its previous history, hence trajectory information cannot be represented unless higher order Markovian dependencies are used.

### *Hierarchical dependence*

A more powerful approach to multiple frame modelling is to explicitly specify a model of the parameter variation over time. *Variational hyperparameters* are employed to describe the evolution of the parameters over time (figure 4.8). These may reflect prior knowledge of the parameter variation in terms of a static prior distribution, or knowledge of the time-varying characteristics of the parameter, for instance constant, linear or autoregressive. The parameter space of the model is augmented by these hyperparameters and estimation of the parameters in each frame and also the hyperparameters is performed simultaneously. In many applications these may be of more interest than the parameters themselves as they describe the underlying variation of the parameters. The use of hyperparameters to strengthen the inference for multiple components is known as *parameter tying* in graphical models. A set of  $N_f$  successive frames are grouped together into a *block* and, since these are all estimated jointly, this form of analysis will have a greater latency than the independent or Markovian methods. The next section will be devoted to discussion of this hierarchical modelling scheme for a single component

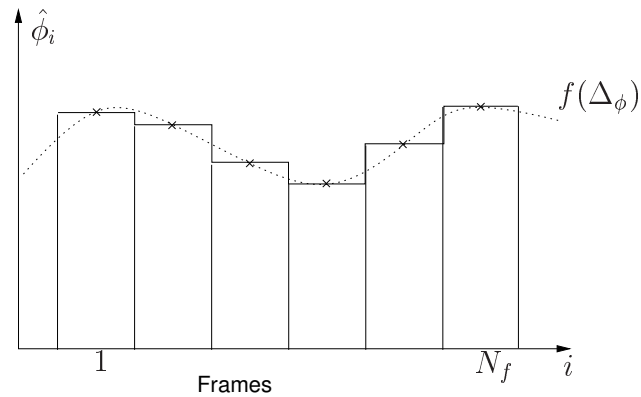


Figure 4.9: The function  $f(\Delta_\phi, i)$  provides a prior estimate of the parameter  $\phi$  in frame  $i$ .

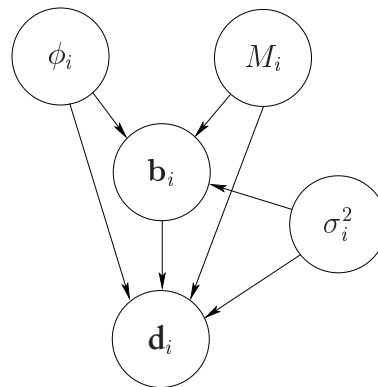
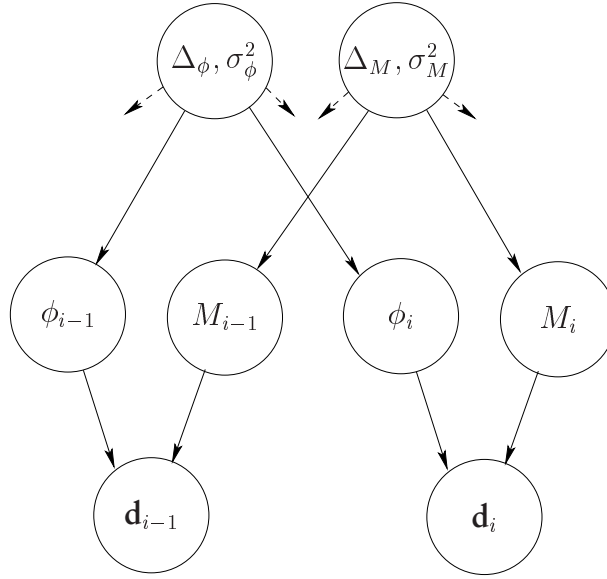


Figure 4.10: Graphical model for a single component GLM employing a  $g$ -prior for  $\mathbf{b}$ .

general linear model.

#### 4.6.4 Hierarchical multiple frame models

There is a great deal of flexibility regarding the specification of prior information about the variation of the model parameters. One option which is quite generic is to envisage a predictor function  $f(\Delta_\phi, i)$  which generates estimates  $\hat{\phi}_i$  of the parameter  $\phi_i$  as a function of frame number  $i$  (figure 4.9). The prior distribution of  $\phi_i$  is a distribution



**Figure 4.11:** Graphical model for a single component multiple frame GLM with marginalised amplitudes and error variances.

centred upon  $f(\Delta_\phi, i)$  with spread parameter  $\sigma_\phi^2$ , for instance a Gaussian,

$$\begin{aligned} p(\phi_i | \Delta_\phi, \sigma_\phi^2) &= \text{N}(\phi_i; f(\Delta_\phi, i), \sigma_\phi^2). \\ &= \frac{1}{(2\pi\sigma_\phi^2)^{\frac{1}{2}}} \exp \left[ -\frac{(\phi_i - f(\Delta_\phi, i))^2}{2\sigma_\phi^2} \right]. \end{aligned} \quad (4.50)$$

The variational hyperparameters  $\Delta_\phi$  are chosen according to the type of the predictor function. The simplest case is for a prior which is constant over the block,  $f(\Delta_\phi, i) = k_\phi$  corresponding to  $\Delta_\phi = \{k_\phi\}$ . A linear variation could be represented by  $f(\Delta_\phi, i) = k_\phi(1 + m_\phi i)$  where  $\Delta_\phi = \{k_\phi, m_\phi\}$ .

In many cases, specific knowledge of the absolute values of the parameters will not be available. More commonly though, knowledge about the likely variation will be available, for instance that the rate of variation of the parameters is constrained to be low. In this instance the hyperparameters describing  $\Delta_\phi$  can be regarded as unknowns to be determined along with the other parameters.

A key assumption embedded within this model structure is that the error variance  $\sigma_{e_i}^2$  within each frame is *a priori* independent. This is important as it means that the frames

are conditionally independent given the variational hyperparameters  $\{\Delta_\theta, \sigma_\theta^2\}$  and also the error variance can be marginalised in each frame. The likelihood of all frames is the product of the likelihood of each frame and the likelihood is not dependent upon the variational hyperparameters,

$$p(\{\mathbf{d}_i\} | \{\theta_i, \sigma_{e_i}^2\}, \Delta_\theta, \sigma_\theta^2) = \prod_{i=1}^{N_f} p(\mathbf{d}_i | \theta_i, \sigma_{e_i}^2). \quad (4.51)$$

The joint posterior is hence obtained

$$\begin{aligned} p(\{\theta_i, \sigma_{e_i}^2\}, \Delta_\theta, \sigma_\theta^2 | \{\mathbf{d}_i\}) &\propto p(\{\mathbf{d}_i\} | \{\theta_i, \sigma_{e_i}^2\}, \Delta_\theta, \sigma_\theta^2) p(\{\theta_i, \sigma_{e_i}^2\}, \Delta_\theta, \sigma_\theta^2) \\ &= \prod_{i=1}^{N_f} (p(\mathbf{d}_i | \theta_i, \sigma_{e_i}^2) p(\theta_i | \Delta_\theta, \sigma_\theta^2, \sigma_{e_i}^2) p(\sigma_{e_i}^2)) \times p(\Delta_\theta, \sigma_\theta^2) \end{aligned} \quad (4.52)$$

### General linear model

Considering the general linear model form introduced earlier, the graphical model for a single frame is depicted in figure 4.10. The amplitudes and error variance can be marginalised as before to obtain the model structure shown in figure 4.11. The basis function parameters  $\phi_i$  are assigned a Gaussian prior whilst the model order prior  $p(M_i | \Delta_M, \sigma_M^2)$  must be a discrete distribution, for instance Poisson or a truncated, discretised Gaussian.

It is assumed that the variational hyperparameters for each component are *a priori* independent,

$$p(\Delta_\phi, \sigma_\phi^2, \Delta_M, \sigma_M^2) = p(\Delta_\phi) p(\sigma_\phi^2) p(\Delta_M) p(\sigma_M^2). \quad (4.53)$$

In order to simulate a Markov chain with the joint posterior of all block parameters as the stationary distribution, the Metropolis-Hastings algorithm is once again employed. Its efficient implementation requires careful choice of transition kernels and proposal densities. Initially, attention is turned to the full conditional distributions of the model parameters.

The full conditional distributions for the model parameters and hyperparameters can be obtained simply by appealing to the structure of the graphical model (from eq. (4.49)).

The number of terms in the conditional can be reduced by considering only those in the Markov blanket of the parameter of interest. Firstly for the parameters in each frame,

$$\begin{aligned} p(\phi_i|\bullet) &= p(\phi_i | \mathbf{d}_i, M_i, \Delta_\phi, \sigma_\phi^2) \\ &\propto p(\mathbf{d}_i | \phi_i, M_i) p(\phi_i | \Delta_\phi, \sigma_\phi^2) \end{aligned} \quad (4.54)$$

$$\begin{aligned} p(M_i|\bullet) &= p(M_i | \mathbf{d}_i, \phi_i, \Delta_M, \sigma_M^2) \\ &\propto p(\mathbf{d}_i | \phi_i, M_i) p(M_i | \Delta_M, \sigma_M^2) \end{aligned} \quad (4.55)$$

and for their hyperparameters

$$\begin{aligned} p(\Delta_\phi|\bullet) &= p(\Delta_\phi | \{\phi_i\}, \sigma_\phi^2) \\ &\propto p(\{\phi_i\} | \Delta_\phi, \sigma_\phi^2) p(\Delta_\phi) \end{aligned} \quad (4.56)$$

$$\begin{aligned} p(\sigma_\phi^2|\bullet) &= p(\sigma_\phi^2 | \{\phi_i\}, \Delta_\phi) \\ &\propto p(\{\phi_i\} | \Delta_\phi, \sigma_\phi^2) p(\sigma_\phi^2). \end{aligned} \quad (4.57)$$

The full conditionals for  $\Delta_M$  and  $\sigma_M^2$  take on a similar form.

### *Choice of predictor function*

For many types of signal, it is likely that the variation of parameters with time will be sufficiently slow to be adequately modelled by a constant or linear prediction function. Naturally, if the longer term behaviour is well structured, more esoteric prediction functions may be appropriate. Other potential forms include sinusoids, AR models and splines, but usually the motivation for a particular form comes from knowledge of the underlying physical behaviour of the system. In some applications, it may be that the variational hyperparameters are in fact of direct interest in the inference, *e.g.*, for detecting trends. The constant and linear prediction functions are now considered, and it is shown how they may be integrated into the MCMC simulation scheme.

### *Constant prediction function*

The prediction function is taken to be  $f(\Delta_\phi, i) = k_\phi$ , with hyperparameter  $\Delta_\phi = \{k_\phi\}$ . The *a priori* distribution for the hyperparameter  $k_\phi$  could be perhaps uniform over a

range of expected values or some other distribution centred upon (for instance) the  $k_\phi$  estimate from the previous block. Assuming a Gaussian prior for  $\phi$ , the full conditional for  $\Delta_\phi$  is

$$p(\Delta_\phi | \{\mathbf{d}_i\}, \bullet) \propto \frac{p(k_\phi)}{(2\pi\sigma_\phi^2)^{\frac{N_f}{2}}} \exp \left[ -\frac{\sum_{i=1}^{N_f} (\phi_i - k_\phi)^2}{2\sigma_\phi^2} \right] \quad (4.58)$$

The exponential term can be written as a Gaussian distribution of  $k_\phi$ ,

$$\begin{aligned} \frac{\sum_{i=1}^{N_f} (\phi_i - k_\phi)^2}{2\sigma_\phi^2} &= \frac{N_f (k_\phi - b)^2 + c}{2\sigma_\phi^2} \\ b &= \frac{1}{N_f} \sum_{i=1}^{N_f} \phi_i \quad c = \sum_{i=1}^{N_f} (\phi_i)^2 - \frac{1}{N_f} \left( \sum_{i=1}^{N_f} \phi_i \right)^2 \end{aligned} \quad (4.59)$$

hence the full conditional for  $k_\phi$  is

$$p(k_\phi | \{\mathbf{d}_i\}, \bullet) \approx p(k_\phi) \times \mathcal{N} \left( k_\phi; \sum_{i=1}^{N_f} \frac{\phi_i}{N_f}, \frac{\sigma_\phi^2}{N_f} \right). \quad (4.60)$$

If the prior is diffuse then the mode of this distribution is  $\mathbb{E}[\phi_i]$ .

If an inverse gamma distribution is adopted for the prior  $p(\sigma_\phi^2) = \text{IG}(\sigma_\phi^2; \alpha_\phi, \beta_\phi)$ , then this yields a full conditional which is also inverse gamma

$$p(\sigma_\phi^2 | \bullet) = \text{IG} \left( \alpha_\phi + \frac{N_f}{2}, \beta_\phi + \frac{1}{2} \sum_{i=1}^{N_f} (\phi_i - f(\Delta_\phi, i))^2 \right). \quad (4.61)$$

The mode of this distribution is close to  $\text{Var}[\phi_i] = \sum_i (\phi_i - \mathbb{E}[\phi_i])^2 / N_f$ .

### *Linear prediction function*

An alternative to a constant predictor is a linear predictor, which has a number of benefits. Having an extra degree of freedom, the function would naturally produce a better fit to the parameters than a constant predictor without excessive overfitting. It also means that the parameter values over each block can be made piecewise linear continuous (PLC) such that the values over subsequent blocks of  $N_f$  frames can be aligned. Constraints could potentially be put on the prior probability distributions of

the hyperparameters to reflect this extra information — this step is simply extrapolating from the assumption of predictable behaviour over  $N_f$  frames to longer time-scales.

The prediction function is written as  $f(\Delta_\phi, i) = k_\phi + m_\phi(i - \bar{N})$ , with hyperparameters  $\Delta_\phi = \{k_\phi, m_\phi\}$  and  $\bar{N} = (N_f + 1)/2$ . This parametrisation is chosen since it makes the conditional posteriors of  $k_\phi$  and  $m_\phi$  conditionally independent; this occurs since changes in  $m_\phi$  cause a rotation about  $(k_\phi, \bar{N})$  rather than about the origin. Expanding the quadratic term in the full conditional firstly as a quadratic in  $k_\phi$

$$\begin{aligned} & \sum_{i=1}^{N_f} (\phi_i - k_\phi - m_\phi(i - \bar{N}))^2 \\ &= \sum_{i=1}^{N_f} (\phi_i - m_\phi(i - \bar{N}))^2 - 2k_\phi \sum_{i=1}^{N_f} (\phi_i - m_\phi(i - \bar{N})) + N_f k_\phi^2, \quad (4.62) \\ &= N_f \left( k_\phi - \sum_{i=1}^{N_f} \frac{\phi_i}{N_f} \right)^2 + C \end{aligned}$$

and so the full conditional for  $k_\phi$  is a product of its prior and a Gaussian as in the constant predictor case

$$p(k_\phi | \{\mathbf{d}_i\}, \bullet) \propto p(k_\phi) \times \mathcal{N} \left( k_\phi; \sum_{i=1}^{N_f} \frac{\phi_i}{N_f}, \frac{\sigma_\phi^2}{N_f} \right). \quad (4.63)$$

Similarly expanding as a quadratic in  $m_\phi$

$$\begin{aligned} & \sum_{i=1}^{N_f} (\phi_i - k_\phi - m_\phi(i - \bar{N}))^2 \\ &= \sum_{i=1}^{N_f} (\phi_i - k_\phi)^2 - 2 \sum_{i=1}^{N_f} m_\phi (\phi_i - k_\phi)(i - \bar{N}) + m_\phi^2 \sum_{i=1}^{N_f} (i - \bar{N})^2, \quad (4.64) \\ &= \sum_{i=1}^{N_f} (i - \bar{N})^2 \left( m_\phi - \sum_{i=1}^{N_f} \frac{\phi_i(i - \bar{N})}{\sum_{i=1}^{N_f} (i - \bar{N})^2} \right)^2 + C \end{aligned}$$

and so the full conditional becomes

$$\begin{aligned} p(m_\phi | \{\mathbf{d}_i\}, \bullet) &\approx p(m_\phi) \times \mathcal{N} \left( m_\phi; \frac{\sum_{i=1}^{N_f} \phi_i(i - \bar{N})}{\sum_{i=1}^{N_f} (i - \bar{N})^2}, \frac{\sigma_\phi^2}{\sum_{i=1}^{N_f} (i - \bar{N})^2} \right) \\ \sum_{i=1}^{N_f} (i - \bar{N})^2 &= \frac{N_f(N_f + 1)(N_f - 1)}{12} \end{aligned} \quad (4.65)$$

There is therefore little extra computation penalty in using a linear prediction function model, since the sampling step would only require two draws from univariate Gaussians, which can be done efficiently. When the prior isn't very diffuse and the Gaussians are used as proposals for a Metropolis-Hastings step, acceptance of the proposal is a function of the ratio of full conditionals which can be calculated much more efficiently than the full joint posterior.

## 4.7 Transition kernels for multiple frame methods

A naïve implementation of the Metropolis-Hastings algorithm for the multiple frame model would simply iterate through the update moves for each parameter and hyperparameter in turn. The effect is that each frame samples its parameters using conditional independence, independence and perturbation steps as before (see §3.4.3), whilst the hyperparameter updates adapt to the mean of the values in each frame.

Convergence for this algorithm could be slow if there is an outlier or a second interfering component in one of the frames, as this will distort the value of  $\Delta_\phi$  from the mean of the parameter values which are close together. Moves towards the global maximum may in fact be rejected since there are high posterior correlations between  $\{M_i\}$  and  $\Delta_M$  and between  $\{\phi_i\}$  and  $\Delta_\phi$ , so the proposal of a new parameter value should be accompanied by an update of its hyperparameters. It is also inefficient as the assumption of slowly varying parameters is not being fully exploited. A more efficient scheme is possible by recognising the dependence between the parameters  $\{\phi_i\}$  and their hyperparameters  $\{\Delta_\phi, \sigma_\phi^2\}$  such that a move is proposed jointly for them.

For the purposes of producing an efficient proposal distribution, the correlation between the parameters is made more explicit. A reparameterisation is employed  $\phi_i = f(\Delta_\phi, i) + \xi_i$  which represents the value of the parameter in each frame as a deviation from the predictor  $f(\Delta_\phi, i)$ . The likelihood in each frame is now a function of  $\Delta_\phi$  and  $\xi_i$ . The full conditional for  $\Delta_\phi$  therefore is dependent upon all the observations  $\{\mathbf{d}_i\}$ ,

$$p(\Delta_\phi | \{\mathbf{d}_i, \xi_i\}, \bullet) \propto p(\Delta_\phi) \prod_i p(\mathbf{d}_i | \Delta_\phi, \xi_i, \bullet) \quad (4.66)$$

A conditional independence sampling distribution for  $\Delta_\phi$  can be produced from this full conditional by evaluating it for a range of values of  $\Delta_\phi$  with the deviations  $\xi_i$  set to zero.

A proposal distribution  $q(\Delta_\phi^*)$  is therefore formed

$$q(\Delta_\phi^*) \propto p(\Delta_\phi^*) \prod_i [\mathbf{d}_i^t \mathbf{P}_i \mathbf{d}_i + 2\beta_e]^{-\epsilon} \quad (4.67)$$

which is conditionally independent upon the current model order terms  $\{M_i\}$ . The proposal value is sampled from it

$$\Delta_\phi^* \sim q(\Delta_\phi^*) \quad (4.68)$$

Proposals for the other parameters can be generated from their full conditionals using the proposed hyperparameter but with a fixed variance<sup>6</sup>  $\sigma_{\text{prop}}^2$

$$\begin{aligned} \phi_i^* &\sim q(\phi_i^*; \Delta_\phi^*) \\ q(\phi_i^*; \Delta_\phi^*) &= p(\phi_i^* | \Delta_\phi^*, \sigma_{\text{prop}}^2) p(\mathbf{d}_i | \phi_i^*, M_i^k) \end{aligned} \quad (4.69)$$

Finally an update move for  $(\sigma_\phi^2)^*$  is proposed

$$(\sigma_\phi^2)^* \sim q((\sigma_\phi^2)^*; \Delta_\phi^*, \{\phi_i^*\}) \quad (4.70)$$

where  $q((\sigma_\phi^2)^*; \bullet)$  is the full conditional (4.61). This joint updating scheme is summarised in algorithm 4.2.

The main computational burden of this method is the generation of candidate values for the  $\Delta_\phi^*$  proposal (4.67). This requires the marginalised likelihood  $p(\mathbf{d}_i | \phi_i, M_i)$  to be evaluated for each frame. This expression also appears in the full conditional of each  $\phi_i$  and so great computational savings can be made by firstly calculating the marginalised likelihood separately for each frame. The hyperparameter proposal distribution is then formed from the product of the likelihoods and the hyperparameter prior and then the parameter proposals can be produced from the product of the likelihoods and their priors.

Major efficiency savings may also be made if the distribution  $q(\Delta_\phi^*)$  is calculated initially using a value  $M_i$  which is held constant over all frames, perhaps taken from the mean

---

<sup>6</sup>This is necessary because the basis function parameters should be independent of the current state, and so  $(\sigma_\phi^2)^k$  should not be used. A scheme which selects from several fixed values of  $\sigma_{\text{prop}}^2$  could also be employed, as described in §4.3.

of the prior. The proposal could be evaluated for a range of values over the support of  $\phi_i$ . If it is expensive to calculate a large number of points then it may be possible to use fewer points and then create a piecewise linear distribution, which can subsequently be sampled with a very small overhead. The ratio of transition probabilities required for evaluation of the M-H acceptance function is

$$\frac{q(\Delta_\phi^k) q((\sigma_\phi^2)^k; \Delta_\phi^k, \{\phi_i^k\}) \prod_i q(\phi_i^k; \Delta_\phi^k)}{q(\Delta_\phi^*) q((\sigma_\phi^2)^*; \Delta_\phi^*, \{\phi_i^*\}) \prod_i q(\phi_i^*; \Delta_\phi^*)}. \quad (4.71)$$

---

**Algorithm 4.2:** Joint update move for basis function parameters

```

function: Block update: { { $\phi_i$ },  $\Delta_\phi$ ,  $\sigma_\phi^2$  }
  Generate proposal distribution  $q(\Delta_\phi^*)$  for current { $M_i$ }
   $\Delta_\phi^* \sim q(\Delta_\phi^*)$ 
  for  $i = 1 \dots N_f$  do
     $\phi_i^* \sim q(\phi_i^*; \Delta_\phi^*)$ 
  end for
   $(\sigma_\phi^2)^* \sim q((\sigma_\phi^2)^*; \Delta_\phi^*, \{\phi_i^*\})$ 
  MH-accept(  $\theta^*$ ,  $\theta^k$  )

```

---

The update move for the model order parameters follows similar lines. For given  $\{\phi_i\}$  a proposal distribution  $q(\Delta_M^*)$  can be generated using several values of  $\Delta_M^*$ . Matrix partitioning techniques for efficiently evaluating extra basis functions may be useful here, as discussed in appendix A. Perturbation steps for  $\{M_i\}$  from a distribution  $q(M_i^*; \Delta_M^*)$  and an update of  $(\sigma_M^2)^*$  then follow.

These block update moves from the parameters and model order may also be combined with perturbation moves separately for each  $M_i$  and  $\phi_i$ . This allows the parameters to adapt to the local variation in each frame. On each iteration there is a probability  $\lambda_{\text{block}}$  that the block update move is chosen, otherwise local updates are employed. The local updates are perturbation steps where a candidate value is sampled from a distribution centred upon the current value and then accepted according to the M-H

acceptance function. The hyperparameters are then updated by sampling from their full conditionals. This method is summarised in algorithm 4.3.

---

**Algorithm 4.3:** Single component multiple frame Metropolis-Hastings scheme with local and joint update moves

```

for  $n_i = 1 \dots N_{\text{iter}}$  do
   $u \sim \mathbb{U}_{[0,1]}$ 
  if ( $u < \lambda_{\text{block}}$ ) then
    Block update  $\{ \{\phi_i\}, \Delta_\phi, \sigma_\phi^2 \}$ 
    Block update  $\{ \{M_i\}, \Delta_M, \sigma_M^2 \}$ 
  else
    for  $i = 1 \dots N_f$  do
      Perturbation update  $\phi_i$ 
      Perturbation update  $M_i$ 
    end for
    Update  $\Delta_\phi, \sigma_\phi^2$ 
    Update  $\Delta_M, \sigma_M^2$ 
  end if
end for

```

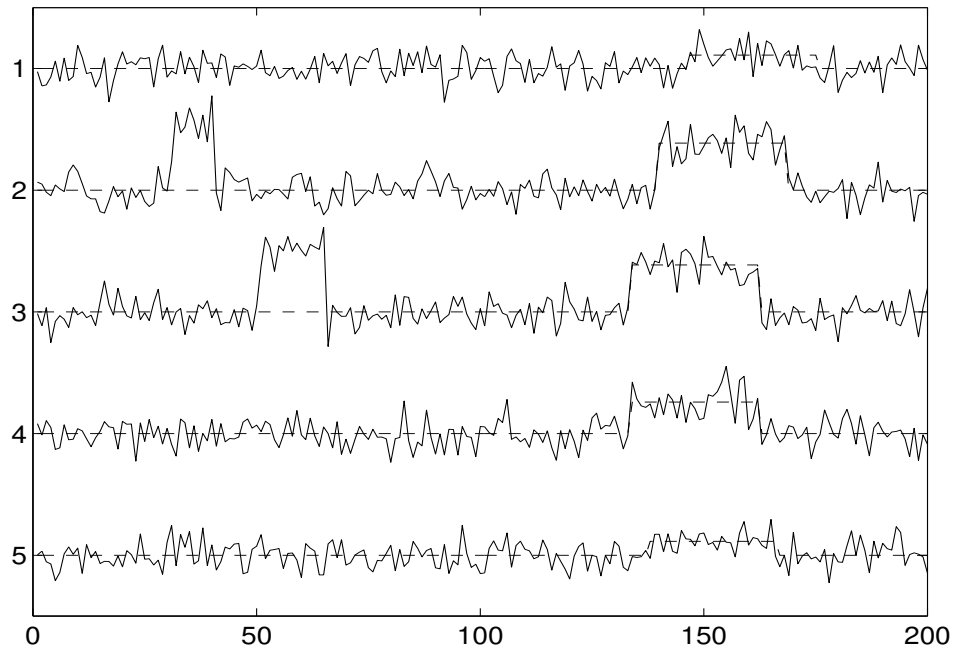
---

#### 4.7.1 Multiple frame simulation

An example is now presented to illustrate the use of the multiple frame model. The model is used to represent a rectangular pulse in white Gaussian noise whose width  $y$  is unknown and whose location in each frame  $x_i$  is known to be approximately constant with mean  $c$  and variance  $\sigma_c^2$  which are also unknown. Extra rectangular pulses are added in two frames of data to simulate a typical application where the signal is prone to crosstalk or echoes. The estimation task is to find the MAP estimates of  $\{c, \sigma_c^2, y, \mathbf{b}, \{x_i\}_{N_f}\}$ .<sup>7</sup>

---

<sup>7</sup>The order of this model is fixed,  $M_i = 1, \forall i$ .



**Figure 4.12:** Rectangular pulse data used for multiple frame example. Frames 2 and 3 also contain interfering pulse components. The MAP estimate model reconstructions are shown as dotted lines.

The basis function of the rectangular pulse is

$$g(x_i) = \mathbb{I}_{[x_i, x_i+y]}(x_i). \quad (4.72)$$

The projection term  $\mathbf{d}_i^t \mathbf{P}_i \mathbf{d}_i$  in the integrated likelihood (4.44) can be considerably simplified for this basis function

$$\mathbf{d}_i^t \mathbf{P}_i \mathbf{d}_i = \|\mathbf{d}_i\|^2 - \frac{\delta^2}{(1 + \delta^2)} \frac{\|\sum_{n=x_i}^{x_i+y} \mathbf{d}_i(x)\|^2}{y}. \quad (4.73)$$

The projection can also be calculated for candidate values of  $x_i = 1 \dots N$  for a given  $y$  by performing the convolution of a vector of ones of length  $y$  with the data, as described in §4.3.1. The method of algorithm 4.3 is employed using a combination of joint updates over the entire block (for  $\{c, \sigma_c^2, \{x_i\}_{N_f}\}$ ) and perturbations of  $\{x_i\}_{N_f}$  followed by hyperparameter updates.

The data used in this example is shown in figure 4.12. A set of rectangular pulses with start locations centred upon a mean value  $c = 139$  were generated and white Gaussian

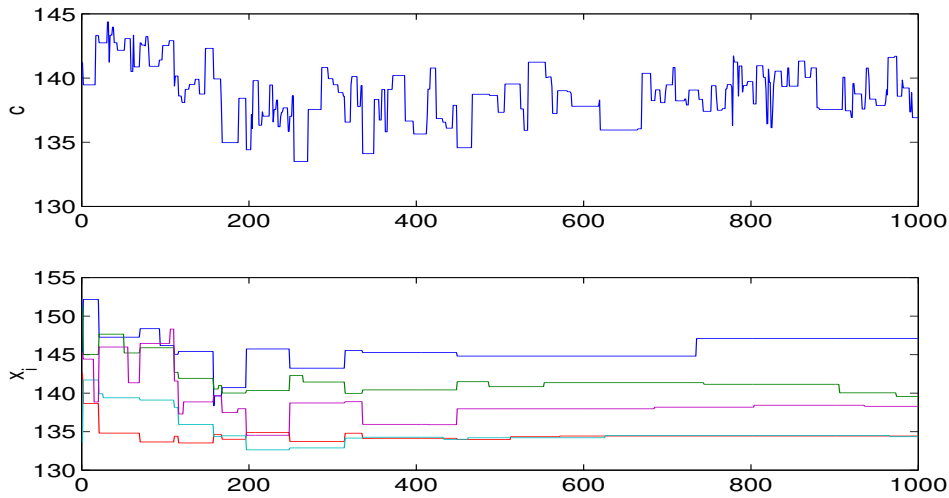


Figure 4.13: Output of the Markov chain for the hyperparameter  $c$  and location parameters  $x_i$ .

noise was added. The signal-to-noise ratio in each frame was approximately -3dB. Interference pulses of similar amplitude were added in two of the frames. The output of the Markov chain for the hyperparameter  $c$  and parameters  $x_i$  is shown in figure 4.13. The joint update move executed on the first iteration succeeds in locating the correct region of the posterior distribution for  $c$  and  $\{x_i\}$ . Detection of the interfering terms is suppressed by virtue of a sufficiently low choice for  $\sigma_{\text{prop}}^2$  (25 in this instance). A method which estimates each frame independently or even with a Markovian dependence may be likely to detect the strong interfering components of frames 2 and 3. This method however, which explicitly assumes that the pulses occur at similar locations in each frame, detects the correct locations jointly. The MAP estimates for the model parameters were taken from the state of the Markov chain with the highest posterior probability. The reconstructions formed from these estimates are shown in figure 4.12 superimposed over the data with dotted lines. All were within one sample of the correct values.

## 4.8 Conclusions

The single component general linear model is a flexible model for many applications. This chapter has detailed its formulation and simulation-based techniques for parameter estimation. Complex models can be computationally expensive to simulate and so particular attention has been paid to the choice of efficient transition kernels within a Metropolis-Hastings algorithm. Prior knowledge of the model structure is employed to produce efficient blocking schemes and transition kernels.

The extension of the basic model to deal with time-varying signals has been addressed, exploiting the high correlations which typically occur between adjacent frames of data. A hierarchical graphical modelling framework was introduced to express the functional form of the time variation of the parameters. On the assumption that the parameters vary slowly over time, an efficient joint update move can be implemented for the correlated parameters. It is often the time-varying behaviour of signal parameters which is of interest, and in this framework it is possible to model this behaviour explicitly rather than conduct independent-frame analyses and then make inferences on the results. Even longer term behaviour may be modelled by specifying Markovian hyperpriors such that the variational hyperparameters are *a priori* dependent upon the inferred value in the previous block.

The common assumption of a Gaussian error is not always justifiable in real-world applications and there may be ‘intelligible’ noise in the signal which is of a similar form to the desired component. There are several solutions to this limitation: use of a more appropriate error model, use of a multiple component model (discussed in chapter 5), or use of a single component model robust to this form of error. An example has shown how this may be achieved for a single component multiple frame model. Knowledge of the expected time-variation of the parameters is exploited to specify proposal densities for joint update moves.

A final benefit of a multiple frame approach is that it can actually be substantially more efficient than a method which analyses frames independently. The multiple frame technique exploits the redundancies inherent in data whose characteristics have high correlations by searching explicitly for the commonality rather than estimating parameters first and then finding common characteristics as a post-processing step.



# *Detection and Estimation of Multiple Component Models*

---

# 5

## *5.1 Introduction*

In chapter 4 a modelling framework for a single component signal was presented. In this chapter, the model is generalised to signals which consist of an additive mixture of a number of different components. There are two motivations for this more general model. Firstly, the detection and estimation of multiple signal components within a mixture is an important signal processing problem with applications in audio processing, communications and image processing. The second motivation is that many ostensibly single component signals found in practical applications deviate substantially from the assumptions inherent in a single component model, particularly that the observation can be modelled as a single component with a Gaussian error. There are many effects in ‘real-world’ signals which cause this deviation from ideal behaviour, of which some of the most common are:

- **Interfering components:** one component is prominent in the mixture but others are present at a lower level, *e.g.*, due to crosstalk or the inability at the source to isolate a single component entirely, as with desktop microphones picking up nearby speakers
- **Non-linearity:** distortion in the signal path or the characteristics of a transducer may introduce extra components which are correlated with the component of interest but which are not represented in the model
- **Modelling error:** the model is generally idealised and will have made a number of simplifying assumptions about the nature of the data

The effect of all of these is that the assumption of a Gaussian i.i.d. error term may no longer be tenable. The multiple frame methods of chapter 4 show how the non-ideal conditions of real-world signal processing may be countered by exploitation of the prior expectations of the time-varying behaviour of the signal. In many cases, however, it is preferable to recover all of the interesting signal components from a mixture and perform inference at a higher level as to whether the component is really of interest. This chapter describes methods suited to the analysis of signals which are composed of linear mixtures of multiple homogeneous or heterogeneous components. The number of each type of component (if any), their model orders and parameter values are all unknown *a priori*.

