BAYESIAN SOURCE MODELLING FOR SINGLE-CHANNEL AUDIO SEPARATION

by

Onur Dikmen

B.S. in Computer Engineering, Boğaziçi University, 2000

M.S. in Computer Engineering, Boğaziçi University, 2002

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

Graduate Program in Computer Engineering

Boğaziçi University

2009

ACKNOWLEDGEMENTS

My gratitude for Prof. Lale Akarun is deep and unbounded. She helped me with every little detail of this thesis and did not discourage me when I come up with unorthodox ideas. I will be the happiest if I ever get a chance to supervise a student of mine like she did. A. Taylan Cemgil has been a friend under the cloak of a supervisor. Throughout this doubly-oriental voyage, we worked hard while having fun. I would very much love to be collaborating with these two people.

This thesis was supported by the Scientific and Technolological Research Council of Turkey (TUBITAK) under the project number 107E050. I would like to thank Prof. Ayşın Ertüzün for being the perfect project leader and for her comments about my work. I also would like to thank professors Müjdat Çetin, Bilge Günsel, Bülent Sankur and Fikret Gürgen for their comments and future directions. This thesis was also partially supported by The State Planning Organization of Turkey (DPT), under grant number DPT 07K120610.

This section will be incomplete if I do not name names of those who are involved in this thesis in some way. The inner circle: E. Itır Karaç, Özcan Vardar, and my family. My dear colleagues: A. Haydar Özer, Mehmet Gönen, İsmail Arı, Oya Aran, Atay Özgövde, Derya Çavdar, Furkan Kıraç, and Rabun Koşar. Cambridge crew: Stelios Karagiorgis, Nikolas Kantas, and Paris Kaimakis. Artistic breakfast club aka ÇOT!: Çetin Meriçli, Ahmet Yıldırım, Yunus Durmuş, and Tekin Meriçli. Those who were there in the last eighty minutes: Berk Gökberk, Neşe Alyüz, and Pınar Santemiz. Halflings: Y. Emre Kara and Gaye Genç. Gods: Peter Hammill and Ian Anderson. I hail Okan İrfanoğlu from overseas. And finally, a special “Hi!” to Berk Özen.

ABSTRACT

BAYESIAN SOURCE MODELLING FOR SINGLE-CHANNEL AUDIO SEPARATION

In many audio processing tasks, such as source separation, denoising or compression, it is crucial to construct realistic and flexible models to capture the physical properties of audio signals. This can be accomplished in the Bayesian framework through the use of appropriate prior distributions. In this thesis, we describe two prior models, Gamma Markov chains (GMCs) and Gamma Markov random fields (GMRFs) to model the sparsity and the local dependency of the energies of time-frequency expansion coefficients. We build two audio models where the variances of source coefficients are modelled with GMCs and GMRFs, and the source coefficients are Gaussian conditioned on the variances. The application area of these models are not limited to variance modelling of audio sources. They can be used in other problems where there is dependency between variables, such as the Poisson observation models. In single-channel source separation using non-negative matrix factorisation (NMF), we make use of GMCs to model the dependencies in frequency templates and excitation vectors.

A GMC model defines a prior distribution for the variance variables such that they are correlated along the time or frequency axis, while a GMRF model describes a non-normalised joint distribution in which each variance variable is dependent on all the adjoining variance variables. In our audio models, the actual source coefficients are independent conditional on the variances and distributed as zero-mean Gaussians. Our construction ensures a positive coupling between the variance variables, so that signal energy changes smoothly over both axes to capture the temporal and/or spectral continuity. The coupling strength is controlled by a set of hyperparameters.

Inference on the overall model, i.e., GMC or GMRF coupled with a Gaussian or

Poisson observation model, is convenient because of the conditional conjugacy of all of the variables in the model, but automatic optimisation of hyperparameters is crucial to obtain better fits. In GMCs, hyperparameter optimisation can be carried out using the Expectation-Maximisation (EM) algorithm, with the E-step approximated with the posterior distribution estimated by the inference algorithm. In this optimisation, it is important for the inference algorithm to estimate the covariances between the variables inferred, because the hyperparameters depend on them.