In section 5.2 the multiple component general linear model is formulated and posed in a Bayesian framework. Section 5.3 describes how MCMC methods may be applied to the parameter estimation problem, and in particular the use of reversible jump techniques for dealing with variable numbers of components. A major emphasis is put on the development of efficient techniques for multiple component detection and estimation to make real-time MCMC simulation viable. A heterogeneous mixture example follows in section 5.4. Section 5.5 extends the multiple component model to time-varying signals, expressed in terms of a Bayesian graphical model extending across multiple frames with correlated parameters.

---

## 5.2 *The multiple component general linear model*

### 5.2.1 *Model formulation*

The general linear model of section 4.2 can be generalised for a multiple component signal,

$$\mathbf{d} = \sum_{q \in \mathcal{Q}} \mathbf{G}^q \mathbf{b}^q + \mathbf{e}. \quad (5.1)$$

There are a maximum of  $Q$  components, each of which can be ‘switched into’ the model when a binary indicator variable  $\Gamma^q$  is true. The set of all component indices which are

switched on is  $\mathbb{Q} = \{q : \Gamma^q = 1, q = 1 \dots Q\}$ . The corresponding model composition is written as  $\mathcal{M}_{\mathbb{Q}}$  such that  $p(\mathcal{M}_{\mathbb{Q}}) = p(\{\Gamma^q\})$ . The parameters of this model are  $\{\theta^q\}_{\mathbb{Q}}$  where  $\theta^q = \{\mathbf{b}^q, \phi^q, M^q\}$ .<sup>1</sup>

For compactness of notation therefore, the joint posterior distribution over all candidate models is written as  $p(\mathcal{M}_{\mathbb{Q}}, \{\theta^q\}_{\mathbb{Q}}, \sigma_e^2 | \mathbf{d})$ . This representation follows that of Green's reversible jump [58] (amongst others) where a posterior probability distribution is defined over the union of the products of each candidate model and its parameters. All inference is then based upon this distribution.

It may be noted that the sum of a number of GLMs can itself be written as a single GLM by extending the number of columns in the basis matrix since a GLM is defined as the sum of a number of basis functions. Therefore the question arises as to how to determine the best grouping of basis functions into components. The intuitive solution is to group the basis functions by their statistical or logical dependencies. Gestalt grouping principles (see §2.4.4) may be employed for data which has a perceptual significance. Such principles include similarity (proximity in time, space or parameters), common fate (characteristics common to a number of elements), and good continuation (favouring smooth variation over large jumps).

For instance if a model is created to represent a stream of rectangular pulses then this would be regarded as a single signal component if the pulses have a common shape function and regular spacing. Conversely, a stream of pulses with different shapes and irregular spacing share few attributes and so would be considered as separate components. In theory any grouping of the parameters could be chosen, but much more efficient algorithms can be produced if the parameters with high posterior correlations are grouped together.

For the model composition  $\mathcal{M}_{\mathbb{Q}}$  the model equation (5.1) is rewritten in terms of a single GLM by forming the *composite basis matrix*  $\mathbf{G}^c$  and corresponding vector of linear coefficients  $\mathbf{b}^c$ , each constructed as the concatenation of the components  $q \in \mathbb{Q}$ . The indexing variable  $l$  is introduced, such that  $\mathbb{Q} = \{l_1, l_2, \dots, l_{N_{\mathbb{Q}}}\}$ ,  $l_1 < l_2 < \dots < l_{N_{\mathbb{Q}}}$ ,

---

<sup>1</sup>The following shorthand is used to denote the parameters of all model components currently included,  $\{\theta^q\}_{\mathbb{Q}} = \{\theta^q : q \in \mathbb{Q}\}$ .

where  $N_Q = \#\{\mathbb{Q}\} = \sum_{q=1}^Q \Gamma^q$ .

$$\begin{aligned} \mathbf{G}^c &= [\mathbf{G}_{l_1} \mathbf{G}_{l_2} \dots \mathbf{G}_{l_{N_Q}}] \\ \mathbf{b}^c &= [\mathbf{b}_{l_1}^t \mathbf{b}_{l_2}^t \dots \mathbf{b}_{l_{N_Q}}^t]^t \end{aligned} \quad (5.2)$$

The data is therefore expressed as

$$\mathbf{d} = \mathbf{G}^c \mathbf{b}^c + \mathbf{e}. \quad (5.3)$$

which for a fixed model composition is a function of only the parameters in  $\mathcal{M}_{\mathbb{Q}}$ , that is  $\{\theta^q\}_{\mathbb{Q}}$ .

### 5.2.2 Bayesian formulation

Predictably, the estimation task is much more difficult than it is for a single component model. It is desired to determine which of the  $Q$  components are present in the model (*i.e.*, a subset selection problem), and for each of these components, the parameters of  $\mathbf{G}^q$ , the amplitudes of each basis vector  $\mathbf{b}^q$  and the number of columns  $M^q$ . The estimation task now encompasses model subset selection, model order selection and parameter estimation. In the words of Richardson and Green [120], “the number of things you don’t know is one of the things you don’t know”.

The size of the parameter space is now much larger than the single component GLM model — there are  $2^Q$  candidate models, each with unknown parameters and model size, and so an exhaustive search of the parameter space is clearly infeasible. In this section an expression is produced for the joint posterior distribution which will subsequently be employed in the parameter estimation.

The likelihood expression for the multiple component model is written in terms of the error  $\mathbf{e}$

$$\begin{aligned} p(\mathbf{d} | \mathcal{M}_{\mathbb{Q}}, \{\theta^q\}_{\mathbb{Q}}, \sigma_e^2) &= (2\pi\sigma_e^2)^{-\frac{N}{2}} \exp \left[ -\frac{\|\mathbf{e}\|^2}{2\sigma_e^2} \right] \\ \mathbf{e} &= \mathbf{d} - \sum_{q \in \mathbb{Q}} \mathbf{G}^q \mathbf{b}^q. \end{aligned} \quad (5.4)$$

An expression for the joint posterior can be obtained from Bayes’ theorem as

$$p(\mathcal{M}_{\mathbb{Q}}, \{\theta^q\}_{\mathbb{Q}}, \sigma_e^2 | \mathbf{d}) \propto p(\mathbf{d} | \mathcal{M}_{\mathbb{Q}}, \{\theta^q\}_{\mathbb{Q}}, \sigma_e^2) p(\{\theta^q\}_{\mathbb{Q}}, \mathcal{M}_{\mathbb{Q}}, \sigma_e^2) \quad (5.5)$$

and the adopted prior structure is

$$p(\{\theta^q\}_{\mathbb{Q}}, \mathcal{M}_{\mathbb{Q}}, \sigma_e^2) = p(\{\theta^q\}_{\mathbb{Q}} | \mathcal{M}_{\mathbb{Q}}, \sigma_e^2) p(\mathcal{M}_{\mathbb{Q}}) p(\sigma_e^2) \quad (5.6)$$

The dependence upon the state of prior knowledge  $\mathcal{I}$  is implicit and omitted for brevity (see §3.3). The error variance is given an inverse gamma prior,

$$p(\sigma_e^2) = \text{IG}(\sigma_e^2; \alpha_e, \beta_e). \quad (5.7)$$

It is convenient to partition the parameters of each component as  $\theta^q = \{\tilde{\theta}^q, \mathbf{b}^q\}$  where  $\tilde{\theta}^q = \{\phi^q, M^q\}$  since the amplitudes are to be marginalised. The basis function parameters  $\tilde{\theta}^q$  are assumed *a priori* independent,

$$p(\{\theta^q\}_{\mathbb{Q}} | \mathcal{M}_{\mathbb{Q}}, \sigma_e^2) = p(\{\mathbf{b}^q\}_{\mathbb{Q}} | \{\tilde{\theta}^q\}_{\mathbb{Q}}, \mathcal{M}_{\mathbb{Q}}, \sigma_e^2) \prod_{q \in \mathbb{Q}} p(\tilde{\theta}^q) \quad (5.8)$$

From the composite general linear model (5.3) a prior for the amplitudes is defined for the composite amplitude vector  $\mathbf{b}$ . A  $g$ -prior is used which is relatively uninformative and has useful analytical properties, as described in the previous chapter.

$$\begin{aligned} p(\mathbf{b}^c | \{\tilde{\theta}^q\}_{\mathbb{Q}}, \sigma_e^2) &= \text{N}(\mathbf{0}, \sigma_e^2 \Sigma_c) \\ \Sigma_c &= \delta^2 (\mathbf{G}^{ct} \mathbf{G}^c)^{-1} \end{aligned} \quad (5.9)$$

If the basis matrices are orthogonal to each other, *i.e.*, where  $\mathbf{G}^{it} \mathbf{G}^j = \mathbf{0}_{[M_i \times M_j]}$ ,  $i \neq j$  then this is equivalent to ascribing independent priors  $p(\mathbf{b}^q | \tilde{\theta}^q, \sigma_e^2) = \text{N}(\mathbf{0}, \sigma_e^2 \Sigma_q)$  where  $\Sigma_q = \delta^2 (\mathbf{G}^{qt} \mathbf{G}^q)^{-1}$  for  $q \in \mathbb{Q}$ . Following the marginalisation of this composite amplitude vector and also the error variance  $\sigma_e^2$ , the marginal joint posterior is obtained. The marginalisation is detailed in appendix B.1.

$$p(\mathcal{M}_{\mathbb{Q}}, \{\tilde{\theta}^q\}_{\mathbb{Q}} | \mathbf{d}) \propto \frac{p(\{\tilde{\theta}^q\}_{\mathbb{Q}} | \mathcal{M}_{\mathbb{Q}}) p(\mathcal{M}_{\mathbb{Q}})}{(1 + \delta^2)^{\frac{M}{2}} [\mathbf{d}^t \mathbf{P}^c \mathbf{d} + 2\beta_e]^\varepsilon} \quad (5.10)$$

$$\begin{aligned} M &= \sum_{q \in \mathbb{Q}} M^q & \varepsilon &= \frac{N}{2} + \alpha_e \\ \mathbf{P}^c &= \mathbf{I}_N - \frac{\delta^2}{1 + \delta^2} \mathbf{G}^c (\mathbf{G}^{ct} \mathbf{G}^c)^{-1} \mathbf{G}^{ct}. \end{aligned} \quad (5.11)$$

The prior probability of the model composition  $\mathcal{M}_{\mathbb{Q}}$  is

$$p(\mathcal{M}_{\mathbb{Q}}) = p(\{\Gamma^q\}) \quad (5.12)$$

In the instance that all components are of the same type and are *a priori* independent, this simplifies to  $p(\mathcal{M}_{\mathbb{Q}}) = \prod_{q=1}^Q p(\Gamma^q)$ . Typically a Bernoulli prior is employed,  $p(\Gamma^q) = (\alpha_{\Gamma^q})^{\Gamma^q} (1 - \alpha_{\Gamma^q})^{(1-\Gamma^q)}$  as this allows us to assign an independent prior probability for the inclusion of each component. A Poisson prior on the number of components is more commonly found in the literature (*e.g.*, [5, 23]). This model is intended to be more generic insofar as heterogeneous as well as homogeneous component mixtures may be represented. For heterogeneous mixtures it is generally not meaningful to reason *a priori* about the total number of different components. A further reason for the choice of Bernoulli priors is that each component may have established a ‘context’ in previous observations and it is desired to track the evolution of individual components over time. Hence it is more important to reason about the evolution of a collection of individual components rather than the evolution of their number.

### 5.3 Simulation and inference

The posterior distribution is defined over the union of the product of each model and its parameters, for notational simplicity, the shorthand  $\psi = \{\mathcal{M}_{\mathbb{Q}}, \{\tilde{\theta}^q\}_{\mathbb{Q}}\}$  is introduced to refer to the model state. A Markov chain is simulated with the desired posterior distribution as its stationary distribution. The method employed is in the spirit of Green’s reversible jump [58] and Carlin and Chib’s composite model space [17]. The algorithm follows similar lines to that of Godsill’s variable selection [51] and the sinusoidal estimation technique of Andrieu and Doucet [5], where transitions are performed between different subspaces according to the number of components included within the model.

As the parameters of each component are assumed to be more highly correlated with each other than with other components (reflected in the choice of independent or conditionally independent priors), the blocking scheme adopted is that parameters of each component are updated together. More sophisticated update moves to split or combine pairs of components could also be incorporated (for instance see [120, 5]).

The basic method, employing the Metropolis-Hastings algorithm, for a multiple signal component model is shown in algorithm 5.1. Upon each iteration all components are updated. With probability  $\lambda_{\text{switch}}$  the model composition is updated by trying to switch the component on or off (discussed later) and with probability  $(1 - \lambda_{\text{switch}})$  the component parameters  $\tilde{\theta}^q$  are updated. This section and the next will discuss this parameter update move.

---

**Algorithm 5.1:** Metropolis-Hastings algorithm for multiple component signals

```

initialise  $\psi^0 = \{\mathcal{M}_{Q^0}^0, \{\tilde{\theta}\}_{Q^0}^0\}$ 
for iteration  $k = 1 \dots N_{\text{iter}}$  do
  for component  $q = 1 \dots Q$  do
     $u \sim \mathbb{U}_{[0,1]}$ 
    if  $u < \lambda_{\text{switch}}$  then
      Birth-Death(  $q$  )
    else
      if  $\Gamma^q = 1$  then
        Update  $\tilde{\theta}^q$ 
      end if
    end if
  end for
end for

```

---

The form of the  $\tilde{\theta}^q$  update move is shown in algorithm 5.2. There is a non-deterministic choice between three types of transition: an independence step, conditional independence step and perturbation step. The motivation for using each of these transition types is discussed in §3.5 and some specific examples for parameters of the general linear model are highlighted in §4.3. The independence sampling step draws a proposal from a distribution which is independent of the current state and the perturbation step draws a sample from a distribution centred upon the previous state. There is a great deal more flexibility in the treatment of the conditional independence step, which is now considered.

**Algorithm 5.2:** Parameter update move for the multiple component algorithm

**function:** Update  $\tilde{\theta}^q$

$u \sim \mathbb{U}_{[0,1]}$

**if**  $u < \lambda_{\text{ind}}$  **then**

$\tilde{\theta}^{q*} \sim q_{\text{ind}}(\tilde{\theta}^{q*})$

**else if**  $u < \lambda_{\text{ind}} + \lambda_{\text{cond}}$  **then**

$\tilde{\theta}^{q*} \sim q_{\text{cond}}(\tilde{\theta}^{q*}; \{\tilde{\theta}^q\}_{-q}^k)$

**else**

$\tilde{\theta}^{q*} \sim q_{\text{pert}}(\tilde{\theta}^{q*}; \tilde{\theta}^{q^k})$

**end if**

MH-accept(  $\psi^*, \psi^k$  )

**function:** MH-accept(  $\psi^*, \psi^k$  )

evaluate  $Q(\psi^*, \psi^k) = \min \left( 1, \frac{p(\psi^* | \mathbf{d}) q(\psi^k; \psi^*)}{p(\psi^k | \mathbf{d}) q(\psi^*; \psi^k)} \right)$

draw  $v \sim \mathbb{U}_{[0,1]}$

**if**  $v < Q(\psi^*, \psi^k)$  **then**

accept proposal  $\psi^{k+1} = \psi^*$

**else**

keep old state  $\psi^{k+1} = \psi^k$

**end if**

By separating the contributions of each component to the model, it is possible to draw a proposal from the full conditional distribution of each component

$$\tilde{\theta}^{q*} \sim p(\tilde{\theta}^q | \{\tilde{\theta}^q\}_{-\{q\}}, \mathcal{M}_{\mathbb{Q}}, \mathbf{d}). \quad (5.13)$$

This distribution is conditional upon the current state of the other components  $\{\tilde{\theta}^q\}_{-\{q\}}$  and the current model composition  $\mathcal{M}_{\mathbb{Q}}$ . The full conditional can be obtained from the joint posterior

$$p(\tilde{\theta}^q | \{\tilde{\theta}^q\}_{-\{q\}}, \mathcal{M}_{\mathbb{Q}}, \mathbf{d}) = \frac{p(\{\tilde{\theta}^q\}_{\mathbb{Q}}, \mathcal{M}_{\mathbb{Q}} | \mathbf{d})}{\int p(\{\tilde{\theta}^q\}_{\mathbb{Q}}, \mathcal{M}_{\mathbb{Q}} | \mathbf{d}) d\tilde{\theta}^q} \quad (5.14)$$

but since the denominator is not a function of  $\tilde{\theta}^q$ , it will cancel out in the MH acceptance function, so the full conditional is proportional to the joint posterior. In fact, it is only required to consider the terms of the joint posterior which are functions of  $\tilde{\theta}^q$ . This can potentially result in more efficiency savings by eliminating unnecessary calculations. However, in this instance a large matrix inversion is still required to calculate the joint posterior, eq. (5.10). The matrix  $\mathbf{P}^c$  is a function of  $\{\tilde{\theta}^q\}_{\mathbb{Q}}$  and its construction requires the inversion of  $\mathbf{G}_c^t \mathbf{G}_c$  where  $\mathbf{G}_c$  is the composite basis matrix discussed earlier. Some efficiency gains can be made by partitioning  $\mathbf{G}_c$  (see appendix A), but this still involves a large matrix inversion. Much greater savings can be achieved by a reformulation of the model.

### 5.3.1 Residual methods for mixtures of GLMs

An intuitive approach to the analysis of a multi-component signal would be to produce iteratively estimates for each component in turn, subtract the estimated component from the data, and then continue for the next component using the resulting *residual*. This is the spirit of the approach described in this section, except that the Metropolis-Hastings algorithm is more sophisticated than a simple iterative algorithm — such a method would perhaps optimise the parameters one at a time and would be likely to get stuck in local maxima. Stochastic methods such as MCMC help to overcome this problem by making the entire parameter space accessible rather than simply heading towards local maxima. Li and Djurić [80] describe an alternative deterministic scheme that iteratively evaluates the conditional distributions for each component, circumventing the need for integrations over high dimensions.

When considering component  $q$ , the residual  $\mathbf{r}^q$  is formed from all the other components currently included in the model, and express the error in terms of the component  $q$  parameters and its residual,

$$\begin{aligned}\mathbf{r}^q &= \mathbf{d} - \sum_{\substack{q' \in \mathbb{Q} \\ q' \neq q}} \mathbf{G}^{q'} \mathbf{b}^{q'} \\ \therefore \mathbf{e} &= \mathbf{r}^q - \mathbf{G}^q \mathbf{b}^q\end{aligned}\tag{5.15}$$

such that a single GLM is formed. This last equation is used as the model equation, and marginalisation of  $\mathbf{b}^q$  and  $\sigma_e^2$  is performed to obtain an expression for the conditional posterior

$$p(\tilde{\theta}^q | \mathbf{r}^q, \mathcal{M}_{\mathbb{Q}}) \propto (1 + \delta^2)^{-\frac{M^q}{2}} [\mathbf{r}^{q^t} \mathbf{P}^q \mathbf{r}^q + 2\beta_e]^{-\varepsilon} p(\tilde{\theta}^q).\tag{5.16}$$

Ideally, it is desired to use this expression in place of the joint posterior (5.10) since the residual formulation can be calculated much more efficiently — the resulting matrix inversions will be  $\mathcal{O}((M^q)^3)$  rather than  $\mathcal{O}((\sum_{q \in \mathbb{Q}} M^q)^3)$ . However, the above expression corresponds to a different model from the one which was initially of interest, since  $\mathbf{b}^q$  has been marginalised independently for each  $q$ , rather than marginalising a composite vector  $\mathbf{b}^c$  for all components simultaneously. It transpires that, if the basis matrices are approximately orthogonal with respect to each other, the following approximation may be made

$$p(\tilde{\theta}^q | \{\tilde{\theta}^q\}_{-q}, \mathcal{M}_{\mathbb{Q}}, \mathbf{d}) \approx p(\tilde{\theta}^q | \mathbf{r}^q, \mathcal{M}_{\mathbb{Q}}).\tag{5.17}$$

so that the conditional posterior expression (5.16) may be used, instead of evaluating the entire joint distribution. The justification for this relation is detailed in appendix B.3; essentially it exploits the orthogonality of the components since the energy of the sum of components is equal to the sum of the energies of the individual components.

Furthermore, if the columns of the basis matrix  $\mathbf{G}^q$  are approximately orthogonal and  $N$  is large such that  $(\mathbf{G}^{q^t} \mathbf{G}^q)^{-1} \approx k_G \mathbf{I}_M$  then

$$\mathbf{r}^{q^t} \mathbf{P}^q \mathbf{r}^q \approx \mathbf{r}^{q^t} \mathbf{r}^q - \left( \frac{k_G \delta^2}{1 + \delta^2} \right) \left\| \mathbf{G}^{q^t} \mathbf{r}^q \right\|^2.\tag{5.18}$$

This expression can be even more efficient to calculate as it doesn't require a matrix inversion, and can be of great use for the design of efficient transition kernels. If a fast

method of calculating the projection of a basis function onto data is available, this may be employed to calculate  $\|\mathbf{G}^{q^t} \mathbf{r}^q\|^2$  for a range of parameter values (this is illustrated for several types of basis function in §4.3 and §5.4). Hence a piecewise linear continuous approximation to the conditional posterior may be made very efficiently which could then be used as an independence sampling distribution. Independence samplers [150] are very beneficial as they improve the mixing of the Markov chain by reducing serial correlations and can produce rapid convergence to the posterior distribution. Discrete Fourier Transforms have been harnessed to perform the projection of data onto sinusoidal basis functions for the purposes of constructing independence sampling distributions in [5, 159, 162]. Section 4.3 shows how these distributions may be calculated efficiently for different types of basis function parameters. Chapter 6 illustrates the use of harmonic basis functions.

There are two ways in which this approximation may be used to produce a more efficient simulation. In the first method, the approximation to the full conditional and the projection assumption above are employed to produce a proposal distribution for a range of candidate parameter values. A proposal is sampled from this distribution and is accepted according to the M-H acceptance function, so the sample produced is an exact sample from the posterior. The second method generates a proposal in the same fashion but substitutes the ratio of posteriors in the M-H acceptance function with the ratio of full conditionals. The sample produced is not an exact sample from the posterior distribution, but where the orthogonality assumption is justified, this slight loss of accuracy may be tolerable, considering the major efficiency improvements. The cost of evaluating the joint posterior is  $\mathcal{O}(M^3 Q^3)$  compared to  $\mathcal{O}(M^3)$  for the approximation.

### 5.3.2 Residual-dependent and signal-dependent kernels

This section distinguishes between two types of transition kernel — those which use independence sampling distributions and those which use conditional independence sampling distributions. In the former the proposal is independent of the current state of the Markov chain as it is formed as a function of the original observation  $\mathbf{d}$ . For convenience, this is termed a *signal-dependent kernel*. The conditional independence sampling distribution is independent of the current state of the component under consideration, but dependent upon the state of the other components. This will be termed the *residual-*

*dependent kernel* as the proposal distribution is a function of the current residual (5.15) and is therefore dependent upon all the other components. Both of these types of transition kernel are very important in the operation of the Metropolis-Hastings algorithm. The signal-dependent kernel may be useful for identifying the dominant modes of the posterior distribution but is unsuitable for the detection of weaker components. Sampling with this kernel reduces correlations in the Markov chain and is good for exploring different regions of the posterior distribution, allowing the escape from local maxima.

The residual-dependent kernel is good for rapid convergence of the Markov chain once the dominant components have been detected. These are subtracted from the original signal and a proposal is generated based upon the residual waveform. In this manner, weaker components can also be detected rapidly. These two types of transition kernel can also be combined with perturbation steps to perform local exploration of posterior modes, as shown in algorithm 5.2.

The full conditional for the parameters of the  $a^{\text{th}}$  component  $\tilde{\theta}^a$  for the case that  $\Gamma^a = 1$  is

$$p(\tilde{\theta}^a | \{\tilde{\theta}^q\}_{-a}, \mathcal{M}_{\mathbb{Q}}, \mathbf{d}) \propto \frac{p(\tilde{\theta}^a)}{(1 + \delta^2)^{\frac{M}{2}} [\mathbf{d}^t \mathbf{d} - \frac{\delta^2}{1 + \delta^2} \mathbf{d}^t \mathbf{G}^c (\mathbf{G}^{c t} \mathbf{G}^c)^{-1} \mathbf{G}^{c t} \mathbf{d}]^\varepsilon} \quad (5.19)$$

It is desired to sample from this distribution, but the evaluation of the term  $(\mathbf{G}^{c t} \mathbf{G}^c)^{-1}$  will be expensive, particularly for models with many components. In the case of signal-dependent kernels, a proposal distribution can be calculated for each type of component once at the start of the algorithm. A single component model is constructed for each component type, *i.e.*,  $\mathbf{G}^c = \mathbf{G}^a$ . Therefore an independence sampling distribution can be constructed as

$$q_{\text{ind}}(\tilde{\theta}^{a*}) \propto \frac{p(\tilde{\theta}^{a*})}{(1 + \delta^2)^{\frac{M^{a*}}{2}} [\mathbf{d}^t \mathbf{d} - \frac{\delta^2}{1 + \delta^2} \mathbf{d}^t \mathbf{G}^a (\mathbf{G}^{a t} \mathbf{G}^a)^{-1} \mathbf{G}^{a t} \mathbf{d}]^\varepsilon} \quad (5.20)$$

For conditional independence steps, some of the techniques of §5.3.1 can be employed to simplify this expression and produce approximations to it. The use of the residual is particularly beneficial as it is no longer necessary to calculate the composite basis matrix  $\mathbf{G}^c$ . In this instance, the full conditional can be written

$$p(\tilde{\theta}^a | \mathcal{M}_{\mathbb{Q}}, \mathbf{r}^a) \propto \frac{p(\tilde{\theta}^a | \mathcal{M}_{\mathbb{Q}})}{(1 + \delta^2)^{\frac{M}{2}} [\mathbf{r}^{a t} \mathbf{r}^a - \frac{\delta^2}{1 + \delta^2} \mathbf{r}^{a t} \mathbf{G}^a (\mathbf{G}^{a t} \mathbf{G}^a)^{-1} \mathbf{G}^{a t} \mathbf{r}^a]^\varepsilon} \quad (5.21)$$

and so

$$q_{\text{cond}}(\tilde{\theta}^{a*}; \{\tilde{\theta}^a\}_{-\{a\}}^k) = p(\tilde{\theta}^{a*} | \mathcal{M}_{\mathbb{Q}}, \mathbf{r}^a) \quad (5.22)$$

Since the residual  $\mathbf{r}^a$  is dependent on the current state of the other components the proposal distribution must be recalculated each time it is used, so it is necessary to make its calculation very efficient.

### 5.3.3 Sampling for indicator variables

Some efficient methods for sampling the parameters of the model components in terms of *update* moves have been described in the previous sections. Transition kernels are now considered which traverse different subspaces of the posterior distribution in order to effectively update the model composition  $\mathcal{M}_{\mathbb{Q}}$ .

The simplest transition between different subspaces of the composite model space is via a birth or death move. In these moves a single component is switched into or out of the model, keeping all other component parameters constant. The birth move is described first — in satisfying the reversibility requirements of the Markov chain, the probability of the corresponding death proposal must also be evaluated. When a component is switched on, the subspace of the current state is extended to include that component's parameter space. A proposal density generates a proposal in this new parameter subspace. It is important to use a proposal density which is capable of generating moves into high probability regions, otherwise the Markov chain will be slow to explore the high posterior subspaces. Where quite compelling prior information is available in the form of a low variance prior density, then sampling a proposal value from the prior may be attractive. Such specific prior information will rarely be available and the prior is often diffuse compared with the likelihood. Hence in this section attention is directed towards the specification of more efficient transition kernels for birth and death moves.

Consider the transition for the birth ('switching on') of component  $a$ . The current and proposal states are written as  $\psi^k = \{\mathcal{M}_{\mathbb{Q}^k}, \{\tilde{\theta}^q\}_{\mathbb{Q}^k}^k\}$  and  $\psi^* = \{\mathcal{M}_{\mathbb{Q}^*}, \{\tilde{\theta}^q\}_{\mathbb{Q}^*}^*\}$  respectively. In terms of the indicator variables, the birth proposal requires that  $(\Gamma^a)^k = 0$  and  $(\Gamma^a)^* = 1$ . Hence the proposed model composition is  $\mathbb{Q}^* = \mathbb{Q}^k \cup \{a\}$ . The proposal parameter space now also encompasses the parameters for the component  $\tilde{\theta}^a$ , therefore

the transition kernel must propose a value for  $\tilde{\theta}^{a^*}$ . A conditional independence distribution is employed,  $\tilde{\theta}^{a^*} \sim q_{\text{cond}}(\tilde{\theta}^{a^*}; \{\tilde{\theta}^q\}_{\mathbb{Q}^k}^k, \mathcal{M}_{\mathbb{Q}^*})$ . This reversible jump birth/death move is highlighted in algorithm 5.3.

---

**Algorithm 5.3:** Birth-Death reversible jump move for multiple component model

```

function: Birth-Death(  $q$  )
  if  $\Gamma^{q^k} = 1$  then
     $\Gamma^{q^*} = 0$ 
  else
     $\Gamma^{q^*} = 1$ 
     $\tilde{\theta}^{q^*} \sim q_{\text{cond}}(\tilde{\theta}^{q^*}; \{\tilde{\theta}^q\}_{\mathbb{Q}^k}^k, \mathcal{M}_{\mathbb{Q}^*})$ 
  end if
  MH-accept(  $\psi^*, \psi^k$  )

```

---

The ratio of transition probabilities is required to calculate the Metropolis-Hastings acceptance function (algorithm 5.2). For the birth move of component  $q$  this is

$$\text{TR}_{\text{birth}}(\psi) = \frac{q(\Gamma^k)}{q(\Gamma^*) q_{\text{cond}}(\tilde{\theta}^{q^*}; \{\tilde{\theta}^q\}_{\mathbb{Q}^k}^k, \mathcal{M}_{\mathbb{Q}^*})} \quad (5.23)$$

and for the death move,

$$\text{TR}_{\text{death}}(\psi) = \frac{q(\Gamma^k) q_{\text{cond}}(\tilde{\theta}^{q^k}; \{\tilde{\theta}^q\}_{\mathbb{Q}^*}^*, \mathcal{M}_{\mathbb{Q}^k})}{q(\Gamma^*)} \quad (5.24)$$

where  $q_{\text{cond}}(\tilde{\theta}^{q^k}; \{\tilde{\theta}^q\}_{\mathbb{Q}^*}^*, \mathcal{M}_{\mathbb{Q}^k})$  is the probability density function for the proposal of the previous state  $\tilde{\theta}^{q^k}$  from the corresponding birth move.

The conditional independence distribution may be the same one used by the parameter update step, *i.e.*, a residual-dependent distribution as described in §5.3.2,

$$q_{\text{cond}}(\tilde{\theta}^{q^*}; \{\tilde{\theta}^q\}_{\mathbb{Q}^k}^k, \mathcal{M}_{\mathbb{Q}^*}) = p(\tilde{\theta}^{q^*} | \mathcal{M}_{\mathbb{Q}^*}, \mathbf{r}^q). \quad (5.25)$$

It is very important that this distribution is normalised since unlike the parameter update moves the normalisation constant will not cancel out of the transition probability ratio.

### 5.3.4 Non-marginalised amplitudes

It is possible to capitalise upon the more efficient residual-dependent kernels described in §5.3.1 whilst still producing an exact sample from the posterior distribution. Major efficiency achievements are possible by introducing an approximation to the full conditional which is used in the generation of proposal distributions. This approximation may also be used in the ratio of posteriors of the M-H acceptance function, resulting in a speed increase at the expense of approximate sampling from the posterior. An alternative approach may be to suppress the marginalisation of the basis function amplitudes and incorporate sampling for the amplitudes into the simulation scheme. The approximate form of the full conditional may still be employed, however, to generate efficient proposals for  $\tilde{\theta}^q$ . Using the sampled value of  $\tilde{\theta}^{q*}$ , an update move for  $\mathbf{b}^q$  is executed by sampling from a distribution approximately equal to its full conditional (4.46) (as described in §4.5.1),

$$\begin{aligned}\hat{\mathbf{b}}^{q*} &= (\mathbf{G}^{q*t} \mathbf{G}^{q*})^{-1} \mathbf{G}^{q*t} \mathbf{d} \\ \mathbf{b}^{q*} &\sim \text{N} \left( \mathbf{b}^{q*}; \frac{\delta^2}{1 + \delta^2} \hat{\mathbf{b}}^{q*}, \frac{\delta^2}{1 + \delta^2} \sigma_e^2 (\mathbf{G}^{q*t} \mathbf{G}^{q*})^{-1} \right).\end{aligned}\quad (5.26)$$

The proposal is accepted according to the M-H acceptance function and therefore is an exact sample from the posterior achieved with similar computational overhead as the approximate method. There are several variations possible on this scheme. This is a signal-dependent kernel; the observation  $\mathbf{d}$  could be substituted for the residual  $\mathbf{r}^q$  for a residual-dependent kernel. Alternatively the move could be a joint update for the composite amplitude vector  $\mathbf{b}^c$  in which case, the above step would be an exact sample from the full conditional, but the matrix inversion makes the update very inefficient. Where applicable, orthogonality assumptions may be employed to simplify (5.26).

## 5.4 Multiple component example

To illustrate the design of an algorithm for detection and estimation of multiple components, an example is presented where a signal is composed of an unknown number of sinusoids, Gaussians and step functions. Firstly the formulation of the model and

its prior structure is introduced, and then efficient proposal distributions for each of the component types are described. Techniques for Bayesian spectrum estimation with multiple sinusoids are described in [5, 32]. The characterisation of gamma-ray emission spectra in terms of multiple Gaussian peaks is described in [23]. The representation of musical signals as the sum of harmonic sinusoids is discussed in chapter 6.

#### 5.4.1 Model formulation and priors

##### Step function

The step location  $h$  is bounded  $1 \leq h \leq N$  and the basis matrix defined as

$$\mathbf{G}_{\text{st}} = [g_{\text{st}}(1) \ g_{\text{st}}(2) \ \dots \ g_{\text{st}}(N)]^t$$

$$g_{\text{st}}(i) = \begin{cases} 0 & \text{if } i < h \\ 1 & \text{if } i \geq h \end{cases} \quad (5.27)$$

For discrete  $h$  the prior is  $h \in \{1, \dots, N\}$ ,

$$p(h) = 1/N. \quad (5.28)$$

##### Sinusoid

A sinusoid can be represented using an in-phase and quadrature form so that for large  $N$  it can be expressed in terms of two approximately orthogonal basis vectors

$$\mathbf{G}_{\text{sin}} = \begin{bmatrix} \sin(0) & \cos(0) \\ \sin(\omega dT) & \cos(\omega dT) \\ \vdots & \vdots \\ \sin((N-1)\omega dT) & \cos((N-1)\omega dT) \end{bmatrix} \quad (5.29)$$

For continuous frequency  $\omega$  a uniform prior is used

$$p(\omega) = 1/(\pi F_s) = dT/\pi. \quad (5.30)$$

### Gaussian

The Gaussian basis matrix is defined as

$$\mathbf{G}_G = [g_G(1) g_G(2) \dots g_G(N)]^t$$

$$g_G(i; \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left[-\frac{(i - \mu)^2}{2\sigma^2}\right] \quad (5.31)$$

For discrete  $\mu \in \{1, \dots, N\}$  and continuous  $\sigma \in [\sigma_{lo}, \sigma_{hi}]$  the following independent priors are adopted

$$p(\mu) = 1/N$$

$$p(\sigma) = 1/(\sigma_{hi} - \sigma_{lo}). \quad (5.32)$$

#### 5.4.2 Proposal distributions

For each of the three types of component under consideration, simplifications or approximations to the full conditional densities are produced in order to obtain a set of efficient transition kernels. In each case,  $\mathbf{d}$  may refer to the signal vector or to the residual depending on the type of kernel required. In most cases, it is desired to simplify the projection term  $(\mathbf{d}^t \mathbf{G} (\mathbf{G}^t \mathbf{G})^{-1} \mathbf{G}^t \mathbf{d})$  for the particular structure of each  $\mathbf{G}$ .

#### Step function

For the step function,  $\mathbf{G}_{st}^t \mathbf{G}_{st}$  is a scalar,

$$\mathbf{G}_{st}^t \mathbf{G}_{st} = N - h + 1. \quad (5.33)$$

Projection of data onto the basis matrix is simply achieved for a given  $h$ :

$$\|\mathbf{G}_{st}(h)^t \mathbf{d}\|^2 = \left( \sum_{i=h}^N \mathbf{d}(i) \right)^2 \quad (5.34)$$

which can be calculated efficiently for all  $h$  values using a cumulative sum, and hence the proposal distribution can be evaluated as

$$q(h) \propto \left[ \mathbf{d}^t \mathbf{d} - \frac{\delta^2}{1 + \delta^2} \frac{(\sum_{i=h}^N \mathbf{d}(i))^2}{N - h + 1} \right]^{-\epsilon} p(h). \quad (5.35)$$

**Sinusoid**

If  $N$  is large and  $\omega \gg 2\pi/(N dT)$  then the columns of  $\mathbf{G}_{\sin}$  are approximately orthogonal,

$$\mathbf{G}_{\sin}^t \mathbf{G}_{\sin} \approx \frac{N}{2} \mathbf{I}_2. \quad (5.36)$$

An expression for the energy of the projection  $\|\mathbf{G}_{\sin}^t \mathbf{d}\|^2$  for a given  $\omega$  needs to be found

$$\begin{aligned} \mathbf{G}_{\sin}(\omega)^t \mathbf{d} &= \begin{bmatrix} \sum_{i=1}^N \mathbf{d}(i) \sin(\omega(i-1)dT) \\ \sum_{i=1}^N \mathbf{d}(i) \cos(\omega(i-1)dT) \end{bmatrix} \\ \therefore \|\mathbf{G}_{\sin}(\omega)^t \mathbf{d}\|^2 &= \left( \sum_{i=1}^N \mathbf{d}(i) \sin(\omega(i-1)dT) \right)^2 + \left( \sum_{i=1}^N \mathbf{d}(i) \cos(\omega(i-1)dT) \right)^2 \end{aligned} \quad (5.37)$$

This form appears similar to the magnitude of the DFT

$$\begin{aligned} \mathbf{X}(k) &= \sum_{n=1}^N \mathbf{x}(n) \exp \left[ -\frac{2\pi j(k-1)(n-1)}{N} \right] \\ &= \sum_{n=1}^N \mathbf{x}(n) \cos \left( \frac{2\pi(k-1)(n-1)}{N} \right) - j \sum_{n=1}^N \mathbf{x}(n) \sin \left( \frac{2\pi(k-1)(n-1)}{N} \right) \\ |\mathbf{X}(k)|^2 &= \left( \sum_{n=1}^N \mathbf{x}(n) \cos \left( \frac{2\pi(k-1)(n-1)}{N} \right) \right)^2 + \left( \sum_{n=1}^N \mathbf{x}(n) \sin \left( \frac{2\pi(k-1)(n-1)}{N} \right) \right)^2 \end{aligned} \quad (5.38)$$

Hence with  $\mathbf{x} \equiv \mathbf{d}$  and  $n \equiv i$ ,

$$\omega = \frac{2\pi(k-1)}{N dT} \quad (5.40)$$

This allows us to produce a proposal distribution for a discrete set of values  $\omega$  spaced at intervals of  $2\pi/(N dT)$ . A high resolution FFT might be employed for the signal-dependent kernel whilst a lower resolution might be used for the residual-dependent kernel as it will be executed many times. This proposal distribution is discrete whilst the parameter  $\omega$  is continuous so it is necessary to either introduce a small random perturbation to the sampled value or to sample from an interpolated distribution. The expression for the *discrete* proposal distribution is then

$$q(\omega') \propto \left[ \|\mathbf{d}\|^2 - \frac{2\delta^2}{N(1+\delta^2)} |\mathbf{D}(\omega') \bullet \mathbf{D}(\omega')|^2 \right]^{-\epsilon} p(\omega') \quad (5.41)$$

where  $\mathbf{D} = \text{fft}(\mathbf{d})$

### Gaussian

To calculate  $\mathbf{G}_G^t \mathbf{G}_G$  it is observed that the square of a Gaussian function  $N(\mu, \sigma^2)$  is also Gaussian  $N(\mu, \sigma^2/2)$  and calculation of the normalisation factor leads to (for large  $N$ )

$$\begin{aligned} \mathbf{G}_G^t \mathbf{G}_G &= \sum_{i=1}^N (g_G(i))^2 \\ &\approx (4\pi\sigma^2)^{-\frac{1}{2}} \end{aligned} \quad (5.42)$$

As the shape of the Gaussian is a function of the variance only, if the variance is held constant then the projection of the basis function onto the data can be performed for a range of  $\mu$  values by convolution.

$$\begin{aligned} \mathbf{G}_G^t(\mu) \mathbf{d} &= \sum_{i=1}^N \mathbf{d}(i) g_G(i; \mu, \sigma^2) \\ &= \sum_{i=1}^N \mathbf{d}(i) g_G(\mu - i; 0, \sigma^2) \end{aligned} \quad (5.43)$$

A proposal distribution for discrete  $\mu$  can now be constructed

$$\begin{aligned} q(\mu; \sigma^2) &\propto \left[ \|\mathbf{d}\|^2 - \frac{\delta^2}{1 + \delta^2} \sqrt{4\pi\sigma^2} (\mathbf{D}(\mu))^2 \right]^{-\varepsilon} p(\mu) \\ \mathbf{D} &= \mathbf{d} * \mathbf{g}_G(\sigma^2) \end{aligned} \quad (5.44)$$

where  $\mathbf{g}_G(\sigma^2)$  is a zero mean Gaussian. This proposal distribution is dependent on the chosen value of  $\sigma^2$ . The distribution could therefore be evaluated for several values of  $\sigma^2$  to form a joint proposal distribution  $q(\mu, \sigma^2) = q(\mu; \sigma^2) q(\sigma^2)$ .

### 5.4.3 Results

Three experiments were performed using different numbers of components, signal to noise ratios and iteration counts. In each case the maximum number of components of each type  $Q$  was specified and then the data was generated randomly. The other parameters were generated randomly from uniform distributions. White Gaussian noise was added to each dataset. The frame length was  $N = 500$  samples and the first 20% of all iterations were discarded as burn-in. The three experimental runs had the following parameter values:

- Dataset 1:  $Q = 2$ , SNR=20dB,  $N_{iter} = 200$ .
- Dataset 2:  $Q = 4$ , SNR=20dB,  $N_{iter} = 500$ .
- Dataset 3:  $Q = 4$ , SNR=10dB,  $N_{iter} = 500$ .

Three graphs are shown for each dataset. The top graph shows the original noisy data vector with the MAP estimate reconstructions superimposed. There are two estimates: one obtained from histogramming the Markov chain output and another from the highest posterior state emitted by the Markov chain. The second plot shows the value of the indicator variable over the run of the Markov chain — the first  $Q$  components are Gaussians, the next  $Q$  are sinusoids and the last  $Q$  are step functions. The final plot shows the log-posterior for each iteration of the Markov chain.

For the first dataset, the Markov chain detects the stronger components within the first few iterations, then refines the estimates, getting very close to the MAP state within 35 iterations. The second run increases the size of the model to a maximum of  $Q = 4$  of each type of component. Once again, the significant components are detected in the first few iterations and after 80 iterations a near optimal state is reached, where the weaker components in the mixture have also been identified. The plot of the indicator variables shows how the main components (5,8,11) are detected with a high (marginal) posterior probability  $p(\Gamma^q|\mathbf{d})$  and how overfitting is avoided as components close to the threshold of detection are switched on for fewer iterations. The third run shows the effect of reducing the SNR. The main components are detected rather more slowly than in the previous examples and the model uncertainty is reflected in the large number of transitions in the switch variable plot. More iterations are clearly needed to get reliable parameter estimates when the SNR is low.

The model used for these examples does not incorporate any identifiability criterion for the ordering of the signal components. The intention of the model structure is that over longer timescales each component has its own *context*, and identifiability is established through the parameter priors: each component has a prior centred upon some predicted value according to its previous history. The multiple frame methods of the next section reinforce the concept of signal context by joint modelling of several frames of data with Markovian dependencies between blocks.

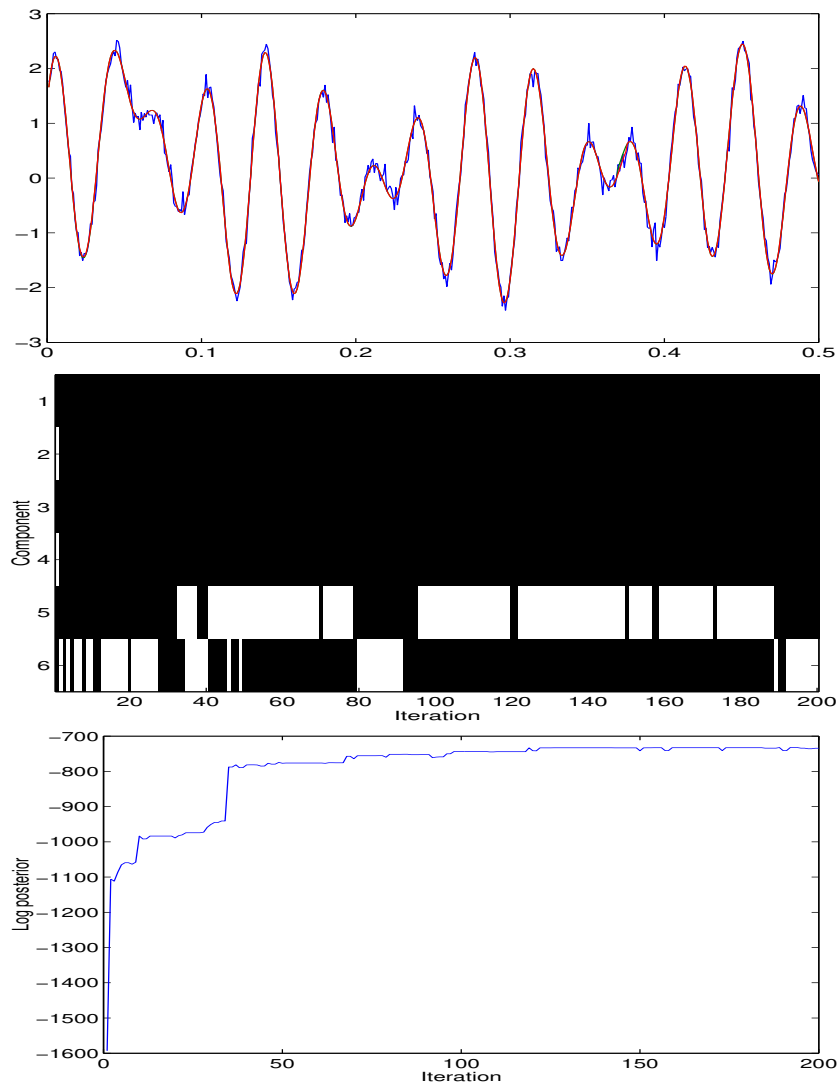


Figure 5.1: Dataset 1:  $Q = 2$ ,  $\text{SNR}=20\text{dB}$ ,  $N_{\text{iter}} = 200$ . Components 1–2 are Gaussians, 3–4 are sinusoids and 5–6 are step functions.

## 5.5 Multiple component, multiple frame models

In the previous sections of this chapter the multiple component signal model has been presented. In chapter 4, a technique for the detection and estimation of time-varying signals was presented which employs hyperparameters for the basis function parameter

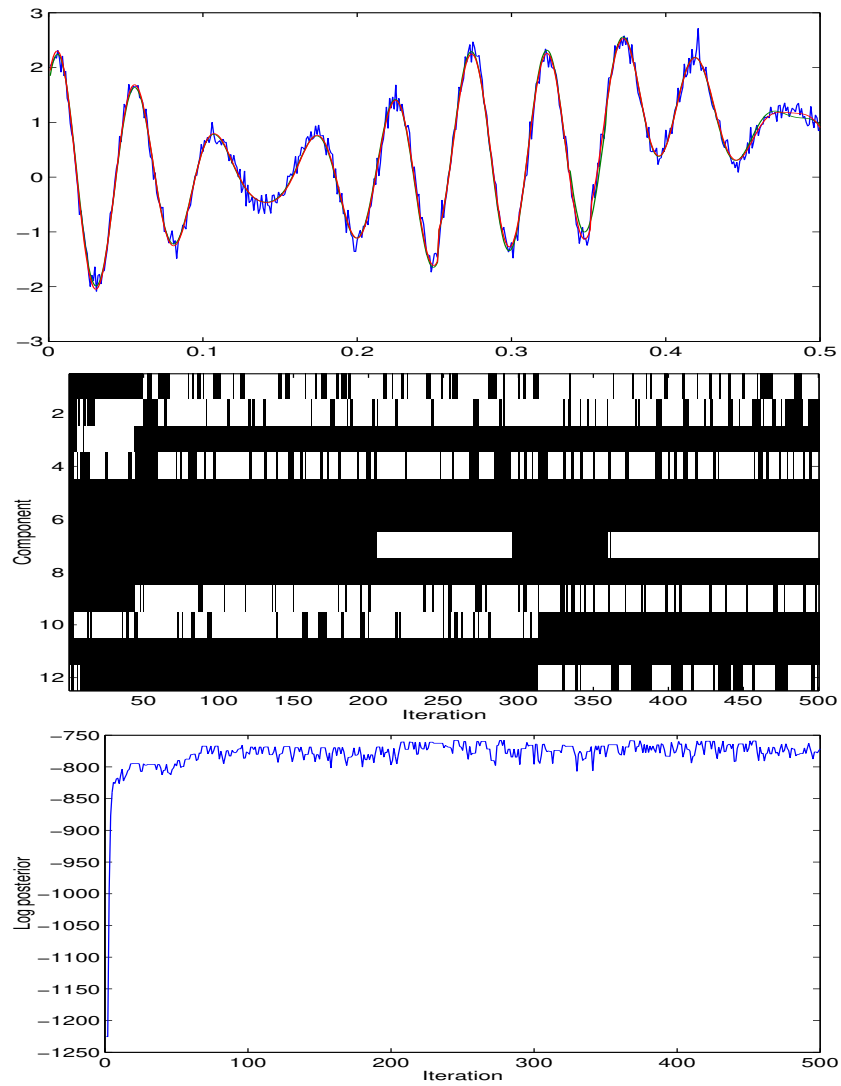


Figure 5.2: Dataset 2:  $Q = 4$ ,  $\text{SNR}=20\text{dB}$ ,  $N_{\text{iter}} = 500$ . Components 1–4 are Gaussians, 5–8 are sinusoids and 9–12 are step functions.

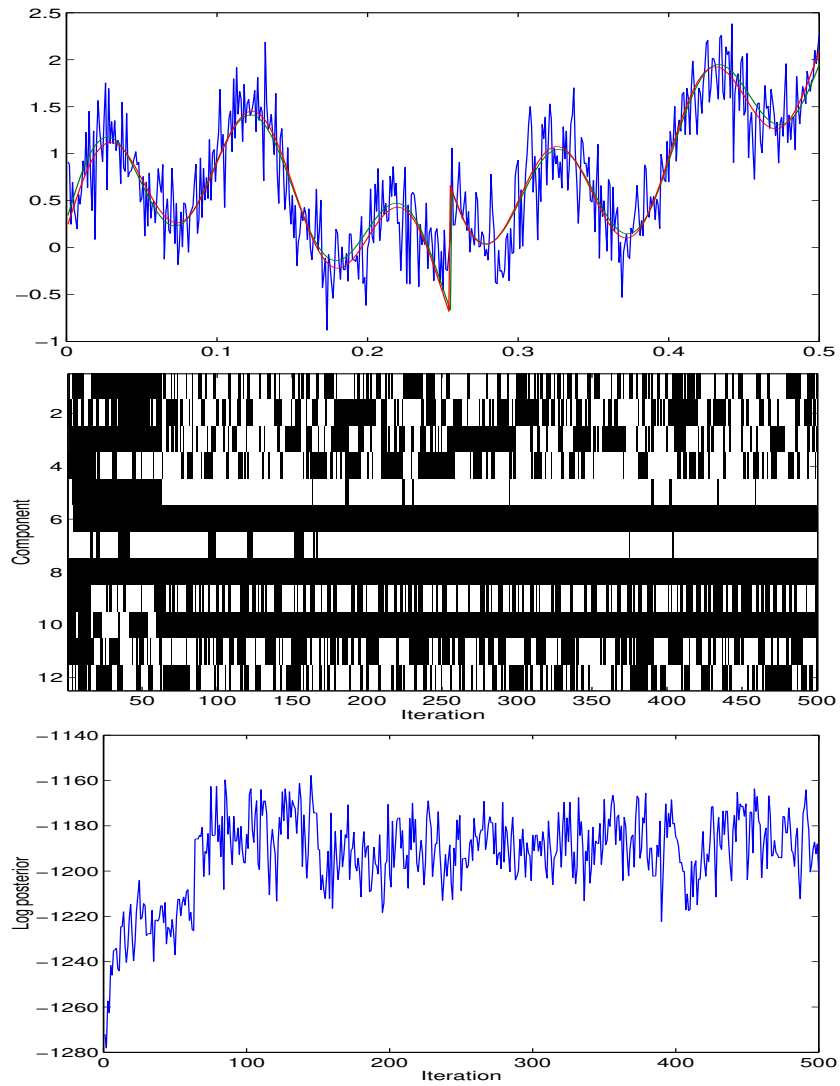


Figure 5.3: Dataset 3:  $Q = 4$ , SNR=10dB,  $N_{\text{iter}} = 500$ . Components 1–4 are Gaussians, 5–8 are sinusoids and 9–12 are step functions.

priors in order to perform a joint estimation over multiple frames. In this section, these two schemes are combined to produce a highly general technique for the detection and estimation of time-varying multiple component signals.

It is assumed that the basis function parameters and model order varies either slowly over time or according to a known functional form (*e.g.*, a linear variation). The parameter dependencies are expressed in terms of a Bayesian graphical model, and the structure, much simplified, is shown in figure 5.4. Once again, the basis function amplitudes  $\mathbf{b}_i^q$  and the error variances  $\sigma_{e_i}^2$  have been marginalised.

Variational hyperparameters have been specified for the basis function parameters  $\{\Delta_{\phi^q}, \sigma_{\phi^q}^2\}$  and the model order parameters  $\{\Delta_{M^q}, \sigma_{M^q}^2\}$ . Each component  $q \in \mathbb{Q}$  has its own set of variational hyperparameters across the block (a block is the collection of  $N_f$  successive frames). The use of multiple components is indicated by the ‘plates’ in the figure, such that each hyperparameter plate is a dependency of the corresponding plates for each frame  $i$ . For notational simplicity the variational hyperparameters are grouped as  $\bar{\Delta} = \{\Delta_{\phi^q}, \sigma_{\phi^q}^2, \Delta_{M^q}, \sigma_{M^q}^2\}$  and the parameters of the model with composition  $\mathbb{Q}$  as  $\bar{\theta} = \{\{\tilde{\theta}_i^q\}_{\mathbb{Q}, N_f}\}$ .<sup>2</sup>

The joint posterior of this model is

$$p(\bar{\theta}, \bar{\Delta}, \mathcal{M}_{\mathbb{Q}} | \{\mathbf{d}_i\}) = \prod_{i=1}^{N_f} [p(\mathbf{d}_i | \bar{\theta})] p(\bar{\theta} | \bar{\Delta}, \mathcal{M}_{\mathbb{Q}}) p(\bar{\Delta} | \mathcal{M}_{\mathbb{Q}}) p(\mathcal{M}_{\mathbb{Q}}) \quad (5.45)$$

Appealing to the conditional independence structure of figure 5.4 these can be written in terms of their dependencies,

$$p(\mathbf{d}_i | \bar{\theta}) = p(\mathbf{d}_i | \{\tilde{\theta}_i^q\}_{\mathbb{Q}}) \quad (5.46)$$

$$p(\bar{\theta} | \bar{\Delta}, \mathcal{M}_{\mathbb{Q}}) = \prod_{i=1}^{N_f} \prod_{q \in \mathbb{Q}} p(\tilde{\theta}_i^q | \Delta_{\phi^q}, \sigma_{\phi^q}^2) p(M_i^q | \Delta_{M^q}, \sigma_{M^q}^2) \quad (5.47)$$

$$p(\bar{\Delta} | \mathcal{M}_{\mathbb{Q}}) = \prod_{q \in \mathbb{Q}} p(\Delta_{\phi^q}) p(\sigma_{\phi^q}^2) p(\Delta_{M^q}) p(\sigma_{M^q}^2) \quad (5.48)$$

$$p(\mathcal{M}_{\mathbb{Q}}) = \prod_{q \in \mathbb{Q}} p(\Gamma^q). \quad (5.49)$$

The dependencies of the basis function parameters and the variational hyperparameters are shown more explicitly in figures 5.5 and 5.6. The model is structured so that the model composition  $\mathcal{M}_{\mathbb{Q}}$  is the same across the entire block, hence an indicator variable with value  $\Gamma^q = 1$  signifies that  $q \in \mathbb{Q}$  for frames  $i = 1 \dots N_f$ .

<sup>2</sup>The double subscript notation is a shorthand such that  $\{\tilde{\theta}_i^q\}_{\mathbb{Q}, N_f} = \{\{\tilde{\theta}_i^q\}_{q \in \mathbb{Q}}\}_{i=1:N_f}$ .

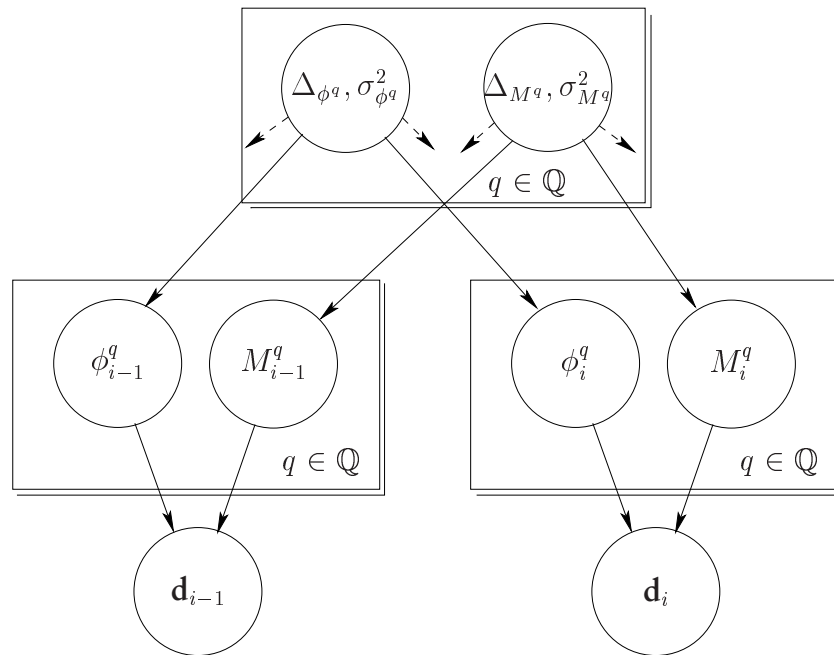


Figure 5.4: Simplified graphical model for a multiple component multiple frame GLM with marginalised amplitudes and error variances.

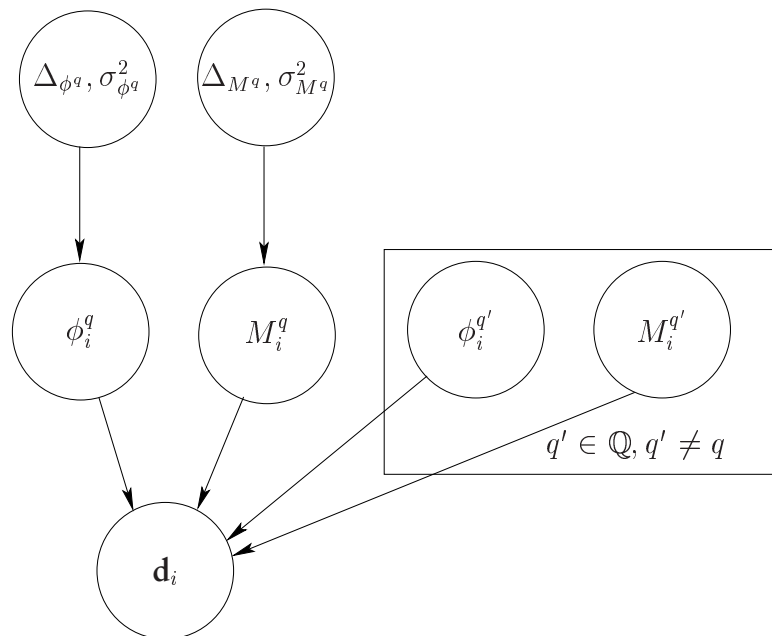
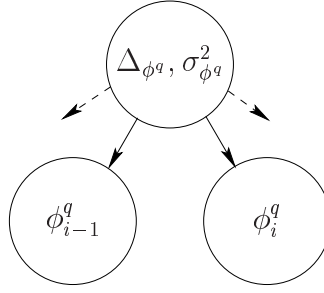


Figure 5.5: Graphical model showing the dependencies of  $\phi_i^q$  and  $M_i^q$ .



**Figure 5.6:** Graphical model showing the dependencies of  $\Delta_{\phi^q}$  and  $\sigma_{\phi^q}^2$ . The dependencies of the model order hyperparameters are identical with  $M$  substituted for  $\phi$ .

A method of simulating a Markov chain with the above joint posterior distribution is shown in algorithm 5.4. The method exploits the high correlations between parameters of a single component across multiple frames and also the prior independence between different components.

**Algorithm 5.4:** Multiple component multiple frame Metropolis-Hastings scheme with local and joint update moves (single iteration shown)

```

for  $q = 1 \dots Q$  do
   $v \sim \mathbb{U}_{[0,1]}$ 
  if  $v < \lambda_{\text{switch}}$  then
    Birth-Death(  $q$  )                                     {Update model composition}
  else if  $\Gamma^q = 1$  then
     $u \sim \mathbb{U}_{[0,1]}$ 
    if ( $u < \lambda_{\text{block}}$ ) then
      Block update {  $\{\phi_i^q, M_i^q\}_{N_f}, \Delta_{\phi^q}, \sigma_{\phi^q}^2, \Delta_{M^q}, \sigma_{M^q}^2$  }   {Update block parameters}
    else
      Perturbation update(  $q$  )
    end if
  end if
end for

```

On each iteration, each component is updated in turn. A non-deterministic choice is made between a subspace transition (birth/death move) or a parameter update move (if the current component is included in the model). The parameter update may be of two types. It may be a joint block update move where all the parameters and hyperparameters of one component are updated together or it may be comprised of individual perturbations of the frame parameters followed by an update of the frame hyperparameters from their full conditionals. Each of these transitions will now be discussed.

### 5.5.1 Component birth/death move

The birth/death move for a component is now more complex than in the single frame case as it is necessary to propose values for the component in each frame and also for the variational hyperparameters. For a birth move the improvement in model fit provided by the extra component must outweigh the extra cost of increasing the size of the parameter space, in accordance with Ockham's razor. The reversible-jump birth move is described below.

A value for  $\Delta_{M^q}^*$  is sampled from an independent proposal distribution which may be equivalent to its prior. If successive blocks are modelled with Markovian hyperpriors linking  $\Delta_{M^q}$  in each block then this distribution will be able to take advantage of previous estimations. Values of  $M_i^{q*}$  are sampled from a distribution centred upon  $\Delta_{M^q}^*$ .

$$\Delta_{M^q}^* \sim q(\Delta_{M^q}^*) \quad (5.50)$$

$$M_i^{q*} \sim q(M_i^{q*}; \Delta_{M^q}^*) \quad (5.51)$$

Using these model order values, a proposal distribution is formed for the basis function parameters  $\Delta_{\phi^q}^*$  in the same manner as §4.7. In the basis functions the substitution  $\phi_i^q = f(\Delta_{\phi^q}, i)$  is made, corresponding to the assumption that the deviation from the predictor value will be small in each frame,

$$q(\Delta_{\phi^q}^*; \{M_i^q\}_{N_f}) \propto p(\Delta_{\phi^q}^*) \prod_i [\mathbf{d}_i^t \mathbf{P}_i^c \mathbf{d}_i + 2\beta_e]^{-\epsilon}. \quad (5.52)$$

$$\Delta_{\phi^q}^* \sim q(\Delta_{\phi^q}^*; \{M_i^q\}_{N_f}) \quad (5.53)$$

where  $\mathbf{P}_i^c$  is as defined in (5.11). If the basis functions are approximately orthogonal then major efficiency savings can be made by using the residual vector rather than the

data vector as described in §5.3.1. A proposal for  $\Delta_{\phi^q}^*$  is generated from the above distribution and proposals for  $\{\phi_i^{q*}\}_{N_f}$  are obtained from their full conditionals using the proposed  $\Delta_{\phi}^*$  value, a fixed hyperparameter variance (see §4.7) and the proposed model order.<sup>3</sup>

$$\begin{aligned} q(\phi_i^{q*}; \Delta_{\phi^q}^*, M_i^{q*}) &\propto p(\phi_i^{q*} | \Delta_{\phi^q}^*, \sigma_{\text{prop}}^2) p(\mathbf{d}_i | \{\theta_i^{qk}, M_i^{qk}\}_{-q}, \theta_i^{q*}, M_i^{q*}) \\ \phi_i^{q*} &\sim q(\phi_i^{q*}; \Delta_{\phi^q}^*, M_i^{q*}) \end{aligned} \quad (5.54)$$

Finally updates for  $(\sigma_{\phi^q}^2)^*$  and  $(\sigma_{M^q}^2)^*$  are proposed which are samples from the full conditionals (4.61),

$$(\sigma_{\phi^q}^2)^* \sim q((\sigma_{\phi^q}^2)^*; \Delta_{\phi^q}^*, \{\phi_i^{q*}\}_{N_f}) \quad (5.55)$$

$$(\sigma_{M^q}^2)^* \sim q((\sigma_{M^q}^2)^*; \Delta_{M^q}^*, \{M_i^{q*}\}_{N_f}). \quad (5.56)$$

The ratio of transition probabilities for the birth move is

$$\frac{q(\bar{\theta}^k, \bar{\Delta}^k, \mathcal{M}_{\mathbb{Q}^k}; \bar{\theta}^*, \bar{\Delta}^*, \mathcal{M}_{\mathbb{Q}^*})}{q(\bar{\theta}^*, \bar{\Delta}^*, \mathcal{M}_{\mathbb{Q}^*}; \bar{\theta}^k, \bar{\Delta}^k, \mathcal{M}_{\mathbb{Q}^k})} = \frac{1}{q(\Delta_{M^q}^*) q(\Delta_{\phi^q}^*; \{M_i^{q*}\}_{N_f}) \prod_i q(M_i^{q*}; \Delta_{M^q}^*) q(\phi_i^{q*}; \Delta_{\phi^q}^*)}. \quad (5.57)$$

The transition probability ratio for the death move is the reciprocal of the above with the current and proposal states swapped. This update move is summarised in algorithm 5.5.

### 5.5.2 Joint block update move

The joint block update move is a conditional independence sampling step which generates a proposal for all the parameters pertaining to one component. The proposal is independent of the current component state but conditional upon the current state of the other components. The purpose of this move is an attempt to locate a region of high posterior probability which accounts for a high proportion of the energy of the residual formed by the other components.

<sup>3</sup>Note that these distributions are implicitly conditioned upon the state of the other components.