The marginal likelihood of the GMRF model is not available because of the intractable normalising constant. Thus, the hyperparameters of a GMRF cannot be optimised using maximum likelihood estimation. There are methods to estimate the optimal hyperparameters in these cases, such as pseudolikelihood, contrastive divergence and score matching. However, only contrastive divergence is readily applicable to models with latent variables. We optimised the hyperparameters of our GMRF-based audio model using contrastive divergence.

We tested our audio models that are based on GMC and GMRF models in denoising and single-channel source separation problems where all the hyperparameters are jointly estimated given only audio data. Both models provided promising results, but the reconstructed signals by the GMRF model were slightly better and more natural sounding.

Our third model makes use of Gamma and GMC prior distributions in an NMF setting for single-channel source separation. The hyperparameters are again optimised during the inference phase and the model needs almost no other design decisions. This model performs substantially better than the previous two models. In addition, it is less demanding in terms of computational power. However, it is designed only for source separation, i.e., it is not a general audio model as the previous two models.

ÖZET

SES SİNYALLERİNİN TEK KANALDAN AYRIŞTIRILMASINDA BAYESÇİ MODELLER

Kaynak ayrıştırma veya gürültü temizleme gibi ses işleme problemlerinde ses sinyallerinin fiziksel özelliklerini yansıtabilecek modellere ihtiyaç vardır. Bayesçi yaklaşımda, bu, gerçekçi önsel dağılımlar tanımlamayarak gerçekleştirilebilir. Biz, bu tezde, ses sinyallerinin zaman-frekans bölgesi gösterimlerindeki yerel ilintileri içerecek iki model geliştirdik: Gamma Markov zincirleri (GMZ) ve Gamma Markov rasgele alanları (GMRA). Önerdiğimiz ses modellerinde, zaman-frekans katsayılarının değişintileri bu yapılar kullanılarak birbirlerine bağlı olarak modellenirken, katsayılar bu değişintilere koşullu olarak, bağımsız Gauss dağılımlarından gelmektedir. GMZ ve GMRA modellerinin kullanım alanı, ses kaynaklarının değişintilerinin modellenmesiyle sınırlı değildir. Değişkenler arasında bağımlılık olan herhangi bir problemde, mesela Poisson serilerinde, de kullanılabilirler. Bunu göstermek için, negatif olmayan matris ayrıştırma (NOMA) kullanarak tek kanaldan kaynak ayrıştırma probleminde, frekans şablonları ve uyarma vektörlerindeki bağımlılığı modellemek için GMZ'leri kullandık.

GMZ'ler ile değişinti değişkenlerinin sadece zaman ya da frekans ekseni boyunca olan bağımlılıklarını modelleyebiliriz. GMRA'lar ise değişkenlerin tüm komşularına bağımlı olduğu düzgelenmemiş bir dağılım tanımladıkları için iki yöndeki bağımlılıkları da içerebilir. İki model de değişinti değişkenleri arasında pozitif ilinti olacak şekilde tanımlanmıştır. Böylece, sinyalin enerjisi hem zaman hem de frekans ekseni boyunca yavaşça değişmektedir. Değişkenler arasındaki ilintinin büyüklüğü ise modelin hiper parametreleri ile belirlenmektedir.

Bu modelleri kullanan hem Gauss hem de Poisson gözlem modellerinde, değişkenler koşullu eşlenik oldukları için kestirim kolay yapılabilmektedir. Ancak, daha başarılı

sonuçlar elde etmek için, hiper parametrelerin eniyilenmesi de gerekmektedir. GMZ'lerin hiper parametreleri beklenti-enbüyütme algoritmasıyla gerçekleştirilebilmektedir. Burada, olabilirlik, kestirim sırasında elde edilen istatistikler ile yakınsanmaktadır. Bu yüzden, kestirim metodunun değişkenler arasındaki ilintiyi de yakınsayabilmesi önem taşımaktadır.