---

**Algorithm 5.5:** Birth-Death reversible jump move for multiple frame, multiple component model

```

function: Birth-Death(  $q$  )
  if  $\Gamma^{q^k} = 1$  then
     $\Gamma^{q^*} = 0$  {Death move}
  else
     $\Gamma^{q^*} = 1$  {Birth move}
     $\Delta_{M^q}^* \sim q(\Delta_{M^q}^*)$  {Sample model order terms}
    for  $i = 1 \dots N_f$  do
       $M_i^{q^*} \sim q(M_i^{q^*}; \Delta_{M^q}^*)$ 
    end for
     $\Delta_{\phi^q}^* \sim q(\Delta_{\phi^q}^*; \{M_i^{q^*}\}_{N_f})$  {Sample basis function parameters}
    for  $i = 1 \dots N_f$  do
       $\phi_i^{q^*} \sim q(\phi_i^{q^*}; \Delta_{\phi^q}^*)$ 
    end for
     $(\sigma_{\phi^q}^2)^* \sim q((\sigma_{\phi^q}^2)^*; \Delta_{\phi^q}^*, \{\phi_i^{q^*}\}_{N_f})$  {Sample hyperparameters}
     $(\sigma_{M^q}^2)^* \sim q((\sigma_{M^q}^2)^*; \Delta_{M^q}^*, \{M_i^{q^*}\}_{N_f})$ 
  end if
  MH-accept(  $\{\bar{\theta}^*, \bar{\Delta}^*, \mathcal{M}_{\mathbb{Q}^*}\}, \{\bar{\theta}^k, \bar{\Delta}^k, \mathcal{M}_{\mathbb{Q}^k}\}$  )

```

---

The update move employs the same proposal distributions as in §5.5.1, generating independent samples for  $\Delta_{M^q}^*$  from the prior and then sampling  $\{M_i^{q*}\}_{N_f}$  from perturbations about  $\Delta_{M^q}^*$ . Using these model order values a proposal distribution for  $\Delta_{\phi^q}^*$  is obtained from (5.52) and then  $\{\phi_i^{q*}\}_{N_f}$  are sampled from their full conditionals using the new hyperparameter value. The variance hyperparameters are also updated by sampling from their full conditionals. This joint updating scheme is summarised in algorithm 5.6.

---

**Algorithm 5.6:** Joint update move for basis function parameters

**function:** Block update:  $\{ \{ \phi_i^q, M_i^q \}_{N_f}, \Delta_{\phi^q}, \sigma_{\phi^q}^2, \Delta_{M^q}, \sigma_{M^q}^2 \}$

$\Delta_{M^q}^* \sim q(\Delta_{M^q}^*)$  {Sample model order terms}

**for**  $i = 1 \dots N_f$  **do**

$M_i^{q*} \sim q(M_i^{q*}; \Delta_{M^q}^*)$

**end for**

$\Delta_{\phi^q}^* \sim q(\Delta_{\phi^q}^*; \{M_i^{q*}\}_{N_f})$  {Sample basis function parameters}

**for**  $i = 1 \dots N_f$  **do**

$\phi_i^{q*} \sim q(\phi_i^{q*}; \Delta_{\phi^q}^*, M_i^{q*})$

**end for**

$(\sigma_{\phi^q}^2)^* \sim q((\sigma_{\phi^q}^2)^*; \Delta_{\phi^q}^*, \{\phi_i^{q*}\}_{N_f})$  {Sample hyperparameters}

$(\sigma_{M^q}^2)^* \sim q((\sigma_{M^q}^2)^*; \Delta_{M^q}^*, \{M_i^{q*}\}_{N_f})$

MH-accept(  $\{\bar{\theta}^*, \bar{\Delta}^*\}, \{\bar{\theta}^k, \bar{\Delta}^k\}$  )

---

### 5.5.3 Perturbation update

The perturbation update move is shown in algorithm 5.7. Non-deterministically a frame is chosen to be updated. A further non-deterministic choice is made whether to update the basis function parameters or the model order parameters. In either case, the frame parameter proposal is generated from a perturbation about the current value and then the variational hyperparameters are sampled from their full conditionals (§4.6.4).

---

**Algorithm 5.7:** Perturbation update for multiple frame multiple component model

**function:** Perturbation update(  $q$  )

$n \sim \mathbb{U}_{\{1 \dots N_f\}}$  {Select a frame to update}

$u \sim \mathbb{U}_{[0,1]}$

**if**  $u < \lambda_{\text{basis}}$  **then**

$\phi_n^{q*} \sim q(\phi_n^{q*}; \phi_n^{qk})$  {Update basis function parameters}

$(\Delta_{\phi^q}^2)^* \sim q((\Delta_{\phi^q}^2)^*; \phi_n^{q*}, \{\phi_n^{qk}\}_{-\{n\}})$

$(\sigma_{\phi^q}^2)^* \sim q((\sigma_{\phi^q}^2)^*; \Delta_{\phi^q}^*, \phi_n^{q*}, \{\phi_n^{qk}\}_{-\{n\}})$

**else**

$M_n^{q*} \sim q(M_n^{q*}; M_n^{qk})$  {Update model order parameters}

$(\Delta_{M^q}^2)^* \sim q((\Delta_{M^q}^2)^*; M_n^{q*}, \{M_n^{qk}\}_{-\{n\}})$

$(\sigma_{M^q}^2)^* \sim q((\sigma_{M^q}^2)^*; \Delta_{M^q}^*, M_n^{q*}, \{M_n^{qk}\}_{-\{n\}})$

**end if**

MH-accept(  $\{\bar{\theta}^*, \bar{\Delta}^*\}, \{\bar{\theta}^k, \bar{\Delta}^k\}$  )

---

## 5.6 Conclusions

This chapter has presented a generic framework for the analysis of homogeneous or heterogeneous multiple component signals. A Bayesian formulation based upon general linear models affords a powerful representation which allows prior information to be incorporated about likely model parameters. A Metropolis-Hastings algorithm employing reversible jumps allows the construction of a Markov chain that can jump across parameter subspaces, since the joint posterior distribution is of variable dimensionality. An algorithm is presented which employs several different types of transition kernel based upon the original data vector and upon the residual to efficiently detect both strong and weak components in a mixture. The model is extended for time-varying signals that have highly correlated parameters between frames, and transition kernels exploiting this structure are created. The application of the multiple component model to musical signals is detailed in the next chapter.



# *Application to Monophonic Musical Signals*

---

# 6

## *6.1 Introduction*

This chapter unites some of the most important aspects of the previous chapters for the construction of a model suitable for the analysis of musical signals. Some of the important considerations specific to musical signals, which have been described in chapter 2, will be applied in this chapter. The findings of psychoacoustic and perceptual psychological research are very significant for musical signals: the nature of pitch perception and Gestalt grouping principles are of particular interest. Musical modelling techniques are important for the highest levels of modelling and inference. Signal processing techniques provide the calculus for the description of signals and methods for the physical modelling of musical instruments.

Chapter 3 has reviewed the use of Bayesian techniques for the parameter estimation of complex models. Chapters 4 and 5 have described an hierarchical Bayesian approach to signal modelling for signals that can be represented in terms of a single component or multiple components, with parameters that vary over time.

This chapter applies these techniques to the problem of pitch estimation for monophonic musical signals. Monophonic musical signals are modelled as a set of time-varying harmonically-related sinusoids. Parameter estimation is performed using MCMC methods and the structure of harmonic sinusoids model is exploited to obtain efficient proposal densities. Various aspects of the material in this chapter have been presented in [159, 160, 161, 162].

Section 6.2 introduces the motivation for the harmonic model and its formulation. It is

shown how proposal distributions can be formed from the periodogram for an efficient MCMC implementation. Section 6.3 describes the construction of a monophonic pitch detector, with particular attention paid to the construction of efficient transition kernels that exploit the structure of the posterior distribution. Harmonic transforms are employed to obtain independence sampling proposal distributions for a large number of candidate frequencies at low computational cost. Harmonic transition kernels are employed to explore related modes of the posterior distribution. Section 6.4 constructs a more robust monophonic model by jointly modelling the signal over multiple frames, under the assumption of slow parameter variation. Transition kernels exploiting this assumption allow proposal distributions to be constructed that quickly locate frequency regions which have significant energy over all time frames. Some pitch estimation examples are shown, along with an illustration of the different harmonic character of several musical instruments.

The main emphasis of this chapter is the construction of an efficient parameter estimation scheme. Conventionally, MCMC simulation-based methods operate many thousands of times slower than real-time. By carefully exercising certain approximations and assumptions about the nature of musical signals, simplifications can be introduced that drastically reduce the computational load. Furthermore, with careful coding and good estimates as initial Markov chain values, performance approaching the order of real-time is a viable prospect.

## 6.2 *Harmonic modelling*

One of the most popular interpretations of pitch perception is that pitch is generally perceived to be the fundamental frequency of a harmonic set of pure tones (*i.e.*, sinusoids). The sinusoids need not be perfectly harmonic — a distinct pitch can be perceived even with frequency deviations of a few percent [99, 100] (see also §2.2.2). Nor is it required that all the harmonic frequencies be present; the fundamental and lower harmonics can be removed whilst still evoking the same perception of pitch (called *low pitch*, see §2.2.1). In their steady state, musical instruments that evoke a definite sensa-

tion of pitch<sup>1</sup> can be modelled well by a harmonic series, since the periodic signals that are produced can be represented in terms of a harmonic set of sinusoids. In contrast to some of the techniques in the literature, *e.g.*, [39, 91, 141], the signal is modelled explicitly as harmonic, rather than firstly extracting a set of sinusoids and then applying a harmonic grouping as a subsequent step. This is consistent with the principle of exploiting prior expectations of the structure of the data in order to construct a suitable model.

### 6.2.1 Formulation

A single musical note is constructed from the sum of a set of  $H$  harmonically related sinusoids with fundamental frequency  $\omega$ , and can be represented in terms of a General Linear Model. The basis matrix for the note is formed from in-phase and quadrature sinusoidal components,

$$\begin{aligned}\mathbf{G} &= [\mathbf{s}(\omega) \dots \mathbf{s}(H\omega) \mathbf{c}(\omega) \dots \mathbf{c}(H\omega)] \\ \mathbf{s}(\omega) &= [\sin(\omega t_1) \sin(\omega t_2) \dots \sin(\omega t_N)]^t \\ \mathbf{c}(\omega) &= [\cos(\omega t_1) \cos(\omega t_2) \dots \cos(\omega t_N)]^t \\ t_i &= i/f_s.\end{aligned}\tag{6.1}$$

The harmonic amplitudes are represented by a vector  $\mathbf{b}$  of length  $2H$ . The GLM formulation allows the amplitudes to be marginalised; if required, estimates may be obtained from the least-squares value  $\hat{\mathbf{b}}$ , or from the full conditional. For large  $N$  and  $\omega \gg 2\pi f_s/N$ , where  $f_s$  is the sampling frequency, the basis matrix is approximately orthogonal and independent of  $\omega$ ,

$$\mathbf{G}^t \mathbf{G} \approx \frac{N}{2} \mathbf{I}_N,\tag{6.2}$$

such that least-squares estimates for  $\mathbf{b}$  can be efficiently obtained from the projections of each of the basis functions onto the observation,  $\hat{\mathbf{b}} \approx (2/N) \mathbf{G}^t \mathbf{d}$ .

### 6.2.2 Periodogram estimator

Suppose that the model is used to represent a single sinusoid of arbitrary phase, *i.e.*,  $H = 1$ . If a  $g$ -prior is used for  $\mathbf{b}$  and an inverse gamma prior used for the error variance

<sup>1</sup>To make the distinction with unpitched percussion, for instance.

then the marginal posterior for  $\omega$  is (from 4.44)

$$p(\omega | \mathbf{d}) \propto (1 + \delta^2)^{-1} \left[ \|\mathbf{d}\|^2 - \frac{\delta^2}{1 + \delta^2} \|\mathbf{f}\|^2 + 2\beta_e \right]^{-\varepsilon} p(\omega) \quad (6.3)$$

Further, if a uniform prior is employed for  $\omega$ , the MAP estimate for the frequency is that which maximises  $\|\mathbf{f}\|^2$ . Owing to the orthogonality of in-phase and quadrature components this can be expressed as

$$\|\mathbf{f}\|^2 = \frac{2}{N} [\|\mathbf{c}^t \mathbf{d}\|^2 + \|\mathbf{s}^t \mathbf{d}\|^2] \quad (6.4)$$

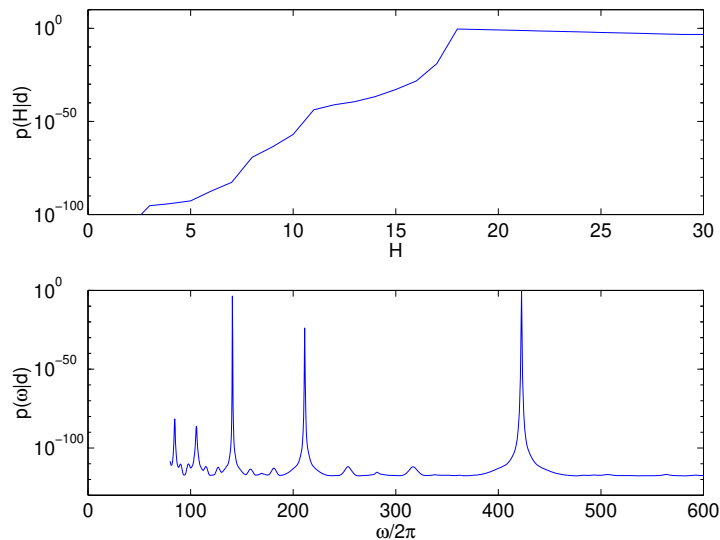
which is the magnitude of the Fourier spectrum for the observation  $\mathbf{d}$  (see also §5.4.2). Hence the frequency that maximises the Fourier spectrum is the optimal estimator for a single sinusoid in white Gaussian noise; a result observed by Bretthorst [12], who termed the projection energy the Schuster periodogram (after [130]). This interpretation is interesting as it shows the rôle of the Fourier transform as an estimator rather than as a transformation. It provides a vindication for spectrum peak-picking as a frequency estimation scheme, but also draws attention to the limitation that the estimator is only valid for an observation consisting of a single sinusoid.

### 6.2.3 Posterior distribution

For a single note (monophonic) model, the posterior distribution for  $\{\omega, H\}$  is

$$p(\omega, H | \mathbf{d}) \propto (1 + \delta^2)^{-H} \left[ \|\mathbf{d}\|^2 - \frac{\delta^2}{1 + \delta^2} \|\mathbf{f}\|^2 + 2\beta_e \right]^{-\varepsilon} p(\omega) p(H). \quad (6.5)$$

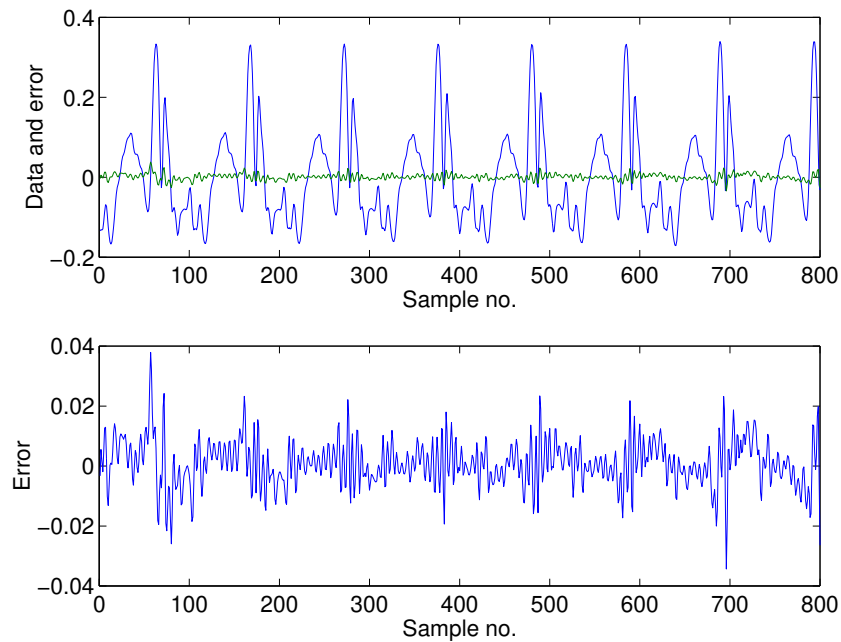
Figure 6.1 shows the marginal posterior densities for  $H$  and  $\omega$  for a frame of monophonic saxophone data. The following prior distributions are used:  $p(\omega) = \mathbb{U}(\omega)$ ,  $p(H) = \text{Poisson}(5)$ ,  $p(\sigma^2) = \text{IG}(1, 0.01)$  and the  $g$ -prior parameter for  $p(\mathbf{b})$  is  $\delta^2 = 100$ . The frequency posterior has several marked peaks (note the logarithmic scale) at the apparent fundamental frequency of the harmonic series, and also at fractions and integer multiples of that frequency. This phenomenon is a common cause of *octave errors*, also known as *pitch period doubling*; it arises since all the harmonics of  $\omega$  are also harmonics of  $\omega/n$ , for integer  $n$ . The model order penalisation in the posterior ensures that the most parsimonious model is chosen. The frequency posterior distribution has very narrow modes (in the case of figure 6.1 a frequency grid of around 2Hz is required



**Figure 6.1:** Marginal posteriors for  $H$  and  $\omega$  for a frame of monophonic saxophone data, drawn on a logarithmic probability scale. The fundamental frequency of the data and perceived pitch of the sound is 423Hz.

to resolve the peaks), and hence a large number of points would be required to ensure detection of the global maximum, which is computationally expensive. Another issue arising is that, due to the extreme orders of magnitude between the peaks of the posterior, the influence of parameter priors will be minimal unless the prior is very informative; this may be significant if the prior structure is in some way compensating for modelling errors.

The MAP estimate for this data is  $\omega/(2\pi) = 423\text{Hz}$ ,  $H = 22$ . The MAP reconstruction is shown in figure 6.2; in the top plot, the observed signal and the residual error, obtained using the MAP parameter estimates, are shown. This demonstrates how the harmonic model is capable of a high quality reconstruction for steady state signals. The lower plot shows the residual error, drawn to a different scale. There appears to be slight artifacts in the residual at locations separated by the pitch period. These may be partially attributed to aperiodic excitation; in some manually excited instruments, the excitation is generated as a result of the chaotic oscillations of the lips (brass instruments), reeds (wind instruments) or vocal folds (singing) due to the turbulent streaming of air. This excitation is unlikely to be perfectly periodic, particularly at low frequencies, and leads



**Figure 6.2:** Model fit for saxophone data. The top plot shows the observation and the residual error obtained using the MAP estimate. The lower plot shows the error signal on a different scale.

to an effect termed *vocal fry* [124]. This may be observed in speech when the intonation goes down at the end of a sentence and the glottal pulses start to become irregular as a result of the breakdown of the chaotic oscillation. The magnitudes<sup>2</sup> of the harmonics for the MAP parameter estimates are shown in figure 6.3, where the rich spectral character of the saxophone is apparent. The harmonic characteristics will be explored in the next section.

#### 6.2.4 Harmonic transform

Since the evaluation of the posterior at a finely spaced frequency grid is computationally expensive, particularly for high values of  $H$ , a more efficient method for calculating the energy of the projection for a range of frequencies is required. The method outlined

<sup>2</sup>The harmonic magnitudes  $a$  are defined in terms of the in-phase and harmonic amplitudes as  $a_h = (b_h^2 + b_{h+H}^2)^{1/2}$ ,  $h = 1 \dots H$ .

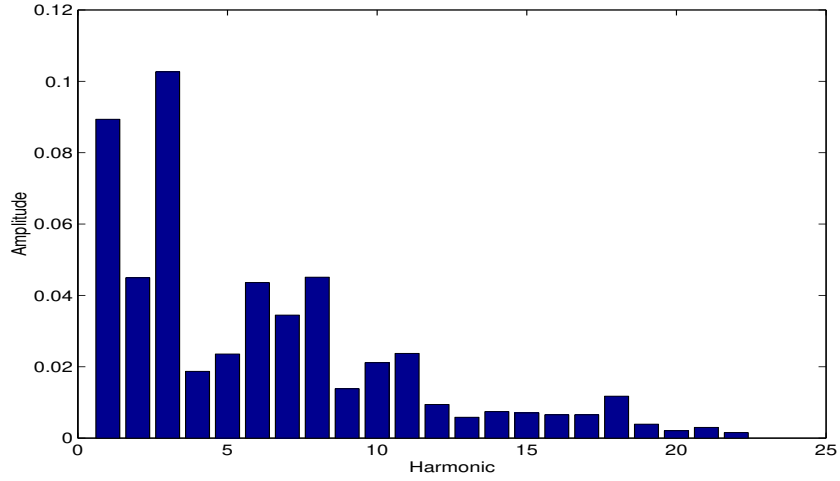


Figure 6.3: Saxophone harmonic magnitudes for the MAP parameter estimates.

in this section builds upon the frequency estimator interpretation of the periodogram described in §6.2.2.

The order  $P$  harmonic transform  $\mathcal{H}_P(\mathbf{x}, l)$  of a signal  $\mathbf{x}$  is defined as,

$$\mathcal{H}_P(\mathbf{x}, l) = \sum_{p=1}^P X_p^*[l] X_p[l], \quad (6.6)$$

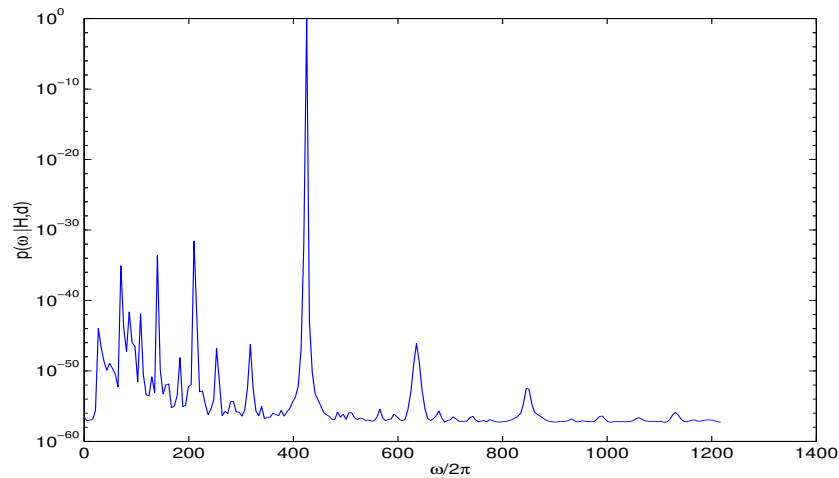
$$X_p[l] = \sum_{n=1}^{N_{\text{fft}}} x[n] \exp \left[ -j \frac{2\pi pl(n-1)}{N_{\text{fft}}} \right]$$

where  $l$  is the frequency bin number ( $l = 1 \dots L$ ,  $L = \lfloor N_{\text{fft}}/P \rfloor$ ), and it is assumed that  $\mathbf{x}$  has been zero-padded to length  $N_{\text{fft}}$ . The transform can be calculated from the FFT of  $\mathbf{x}$ , since  $X_p[l] = X_1[pl]$ , to obtain a frequency grid spacing of  $f_s/N_{\text{fft}}$ . The transform calculates the energy of the frequency components at multiples of  $1 \dots P$  times the frequency of each bin. Due to the Nyquist limit,  $L$  may need to be reduced for higher frequencies to ensure that  $pL < N_{\text{fft}}/2$ .<sup>3</sup>

The energy of the projection for  $\{\omega, H\}$  can then be approximated by

$$\|\mathbf{f}\|^2 \approx \frac{2}{N} \left[ \mathcal{H}_H \left( \mathbf{d}, \left[ \frac{\omega N_{\text{fft}}}{2\pi f_s} \right] \right) \right]. \quad (6.7)$$

<sup>3</sup>The harmonic transform effectively constructs a Schröder histogram [129] and is similar in form to the harmonic sum spectrum and (log) harmonic product spectrum, *e.g.*, see [60].



**Figure 6.4:** Conditional distribution  $p(\omega|H, \mathbf{d})$  using the marginal MAP estimate of  $H$ , using the saxophone data set. The harmonic transform is used to approximately evaluate the posterior over a fine frequency grid.

For  $H = 1$  this reduces to the single sinusoid case described in §5.4.2. In figure 6.4, the conditional posterior for  $\omega$  for the saxophone data is evaluated for the marginal MAP estimate of  $H$  using the harmonic transform. The data, of length  $N = 800$ , is zero-padded to 8192 points to produce a frequency grid spaced at 5Hz intervals, which is fine enough to locate the main peak at 423Hz. The harmonic transform is a useful tool for the generation of proposal distributions, which will be described in subsequent sections in this chapter.

### 6.3 Monophonic pitch detection

In this section the harmonic model is employed for monophonic (single note) pitch estimation. It is shown how MCMC techniques may be used for the parameter estimation, and then that the model can be extended over multiple frames for time-varying signals. This model lays the foundations for the polyphonic detector described in the next section.

### 6.3.1 MCMC techniques

For a monophonic model, the two parameters to be estimated are the fundamental frequency  $\omega$  and the number of harmonics  $H$ . The previous section has highlighted some of the difficulties of attempting to maximise the posterior distribution by evaluating it for a range of values of  $\omega$  and  $H$ . In particular, the requirement of high frequency resolution to ensure that the global maximum is detected is a matter of concern. A local Metropolis-Hastings algorithm (§3.4.3) may be applied instead to produce frequency estimates which are not limited by the resolution of a frequency grid, while still ensuring that the high probability regions of the posterior distribution may be explored efficiently.

Three types of transition kernel are employed that exploit features of the posterior distribution (originally presented in [159]). Each performs a specific type of movement around the posterior distribution, as discussed in §3.5.

#### *Independence sampler*

The first transition kernel is an independence sampling step where a proposal  $\{\omega^*, H^*\}$  is generated that is independent of the current state of the Markov chain. A value for  $H^*$  is first sampled from a distribution  $q(H^*)$ ; the prior could be used, for instance, since this represents the prior expectations for  $H$ . Using this value, a proposal distribution for  $\omega^*$  is constructed using the harmonic transform (6.7).

$$\begin{aligned} H^* &\sim q(H^*) \\ \omega^* &\sim q(\omega^*; H^*) \end{aligned} \tag{6.8}$$

where

$$\begin{aligned} q(\omega^*; H^*) &= \sum_{l=1}^L \mathcal{N}\left(\omega^*; \frac{2\pi f_s l}{N_{\text{fft}}}, \sigma_\omega^2\right) \\ &\quad \times (1 + \delta^2)^{-H^*} \left[ \|\mathbf{d}\|^2 - \frac{2\delta^2}{N(1 + \delta^2)} \mathcal{H}_{H^*}(\mathbf{d}, l) + 2\beta_e \right]^{-\varepsilon} p(\omega^*) \end{aligned} \tag{6.9}$$

The term in square brackets is a function of the difference in energy of the observation and the projection of a harmonic set with  $H^*$  harmonics and a fundamental frequency of

$lf_s/N_{\text{fft}}$ . The distribution is constructed as a Gaussian mixture distribution with amplitudes at each FFT bin frequency determined by the harmonic transform. The variance  $\sigma_\omega^2$  is chosen to be of the order of the square of the frequency bin size. This transition kernel is very effective for rapid convergence of the Markov chain as the dominant modes are easily identified. It may be useful on occasion to make the distribution more diffuse by reducing the exponent  $\varepsilon$  by a factor of 10 or more to encourage other peaks besides the dominant one to be explored. This technique for employing periodogram estimates for independence sampling steps is similar to that presented by Andrieu and Doucet [5], but extends the method to harmonically related sinusoids using the harmonic transform.

### *Harmonic transition kernel*

The second transition kernel is intended to alleviate the problems caused by octave errors. A joint move for  $\{\omega^*, H^*\}$  is proposed that jumps between the related modes of the posterior distribution. The frequencies of the modes are related by a factor of a ratio of integers, and the number of harmonics by the inverse of that amount. This kernel is useful as the harmonic representation is not unique: the signal produced by a given set of harmonic sinusoids can be represented equally well by another harmonic series with half the fundamental frequency of the original and twice the number of harmonics. Since half of these new harmonics will be of zero amplitude, the representation is not efficient, and the complexity penalisation of the posterior ensures that the more economical representation will be favoured. Given the sharply peaked nature of the posterior distribution, it is likely that an independence sampling step (as described above) will propose a move to a mode that is related to the global maximum; this transition kernel will explore the harmonically related modes in search of the global maximum.

A value  $r$  is sampled from a set of ratios of integers  $R$ . The frequency proposal is the factor  $r$  multiplied by the current value and the number of harmonics is reduced by the same factor. To ensure reversibility, a slight modification must be made, such that the  $H^*$  proposal is drawn from a distribution centred upon  $\lfloor H^k/r \rfloor$  (e.g., a Poisson

distribution), to account for the case that  $H^k$  is not exactly divisible by  $r$ .

$$\begin{aligned} r &\sim \mathbb{U}_R \\ \omega^* &= r\omega^k \\ H^* &\sim q(H^*; \lfloor H^k/r \rfloor) \end{aligned} \tag{6.10}$$

The set  $R = \{\frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{3}{2}, 2, 3\}$  works well in practice. Increasing the weighting of  $\frac{1}{2}$  and 2 can be useful, as the most commonly encountered octave errors in practice are factors of two. Care must be taken to ensure that  $\omega H$  does not exceed the Nyquist frequency; this may necessitate truncation of the proposal distribution.

### *Perturbation kernel*

The third transition kernel used is a perturbation kernel, or random walk sampler. This move is used for local exploration of the posterior modes. Perturbations are applied independently for  $\omega$  and  $H$ . For the fundamental frequency parameter, a value is sampled from a narrow distribution centred upon the value of the current state  $\omega^k$ ,

$$\omega^* \sim N(\omega^*; \omega^k, \sigma_\omega^2). \tag{6.11}$$

The standard deviation of the proposal distribution is chosen to be of the same order of magnitude as the modes of the posterior distribution (typically several Hz). The harmonic number parameter is sampled from a distribution centred upon the current value  $H^k$ , e.g., perhaps using a Poisson distribution,

$$H^* \sim q(H^*; H^k). \tag{6.12}$$

### 6.3.2 Monophonic analysis example

The methods of this section are employed for the analysis of a monophonic extract; the dataset sax is used, which is a 5 second solo saxophone melody. A uniform prior is used for  $\omega$  and a Poisson prior with variance of 6 is used for  $H$ . Each frame of data (of length 22ms) is analysed individually, and a hop size of half the frame length is used. The MAP estimates for  $\omega$  are shown in figure 6.5. The MAP estimates are obtained from the point in the Markov chain with the highest probability density; MAP estimates are

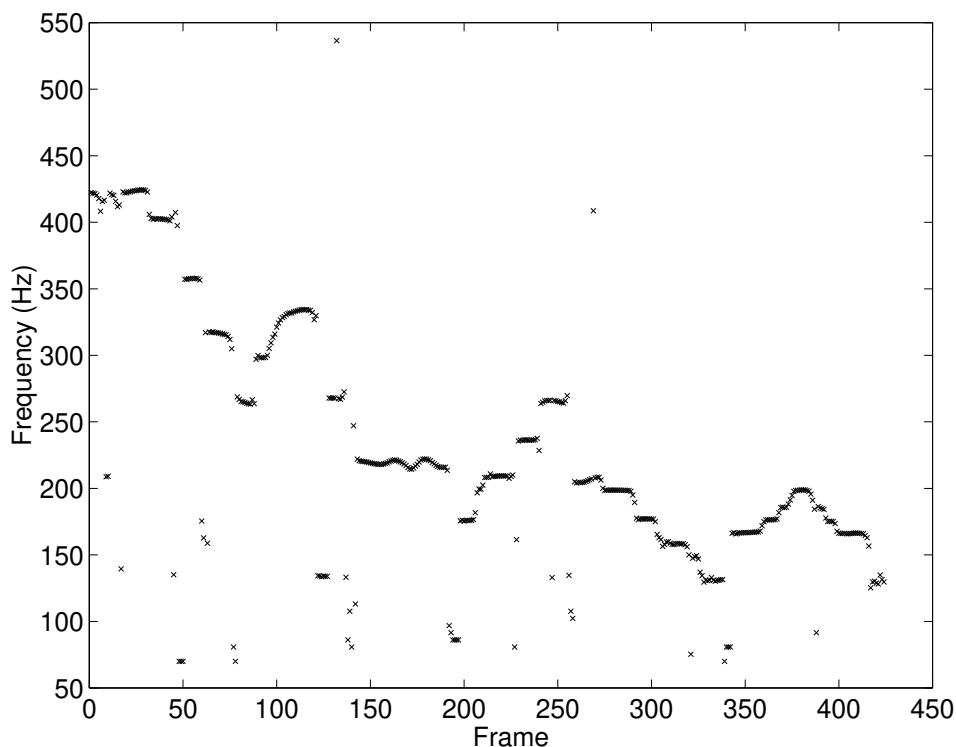


Figure 6.5: Frequency tracks for independent frame model. Dataset is monophonic sample sax.

used since a point pitch estimate is required rather than more detailed inferences about the posterior distribution. The method picks up much of the fine frequency detail of the performance, for instance the abrupt pitch discontinuities at note boundaries (*e.g.*, frame 50), note slurs (frames 80–120) and vibrato (frames 150–200). However, the number of outliers scattered about the plot is striking. Further inspection shows that these almost entirely fall at note boundaries, in the abrupt transition between one note and the next, where the data is likely to be rapidly time-varying. Moreover, virtually all of the outliers are at half the frequency of the underlying variation, and hence are the manifestations of octave errors.

Figure 6.6 shows frame 7 of the dataset sax, which generates an octave error in figure 6.5. The first note apparent from the frequency track is in fact composed of two overlapping notes very close in pitch, and the indistinct region around frames 10–20 happens to be around 20dB lower in power than frames 0–5 and 30–50. In this region, the first note decays, but natural reverberation acts to sustain the sound for a short time

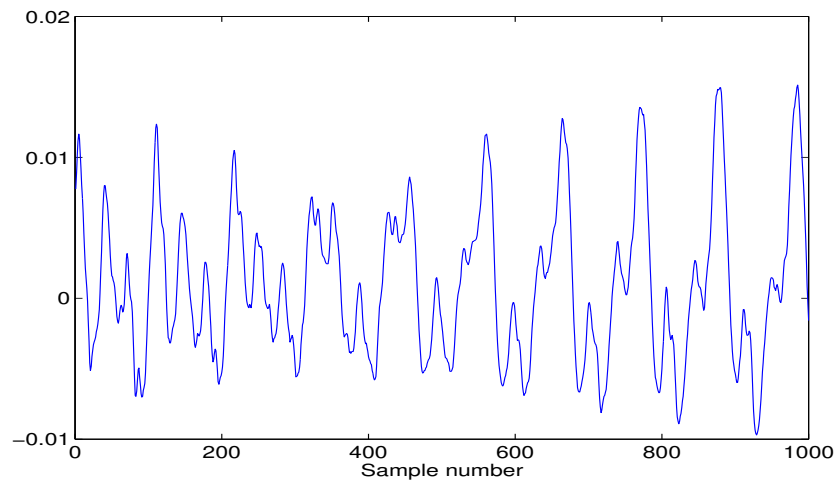


Figure 6.6: Rapidly varying data from frame 7 of dataset sax.

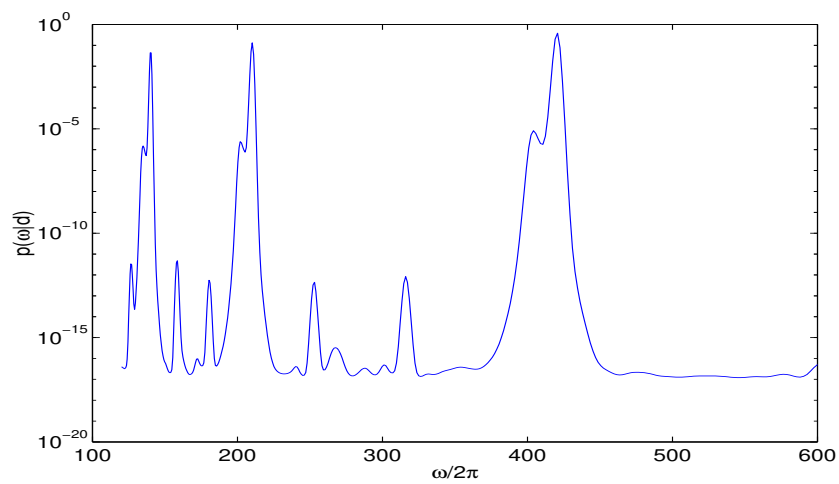


Figure 6.7: Marginal frequency posterior for rapidly varying saxophone data. Note the two sets of peaks in the major modes.

after the excitation has been removed. When the attack for the next note starts, a common characteristic of many manually excited instruments is that the note starts flat (*i.e.*, of lower pitch) and the player increases the pitch until it is in tune — it is fairly rare for players to start sharp of the desired pitch and then go flat. The posterior distribution for this frame of data is shown in figure 6.7. The location of two sets of closely-spaced peaks is quite clear. Although the dominant modes are located around the factors of 420Hz, each major mode has two ‘sub-peaks’ — the higher of the two corresponding to the decay of the previous note whilst the lower corresponds to the onset of the next note which is starting flat. The octave error is likely to have arisen since the reduction of fundamental frequency by a factor of two yields an extra degree of freedom to explain the ‘beating’ of the two close frequencies.

The construction of a more accurate model to handle the abrupt frequency changes and overlap of consecutive notes is arguably an excessively complex solution.<sup>4</sup> The model will seek to explain these problem regions in terms of a harmonic series, which may produce a (mathematically) reasonable fit. However, in a perceptual sense, these estimates may be meaningless. For instance, the few artifacts which appear in figure 6.5 generally occur in regions where there is a transition between notes, or where the signal energy is very low. In these regions, the perception of pitch is itself not well-defined. The human auditory apparatus requires reasonably steady-state stimuli in order to determine a pitch. This suggests that some subsequent processing is required upon the raw output of the Markov chain. Applying a threshold at a low power level generally has the desired effect of suppressing these artifacts. This could be incorporated into an inference scheme by assigning a ‘labelling’ variable to each frame which signifies whether the inference is deemed to be perceptually significant.

The behaviour in the steady state, however, is captured well by the model. Further, the generally smooth variation in pitch makes a multiple frame model well suited to this form of data. The next section considers a multiple frame approach which provides more robustness against problems of the type encountered in this example.

---

<sup>4</sup>Such an approach however might incorporate the detection of parameter changepoints, *e.g.*, as demonstrated by [111, 126], or musically probable note transitions, as demonstrated by [71].

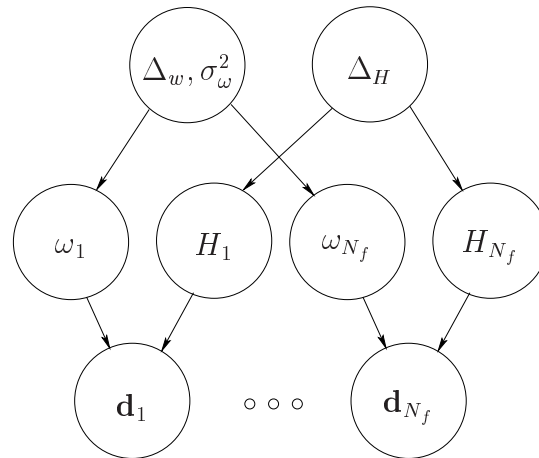


Figure 6.8: Dependencies for monophonic multiple frame model.

## 6.4 Multiple frame model

The monophonic example in the previous section employs an independent analysis for each frame. Great increases in robustness can be obtained by explicitly modelling the slow variation of the fundamental frequency and model order over time. The methods of section 4.7 are drawn upon here to produce efficient simulations for a multiple frame model.

The graphical model showing the parameter dependencies is shown in figure 6.8. The hyperparameters  $\{\Delta_\omega, \sigma_\omega^2\}$  govern the underlying pitch variation across the block, whilst  $\Delta_H$  governs the variation of the number of harmonics (model order).<sup>5</sup> These parameters describe a specific functional form of evolution over time; in the simplest case, a near-constant value across each block is assumed.

The model equation is

$$\mathbf{d}_i = \mathbf{G}_i \mathbf{b}_i + \mathbf{e}_i. \quad (6.13)$$

<sup>5</sup>Since a Poisson distribution is adopted for the prior on  $H_i$ , a spread hyperparameter is not required.

The priors for this model are

$$\begin{aligned}
p(\omega_i | \Delta_\omega, \sigma_\omega^2) &= \text{LN}(\omega_i; \Delta_\omega, \sigma_\omega^2) \\
p(\Delta_\omega) &= \mathbb{U}_{[\omega_{\text{low}}, \omega_{\text{high}}]} \\
p(\sigma_\omega^2) &= \text{IG}(\alpha_\omega, \beta_\omega) \\
p(H_i | \Delta_H) &= \text{Poisson}(H_i; \Delta_H) \\
p(\Delta_H) &= \text{Poisson}(\Delta_H; H_0).
\end{aligned} \tag{6.14}$$

A log-normal distribution is used for the  $\omega_i$  prior. This is chosen because of the logarithmic nature of the frequency axis, and corresponds to  $\log(\omega_i)$  being normally distributed as  $N(\log(\Delta_\omega), \sigma_\omega^2)$ . This also circumvents problems which would arise with a Gaussian frequency prior, for instance, needing to truncate the prior at  $\omega_i = 0$  (since the log-Normal distribution is only defined for  $\omega > 0$ ). For the harmonic transition kernel, where all frequencies are multiplied by a constant, the variance of the frequencies across the block will increase, but the variance of the log frequencies will stay constant. The prior for the frequency hyperparameter  $\Delta_\omega$  is chosen to be a uniform distribution over the expected range of pitches (e.g., 100-1000Hz) but this could be changed to be a Markovian prior centred upon the estimated value from the previous block, to increase the parameter dependencies over longer timescales. The logarithmic frequency characteristic also has musical connotations. The distance between pitches is measured in terms of their ratio rather than their difference, such that a factor of two separation is an octave. The octave scale is split into twelve *semitone* steps, each separated by a factor of  $2^{1/12}$  (roughly a 6% increase per semitone step). The unit of log pitch is sometimes called *height*, which relates to the two-dimensional nature of pitch perception in terms of height and chroma (see §2.4.1).

#### 6.4.1 Transition kernels

The joint posterior for the model is

$$\begin{aligned}
p(\{\omega_i, H_i\}_{N_f}, \Delta_\omega, \sigma_\omega^2, \Delta_H | \{\mathbf{d}_i\}_{N_f}) &\propto \\
p(\Delta_\omega) p(\sigma_\omega^2) p(\Delta_H) &\prod_{i=1}^{N_f} p(\mathbf{d}_i | \omega_i, H_i) p(H_i) p(\omega_i | \Delta_\omega, \sigma_\omega^2). \tag{6.15}
\end{aligned}$$

The method required to simulate a Markov chain from this posterior distribution is essentially that of algorithm 4.3 (in §4.7) with the exception that there are two forms of block update move — one independent and one dependent. The first block update move is an independent sampling step which is intended to locate high probability regions of the posterior for notes which are strong across all frames in the block. A value for the  $\Delta_H$  hyperparameter is sampled from an independent proposal distribution, and then values  $H_i$  in each frame are sampled from distributions centred upon  $\Delta_H^*$

$$\begin{aligned}\Delta_H^* &\sim q(\Delta_H^*) \\ H_i^* &\sim q(H_i^*; \Delta_H^*) \quad i = 1 \dots N_f\end{aligned}\tag{6.16}$$

A proposal distribution is then formed for  $\Delta_\omega$  on the assumption that the frequency deviations in each frame are negligible,

$$\begin{aligned}\Delta_\omega^* &\sim q(\Delta_\omega^*; \{H_i^*\}_{N_f}) \\ q(\Delta_\omega^*; \{H_i^*\}_{N_f}) &\propto \frac{p(\Delta_\omega^*)}{(1 + \delta^2)^{\sum_i H_i^*}} \prod_i \left[ \|\mathbf{d}_i\|^2 - \frac{\delta^2}{1 + \delta^2} \|\mathbf{f}_i^*\|^2 + 2\beta_e \right]^{-\varepsilon}\end{aligned}\tag{6.17}$$

where  $\mathbf{f}_i^*$  is the projection obtained in frame  $i$ , setting  $\omega_i = \Delta_\omega^*$ . This distribution may be evaluated for a large number of  $\Delta_\omega^*$  values using the harmonic transform for each frame, *e.g.*, as shown in (6.9). The frequency for each frame is obtained as a small perturbation about the sampled value of  $\Delta_\omega^*$ ,

$$\omega_i^* \sim q(\omega_i^*; \Delta_\omega^*) \quad i = 1 \dots N_f.\tag{6.18}$$

An update for  $\sigma_\omega^2$  is also performed by sampling a value from its full conditional (4.61),

$$(\sigma_\omega^2)^* \sim p((\sigma_\omega^2)^* | \Delta_\omega^*, \{\omega_i^*\}^*)\tag{6.19}$$

The second block update move is dependent upon the current state and is used to correct for octave errors by finding a more economical model representation. It employs a harmonic transition kernel across the entire block,

$$\begin{aligned}r &\sim \mathbb{U}_R \\ \Delta_\omega^* &= r \Delta_\omega^k \\ \omega_i^* &= r \omega_i^k \quad i = 1 \dots N_f \\ \Delta_H^* &\sim q(\Delta_H^*; \lfloor \Delta_H^k / r \rfloor) \\ H_i^* &\sim q(H_i^*; \lfloor H_i^k / r \rfloor) \quad i = 1 \dots N_f\end{aligned}\tag{6.20}$$

The proposal distributions for  $\Delta_H$  and  $H_i$  are chosen to satisfy the reversibility issues described in the previous section.

The third form of move is a perturbation step, applied non-deterministically to the frequency or harmonic number parameter in one of the frames. It is followed by an update for the corresponding hyperparameters from their full conditionals. For a frequency perturbation,

$$\begin{aligned}
 j &\sim \mathbb{U}_{\{1\dots N_f\}} \\
 \omega_j^* &\sim q(\omega_j^*; \omega_j^k) \\
 \Delta_\omega^* &\sim p(\Delta_\omega^* \mid \omega_j^*, \{\omega_i^k\}_{-j}, (\sigma_\omega^2)^k) \\
 (\sigma_\omega^2)^* &\sim p((\sigma_\omega^2)^* \mid \Delta_\omega^*, \omega_j^*, \{\omega_i^k\}_{-j}).
 \end{aligned} \tag{6.21}$$

Similarly for a harmonic number perturbation,

$$\begin{aligned}
 j &\sim \mathbb{U}_{\{1\dots N_f\}} \\
 H_j^* &\sim q(H_j^*; H_j^k) \\
 \Delta_H^* &\sim q(\Delta_H^*; H_j^*, \{H_i^k\}_{-j})
 \end{aligned} \tag{6.22}$$

#### 6.4.2 Monophonic multiple frame example

Figure 6.9 shows the results of a multiple-frame monophonic analysis on the sax dataset using a block size of  $N_f = 5$  (110ms). A comparison with the independent frames method of the previous section shows that the fine frequency detail of figure 6.5 is preserved, whilst the artifacts have almost entirely disappeared. In the problem regions such as between note transitions and in areas of low power, the signal deviates from its steady-state harmonic behaviour and so the strength of evidence for a particular fundamental frequency will be low. In these instances, prior information compensates for indeterminate data, and the strong evidence provided by the nearby steady-state frames of data will be dominant. Figure 6.10 shows the pitch estimates for the dataset memory, which is a vocal melody. Once again, the fine detail of the performance has been captured, for instance the *vibrato* around frames 420–480 and 530–600, and the slurs between notes, *e.g.*, frames 20–50. There are several regions on the graph where the pitch estimate is rather indistinct. These largely correspond to unvoiced regions where the glottal excitation has stopped or is masked by an unvoiced sound such as sibilants

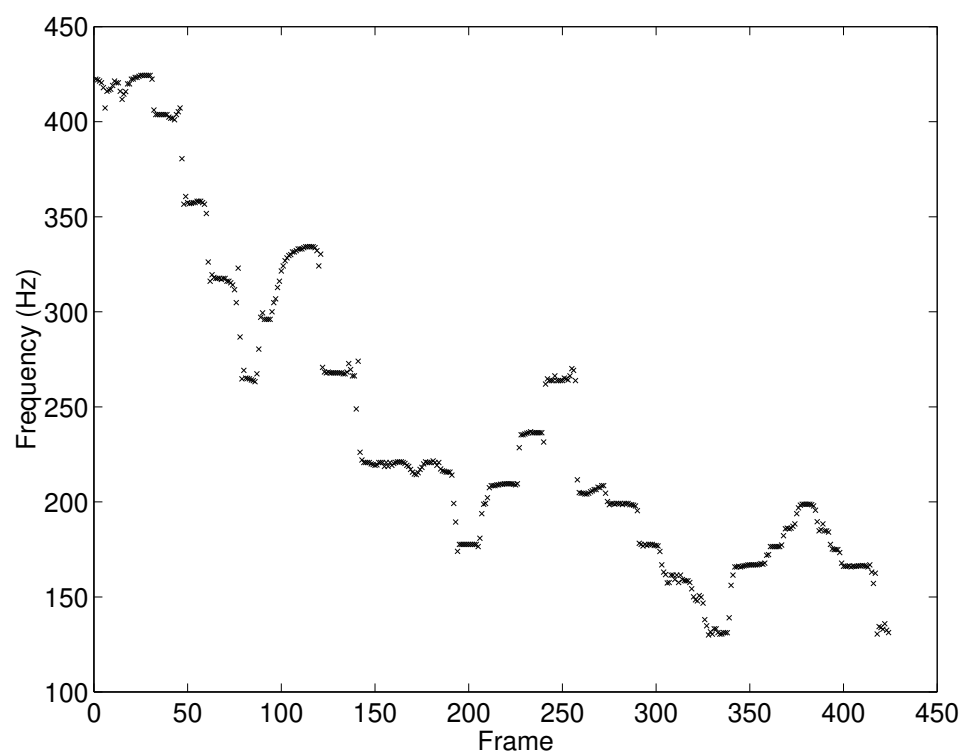


Figure 6.9: Multiple frame model for monophonic data. Dataset: sax.

(as in ‘s’), fricatives (as in ‘f’) or plosives (as in ‘t’). The vocal data has a sibilant sound in frames 300–330, fricatives in frames 130–150 and 520–540 and plosives around frames 75 and 680. In singing, it is generally the case that most of the sung regions are voiced, with a more-or-less continuous glottal excitation, as compared to speech which tends to be more punctuated.

### 6.4.3 Harmonic evolution

In addition to the extraction of a pitch inference over time, the model order and harmonic amplitudes are also estimated. These can be displayed on a harmonic magnitude plot which shows the evolution of the sound over time. Figure 6.11 shows the evolution of dataset sax which is a 5 second monophonic saxophone melody.<sup>6</sup> The rich

<sup>6</sup>No overlap between frames is used in these examples, and so the frame numbers differ by a factor of two from those in figures 6.9 and 6.10.

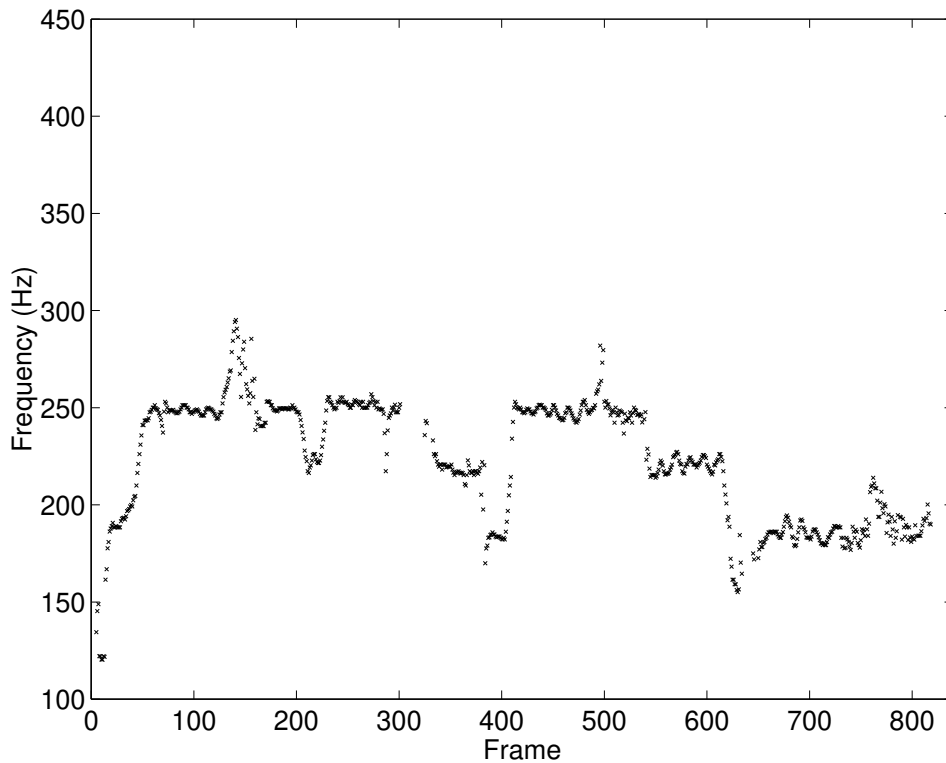


Figure 6.10: Multiple frame model for monophonic vocal data. Dataset: memory.

harmonic character of the saxophone (in common with brass instruments in general) is apparent, with prominent amplitudes for the first 20 harmonics. There is also a strong third harmonic which can just be discerned in the plot.

Figure 6.12 shows the harmonic magnitude plot for dataset memory which is an 8 second vocal melody. Most of the signal energy is contained within the first six harmonics and the amplitudes decrease substantially after the 10th harmonic. The occasional significant amplitudes between harmonics 15–20 arise due to the *singer's formant*, which occurs in male singers; the third and fourth formants lie in the region of 2.5-3kHz, reinforcing the harmonic amplitudes in this frequency range for steady-state voicing.

Figure 6.13 shows the markedly different characteristic of the flute. The flute tone is rather pure with only the first 3 or 4 harmonics having appreciable amplitude. The second harmonic is very strong in this instance, which can sometimes lead to octave errors. There is some very deep *vibrato* in frames 80 to 150 which modulates the

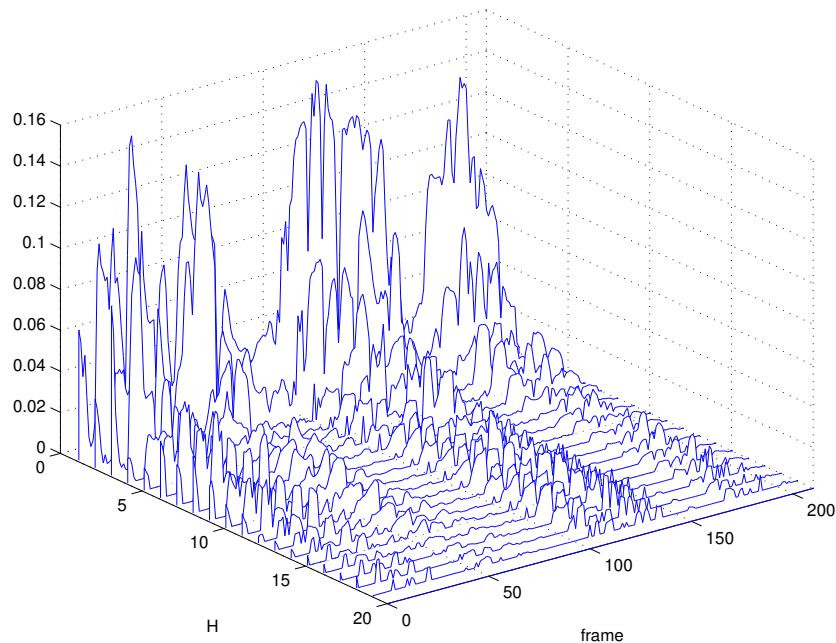


Figure 6.11: Harmonic magnitudes for sax.

amplitudes of the harmonics. This is consistent with the observations of Mellody and Wakefield [95], who noted that the perception of *vibrato* in violin sounds is actually dominated by the extreme amplitude modulation rather than the comparatively small frequency modulation.

## 6.5 Conclusions

This chapter has described the application of single component harmonic models to monophonic musical signals. The characteristics of the harmonic model and its posterior distribution have been described. The harmonic transform allows the conditional frequency distribution to be evaluated for a large number of candidate values at low computational cost, since it can employ highly optimised FFT routines, with extra zero padding leading to more resolution in the proposal distribution. Some new types of transition kernel are presented that exploit the characteristics and redundancy of harmonic models: an independence sampling step employs estimates from the harmonic

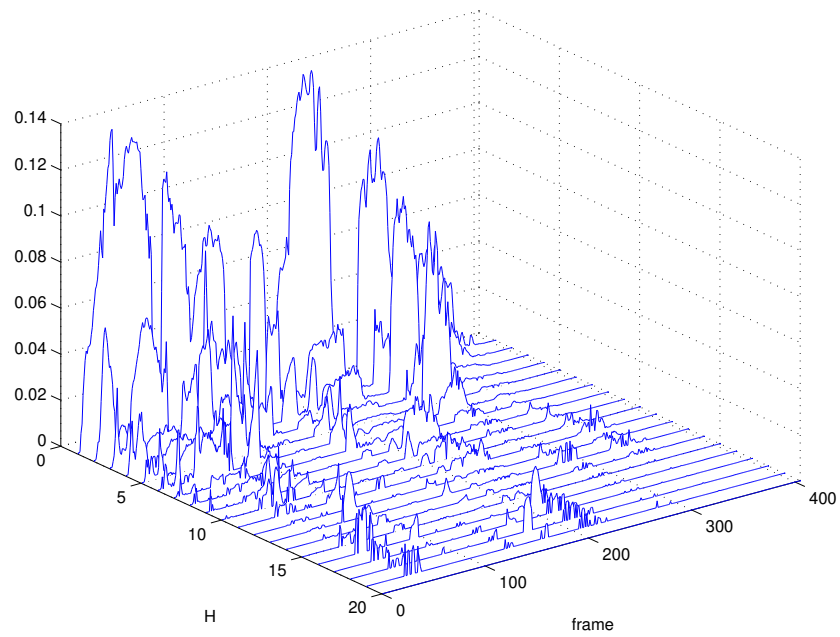


Figure 6.12: Harmonic magnitudes for vocal extract memory.

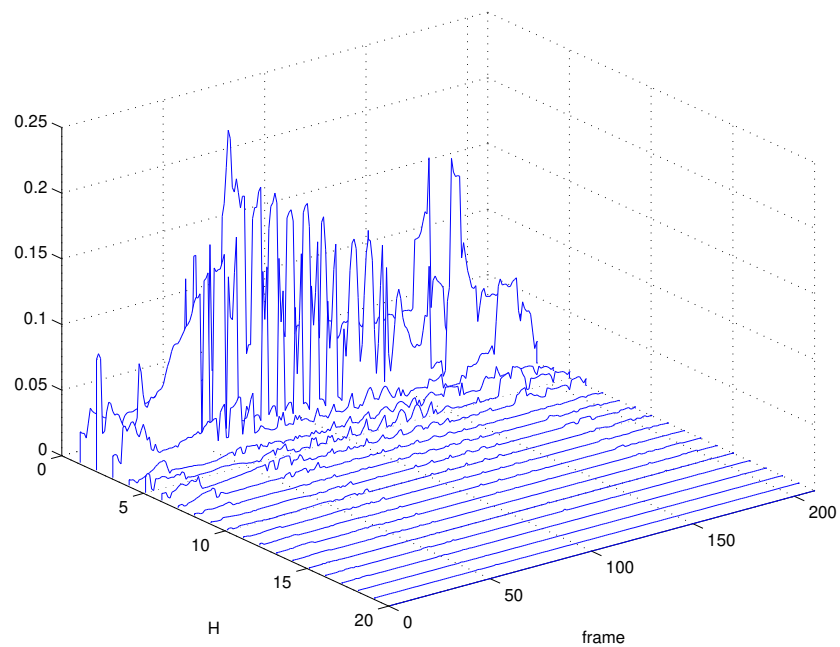


Figure 6.13: Harmonic magnitudes for flute.

---

transform to rapidly find the dominant modes of the posterior distribution, and a harmonic transition kernel explores related modes of the distribution to reduce the incidence of octave errors. Efficient transition for joint multiple frame estimation and detection are shown.



# *Application to Polyphonic Pitch Detection*

---

# 7

## *7.1 Introduction*

In this chapter, polyphonic musical signals are modelled hierarchically as a set of time-varying notes, where each note is comprised of a set of harmonically related sinusoids. The multiple component model of chapter 5 is employed, together with the harmonic model of chapter 6. The hierarchical model incorporates both horizontal and vertical Gestalt grouping mechanisms, as described in chapter 2. The techniques for producing efficient proposal distributions based upon signal- and residual-dependent kernels are employed, using the harmonic transform to evaluate proposals for a wide range of frequencies at low computational cost.

Section 7.2 describes the motivation for the polyphonic model and its formulation. The importance of the interpretation of the model composition and the concept of signal *context* for component identifiability and perceptual streaming are discussed. The transition kernels used in the simulation scheme are also presented, employing reversible jump moves for dealing with the variable sized parameter space of the model. The extension to multiple frame joint detection and estimation is presented in section 7.3 and the transition kernels required for an efficient simulation are described. Some simulation results for synthetic and real polyphonic datasets are shown in section 7.4. Section 7.5 discusses some of the problems encountered in musical signals and describes some of the advantages and limitations of the harmonic model.

Once again in this chapter, emphasis is put on the development of efficient techniques for MCMC simulation by careful choice of approximations and assumptions, and by

exploiting the structure of the harmonic models and its posterior distributions.

## 7.2 Single frame polyphonic model

### 7.2.1 Motivation

A signal comprised of polyphonic musical data may be regarded as a multiple component signal. The constituent notes may arrive from separate sources (*i.e.*, instruments) or from different oscillating systems in a single instrument (*e.g.*, the individual strings of a guitar or piano). Hence it seems reasonable to assume prior independence for the notes, by virtue of their (maybe only approximate) physical independence. The abstraction of a musical note is a sensible one; it effects a logical grouping for the large collection of sinusoids present in the observation. On a physical level, it groups those sinusoids which have come from the same resonant system of one instrument, and on a perceptual level it corresponds to the percept of a note, which, in many applications, is the desired object for inference.<sup>1</sup> The perceptual organisation of sinusoids into musical notes is largely determined by Gestalt grouping cues. The predominant cues are common harmonicity and common onset (see §2.2.2). The multiple component harmonic model provides such a grouping mechanism, where each note constitutes a signal component. Common harmonicity is represented explicitly by the model; common onset is more implicit, and is represented through the variational hyperparameter for the number of harmonics over time in the multiple frame model. The next section considers a single frame approach to polyphonic signal modelling. Section 7.3 describes a multiple frame model which achieves a robust pitch analysis than the independent frame model.

The model equation is

$$\mathbf{d} = \sum_{q \in \mathbb{Q}} \mathbf{G}^q \mathbf{b}^q + \mathbf{e}. \quad (7.1)$$

The multiple component harmonic model assumes prior independence between its components. The parameter space of the model is  $\{\{\omega^q, H^q\}_{\mathbb{Q}}, \mathcal{M}_{\mathbb{Q}}\}$  where the model pa-

<sup>1</sup>By contrast, in an audio coding application this level of inference may not be required; the individual sinusoidal frequency tracks may be of more interest.

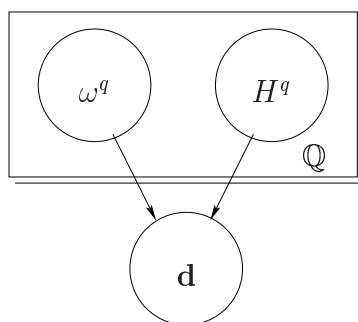


Figure 7.1: Dependencies for the polyphonic single frame model.

rameters and also the model composition are to be estimated; the graphical model for this model is shown in figure 7.1. The simulation scheme for the model employs reversible jump moves to make transitions across subspaces of the model. The techniques described in §5.3.1 for obtaining efficient transition kernels by employing the residual waveform are also applied in this section to reduce the computational requirements significantly and enable processing at speeds approaching the order of real time.

### 7.2.2 Model composition

The model composition  $\mathcal{M}_Q$  (*i.e.*, the number of notes) is unknown and is to be estimated along with the fundamental frequency and number of harmonics of each note. The model can be expressed as a single GLM by forming the composite basis matrix (5.2) for the components that are switched on. The inclusion of each component is controlled by a Boolean indicator variable  $\Gamma^q$ . Hence the problem of finding the notes which are switched on is one of subset selection rather than of model order selection. This interpretation differs from similar techniques in the literature, *e.g.*, [5, 51, 58], who propose reversible jump schemes for detecting the number of components in a mixture by considering the number of components to be an explicit parameter of the model. The use of indicator variables is a different type of parametrisation which allows for the specification of different prior expectations for each component. The example in §5.4 illustrates how this may be employed for the detection and estimation of heterogeneous mixtures. For musical signals, this parametrisation allows each component to form a stream over time, such that a context is built up for each component on the basis of

its past history. This dependence is formed through the parameter priors: for a single frame method the dependence could be Markovian, based upon the behaviour in the previous frame. For a multiple frame model, the dependence is represented between frames by the block hyperparameters, which may be assigned a Markovian dependence between blocks.

This parametrisation reflects the rôle of streaming in human perception (see also §2.4.4). A stimulus must build up a context for itself over time before it is accepted as a meaningful entity such as a note. Once the context for one note has been established, the addition of a second note at a later instant in time will be regarded as a new note, even if the pitch is almost identical. If, however, both notes are presented simultaneously, then they may be perceived as a single note, since a context has not been established for either. Players in an ensemble produce onset asynchronies of the order of 30–100ms [119], which is generally sufficient to perceive the notes individually. By contrast, gamelans are played in pairs and are inharmonic, so that two chimes struck at the same time are perceived as one event, since we can no longer appeal to common harmonicity as a grouping cue [76]. The importance of horizontal and vertical mechanisms of grouping has been appreciated by several authors in the field of musical signal processing, for instance [18, 39, 71].

### 7.2.3 Transition kernels

The simulation scheme employed for multiple notes is based largely on that of algorithm 5.1 in §5.3. A non-deterministic choice is made between a note birth-death move and an update move for a single note on each iteration. The update move is similar to that described in algorithm 5.2 except that four transition kernels are used rather than three.

#### *Independence sampling kernel*

This is a signal-dependent kernel (see §5.3.2) which is calculated before the start of the Metropolis-Hastings algorithm. A proposal distribution is generated by constructing a monophonic harmonic model from the observed signal  $\mathbf{d}$ . A value  $H^{q*}$  is sampled from a distribution centred upon a value  $H_0$  which represents roughly the expected number

of harmonics (and hence may also be the mean of the prior for  $H^q$ ). A value of  $H_0 = 6$  works well for a wide variety of sounds, since many instruments have most of their energy concentrated in the first six harmonics. The first six harmonics are also the ones which have the most influence on pitch perception [99]. A proposal distribution for  $\omega^{q*}$  is also generated using the value  $H_0$ ,

$$\begin{aligned} H^{q*} &\sim q(H^{q*}; H_0) \\ \omega^{q*} &\sim q(\omega^{q*}; H_0). \end{aligned} \tag{7.2}$$

This distribution is approximated using the harmonic transform, since this allows calculation of the proposal distribution over many frequency points in a computationally efficient manner. The proposal of (6.9) is used but with  $H^*$  replaced with  $H_0$  since this distribution can be calculated at the outset of the simulation. A more sophisticated approach might be to generate this distribution for several values of  $H$  and then draw  $\{\omega^{q*}, H^{q*}\}$  jointly from the resulting two-dimensional distribution. This is computationally more expensive than evaluating the distribution for a single  $H$  value, but the FFT only has to be calculated once. It may be useful where it is expected that the data might be comprised of several notes with different numbers of harmonics. The extra computational expense incurred in increasing the amount of zero-padding in the FFT may be worthwhile since this will go some way towards reducing octave errors, due to the very narrow peaks of the posterior distribution.

### *Conditional independence sampling kernel*

The conditional independence sampling kernel generates a proposal distribution based upon the residual. The approach is analogous to an iterative ‘estimate-and-subtract’ method, where the estimates for each note are made using the residual waveform of the original observation minus the reconstructions of the other components. In this scheme, a proposal distribution is created using a single note model from the residual  $\mathbf{r}^q$ ,

$$\mathbf{r}^q = \mathbf{d} - \sum_{\substack{q' \in \mathbb{Q} \\ q' \neq q}} \mathbf{G}^{q'} \mathbf{b}^{q'} \tag{7.3}$$

where the value for  $\mathbf{b}^{q'}$  could be obtained from the least-squares estimate or as a sample from the full conditional. The orthogonality of the harmonic model is exploited<sup>2</sup> to allow the resulting distribution to be used as an approximation to the full conditional,

$$p(\tilde{\theta}^q | \{\tilde{\theta}^q\}_{-q}, \mathcal{M}_{\mathbb{Q}}, \mathbf{d}) \approx p(\tilde{\theta}^q | \mathbf{r}^q, \mathcal{M}_{\mathbb{Q}}) \quad (7.4)$$

where  $\tilde{\theta}^q = \{\omega^q, H^q\}$ . The background behind this relation is described in §5.3.1. A value for  $H^{q*}$  is drawn from an independent proposal distribution  $q(H^{q*})$ . The proposal for  $\omega^{q*}$  is then drawn from a distribution constructed using the residual,

$$\begin{aligned} H^{q*} &\sim q(H^{q*}) \\ \omega^{q*} &\sim q(\omega^{q*}; H^{q*}, \{\omega^{qk}, H^{qk}\}_{-q}). \end{aligned} \quad (7.5)$$

The proposal employs the harmonic transform

$$\begin{aligned} q(\omega^{q*}; \bullet) &= \sum_{l=1}^L \mathcal{N}\left(\omega^{q*}; \frac{2\pi f_s l}{N_{\text{fft}}}, \sigma_\omega^2\right) \\ &\times (1 + \delta^2)^{-H^{q*}} \left[ \|\mathbf{r}^q\|^2 - \frac{2\delta^2}{N(1 + \delta^2)} \mathcal{H}_{H^{q*}}(\mathbf{r}^q, l) + 2\beta_e \right]^{-\varepsilon} p(\omega^{q*}). \end{aligned} \quad (7.6)$$

### *Harmonic transition kernel*

A dependent kernel employing a harmonic transition move is also used. This kernel is used to move around the related modes of the posterior distribution, in the hope that a more efficient representation may be obtained,

$$\begin{aligned} r &\sim \mathbb{U}_R \\ \omega^{q*} &= r\omega^{qk} \\ H^{q*} &\sim q(H^{q*}; \lfloor H^{qk}/r \rfloor). \end{aligned} \quad (7.7)$$

### *Perturbation kernel*

A perturbation kernel allows the Markov chain to explore local modes, which, combined with the harmonic transition kernel, is important for avoiding octave errors. It

<sup>2</sup>The harmonic model is approximately orthogonal as long as  $N$  is large and that no harmonics are shared between notes. The validity of this assumption in the context of musical signals will be discussed later in this chapter.

is decided non-deterministically whether to propose a perturbation for the harmonic number or for the fundamental frequency,

$$\omega^{q*} \sim \text{N}(\omega^{q*}; \omega^{qk}, \sigma_\omega^2) \quad (7.8)$$

or

$$H^{q*} \sim q(H^{q*}; H^{qk}). \quad (7.9)$$

### *Birth-death move*

A birth-death move is proposed which performs a reversible jump between subspaces of the model. The four transition kernels described above are all ‘local’ to the note under consideration, as they update the state for a single note. The birth-death move also proposes a change in the model composition  $\mathcal{M}_{\mathbb{Q}}$ , although the states of the other notes are not affected. The basic birth-death move is described in algorithm 5.3. If the current note  $q$  is switched off, then the birth move is executed, such that the proposal model composition  $\mathcal{M}_{\mathbb{Q}^*}$  now includes the current note,  $\mathbb{Q}^* = \mathbb{Q}^k \cup \{q\}$ . A proposal must be formed for  $\{\omega^{q*}, H^{q*}\}$ . This is done using the conditional independence sampling kernel described in (7.5) and (7.6). A proposal value for  $H^{q*}$  is sampled from an independent distribution that is then used in the proposal distribution for  $\omega^{q*}$ . The distribution (7.6) is constructed with the harmonic transform of the residual  $\mathbf{r}^q$  to ensure that there is a high probability of locating a mode of the posterior distribution that hasn’t already been accounted for by another note. The corresponding death move, proposed if the note is currently included in the model, proposes switching off the current note.

This form of birth-death move is very important for rapid convergence of the Markov chain since it is very likely to propose moves into high probability regions of the posterior distribution. The use of the residual waveform rather than the observation allows weaker components to be detected once the stronger components have been found.

### 7.3 Multiple frame polyphonic model

Polyphonic pitch estimation is a more complex task than finding the fundamental frequencies in successive frames of data. Such a method would be very ‘short-sighted’ in the sense that inferences about the underlying pitches cannot be reliably made over such short time scales. Many sinusoidal analysis techniques, following the lead of McAulay and Quatieri [90, 113] acknowledge the importance of directly incorporating time-varying frequencies and amplitudes, rather than presenting a set of discrete frequency estimates at each instant in time. These techniques are largely biased towards audio coding and transformation applications. In the former, the goal is to obtain a compact representation of the data that can be used to generate a high-quality resynthesis of the signal.

The parametrisation of time-varying frequency tracks exploits the redundancy in the relatively slow variation of frequency over time. Most such techniques rely upon an independent analysis in each frame, *e.g.*, from spectrum peak-picking, and then form frequency tracks as a subsequent step. McAulay and Quatieri find the tracks that minimise the differences between frequencies in successive frames [90]. Serra creates a set of *frequency guides* that are created and destroyed dynamically, but which are allowed to ‘sleep’ for several frames if the evidence for that frequency is temporarily absent [132]. Rodet uses a probabilistic method which is applied globally to the entire set of frequency estimates; the method uses a Hidden Markov Model to find an optimal set of smooth frequency tracks [33, 124].

The conceptual jump between sets of frequencies at discrete instants in time to continuously varying frequency tracks is analogous to the psychological streaming mechanisms used for horizontal grouping. For instance, melodies are heard as a pitch moving in time, rather than a set of separate events [24]. The horizontal and vertical Gestalt grouping mechanisms favour coherent behaviour of the frequency tracks: common harmonicity and common onset for vertical grouping, and similarity and good continuity for horizontal grouping.

Therefore, rather than performing pitch estimation independently in each frame of data, inference is performed over longer time scales. The time-varying nature of musical signals puts an upper limit upon the length of the analysis window, but the parameters of

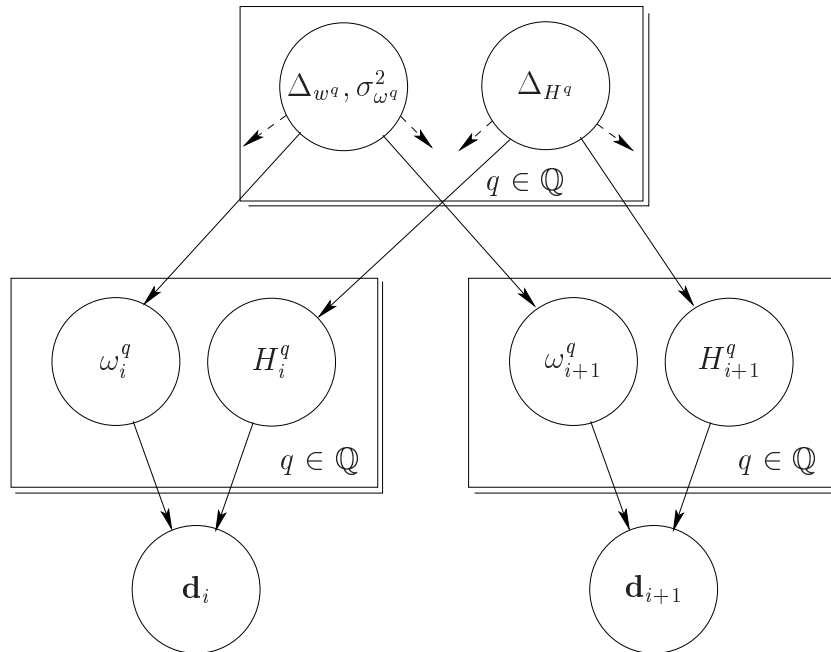


Figure 7.2: Polyphonic multiple frame harmonic graphical model.

the signal are assumed to vary slowly over time and are assumed constant across analysis windows of the order of 20ms. For a sound to be perceived as a note it must have a duration of several frames. A model that assumes notes sound continually across a block of several frames will detect the most ‘stable’ notes. An independent frame analysis invariably generates more pitch hypotheses than are perceived to be present. Extra sinusoids appear in the data to ‘explain’ transient signal components, inharmonicities and non-linearities such as vocal fry.<sup>3</sup> Many of these artifacts do not appear in the inference if the constraint of requiring a component across all frames in the block is present.

A graphical model to represent the multiple frame polyphonic model is shown in figure 7.2. Blocks are formed from  $N_f$  frames (typically 5–10), and each note is associated with an underlying context, with frequency  $\Delta_{\omega^q}$  and  $\Delta_{H^q}$  harmonics. The frequency and number of harmonics in each frame is assumed to be close to the underlying value, and so the priors for the parameters in each frame are dependent on the *variational*

<sup>3</sup>A sinusoids plus residual model may be better for representing these artifacts, if explicit inferences about them are required. [57, 124, 132].

*hyperparameters* of the block. Dependencies at a higher level could also be represented by a Markovian prior on the variational hyperparameters, which would associate the detected note with similar frequencies in previous blocks.

### 7.3.1 Simulation scheme

The priors for each note in the multiple frame model are the same as those listed in §6.4. The joint posterior for the model is

$$\begin{aligned}
 p(\{\{\omega_i^q, H_i^q\}_{N_f}, \Delta_{\omega^q}, \sigma_{\omega^q}^2, \Delta_{H^q}\}_{\mathbb{Q}} | \{\mathbf{d}_i\}_{N_f}) \propto \\
 \prod_{i=1}^{N_f} \left[ p(\mathbf{d}_i | \{\omega_i^q, H_i^q\}_{\mathbb{Q}}) \prod_{q \in \mathbb{Q}} p(\omega_i^q | \Delta_{\omega^q}, \sigma_{\omega^q}^2) p(H_i^q | \Delta_{H^q}) \right] \\
 \times \prod_{q \in \mathbb{Q}} [p(\Delta_{\omega^q}) p(\sigma_{\omega^q}^2) p(\Delta_{H^q})] p(\mathcal{M}_{\mathbb{Q}}). \quad (7.10)
 \end{aligned}$$

The simulation scheme used for the multiple frame polyphonic model combines the multiple component scheme of algorithm 5.4 (in §5.5) with the transition kernels specific to harmonic signals discussed in the previous sections of this chapter. On each iteration of the algorithm, the state for each note is updated using one of a number of transition kernels.

One candidate is a block update move that updates the state of a given note across all frames in the block. It is a conditional independence sampling step that generates its proposals from the residual  $\mathbf{r}^q$ ; it is described in more detail in §5.6. A value for  $\Delta_{H^q}^*$  is proposed from an independent distribution, and then a proposal distribution for  $\Delta_{\omega^q}^*$  is constructed by assuming that the deviation of frequency in each frame will be small; the distribution is of the form of (5.52), which is obtained from the product of the likelihood in all frames after substituting  $\Delta_{\omega^q}$  for  $\omega_i^q$ . The harmonic transform is evaluated in each frame to obtain a distribution that can quickly identify components with similar fundamental frequency across all frames,

$$\begin{aligned}
 q(\Delta_{\omega^q}^*; \bullet) \propto p(\Delta_{\omega^q}^*) \prod_{i=1}^{N_f} \sum_{l=1}^L \mathcal{N} \left( \Delta_{\omega^q}^*; \frac{2\pi f_s l}{N_{\text{fft}}}, \sigma_{\omega}^2 \right) \\
 \times (1 + \delta^2)^{-H_i^{qk}} \left[ \|\mathbf{r}^q\|^2 - \frac{2\delta^2}{N(1 + \delta^2)} \mathcal{H}_{H_i^{qk}}(\mathbf{r}^q, l) + 2\beta_e \right]^{-\epsilon}. \quad (7.11)
 \end{aligned}$$

Proposals for the parameters of each frame are obtained from perturbations of the hyperparameters. A proposal for the hyperparameter variance,  $\sigma_{\omega_q}^2$  is generated from its full conditional distribution (4.61).

Another move is the birth-death move, shown in greater detail in §5.5.1. For the birth move, a proposal for all the parameters of the new note across the entire block is generated using the conditional independence distribution described above. For the death move, this distribution must also be calculated to evaluate the probability that the current state would have been generated from the proposal in the reverse transition.

There are two types of dependent transition kernel used. One is a joint block update which proposes a harmonic transition for the parameters and hyperparameters over the whole block, as shown in (6.20). The second type of move is a perturbation applied non-deterministically to either the fundamental frequency  $\omega_i^q$  or number of harmonics  $H_i^q$  of a randomly chosen note in a randomly chosen frame. This is followed by an update of the hyperparameters of the note, obtained by sampling from their full conditionals; this step is also described in §5.5.3.

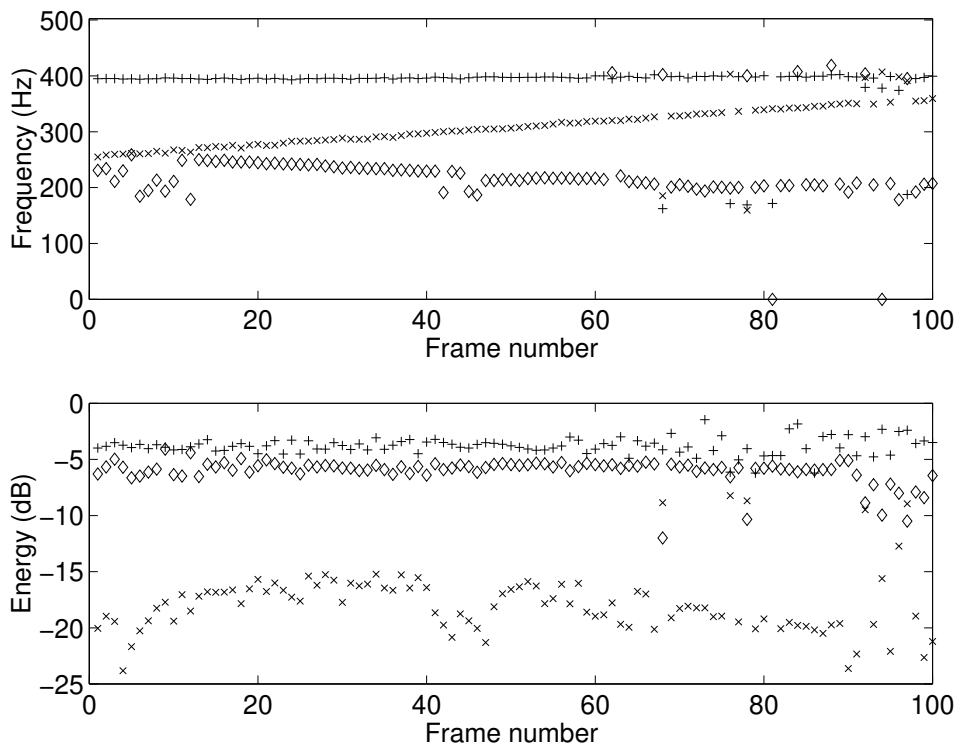
## 7.4 *Simulation results*

### 7.4.1 *Synthetic harmonic data*

A synthetic data set is generated with three sets of harmonic series present. Two sets have similar amplitudes and the third is around 20dB lower. The fundamental frequencies of two components vary linearly over time, whilst the frequency of the other is modulated at around  $7\text{Hz}^4$  and each has between 4 and 6 harmonics. The top plot of figure 7.3 shows the MAP estimates for the fundamental frequencies. The lower plot shows the energy of the components in each frame relative to the signal energy. The general frequency trend of each component is clearly visible. There are several areas however where the detection has failed to pick out all three components. In the initial region of frames 1–10, two of the frequencies cross. This which precludes the detection

---

<sup>4</sup>This corresponds to the typical *vibrato* rate of acoustic instruments.



**Figure 7.3:** Synthetic polyphonic data. Dataset: *synharm*. The top plot shows the frequency tracks, and the lower plot shows the energy of each component. Estimation of the weakest component becomes difficult when its harmonics are close to those of the stronger components.

of the component with the lower amplitude. Around frames 41–45 the frequency track of the lowest component crosses the half-frequency of the highest component. In the area around frame 90 the detection is disrupted by the crossing of harmonics of the lower and middle frequency tracks.

The detection of all components is generally good — the two stronger components are picked up in most frames, whilst the weaker component is found for the frames that do not have a clash of harmonics. These problems occur since the basis matrix is rank deficient: two components share a harmonic and the model is unable to determine how much of the harmonic each component is entitled to have. The simulation was run for 100 iterations in each frame. The algorithm is designed to converge to the high probability regions of the posterior very quickly, which is achieved through the combination of conditional independence sampling using the residual, and harmonic

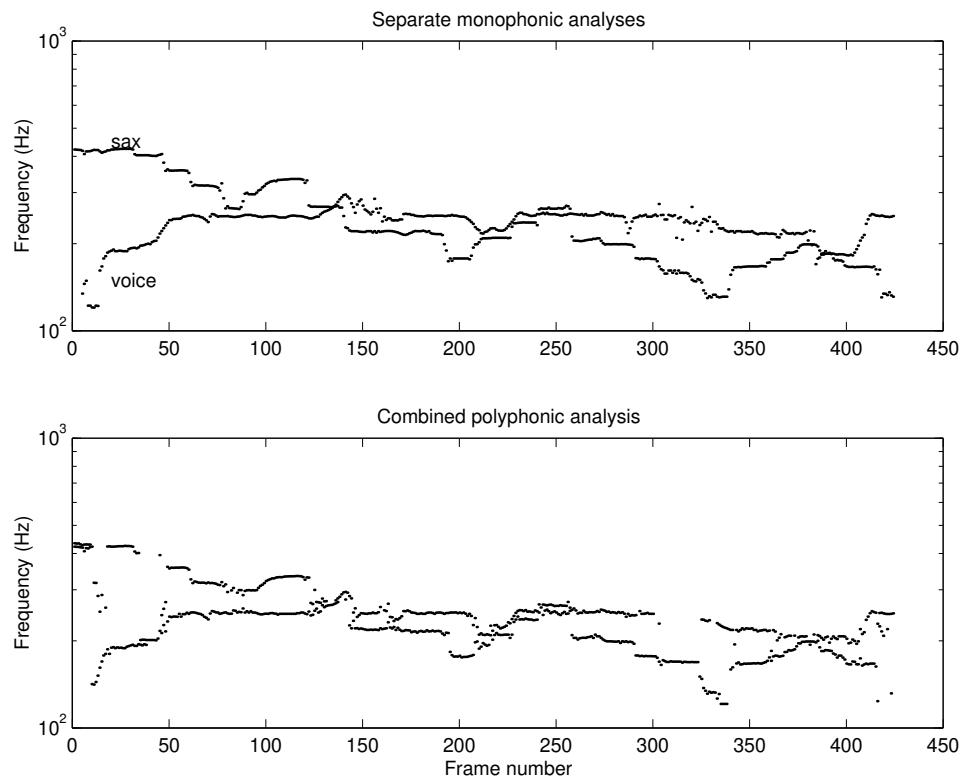


Figure 7.4: Splitting of two independent monophonic sources. Dataset: saxmem.

---

transitions for the elimination of octave errors. With careful implementation, results of the order of real-time are achievable.

#### 7.4.2 Duophonic example

One of the fundamental assumptions behind the polyphonic model is that the components (notes) are *a priori* independent. The motivation for this is that the physical systems creating each note are likely to have a high degree of physical independence, in addition to being a simplifying assumption. The dataset saxmem is created from the addition of two independent monophonic melodies, one from a saxophone (see also figure 6.9) and one vocal sample (the first 430 frames of figure 6.10). The superposition of the pitch contours obtained from the monophonic analysis of each sample separately is shown in the top plot of figure 7.4. An analysis with a two-note (duophonic) model yields a set of pitch estimates shown in the lower plot. Agreement between the two

plots is very good. Most of the two melodies have been detected. The only areas where a pitch hasn't been detected are around frames 40–50, where the vocal part coincides with half of the saxophone pitch, and frames 300–330, which is a sibilant region in the vocal part (and hence there isn't a perceivable pitch in that region).

One limitation of the detection is the inability to follow either melodic line, *i.e.*, the inability to horizontally group the frequency tracks into the correct melodic streams. A prior distribution for the hyperparameters that is based purely on the distance between the pitches in adjacent blocks is inadequate to determine the correct horizontal grouping, since the melodic lines are not continuous (there are occasional jumps) and are not mutually disjoint. Even if trajectory information were to be incorporated into the prior structure, the model would still be unable to resolve many of the problem regions. From visual inspection of the top plot, it is sometimes difficult to discern the correct grouping, particularly around frames 210–240 and 380–400. To achieve good melodic grouping it would also be necessary to impose continuity constraints on the spectral characteristics of the melodic lines (in this instance, the relative harmonic amplitudes are very different for saxophone and voice, see figures 6.11 and 6.12), or to have a better model of pitch variation that can represent the discontinuities inherent in melodies.

### 7.4.3 *Polyphonic piano examples*

Figure 7.5 shows the pitch estimates for a polyphonic piano excerpt with 2–3 note polyphony. The data was created using a piano preset on a synthesiser, and all of the notes were played at a similar level. The frequency axis is on a logarithmic scale, marked with the approximate locations of the note 'C' in each octave, with middle C close to 260Hz. The darkness of each point on the plot is proportional to the log of the energy of each note. The algorithm works well on this synthetic data as the oscillators are very stable and many of the artifacts of real pianos, for instance inharmonicity and non-linearities, are absent.<sup>5</sup> The assumption of slow frequency variation is very well suited to piano tones. The fixed length of piano strings and the constraint of a 12-tone keyboard means that frequencies are constant over the duration of a note.

---

<sup>5</sup>However, many synthesiser manufacturers strive for realism by modelling such effects or sampling real piano sounds.

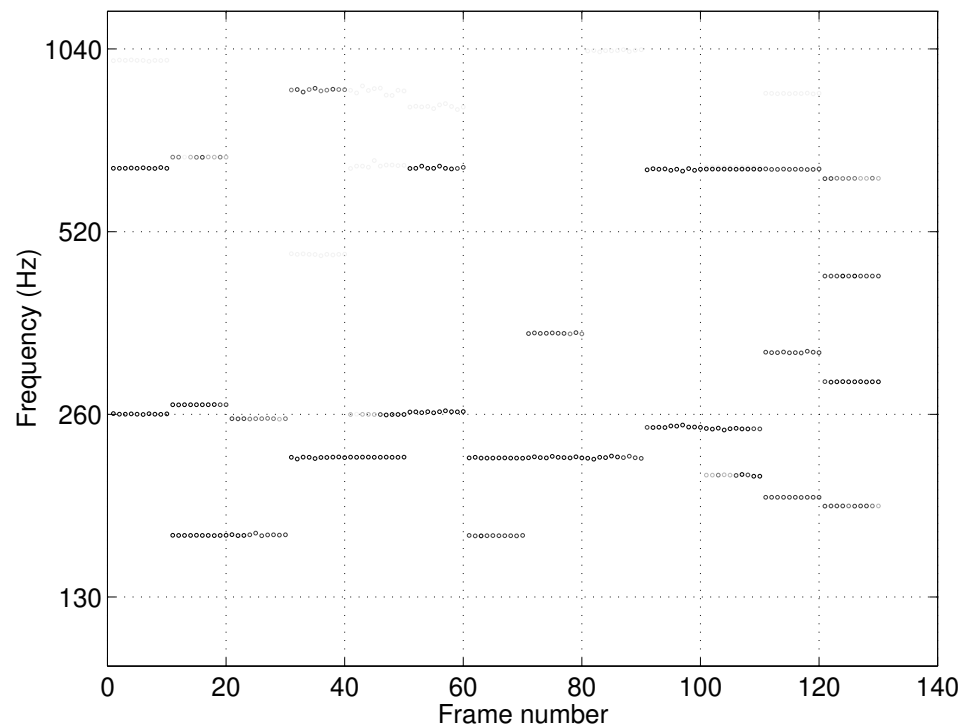


Figure 7.5: Synthesised polyphonic piano example. Dataset: *piantun*.

The pitch estimates for an extract played by an acoustic piano are shown in figure 7.6. The range of note amplitudes is much greater in this instance and the recording is very reverberant. The detection is not as robust as for the synthetic case, and due to the high levels of reverberation repeated notes are not picked up as separate entities as the sound continues across the rest between the notes. A number of very faint lines are just visible, particularly in the top left corner of the plot. Most of these lines fall close to harmonic frequencies of the middle C note that occurs between frames 5–50; lines at the C in the next octave (520Hz), the G above (780Hz) and the E above that (1320Hz) are at multiples of 2, 3, and 5 respectively. Hence these are likely to be caused by slight inharmonicities of the piano, which is more extreme in the lower notes due to the width of the bass strings. The left hand part of the piano melody is a C-E-G-E-C *arpeggio*; the right hand part starts around frame 180.

A second acoustic piano extract is shown in figure 7.7. This tune has a more chordal left hand part, which is clearly visible in the 130–260Hz octave. It is difficult to tell

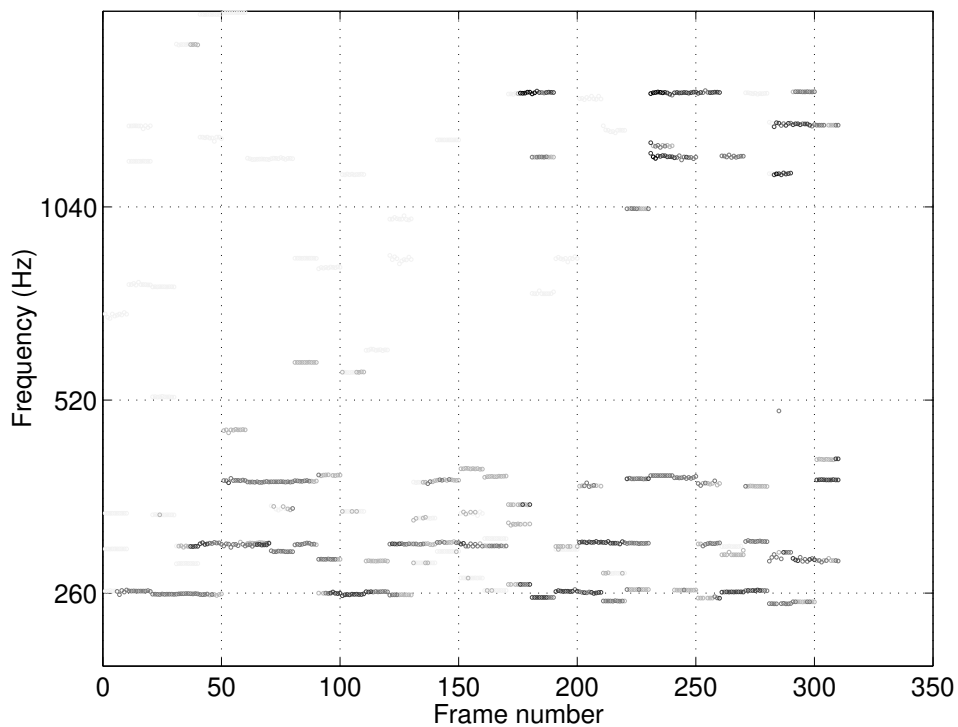


Figure 7.6: Acoustic polyphonic piano example. Dataset: granpre.

whether all of the notes in the octave above are actually being played or whether they are octave errors. The fast right hand *arpeggio* between frames 210–260 is captured well by the model.

## 7.5 Appraisal of the harmonic model

In this section the performance of the polyphonic harmonic model is assessed in terms of its suitability for musical signals, and the validity of the assumptions made in its formulation. Many of the problems that arise in real signals are discussed.

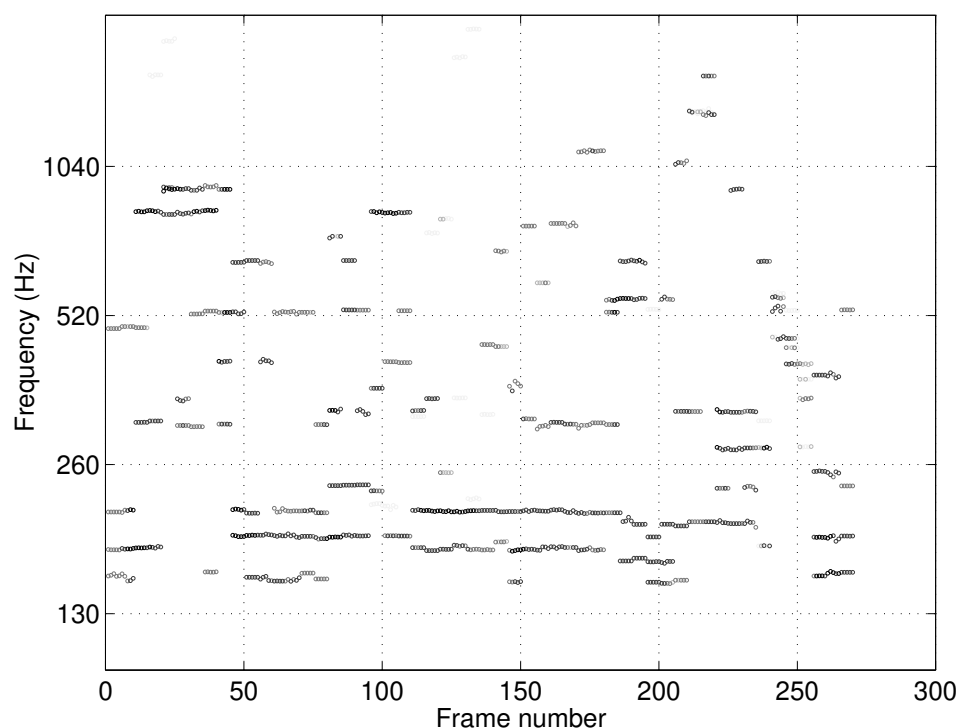


Figure 7.7: Acoustic polyphonic piano example. Dataset: *wow*.

### 7.5.1 Benefits

**Probabilistic model** To the best of the author’s knowledge, this is the first probabilistic treatment of the polyphonic pitch estimation problem, allowing for the incorporation of subjective or statistical information into the prior structure of the model. The use of MCMC methods for the posterior simulation allow a flexible choice of model structure.

**Integrated model** The model jointly deals with low-level (signal) and also high-level (note) aspects, and prior information can be harnessed at each level. It is extensible to higher levels of modelling to represent musical structure such as note timing [116] or note transitions [71]. Encapsulation is achieved through conditional independence and graphical model separation, rather than by using a series of isolated black boxes (*cf.* [141]).

**Good steady state representation** Many (pitched) musical instruments can be modelled

well by a harmonic model since in order to produce a definite, unambiguous sensation of pitch, they must oscillate in a steady state for a period of time of the order of 20–100ms. The draft MPEG-4 standard [68] now has the capability of coding a harmonic series, since this results in a more efficient coding scheme [112].<sup>6</sup> The multiple frame method allows slowly time-varying signals to be modelled by assuming the parameters can be held fixed in each frame.

**Meaningful parametrisation** The parameters of the model map onto the perceptual domain: for most harmonic stimuli, the perceived pitch is equal to the fundamental frequency of the harmonic set of sinusoids, hence the prior structure can explicitly impose prior beliefs upon the pitch variation. This would not be easily possible with (for instance) a model parametrised on AR coefficients or pole locations.

**Parametric model** Since the model is parametric, pitch estimation is just one possible output. The harmonic character of the instruments can also be investigated, the parameters could be used for audio coding, or the signal could be resynthesised, which may be of interest for a signal separation application.

**Resolution** The frequency resolution of the estimation scheme is not limited by DFT resolution. The DFT is harnessed to provide estimates of the high probability regions of the posterior, and then the MCMC simulation locally explores the modes of the distribution. Similarly, the resolution is not limited by a fixed semitone grid as with a musically tempered scale (*cf.* [104, 144]).

**Generality of assumptions** The assumptions made in the model are rather general, such that the model may be applied to a wide range of instruments. If extra knowledge is known about a particular instrument, *e.g.*, that pianos have no frequency deviation, then this may be incorporated into the model. Some polyphonic estimation techniques require each note of one specific instrument to be recorded in isolation, and the detection will only work for that one instrument, which is impractical in many applications [73, 125].

---

<sup>6</sup>However, only a single harmonic series is employed which is only of practical benefit for monophonic signals.

### 7.5.2 Limitations

**Rapid variation** Musical signals generally exhibit a rapid variation, and short steady state regions are punctuated by transient events and time-varying artifacts. Frequencies may vary rapidly, particularly during note attacks, and the pitch contour may have large discontinuities at note boundaries. Notes may start at any point in time rather than at block boundaries. Deep *vibrato* may also lead to modelling errors.

**Non-Gaussian error** In addition to the rapid variations, many instruments, particularly percussive ones such as guitars and pianos, produce transients when struck. The breath noise produced by flautists and bow noise caused by string players are coloured and are essential to the character of the sound. When a model of these sounds is resynthesised without the noise, the result sounds quite unnatural. For coding or separation purposes, the transient and noise components could be separately represented, *e.g.*, as with [132].

**Inharmonicity and aperiodicity** Not all instruments are perfectly harmonic. The piano in particular has distinct inharmonicities for the lower strings, as a result of their finite width and effects of stretching.<sup>7</sup> The excitation of some instruments may have some degree of aperiodicity as a result of the chaotic oscillations causing the excitation, for instance vocal fry and bow friction [124], whereas the harmonic model assumes periodic oscillations.

**Choice of prior values** It is difficult to determine how strong prior information should be, particularly when applying subjective information. It may be possible to determine good values for some priors from statistical observations, but for some parameter priors (for instance the variance of the fundamental frequency hyperparameter) it can be difficult to produce a value that represents the prior beliefs. A major risk is that the prior may be compensating for modelling errors, and the extreme values required to exert an influence over the likelihood may appear intuitively implausible.

**Crisp inferences** In an application such as pitch estimation, crisp inferences are gener-

---

<sup>7</sup>This phenomenon has been exploited by some authors to separate and identify harmonics [74, 125].

ally required; *i.e.*, the desired output is to determine the notes that are present in the observation. It is apparent from some of the polyphonic pitch plots shown in the previous section that many more note hypotheses are produced than are audible. Many of these occur because of modelling errors such as inharmonicity (note that they are more prevalent in the acoustic example of figure 7.6 than the synthetic example of figure 7.5). The extension of the model to include a higher level ‘label’ which would mark the salience or significance of each note could be one way to achieve a more meaningful result for pitch estimation applications.

**Horizontal grouping** The multiple frame method works well for horizontal grouping over short intervals (of the order of 100–200ms), but grouping for longer durations is not as good. This arises partly from the limitations of the pitch variation model, and partly from the difficulty in choosing a suitable value for the prior variance for the frequency hyperparameter.

**Independence and orthogonality** The model assumes prior independence between the components. However, the principles governing Western harmony specify relations between pitches that sound concordant when played together; musical chords are formed from combinations of pitches in particular ratios, *e.g.*, 4:5:6 for major chords. Hence the pitches that may be sounding concurrently are likely to be correlated. This presents another problem: since harmony is dictated by pitches related by ratios of integers, this means that several notes may share harmonics. The composite basis matrix produced to model several such notes now becomes rank deficient and the basis matrices are no longer orthogonal.

What is more likely to happen in practice is that two notes that are harmonically related will be detected as a single note with a fundamental equal to the *harmonic root* of the two note pitches. This accounts for the inability of most algorithms and the human auditory apparatus to hear two simultaneous notes an octave apart. If the notes start at different instants in time then this can help a listener to perceive the notes separately. In this model, if harmonic amplitude information is also included in the multiple frame hyperparameters then this may yield a method of resolving octave ambiguities. There are a few cases in figure 7.7 where chords containing octaves have been detected (*e.g.*, around frames 100 and 150); this is possible partly due to the context which has been established by one of the notes

before the other sounds, and also due to the slight tuning variations that are likely to occur in practice (piano octaves are slightly stretched, as described in §2.2.2), which may be sufficient to make the basis matrix orthogonal once more.

### 7.5.3 *Suitability of MCMC methods*

The model equation is sufficiently complex to preclude a closed-form solution. The model space is sufficiently large and the posterior distribution has fine detail that makes an exhaustive evaluation of the whole parameter space impractical. Consequently, a numerical method employing Markov chain Monte Carlo techniques is employed. The most common criticism of MCMC methods is that they are inherently very slow and require thousands of iterations. In this thesis it is shown that algorithms employing MCMC techniques can be implemented in a much more efficient manner making real-time processing a possibility. In the applications described in this chapter, point estimates are required rather than inferences about the whole of the posterior distribution. As such, the goal of the simulation is to locate the dominant modes of the posterior distribution and then perform a local exploration to find high-resolution frequency estimates. In all the examples shown in this chapter, the best estimates produced after 100 iterations<sup>8</sup> are employed.

The choice of transition kernels is a major contribution to the efficiency of the algorithm. The use of orthogonality assumptions and approximations allows the construction of efficient proposal distributions that are capable of locating high probability modes very quickly, whilst the use of residual-dependent kernels allows the algorithm to find weaker components within the mixture. Many computational savings may be made when implementing the Metropolis-Hastings algorithm, such as caching calculations of residuals, projections and probabilities for each state. The ratios of posteriors calculated in the M-H acceptance step can also be simplified by appealing to conditional independence from the graphical model. The methods in this chapter also initialise the Markov chain in each block with the MAP estimates of the previous block. In many cases this starts the chain off in a region of high probability.

Above implementational and speed considerations, the use of MCMC methods allows a

---

<sup>8</sup>Determined as the states of the Markov chain with the highest posterior probability.

very flexible choice of model structure, particularly for parameter priors. The choice of priors reflects the expected behaviour of the signal, and the imposition of dependences between parameters in adjacent blocks leads to a model which is more robust to modelling spurious transients. The model hierarchy is also extensible, such that higher levels of structure may be imposed and inferences made in the problem domain rather than in the signal domain.

## 7.6 *Conclusions*

This chapter applies the multiple component model of chapter 5 and the harmonic model of chapter 6 to polyphonic audio data. The structure of the harmonic model and the problems that arise in its implementation are discussed. Methods for approximating the posterior distribution using the harmonic transform are described that allow efficient calculation of the distribution for a large number of candidate frequencies. The transition kernels required for an efficient simulation scheme are described; these provide a means of efficiently exploring the posterior distribution by exploiting its features and its redundancies. The model is extended to the joint detection and estimation over multiple frames, associating the hyperparameter of each component over time with a signal context. Typically, no more than around 100 iterations are required for the simulation in each block to produce the polyphonic pitch estimates, which is possible due to the careful choice of proposal distributions that can rapidly explore the posterior distribution. Some of the benefits and limitations of the harmonic model are discussed, along with a discussion of the validity of some of the model's assumptions.

# *Conclusions and Future Research*

---

# 8

## *8.1 Conclusions*

This thesis has presented techniques for the modelling of signals that can be represented in terms of a homogeneous or heterogeneous mixture of time-varying components. This model has been applied to the problem of modelling monophonic and polyphonic musical data, with a particular emphasis on pitch estimation. An hierarchical model has been described that captures the essential characteristics of musical signals in terms of the short-term time variation of a collection of harmonically related sinusoids. These two dimensions of structure correlate with Gestalt psychological grouping principles, which determine the perception of musical signals as a collection of melodic streams. The importance of both high and low levels of modelling has been emphasised, and the contribution of knowledge from areas of signal processing, music and perceptual psychology have been shown to be valuable in the formulation of a musical signal model.

The model is posed within a Bayesian framework, allowing for the specification of prior information about the model parameters. Markov chain Monte Carlo techniques have been employed to simulate a stream of samples from the joint posterior distribution for the purposes of Bayesian inference. The simulation schemes employ efficient transition kernels that are tailored to the models under consideration. Model structure is exploited to yield transition kernels that are capable of rapidly exploring the parameter space. A combination of independent, conditionally independent and perturbation transition kernels have been employed to rapidly detect both strong and weak components in the mixture. Novel forms of transition kernel exploiting the structure of the harmonic model have also been presented. An emphasis on efficiency has yielded techniques that

may allow implementations of MCMC simulations several orders of magnitude faster than are usually achieved.

## 8.2 *Discussion*

### 8.2.1 *Importance of high level modelling*

This thesis has emphasised how many different forms of knowledge are important in the construction of a model for musical signals. A technique based upon signal processing alone will be of limited success. Musical signals have many irregularities, they are generally not produced by stable oscillators, but rather as the acoustical response to instruments excited manually by human beings. Information is not transferred by the signal in a well-determined way (in contrast to communications channels) as the musician is able to vary the performance in a subjective manner, applying his/her own means of expression. No two performances of a particular tune will be identical, and there may be many possible interpretations of the same observation.

Musical signals can be ambiguous in many different ways. For instance if a harmonically related set of sinusoids are staggered in time such that the fundamental is presented first, and then successive harmonics are added a short time later then each will initially be perceived as a discrete pitch, but will gradually fade into the percept of the underlying complex tone pitch [27]. If all harmonics are presented simultaneously then only a single note percept will be formed. This mechanism is also important to the perception of chords, where it is easier to discern the individual pitches in the chord if there is a delay between the note onsets.

Signal processing methods usually operate in the short-term and lack a notion of *context*. Context provides important information about the signal and it is often desired to infer context from a signal as a goal of the estimation task. By way of illustration, consider the problem of tracking the pitches of several instruments playing concurrently (see also figures 7.3 and 7.4). It is highly likely that some of the parts may cross, and so it would appear that trajectory information must be incorporated within the tracking. Musical notes are not continuous, however: the notes in a melody are typically

discontinuous in time and pitch. The identification of the crossing frequency tracks now requires information about amplitudes and timbre, which is a multidimensional attribute of sound and cannot be simply expressed in a mathematical form. Nevertheless, tracking methods are important for the related problems of resolving individual sinusoidal components [33, 90].

There is another major limitation in using signal processing methods for musical signals. Whilst these techniques are able to produce optimal parameter estimates for given models, these parameters are optimal in a mathematical sense, rather than in a perceptual sense. That is to say that an algorithm to calculate the fundamental frequency of a harmonic series will not necessarily produce an estimate of the pitch. *Octave errors* may arise due to the pitch ambiguity of certain sounds where even different subjects may classify the pitch differently. A plot of fundamental frequency over time therefore may generate quite a different picture to the expected visual representation of a tune. These artifacts may also occur due to the violation of some modelling assumptions, *e.g.*, rapid time variation, non-linearity or inharmonicity, but the specification of a more general model may be prohibitively complex. If the goal is to produce an audio coding application then this is probably of little concern. For musical applications, however, the *interpretation* is of much more interest than the raw parameter values. The detection of chords is a further example where the mathematical solution will be different to the perceptual solution, as the mathematical solution will usually be the detection of a single harmonic series with a harmonic root an octave or so below the constituent notes.

Psychoacoustic modelling techniques are necessarily detector-oriented rather than data-oriented, but since the data contains all the information being transmitted from the player then why is it necessary to model the auditory system at all? A degree of interpretation is required to map physical parameters of sound onto their perceptual correlates, which are of greater interest for applications such as musical transcription. The Gestalt principles of grouping and streaming are of great importance for being able to interpret naked sinusoids in terms of meaningful higher level structures of notes, chords and melodies.

Musical models operate at a higher level still, although in much of the literature they are largely extensions of perceptual grouping and streaming models. Musical models have a more explicit recognition of context, in terms of concepts such as key and metre, and

inferences are often made about rhythmic, melodic and harmonic structure. Such models tend to operate upon raw note or note stream data, rather than at the signal level. The signal processing problems of (for instance) calculating the pitches of a polyphonic music signal are assumed to be solved.

### 8.2.2 *Completing the loop*

Each of the above areas contribute very important elements to musical signal processing problems and so it seems obvious that the natural way to incorporate the important aspects of each is to ‘complete the loop’. That is, signal processing methods benefit greatly from the availability of salient prior information and the imposition of structure, whilst psychoacoustic and musical models rely on accurate estimates of signal parameters such as timbre. This could be achieved in two ways — either by feedback of high level inferences to the low level model or by imposing a hierarchical model from the outset. The latter approach has been the preferred one in this thesis. A Bayesian modelling framework is able to make use of prior information at all levels and mediate the effects of external factors through latent variables. It provides a calculus for the probabilistic comparison of different models or model orders, and using MCMC methods, estimation of the correct model and the estimation of its parameters are performed simultaneously. Also significant in this approach is that the highest and lowest level parameters in the model are estimated simultaneously. Hence, there is a bootstrapping effect as all levels exert an influence upon the final model state. For instance, suppose that the likely pitches of each note in a piece of music are dependent on the key, but the key is unknown *a priori*. A model is constructed such that the priors for the pitches are dependent on the key but then the key is considered an unknown parameter and it is assigned a prior that accounts the prior knowledge of likely key distributions. A simulation and subsequently Bayesian inference on the model will provide estimates of both key and individual pitches.

### 8.2.3 *The need for hierarchical modelling and feedback*

Most methods used in actual applications will have parameters which must be ‘tweaked’ to provide an output which is acceptable to the user. It is apparent from the examples

of chapter 7 that the raw output of the model provides much more information than is required. It is generally required to determine some threshold for which the detected ‘notes’ become salient entities, in order to produce a more meaningful and useful inference. It is difficult to specify, *a priori*, parameters such as thresholds of detection as they are dependent on so many factors: the style of music, the instrumentation used, the quality of the recording and the dynamics of the music.

A design parameter such as the detection threshold should be signal-adaptive to respond to changes in the signal energy and noise conditions. However it should only exhibit a slow variation over time and be dependent upon the local context. More intelligent determination of the threshold could be made if we knew *a priori* about the dynamic range of the music and the noise characteristics. These statistics could potentially be incorporated into the model as unknowns and estimated together with the other model parameters. Other latent parameters pertaining to high-level characteristics such as (for instance) dynamic range, note repetition rate, noise floor, *etc.* could be incorporated into the model, *e.g.*, as high level hyperparameters of the Bayesian graphical model.

This modelling paradigm conflicts with a number of approaches found in the literature. Most significantly, few models refer to the specification of design parameters and how they may be elicited from the data. Some models assume a detailed knowledge of the instrument characteristics, for instance isolated recordings of each note, which is generally impractical [73, 125]. The importance of context is unfortunately not acknowledged in many signal processing references.

This paradigm also conflicts with the arguments of Sterian and Wakefield [141] who suggest that components in the processing architecture should be kept separate with no feedback paths, such that the performance of each component can be assessed separately. Whilst this ‘black box’ approach has definite advantages in terms of the simplicity of high level design and interchangeability of components, this assumes that each module has a well defined task that can be objectively assessed. It also assumes that each module can perform its task using solely the data provided by its input stage. A method with separate non-communicating processing modules is unable to make full use of the data by not being able to make use of the inferences produced by other components later in the processing chain. A fundamental tenet of engineering and computer science favours the black box approach to component encapsulation. The box inputs and out-

puts are precisely defined but the internal workings are obscured from view. Knowledge cannot easily be shared between components if they are separated, and the attempt to share knowledge would make the design of independent components difficult.

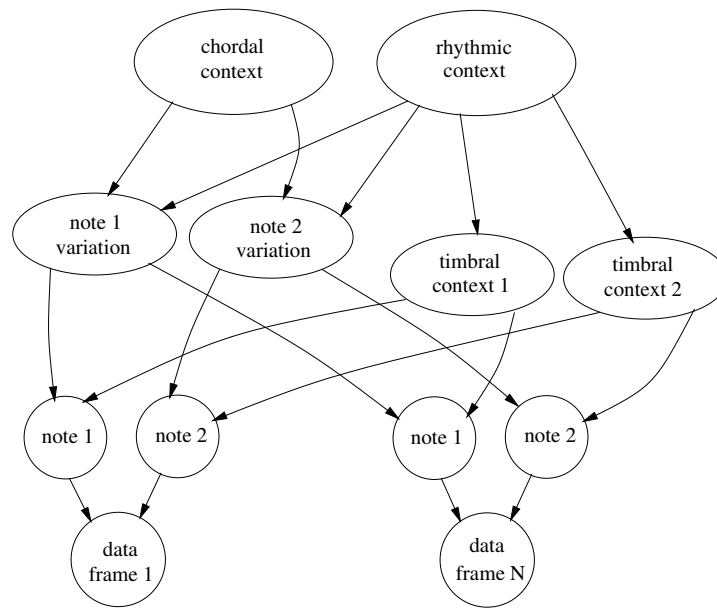
In contrast, knowledge is automatically shared between the different levels of a hierarchical Bayesian model. The process of Bayesian inference is frequently likened to learning, due to the ability to modify prior knowledge in the light of experience; a model that is capable of learning is data-adaptive. As available computing power increases, statistical modelling of signals holds great potential for the next generation of signal processing methods. Models that are capable of representing many levels of structure and that can learn through inference on the highest levels have the potential to revolutionise many signal processing problems as they can produce inferences in the problem domain as well as in the signal domain. It is hoped that the work presented in this thesis contributes towards that goal.

### 8.3 *Future research*

There are a great many outstanding problems concerned with the modelling of polyphonic musical signals. In this section, some possible extensions to the research presented in this thesis are described.

#### 8.3.1 *Timbral context*

The model can be parametrised in terms of the magnitude and phase of the harmonic sinusoids, rather than the in-phase and quadrature representation currently used. It is more intuitive to specify prior information about the harmonic magnitudes than for the in-phase and quadrature amplitudes. Over short time periods ( $\sim 200\text{ms}$ ), the relative magnitudes of the harmonic components vary slowly, and so variational hyperparameters (the *timbral context*) may be specified for them to improve the robustness of the multiple frame joint estimation scheme. This would also allow for improved inter-block horizontal grouping (for the formation of melodic streams) since the relative harmonic magnitudes of different instruments are likely to be markedly different. An



**Figure 8.1:** A high level conceptual graphical model for the representation of musical signals. Low level signal structure is represented by notes whose parameters are dependent on various forms of musical context.

added benefit arises from the ability to identify harmonic roots and octave errors from their distinctive harmonic magnitudes (*i.e.*, many components of zero amplitude).

### 8.3.2 High level modelling

The current model of signal context currently contains very little information about the underlying state of the note, namely the fundamental frequency and number of harmonics. This could be extended to include more high-level musical information as shown in figure 8.1. In addition to the basic variation of the note parameters, the priors of each note over time are also conditioned on their timbral context. For a collection of monophonic instruments, each timbral context may be different and enable identifiability of the melodic streams. For a polyphonic instrument, knowledge of the timbral context may enable the resolution of chords. Higher levels of musical structure impose a conditional independence between components. This could include a chordal context, which specifies prior information about the likely pitch values, and a rhythmic context which specifies the likely instants of note onsets.

### 8.3.3 *Beat inference*

Most of the previous discussion has centred upon pitch estimation, but another important aspect of musical signals is detection of the rhythmic structure. A ‘metrical grid’ is superimposed upon the signal, on the assumption that note onsets are likely to coincide with beat locations, with a lower probability of occurring halfway between beat locations. A general linear model can be constructed to estimate the parameters of the metrical grid — onset location, number of beats and periodicity, using the GLM techniques described in chapters 4 and 5. Knowledge of the metrical grid leads to the ability to predict the likely instants of note onsets in the future, which could be integrated into the high level modelling scheme described above.

### 8.3.4 *Software for multiple component analysis*

The multiple component mixture techniques described in chapter 5 are applicable to a wide range of problems beyond harmonic signals. Work is currently under way to provide a set of C++ classes that encapsulate the characteristics of general linear models and perform highly optimised MCMC simulations for various standard types of basis function such as those with variable scale, offset or periodicity.

# Bibliography

---

- [1] H. Akaike. Information theory and the extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki, editors, *Second International Symposium on Information Theory*, pages 267–281, Budapest, 1973.
- [2] E. Ambikairajah, A. G. Davis, and W. T. K. Wong. Auditory masking and MPEG-1 audio compression. *IEE electronics and communication engineering journal*, 9(4):165–175, August 1997.
- [3] American National Standards Institute. USA standard acoustical terminology, 1960. New York: American National Standards Institute.
- [4] M. S. Andrews, J. Picone, and R. D. Degroat. Robust pitch determination via SVD based cepstral methods. In *Proc. ICASSP*, volume 1, pages 253–256, April 1990.
- [5] C. Andrieu and A. Doucet. Joint Bayesian model selection and estimation of noisy sinusoids via reversible jump MCMC. *IEEE Trans. on Signal Processing*, 47(10):2667–2676, October 1999.
- [6] Christophe Andrieu. *Méthodes MCMC pour l'analyse Bayésienne de modèles de régression paramétrique non-linéaire. Application à l'analyse de raies et à la déconvolution impulsionnelle*. PhD thesis, l'Université de Cergy-Pontoise, 1997.
- [7] J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. John Wiley & Sons, 1994.
- [8] B. Bogert, M. Healy, and J. Tukey. The quefrency analysis of time series for echoes. In M. Rosenblatt, editor, *Proc. Symp. on Time Series Analysis*, pages 209–243. Wiley, New York, 1963.
- [9] G. E. P. Box and G. C. Tiao. *Bayesian Inference in Statistical Analysis*. Addison-Wesley, 1973.
- [10] K. Brandenburg and G. Stoll. ISO-MPEG-1 audio: a generic standard for coding of high-quality digital audio. *JAES*, 4 2(10):761–792, 1994.
- [11] A. S. Bregman. *Auditory Scene Analysis*. MIT Press, 1990.

- [12] G. L. Bretthorst. *Bayesian Spectrum Analysis and Parameter Estimation*. Springer-Verlag, 1989.
- [13] J. C. Brown. Computer identification of musical instruments using pattern recognition with cepstral coefficients. *Journal of the Acoustic Society of America*, 105(3):1933–1941, March 1999.
- [14] E. M. Burns. Intervals, scales and tuning. In *The Psychology of Music* [28], chapter 7.
- [15] E. Cambouropoulos. A formal theory for the discovery of local boundaries in a melodic surface. In *Proceedings of the III Journées d' Informatique Musicale*, Caen, France, 1996.
- [16] E. Cambouropoulos, A. Smaill, and G. Widmer. A clustering algorithm for melodic analysis. In *Diderot Forum on Mathematics and Music*, Vienna, December 1999.
- [17] Bradley P. Carlin and Siddhartha Chib. Bayesian model choice via Markov chain Monte Carlo. *Journal of the Royal Statistical Society, Series B*, 57:473–484, 1995.
- [18] C. Chafe and D. Jaffe. Source separation and note identification in polyphonic music. In *Proc. ICASSP*, volume 2, pages 1289–1292, Tokyo, 1986.
- [19] M. H. Chen. Markov chain Monte Carlo sampling for evaluating multidimensional integrals with application to Bayesian computation. Available as Postscript<sup>1</sup>, 1996.
- [20] I.J. Clarke and G. Spence. Detection and tracking of multi-periodic signals. In *Proc. EUSIPCO*, 1998.
- [21] L. Cohen. Time-frequency distributions. *Proc. IEEE*, 77:941–981, 1989.
- [22] P. R. Cook. Toward physically-informed parametric synthesis of sound effects. In *Proc. IEEE Workshop on Audio and Acoustics*, October 1999.
- [23] R. A. Cook. Bayesian detection and estimation of Gaussian peaks via reversible jump MCMC. First year report. Cambridge University Engineering Department, 1999.
- [24] I. Cross. AI and music perception. In *Artificial Intelligence and the Simulation of Behaviour (AISB) Quarterly*, Edinburgh, UK, April 1999. Available on-line<sup>2</sup>.
- [25] Alain de Cheveigné. Separation of concurrent harmonic sounds: frequency estimation and a time domain cancellation model of auditory processing. *Journal of the Acoustic Society of America*, 93(6):3271–3290, June 1993.

---

<sup>1</sup><http://www.wpi.edu/~mhchen/asaproc96.ps>

<sup>2</sup>[http://www-ext.mus.cam.ac.uk/Music\\_Info/bios96/crossarts/AISB/IRMCAISB.html](http://www-ext.mus.cam.ac.uk/Music_Info/bios96/crossarts/AISB/IRMCAISB.html)

- [26] Ph. Depalle and T. Hélie. Extraction of spectral peak parameters using short-time Fourier transform modelling and no sidelobe windows. In *Proc. IEEE Workshop on Audio and Acoustics*, page 7.1. Proc. IEEE, October 1997.
- [27] D. Deutsch. Grouping mechanisms in music. In *The Psychology of Music* [28], chapter 9.
- [28] D. Deutsch. *The Psychology of Music*. Academic Press, second edition, 1999.
- [29] T. Ding and X. Qian. Processing of musical tones using a combined quadratic polynomial-phase sinusoid and residual (QUASAR) signal model. *J. Audio Eng. Soc.*, 45(7/8), July/August 1997.
- [30] P. M. Djurić, S. J. Godsill, W. J. Fitzgerald, and P. J. W. Rayner. Detection and estimation of signals by reversible jump Markov chain Monte Carlo computations. In *Proc. ICASSP*, 1998.
- [31] M. Dörfler and H. G. Feichtinger. Quantitative description of expression in performance of music, using Gabor representations. In *Diderot Forum on Mathematics and Music*, Vienna, December 1999.
- [32] A. Doucet and C. Andrieu. Robust Bayesian spectral analysis via MCMC sampling. In *Proc. EUSIPCO*, 1998.
- [33] B. Doval and X. Rodet. Fundamental frequency estimation and tracking using maximum likelihood harmonic matching and HMMs. In *Proc. ICASSP*, volume 1, pages 221–224, April 1993.
- [34] D. P. W. Ellis. *Prediction-driven computational auditory scene analysis*. PhD thesis, Massachusetts Institute of Technology, 1996. Available on-line<sup>3</sup>.
- [35] D. P. W. Ellis. Prediction-driven computational auditory scene analysis for dense sound mixtures. In *ESCA workshop on the Auditory Basis of Speech Perception*, Keele, UK, July 1996.
- [36] D. P. W. Ellis and B. L. Vercoe. A perceptual representation of sound for auditory signal separation. In *123rd meeting of the Acoustical Society of America*, Salt Lake City, May 1992.
- [37] G. Evangelista. Wavelets that we can play: an update. In *Diderot Forum on Mathematics and Music*, 1999.

---

<sup>3</sup><http://dpwe.www.media.mit.edu/~dpwe/pdcasa/>

- [38] G. Fant. *Acoustic Theory of Speech Production*. Mouton, 1970.
- [39] P. Fernández-Cid and F. J. Casajús-Quirós. Multi-pitch estimation for polyphonic musical signals. In *Proc. ICASSP*. Proc. IEEE, 1998. #1402.
- [40] J. L. Flanagan and R. M. Golden. Phase vocoder. *Bell Syst. Tech. J.*, 45:1493–1509, 1966.
- [41] H. Fletcher. Auditory patterns. *Rev. Mod. Phys.*, 12:47–65, 1940.
- [42] N. H. Fletcher and T. D. Rossing. *The Physics of Musical Instruments*. Springer, second edition, 1998.
- [43] W. Fucks. Mathematical analysis of formal structure of music. *IRE Trans. on Information Theory*, 1962.
- [44] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.
- [45] E. I. George and R. E. McCulloch. Stochastic search variable selection. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors, *Markov chain Monte Carlo in Practice*. Chapman and Hall, 1996.
- [46] W. R. Gilks. Full conditional distributions. In Gilks et al. [47], chapter 5.
- [47] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors. *Markov chain Monte Carlo in practice*. Chapman and Hall, 1996.
- [48] W. R. Gilks and G. O. Roberts. Strategies for improving MCMC. In Gilks et al. [47], chapter 6.
- [49] S. J. Godsill. Recursive restoration of pitch variation defects in musical recordings. In *Proc. ICASSP*, volume 2, pages 233–236, Adelaide, April 1994.
- [50] S. J. Godsill. Robust modelling of noisy ARMA signals. In *Proc. ICASSP*, April 1997.
- [51] S. J. Godsill. On the relationship between MCMC model uncertainty methods. *J. Comp. Graph. Stats.*, (to appear), 2000.

- [52] S. J. Godsill and P. J. W. Rayner. Robust reconstruction and analysis of autoregressive signals in impulsive noise using the Gibbs sampler. Technical Report CUED/F-INFENG/TR.233, Cambridge University Engineering Department, Cambridge, England, November 1995. Available as Postscript<sup>4</sup>.
- [53] S. J. Godsill, P. J. W. Rayner, and O. Cappé. Digital audio restoration. In K. Brandenburg and M. Kahrs, editors, *Applications of Digital Signal Processing to Audio and Acoustics*. Kluwer Academic Publishers, 1996. *(To appear)*.
- [54] S.J. Godsill and P.J.W. Rayner. *Digital Audio Restoration*. Springer-Verlag, September 1998.
- [55] J. L. Goldstein. An optimum processor theory for the central formation of the pitch of complex tones. *Journal of the Acoustic Society of America*, 54(6):1496–1516, 1973.
- [56] M. Goodwin. Residual modelling in music analysis-synthesis. In *Proc. ICASSP*, Atlanta, May 1996.
- [57] M. Goodwin. *Adaptive signal models: theory, algorithms and audio applications*. Kluwer Academic Publishers, 1998.
- [58] P. J. Green. Reversible jump Markov-chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.
- [59] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.
- [60] W. Hess. *Pitch determination of speech signals*. Springer, 1983.
- [61] J. R. Hopgood and P. J. W. Rayner. Bayesian single channel blind deconvolution using parametric signal and channel models. In *Proc. IEEE Workshop on Audio and Acoustics*, pages 151–154, 1999.
- [62] R. L. Horton. *The General Linear Model*. McGraw-Hill, 1978.
- [63] A. J. M. Houtsma. Pitch perception. In B. C. J. Moore, editor, *Hearing*, chapter 8. Academic Press, 1995.
- [64] M. Imai and S. Inokuchi. Frequency identification by complex spectrum. In *Proc. ICASSP*, volume 1, pages 3.10.1–4, 1986.

---

<sup>4</sup><http://www-sigproc.eng.cam.ac.uk/~sjg/papers/95/impulse.ps>

- [65] E. T. Jaynes. Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, SSC-4:227–241, September 1968.
- [66] E. T. Jaynes. Marginalization and prior probabilities. In R. D. Rosenkrantz, editor, *Papers on Probability, Statistics and Statistical Physics*, chapter 12. Kluwer Academic, 1982.
- [67] H. Jeffreys. *Theory of Probability*. Oxford University Press, 1939.
- [68] ISO/IEC JTC1/SC29/WG11. Overview of the MPEG-4 standard, October 1999. N2995. Available on-line<sup>5</sup>.
- [69] M. Kahrs and K. Brandenburg, editors. *Applications of Digital Signal Processing to Audio and Acoustics*. Kluwer Academic, 1998.
- [70] M. Karjalainen and T. Tolonen. Multi-pitch and periodicity analysis model for sound separation and auditory scene analysis. In *Proc. ICASSP*, Phoenix, May 1999.
- [71] K. Kashino and H. Murase. Music recognition using note transition context. In *Proc. ICASSP*. Proc. IEEE, 1998. #2234.
- [72] W. Kausel. Computer optimization of brass wind instruments. In *Diderot Forum on Mathematics and Music*, pages 227–242, 1999.
- [73] A. Klapuri. Number theoretical means of resolving a mixture of several harmonic sounds. In *Proc. EUSIPCO*, 1998.
- [74] A. Klapuri. Pitch estimation using multiple independent time-frequency windows. In *Proc. IEEE Workshop on Audio and Acoustics*, October 1999.
- [75] A. Kokaram. *Motion Picture Restoration*. Springer-Verlag, 1998.
- [76] L. T. Lalli. Mathematics and musics: a mathematical model for bronze musical instruments. In *Diderot Forum on Mathematics and Music*, pages 325–333, Vienna, December 1999.
- [77] F. Lerdahl and R. Jackendoff. *A generative theory of tonal music*. MIT Press, 1983.
- [78] S. N. Levine, T. S. Verma, and J. O. Smith III. Alias-free multiresolution sinusoidal modelling for polyphonic, wideband audio. In *Proc. IEEE Workshop on Audio and Acoustics*, page 7.3. Proc. IEEE, October 1997.

---

<sup>5</sup><http://www.cselt.it/mpeg/standards/mpeg-4/mpeg-4.htm>

- [79] G. Li and L. Qiu. Speech analysis and synthesis using instantaneous amplitudes. In *Proc. EUSIPCO*, 1998.
- [80] H.-T. Li and P. Djurić. An iterative procedure for joint Bayesian spectrum and parameter estimation of harmonic signals. In *IEEE International Symposium on Circuits and Systems*, volume 2, pages 513–516, Atlanta, May 1996. Proc. IEEE.
- [81] M. D. Macleod. Fast nearly-ML estimation of the parameters of real or complex single tones or resolved multiple tones. *IEEE Trans. Acoustics, Speech and Signal Processing*, 46(1):141–148, January 1998.
- [82] M. D. Macleod. High resolution nearly-ML estimation of sinusoids in noise using a fast frequency domain approach. In *Proc. EUSIPCO*, 1998.
- [83] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 1998.
- [84] J. D. Markel and A. H. Gray. *Linear Prediction of Speech*. Springer-Verlag, 1976.
- [85] K. D. Martin. A blackboard system for automatic transcription of simple polyphonic music. Technical Report No. 385, MIT Media Lab. perceptual computing section, July 1996. Available on-line<sup>6</sup>.
- [86] K. D. Martin. Toward automatic sound source recognition: Identifying musical instruments. Presented at the NATO Computational Hearing Advanced Study Institute, Il Ciocco, Italy, 1-12 July 1998. Available on-line<sup>7</sup>.
- [87] K. D. Martin and Y. E. Kim. Musical instrument identification: a pattern recognition approach. In *Presented at the 136th meeting of the Acoustical Society of America*, October 1998.
- [88] D. C. Massie. Wavetable sampling synthesis. In Kahrs and Brandenburg [69], chapter 8.
- [89] S. McAdams. Segregation of concurrent sounds I: Effects of frequency modulation coherence. *Journal of the Acoustic Society of America*, 86(6):2148–2159, December 1989.
- [90] R. J. McAulay and T. F. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Trans. Acoustics, Speech and Signal Processing*, ASSP-34(4):744–754, 1986.

---

<sup>6</sup><ftp://sound.media.mit.edu/pub/Papers/kdm-TR385.ps.gz>

<sup>7</sup><ftp://sound.media.mit.edu/pub/Papers/kdm-comhear98.pdf>

- [91] R. J. McAulay and T. F. Quatieri. Pitch estimation and voicing detection based on a sinusoidal speech model. In *Proc. ICASSP*, volume 1, pages 249–252, 1990.
- [92] C. M. McIntyre and D. A. Dermott. A new fine-frequency estimation algorithm based on a parabolic regression. In *Proc. ICASSP*, volume 2, pages 541–544, 1992.
- [93] R. Meddis and M. J. Hewitt. Virtual pitch and phase sensitivity of a computer model of the auditory periphery. i: Pitch identification. *Journal of the Acoustic Society of America*, 89(6):2866–2882, 1991.
- [94] D. K. Mellinger. *Event Formation and Separation in Musical Sound*. PhD thesis, Center for Computer Research in Music and Acoustics, Stanford University, 1991.
- [95] M. Mellody and G. H. Wakefield. The time-frequency characteristics of violin vibrato: modal distribution analysis and synthesis. *Journal of the Acoustic Society of America*, 107(1):598–611, January 2000.
- [96] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1091, 1953.
- [97] B. C. J. Moore. *An Introduction to the Psychology of Hearing*. Academic Press, fourth edition, 1997.
- [98] B. C. J. Moore and B. R. Glasberg. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustic Society of America*, 74(3):750–753, 1983.
- [99] B. C. J. Moore, B. R. Glasberg, and R. W. Peters. Relative dominance of individual partials in determining the pitch of complex tones. *Journal of the Acoustic Society of America*, 77(5):1853–1860, May 1985.
- [100] B. C. J. Moore, R. W. Peters, and B. R. Glasberg. Thresholds for the detection of inharmonicity in complex tones. *Journal of the Acoustic Society of America*, 77(5):1861–1867, May 1985.
- [101] T. Nakatani, T. Kawabata, and H. G. Okuno. A computational model of sound stream segregation with multi-agent paradigm. In *Proc. ICASSP*, volume 4, pages 2671–2674, 1995.

- [102] R. M. Neal. Probabilistic inference using MCMC methods. Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto, Canada, September 1993. Available as Postscript<sup>8</sup>.
- [103] D. E. Newland. Harmonic wavelet analysis. *Proc. Royal Society of London*, 443:203–225, 1993.
- [104] D. E. Newland. Harmonic and musical wavelets. *Proc. Royal Society of London*, 444:605–620, 1994.
- [105] T. Niihara and S. Inokuchi. Transcription of sung song. In *Proc. ICASSP*, volume 2, Tokyo, 1986.
- [106] T. W. Parsons. Separation of speech from interfering speech by means of harmonic selection. *Journal of the Acoustic Society of America*, 60(4), March 1976.
- [107] R. D. Patterson and J. Holdsworth. A functional model of neural activity patterns and auditory images. *Advances in Speech, Hearing and Language Processing*, 3(B):547–563, 1996.
- [108] J. R. Pierce. The nature of musical sound. In *The Psychology of Music* [28].
- [109] R. Plomp. The ear as a frequency analyzer. *Journal of the Acoustic Society of America*, 36:1628–1636, 1964.
- [110] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C, Second Edition*. CUP, 1992.
- [111] O. Punskeya, C. Andrieu, A. Doucet, and W. J. Fitzgerald. Bayesian segmentation of piecewise constant autoregressive processes using MCMC. Technical Report CUED/F-INFENG/TR 344, Cambridge University Engineering Department, 1999.
- [112] H. Purnhagen. Advances in parametric audio coding. In *Proc. IEEE Workshop on Audio and Acoustics*, October 1999.
- [113] T. F. Quatieri and R. J. McAulay. Speech transformations based on a sinusoidal representation. *IEEE Trans. Acoustics, Speech and Signal Processing*, ASSP-34(6):1449–1463, 1986.
- [114] J. J. Rajan. *Time Series Classification*. PhD thesis, University of Cambridge, 1994.

---

<sup>8</sup><ftp://ftp.cs.utoronto.ca/pub/radford/review.ps.Z>

- [115] J. J. Rajan, P. J. W. Rayner, and S. J. Godsill. A Bayesian approach to parameter estimation and interpolation of time-varying autoregressive processes using the Gibbs sampler. *To be published*, 1996.
- [116] C. Raphael. Synthesizing musical accompaniments with Bayesian belief networks. In *Diderot Forum on Mathematics and Music*, Vienna, December 1999.
- [117] R. Rasch and R. Plomp. The perception of musical tones. In *The Psychology of Music* [28], chapter 4.
- [118] R. A. Rasch. The perception of simultaneous notes such as in polyphonic music. *Acustica*, 40, 1978.
- [119] R. A. Rasch. Synchronization in performed ensemble music. *Acustica*, 43, 1979.
- [120] S. Richardson and P. J. Green. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, 59(4):731–792, 1997. Available as Postscript<sup>9</sup>.
- [121] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.
- [122] J.-C. Risset and D. L. Wessel. Exploration of timbre by analysis and synthesis. In *The Psychology of Music* [28], chapter 5.
- [123] G. O. Roberts. Markov chain concepts related to sampling algorithms. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors, *Markov chain Monte Carlo in practice*, pages 45–57. Chapman and Hall, 1996.
- [124] X. Rodet. Musical sound signal analysis/synthesis: Sinusoidal + residual and elementary waveform models. *Applied Signal Processing*, Summer 1998.
- [125] L. Rossi, G. Girolami, and M. Leca. Identification of polyphonic piano signals. *Acustica*, 83:1077–1084, 1997.
- [126] J. J. K. Ó Ruanaidh and W. J. Fitzgerald. *Numerical Bayesian Methods Applied to Signal Processing*. Springer-Verlag, 1996.
- [127] E. D. Scheirer. Music perception systems. Technical report, MIT Media Laboratory, October 1998. Available on-line<sup>10</sup>.

---

<sup>9</sup><http://www.stats.bris.ac.uk/pub/reports/MCMC/mix.ps.gz>

<sup>10</sup><http://sound.media.mit.edu/~eds/papers/phdprop/>

- [128] B. Schmeiser and M.-H. Chen. General hit-and-run Monte Carlo sampling for evaluating multidimensional integrals. Technical report, School of Industrial Engineering, Purdue University, 1991.
- [129] M. R. Schröder. Period histogram and product spectrum: new methods for fundamental frequency measurement. *Journal of the Acoustic Society of America*, 43:829–834, 1968.
- [130] A. Schuster. The periodogram and its optical analogy. *Proc. R. Soc. Lond.*, 77:136, 1905.
- [131] G. Schwartz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461, 1978.
- [132] X. Serra. *A System for Sound Analysis/Transformation/Synthesis Based on a Deterministic Plus Stochastic Decomposition*. PhD thesis, Stanford University, 1990.
- [133] X. Serra. Musical sound modeling with sinusoids plus noise. In C. Roads, S. Pope, A. Picialli, and G. De Poli, editors, *Musical Signal Processing*. Swets and Zeitlinger, 1997. Available as Postscript<sup>11</sup>.
- [134] R. N. Shepard. Circularity of judgments of relative pitch. *Journal of the Acoustic Society of America*, 36:2345–2353, 1964.
- [135] M. Slaney, D. Naar, and R. F. Lyon. Auditory model inversion for sound separation. In *Proc. ICASSP*, volume 2, pages 77–80, 1994.
- [136] P. Smith and V. Hardman. Fine-grained scalable sound representations for collaborative composition and performance. In *Audio and music technology: the challenge of creative DSP*. IEE colloquium, 1998.
- [137] J. O. Smith III. Principles of digital waveguide models of musical instruments. In Kahrs and Brandenburg [69].
- [138] D. Spiegelhalter, A. Thomas, N. Best, and W. Gilks. *BUGS — Bayesian inference Using Gibbs Sampling*. MRC Biostatistics Unit, Cambridge, August 1996. Available on-line<sup>12</sup>.
- [139] D. Spiegelhalter, A. Thomas, N. Best, and W. Gilks. Hepatitis B: a case study in MCMC methods. In Gilks et al. [47].
- [140] A. Sterian, M. H. Simoni, and G. H. Wakefield. Model-based musical transcription. In *Proc. Int. Computer Music Conf.*, Beijing, China, 1999. Available on-line<sup>13</sup>.

<sup>11</sup><http://www.iaa.upf.es/~xserra/articles/msm/>

<sup>12</sup><ftp://ftp.mrc-bsu.cam.ac.uk>

<sup>13</sup><http://musen.engin.umich.edu/papers/transcription.pdf>

- [141] A. Sterian and G. H. Wakefield. Robust automatic music transcription systems. In *Proc. Int. Computer Music Conf.*, Hong Kong, 1996.
- [142] A. Sterian and G. H. Wakefield. A frequency-dependent bilinear time-frequency distribution for improved event detection. In *Proc. Int. Computer Music Conf.*, Thessaloniki, Greece, 1997.
- [143] S. S. Stevens, J. Volkman, and E. B. Newman. A scale for the measurement of the psychological magnitude of pitch. *Journal of the Acoustic Society of America*, 8:185–190, 1937.
- [144] D. T. Teaney, V. L. Moruzzi, and F. C. Mintzer. The tempered Fourier transform. *Journal of the Acoustic Society of America*, 67(6):2063–2067, 1980.
- [145] E. Terhardt. Fourier transformation of time signals: Conceptual revision. *Acustica*, 57:243–256, 1985.
- [146] E. Terhardt, G. Stoll, and M. Seewann. Algorithms for extraction of pitch and pitch salience from complex tonal signals. *Journal of the Acoustic Society of America*, 71(3):679–688, March 1982.
- [147] E. Terhardt, G. Stoll, and M. Seewann. Pitch of complex signals according to virtual-pitch theory: tests, examples and predictions. *Journal of the Acoustic Society of America*, 71(3):671–678, March 1982.
- [148] C. W. Therrien. *Discrete Random Signals and Statistical Signal Processing*. Prentice-Hall, 1992.
- [149] A. Thomas, D.J. Spiegelhalter, and W.R. Gilks. BUGS: A program to perform Bayesian inference using Gibbs sampling. In J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, editors, *Bayesian Statistics 4*, pages 837–842. Oxford University Press, 1992.
- [150] L. Tierney. Markov chains for exploring posterior distributions (with discussion). *The Annals of Statistics*, 22:1701–1762, 1994. Available as Postscript<sup>14</sup>.
- [151] L. Tierney. Introduction to general state-space Markov chain theory. In Gilks et al. [47], chapter 4.

---

<sup>14</sup><http://www.stat.umn.edu/PAPERS/tech-reports/tr560.ps>

- [152] P. T. Troughton. *Simulation methods for linear and nonlinear time series models with application to distorted audio signals*. PhD thesis, Cambridge University Engineering Department, June 1999.
- [153] P. T. Troughton and S. J. Godsill. A reversible jump sampler for autoregressive time series, employing full conditionals to achieve efficient model space moves. Technical report, Cambridge University Engineering Department, 1997. Technical Report CUED/F-INFENG/TR.304. Available as Postscript<sup>15</sup>.
- [154] L. van Noorden and D. Moelants. Resonance in the perception of musical pulse. *Journal of New Music Research*, 28(1):43–66, 1999.
- [155] S. V. Vaseghi. *Algorithms for Restoration of Archived Gramophone Recordings*. PhD thesis, University of Cambridge, 1988.
- [156] J. Vermaak, M. Niranjana, and S. J. Godsill. Markov-chain Monte-Carlo estimation for the seasonal autoregressive process with application to pitch modelling. Technical Report CUED/F-INFENG/TR312, Cambridge University Engineering Department, March 1998.
- [157] G. von Békésy. The variations of phase along the basilar membrane with sinusoidal vibrations. *Journal of the Acoustic Society of America*, 19:452–460, 1947.
- [158] J. von Neumann. Various techniques used in connection with random digits. *National Bureau of Statistics, Applied Maths Series*, 12(3):36–38, 1951.
- [159] P. J. Walmsley, S. J. Godsill, and P. J. W. Rayner. Multidimensional optimisation of harmonic signals. In *Proc. EUSIPCO*, volume IV, pages 2033–2036, 1998. Available on-line<sup>16</sup>.
- [160] P. J. Walmsley, S. J. Godsill, and P. J. W. Rayner. Bayesian graphical models for polyphonic pitch tracking. In *Diderot Forum*, Vienna, December 1999. Available on-line<sup>17</sup>.
- [161] P. J. Walmsley, S. J. Godsill, and P. J. W. Rayner. Bayesian modelling of harmonic signals for polyphonic music tracking. In *Cambridge Music Processing Colloquium*, Cambridge, UK, September 30 1999. Available on-line<sup>18</sup>.

---

<sup>15</sup><http://www-sigproc.eng.cam.ac.uk/~ptt10/papers/>

<sup>16</sup><http://www-sigproc.eng.cam.ac.uk/~pjw42/ftp/eus98ps.zip>

<sup>17</sup><http://www-sigproc.eng.cam.ac.uk/~pjw42/ftp/didlt.zip>

<sup>18</sup><http://www-sigproc.eng.cam.ac.uk/~pjw42/ftp/camcolloq.zip>

- 
- [162] P. J. Walmsley, S. J. Godsill, and P. J. W. Rayner. Polyphonic pitch tracking using joint Bayesian estimation of multiple frame parameters. In *Proc. IEEE Workshop on Audio and Acoustics*, Mohonk, New York, October 1999.
- [163] M. Weintraub. A computational model for separating two simultaneous talkers. In *Proc. ICASSP*, volume 1, pages 3.1.1–4, 1986.
- [164] P. J. Wolfe and S. J. Godsill. Formalising perceptually motivated approaches to music restoration. In *Diderot Forum on Mathematics and Music*, Vienna, December 1999.
- [165] M.-D. Wu. *Markov chain Monte Carlo methods applied to Bayesian data analysis*. PhD thesis, Cambridge University Engineering Department, 1997.
- [166] A. Zellner. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques*, pages 233–243. Elsevier, Amsterdam, 1986.
- [167] E. Zwicker. Subdivision of the audible frequency range into critical bands (*Frequenzgruppen*). *Journal of the Acoustic Society of America*, 33(2):248, 1961.
- [168] E. Zwicker and H. Fastl. *Psychoacoustics: Facts and Models*. Springer, second edition, 1999.

# Block matrix inversion

# A

---

Many of the update moves for general linear models require a matrix inversion. When the number of basis functions is increased, this results in a large matrix inversion. However, using matrix partitioning techniques, this computational burden can be significantly reduced. Firstly the general case of inverting a partitioned matrix is considered, and then it is applied to the updating of GLM basis matrices.

## A.1 Inverse by partitioning

Given the matrix  $X$

$$X = \begin{bmatrix} P & Q \\ R & S \end{bmatrix} \quad (\text{A.1})$$

the inverse,  $X^{-1}$ , is to be calculated,

$$X^{-1} = \begin{bmatrix} \hat{P} & \hat{Q} \\ \hat{R} & \hat{S} \end{bmatrix} \quad (\text{A.2})$$

It is assumed that  $S$  is square and nonsingular (ie invertible). An efficient calculation scheme is described by Press *et al.* [110, §2.7],

$$A = S^{-1} \quad (\text{A.3})$$

$$B = AR \quad (\text{A.4})$$

$$\hat{P} = (P - Q(B))^{-1} \quad (\text{A.5})$$

$$\hat{Q} = -\hat{P}QA \quad (\text{A.6})$$

$$\hat{R} = -B\hat{P} \quad (\text{A.7})$$

$$\hat{S} = A - B\hat{Q} \quad (\text{A.8})$$

Also, the determinant can be found:

$$\begin{aligned} |X| &= |S| |P - QS^{-1}R| \\ &= \frac{|S|}{|\hat{P}|}. \end{aligned} \quad (\text{A.9})$$

## A.2 Use with basis matrices

To use with a partitioned basis matrix  $\mathbf{G} = [\mathbf{G}_v + \mathbf{G}_c]$ , split into variable and constant (known) components,  $(\mathbf{G}^t\mathbf{G})^{-1}$  is to be calculated,

$$X = \begin{bmatrix} \mathbf{G}_v^t\mathbf{G}_v & \mathbf{G}_v^t\mathbf{G}_c \\ \mathbf{G}_c^t\mathbf{G}_v & \mathbf{G}_c^t\mathbf{G}_c \end{bmatrix} = \begin{bmatrix} P & Q \\ R & S \end{bmatrix} \quad (\text{A.10})$$

$$X^{-1} = (\mathbf{G}^t\mathbf{G})^{-1} \quad (\text{A.11})$$

Calculating  $\hat{P}, \hat{Q}, \hat{R}, \hat{S}$  as above with  $A = \mathbf{G}_c^t\mathbf{G}_c$  (which is precalculated) and  $B = A(\mathbf{G}_c^t\mathbf{G}_v)$ .

$$\hat{\mathbf{b}} = (\mathbf{G}^t\mathbf{G})^{-1}\mathbf{G}^t\mathbf{d} \quad (\text{A.12})$$

$$\begin{bmatrix} \hat{\mathbf{b}}_v \\ \hat{\mathbf{b}}_c \end{bmatrix} = \begin{bmatrix} \hat{P}\mathbf{G}_v^t + \hat{Q}\mathbf{G}_c^t \\ \hat{R}\mathbf{G}_v^t + \hat{S}\mathbf{G}_c^t \end{bmatrix} \mathbf{d} \quad (\text{A.13})$$

$$\mathbf{f} = \mathbf{G}\hat{\mathbf{b}} \quad (\text{A.14})$$

$$= [\mathbf{G}_v\hat{\mathbf{b}}_v + \mathbf{G}_c\hat{\mathbf{b}}_c]. \quad (\text{A.15})$$

Also to compute  $|\mathbf{G}^t\mathbf{G}|$ , use (A.9) and the precalculated matrix  $\mathbf{G}_c^t\mathbf{G}_c$ ,

$$|\mathbf{G}^t\mathbf{G}| = \frac{|\mathbf{G}_c^t\mathbf{G}_c|}{|\hat{P}|}. \quad (\text{A.16})$$

# Marginalisation of amplitude parameters and error variance

# B

## B.1 Amplitude parameters

For the case of a model described by a single GLM, the linear amplitudes  $\mathbf{b}$  and error variance  $\sigma_e^2$  are to be marginalised. The adoption of a  $g$ -prior allows the linear amplitudes to be integrated out since this is a conjugate prior. It also has other advantages, as mentioned in §4.5.

$$\begin{aligned} p(\mathbf{d}|\phi, M, \sigma_e^2, \delta^2) &= \int_{\mathbb{R}^M} p(\mathbf{d}|\phi, M, \mathbf{b}, \sigma_e^2) p(\mathbf{b}|\phi, M, \sigma_e^2, \delta^2) d\mathbf{b} \\ &= (2\pi\sigma_e^2)^{-\frac{N}{2}} |2\pi\sigma_e^2\Sigma|^{-\frac{1}{2}} \int_{\mathbb{R}^M} \exp\left[-\frac{\|\mathbf{d} - \mathbf{G}\mathbf{b}\|^2 + \mathbf{b}^t\Sigma^{-1}\mathbf{b}}{2\sigma_e^2}\right] d\mathbf{b}. \end{aligned} \quad (\text{B.1})$$

The exponential term can be expanded in terms of a Gaussian

$$\begin{aligned} \|\mathbf{d} - \mathbf{G}\mathbf{b}\|^2 + \mathbf{b}^t\Sigma^{-1}\mathbf{b} &= \mathbf{d}^t\mathbf{d} + \mathbf{b}^t\mathbf{G}^t\mathbf{G}\mathbf{b} - 2\mathbf{d}^t\mathbf{G}\mathbf{b} + \mathbf{b}^t\Sigma^{-1}\mathbf{b} \\ &= \mathbf{d}^t\mathbf{d} + (\mathbf{b} - \mathbf{m})^t\mathbf{M}^{-1}(\mathbf{b} - \mathbf{m}) + \lambda \end{aligned} \quad (\text{B.2})$$

where

$$\mathbf{M}^{-1} = \mathbf{G}^t\mathbf{G} + \Sigma^{-1} = \mathbf{G}^t\mathbf{G} (1 + \delta^{-2}) \quad (\text{B.3})$$

$$\mathbf{m} = \mathbf{M}\mathbf{G}^t\mathbf{d} = \frac{\delta^2}{1 + \delta^2} (\mathbf{G}^t\mathbf{G})^{-1} \mathbf{G}^t\mathbf{d} \quad (\text{B.4})$$

$$\lambda = -\mathbf{m}^t\mathbf{M}^{-1}\mathbf{m} = -\mathbf{d}^t\mathbf{G}\mathbf{M}\mathbf{G}^t\mathbf{d} \quad (\text{B.5})$$

and the normalisation term of the Gaussian is  $|2\pi\sigma_e^2\mathbf{M}|^{\frac{1}{2}}$ . Proceeding with the integration,

$$\begin{aligned} p(\mathbf{d}|\tilde{\theta}, \sigma_e^2, \delta^2) &= (2\pi\sigma_e^2)^{-\frac{N}{2}} \frac{|2\pi\sigma_e^2\mathbf{M}|^{\frac{1}{2}}}{|2\pi\sigma_e^2\Sigma|^{\frac{1}{2}}} \exp\left[-\frac{\mathbf{d}^t\mathbf{d}+\lambda}{2\sigma_e^2}\right] \\ &= (2\pi\sigma_e^2)^{-\frac{N}{2}} (1+\delta^2)^{-\frac{M}{2}} \exp\left[-\frac{\mathbf{d}^t\mathbf{P}\mathbf{d}}{2\sigma_e^2}\right] \end{aligned} \quad (\text{B.6})$$

$$\mathbf{P} = \mathbf{I}_N - \mathbf{G}\mathbf{M}\mathbf{G}^t. \quad (\text{B.7})$$

The full conditional for  $\mathbf{b}$  is

$$p(\mathbf{b}|\mathbf{d}, \phi, M, \sigma_e^2, \delta^2) = \mathcal{N}(\mathbf{b}; \mathbf{m}, \sigma_e^2\mathbf{M}) \quad (\text{B.8})$$

which can be used to generate estimates for  $\mathbf{b}$  given the current values of the other parameters. The mean of the full conditional is  $\mathbf{m}$  and  $\lim_{\delta^2 \rightarrow \infty} \mathbf{m} = \hat{\mathbf{b}}_{\text{ls}}$  where  $\hat{\mathbf{b}}_{\text{ls}}$  is the least-squares estimate of  $\mathbf{b}$ .

## B.2 Error variance

From here it is also possible to marginalise the error variance if a conjugate prior is employed, for instance, an inverse gamma distribution,  $p(\sigma_e^2) \sim \text{IG}(\alpha_e, \beta_e)$ , of which the Jeffreys prior [67] is a special case. Jeffreys prior is a popular choice for scale variables such as error variance, since it is invariant to scale changes and is also a maximum entropy prior. However, it is improper and so lower and upper bounds would usually be imposed to make it proper. The inverse gamma distribution is a popular choice for scale parameters as it can often be marginalised easily, it is proper for  $\alpha_e > 0$ ,  $\beta_e > 0$ , and can be sampled from efficiently.

$$p(\mathbf{d}|\phi, M, \delta^2) = (1+\delta^2)^{-\frac{M}{2}} \int_0^\infty p(\sigma_e^2) (2\pi\sigma_e^2)^{-\frac{N}{2}} \exp\left[-\frac{\mathbf{d}^t\mathbf{P}\mathbf{d}}{2\sigma_e^2}\right] d\sigma_e^2 \quad (\text{B.9})$$

$$= (1+\delta^2)^{-\frac{M}{2}} \frac{(2\beta_e)^{\alpha_e} \Gamma(\varepsilon)}{\Gamma(\alpha_e) \pi^{\frac{N}{2}} [\mathbf{d}^t\mathbf{P}\mathbf{d} + 2\beta_e]^\varepsilon}, \quad \varepsilon = \frac{N}{2} + \alpha_e \quad (\text{B.10})$$

Since  $\{\alpha_e, \beta_e, N\}$  are constant for all models then this simplifies to

$$p(\mathbf{d}|\phi, M, \delta^2) \propto (1+\delta^2)^{-\frac{M}{2}} [\mathbf{d}^t\mathbf{P}\mathbf{d} + 2\beta_e]^{-\varepsilon} \quad (\text{B.11})$$

and hence we obtain the expression for the joint marginal posterior of the model parameters for the case of a single component:

$$p(\phi, M | \mathbf{d}, \delta^2) \propto p(\mathbf{d} | \phi, M, \delta^2) p(\phi, M) \quad (\text{B.12})$$

$$= (1 + \delta^2)^{-\frac{M}{2}} [\mathbf{d}^t \mathbf{P} \mathbf{d} + 2\beta_e]^{-\varepsilon} p(\phi, M). \quad (\text{B.13})$$

In this expression,  $\delta^2$ , which we have used as our expected signal to noise ratio, acts as a sensitivity control. The  $M/2$  term is a penalisation term which reduces the posterior as the model order  $M$  increases, and  $\mathbf{d}^t \mathbf{P} \mathbf{d}$  is the squared error. The exponent  $\varepsilon$  increases the weight of the likelihood against the prior as more data is available, and it is apparent that in order for the prior on  $\sigma_e^2$  to be uninformative,  $\alpha_e \ll N/2$  and  $\beta_e \ll \mathbf{d}^t \mathbf{P} \mathbf{d}/2$ .

With a high value of  $\delta^2$ , a model of high order must produce an appreciably lower error term to obtain a higher posterior probability. This is the mechanism by which Bayesian methods prevent overfitting by trading off model fit against model complexity. The posterior is also affected by the parameter priors  $p(\phi, M)$  which will be specific to the model under consideration. The full conditional for the error variance is

$$p(\sigma_e^2 | \mathbf{d}, \phi, M, \delta^2) = \text{IG}(\sigma_e^2; \frac{N}{2} + \alpha_e, \frac{\mathbf{d}^t \mathbf{P} \mathbf{d}}{2} + \beta_e) \quad (\text{B.14})$$

which for small  $\alpha_e, \beta_e$  has its mode at approximately  $\mathbf{d}^t \mathbf{P} \mathbf{d}/(N + 2)$ .

### B.3 Approximation of conditional residual to joint marginal posterior

In §5.3.1 it is shown how the full conditional for  $\tilde{\theta}^q$  can be approximated as a function of the residual. Here, some justification is given for the simplification. The full conditional of interest is  $p(\tilde{\theta}^q | \{\tilde{\theta}^q\}_{-q}, \mathcal{M}_{\mathbb{Q}}, \mathbf{d})$  which is proportional to the joint posterior  $p(\{\tilde{\theta}^q\}_{\mathbb{Q}}, \mathcal{M}_{\mathbb{Q}} | \mathbf{d})$  (all terms which aren't a function of  $\tilde{\theta}^q$  will cancel out in the Metropolis-Hastings acceptance function and can be ignored). The marginal conditional posterior for  $\tilde{\theta}^q$  can be written in a similar form to (5.10) but using the residual  $\mathbf{r}^q$  instead,

$$p(\tilde{\theta}^q | \mathbf{r}^q, \mathcal{M}_{\mathbb{Q}}) \propto (1 + \delta^2)^{-\frac{M^q}{2}} [\mathbf{r}^{q t} \mathbf{P} \mathbf{r}^q + 2\beta_e]^{-\varepsilon} p(\tilde{\theta}^q | \mathcal{M}_{\mathbb{Q}}) \quad (\text{B.15})$$

The key assumption is that the basis matrices of all components are approximately orthogonal with respect to one another, or  $\mathbf{G}^{q^t} \mathbf{G}^{q'} \approx \mathbf{0}_{[M^q \times M^{q}]}$ ,  $\forall q, q' \in \mathbb{Q}$ ,  $q \neq q'$ .<sup>1</sup> A second assumption is that the expected SNR,  $\delta^2$ , is large ( $\delta^2 \gg 1$ ). Expanding the expression  $\mathbf{r}^{q^t} \mathbf{P}^q \mathbf{r}^q$ ,

$$\mathbf{r}^{q^t} \mathbf{P}^q \mathbf{r}^q \approx \mathbf{r}^{q^t} \mathbf{r}^q - \frac{\delta^2}{1 + \delta^2} \mathbf{r}^{q^t} \mathbf{G}^q (\mathbf{G}^{q^t} \mathbf{G}^q)^{-1} \mathbf{G}^{q^t} \mathbf{r}^q \quad (\text{B.16})$$

and writing  $\mathbf{r}^q = \mathbf{d} - \mathbf{z}$ , where  $\mathbf{z} = \sum_{q' \in \mathbb{Q}, q' \neq q} \mathbf{G}^{q'} \mathbf{b}^{q'}$ , and hence  $\mathbf{G}^{q^t} \mathbf{z} \approx \mathbf{0}_{[M^q \times 1]}$

$$\mathbf{r}^{q^t} \mathbf{P}^q \mathbf{r}^q \approx \|\mathbf{d} - \mathbf{z}\|^2 - \frac{\delta^2}{1 + \delta^2} \mathbf{d}^t \mathbf{G}^q (\mathbf{G}^{q^t} \mathbf{G}^q)^{-1} \mathbf{G}^{q^t} \mathbf{d} \quad (\text{B.17})$$

Now expanding the  $\|\mathbf{d} - \mathbf{z}\|^2$  term and substituting for the least squares estimates  $\hat{\mathbf{b}}^q = (\mathbf{G}^{q^t} \mathbf{G}^q)^{-1} \mathbf{G}^{q^t} \mathbf{d}$ , to which the MAP estimates tend for large  $\delta^2$ ,

$$\|\mathbf{d} - \mathbf{z}\|^2 = \mathbf{d}^t \mathbf{d} + \mathbf{z}^t \mathbf{z} - 2 \mathbf{d}^t \mathbf{z} \quad (\text{B.18})$$

$$= \mathbf{d}^t \mathbf{d} + \left\| \sum_{\substack{q' \in \mathbb{Q} \\ q' \neq q}} \mathbf{G}^{q'} \mathbf{b}^{q'} \right\|^2 - 2 \sum_{\substack{q' \in \mathbb{Q} \\ q' \neq q}} \mathbf{d}^t \mathbf{G}^{q'} \mathbf{b}^{q'} \quad (\text{B.19})$$

$$= \mathbf{d}^t \mathbf{d} + \sum \left\| \mathbf{G}^{q'} \mathbf{b}^{q'} \right\|^2 - 2 \sum \mathbf{d}^t \mathbf{G}^{q'} \mathbf{b}^{q'} \quad (\text{B.20})$$

$$= \mathbf{d}^t \mathbf{d} - \sum \mathbf{d}^t \mathbf{G}^{q'} (\mathbf{G}^{q'^t} \mathbf{G}^{q'})^{-1} \mathbf{G}^{q'^t} \mathbf{d} \quad (\text{B.21})$$

For  $\delta^2 \gg 1$ ,  $\delta^2 / (1 + \delta^2) \approx 1$  and hence

$$\mathbf{r}^{q^t} \mathbf{P}^q \mathbf{r}^q \approx \mathbf{d}^t \mathbf{d} - \sum_{q \in \mathbb{Q}} \mathbf{d}^t \mathbf{G}^q (\mathbf{G}^{q^t} \mathbf{G}^q)^{-1} \mathbf{G}^{q^t} \mathbf{d}. \quad (\text{B.22})$$

If the orthogonality assumption is applied to the joint posterior, the composite matrix  $\mathbf{P}^c$  becomes block diagonal, and so (from (5.10)),

$$p(\{\tilde{\theta}^q\}_{\mathbb{Q}}, \mathcal{M}_{\mathbb{Q}} | \mathbf{d}) \propto (1 + \delta^2)^{-\frac{M}{2}} [\mathbf{d}^t \mathbf{P}^c \mathbf{d} + 2\beta_e]^{-\epsilon} p(\{\tilde{\theta}^q\}_{\mathbb{Q}} | \mathcal{M}_{\mathbb{Q}}) \quad (\text{B.23})$$

$$\mathbf{d}^t \mathbf{P}^c \mathbf{d} \approx \mathbf{d}^t \mathbf{d} - \sum_{q \in \mathbb{Q}} \mathbf{d}^t \mathbf{G}^q (\mathbf{G}^{q^t} \mathbf{G}^q)^{-1} \mathbf{G}^{q^t} \mathbf{d} \quad (\text{B.24})$$

therefore  $\mathbf{d}^t \mathbf{P}^c \mathbf{d} \approx \mathbf{r}^{q^t} \mathbf{P}^q \mathbf{r}^q$ . The prior independence of  $\{\tilde{\theta}^q\}$  means that  $p(\{\tilde{\theta}^q\}_{\mathbb{Q}} | \mathcal{M}_{\mathbb{Q}}) = \prod_{q \in \mathbb{Q}} p(\tilde{\theta}^q)$ , so cancelling terms which aren't a function of  $\tilde{\theta}^q$  leads to the result

$$p(\tilde{\theta}^q | \{\tilde{\theta}^q\}_{-q}, \mathcal{M}_{\mathbb{Q}}, \mathbf{d}) \approx p(\tilde{\theta}^q | \mathbf{r}^q, \mathcal{M}_{\mathbb{Q}}). \quad (\text{B.25})$$

<sup>1</sup>This requires that the frequency separation is greater than  $1/Ndt$  to ensure that the matrix  $\mathbf{G}^t \mathbf{G}$  is dominated by the leading diagonal.

This simplification allows more efficient transition kernels to be designed. If the basis matrices are not orthogonal with respect to each other, then this may still yield a useful result which may put support in the high probability regions of the full conditional for  $\tilde{\theta}^q$ . Proposals are still evaluated with respect to the joint posterior by the MH acceptance function, so if the orthogonality assumptions are not valid, then a lower acceptance rate would be expected but the stationary distribution of the Markov chain would still be the joint posterior.



# *Accompanying CD*

# C

---

A CD containing sound files of the monophonic and polyphonic examples in chapters 6 and 7 accompanies this thesis. The enhanced-mode CD contains a CD-ROM data track. The file `index.html` contains the example sounds and accompanying figures.

The track listing of the audio section of the CD is as follows

1. (data track – do not play)
2. sax (monophonic)
3. memory (monophonic)
4. flute (monophonic)
5. synharm (synthetic polyphonic)
6. saxmem (sax and memory superimposed)
7. piantun (synthesised piano tune)
8. granpre (acoustic piano tune)
9. wow (acoustic piano tune)