GMRA'ların hiper parametreleri beklenti-enbüyütme ile yapılamamaktadır çünkü bu modellerde marjinal olabilirlik hesaplanamamaktadır. Sözde olabilirlik, kkarşıtlık ıraksayı, skor eşitleme gibi yöntemler, bu durumlarda eniyileme yapabilmeyi mümkün kılmaktadır. Bu yöntemlerden sadece karşıtlık ıraksayı gizli değişkenlerin olduğu durumlarda da çalışabilmektedir. Ses işleme uygulamalarında, GMRA'lar değişinti değişkenlerini, yani doğrudan gözlemlenemeyen değişkenleri modeller. Bu yüzden, GMRA'ların hiper parametrelerini karşıtlık ıraksayını kullanarak eniyiledik.

Bu tezde, GMZ ve GMRA temelli ses modellerimizi gürültü temizleme ve tek kanaldan kaynak ayrıştırma problemlerinde kullandık. Ayrıca bir öğrenme kümesine ihtiyaç duymadan, sadece gözlemlenen sinyalin varlığında, kestirim ve eniyileme içiçe gerçekleştirilerek tonal ve vurmalı ses kaynakları birbirlerinden ayrılmaktadır. Bu iki modelle, hem gürültü temizleme, hem de kaynak ayrıştırma problemlerinde başarılı sonuçlar elde ettik. GMRA'lara dayalı olan modelle geri çatılan sinyaller hem biraz daha başarılı, hem de daha doğaldır.

Önerdiğimiz üçüncü bir modelle de Gamma ve GMZ önsel dağılımları kullanarak, NOMA ile tek kanaldan kaynak ayrıştırma yaptık. Burada da hiper parametreler kestirim sırasında eniyilenmekte ve kullanıcının hemen hemen hiçbir kritik karar vermesine gerek kalmamaktadır. Bu modelle elde edilen sonuçlar önceki iki modelle elde edilenlerden daha başarılıdır. Ayrıca, bu modelde kestirim ve eniyileme daha hızlı bir şekilde yapılabilmektedir. Buna rağmen, bu model sadece kaynak ayrıştırma problemi için önerildiğinden, önceki iki model gibi genel uygulanabilirliği yoktur.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS/ABBREVIATIONS

a	Vector of hyperparameters
$s_{\nu,\tau}$	Source coefficient at frame τ and frequency bin ν
$\mathbf{s}$	Vector of source coefficients
$v_{\nu,\tau}$	Variance of the source coefficient at frame τ and frequency bin ν
$\mathbf{v}$	Vector of variance variables
$x_{\nu,\tau}$	Coefficient of the observed signal at frame τ and frequency bin ν
$\mathbf{x}$	Vector of observed coefficients
$\mathbf{z}$	Vector of auxiliary variables
Z_θ	Normalising constant
τ	Frame number
ν	Frequency bin number
BSS	Blind Source Separation
CD	Contrastive Divergence
EM	Expectation-Maximisation
FL	Fully-latent (model)
FO	Fully-observed (model)
GMC	Gamma Markov chain
GMRF	Gamma Markov random field
ICA	Independent Component Analysis
ISA	Independent Subspace Analysis
KL	Kullback-Leibler
MCMC	Markov chain Monte Carlo
MDCT	Modified Discrete Cosine Transform
ML	Maximum Likelihood
MMSE	Minimum Mean Square Error

MSE	Mean Squared Error
NMF	Non-negative Matrix Factorisation
pdf	Probability density function
PL	Pseudolikelihood
PO	Partially-observed (model)
SAR	Signal-to-Artefacts Ratio
SDR	Signal-to-Distortion Ratio
SIR	Signal-to-Interference Ratio
SIS/R	Sequential Importance Sampling / Resampling
SM	Score Matching
SMC	Sequential Monte Carlo
SNR	Signal-to-Noise Ratio
VB	Variational Bayes

1. INTRODUCTION

Blind Source Separation (BSS) [1, 2, 3] is the problem of estimating source signals from observed signals without using any information about the nature of the signals, mixing properties and the noise. The most famous instance of this problem is the cocktail party problem, in which people are talking simultaneously in a room and a person is trying to follow one of the conversations. BSS has applicability to many problems in a wide range of disciplines such as electroencephalography (EEG) [4, 5], magnetoencephalography (MEG) [6, 7, 8, 9, 10], functional Magnetic Resonance Imaging (fMRI) [11], seismic monitoring, surveillance, radar and acoustics [12, 13, 14, 15, 16, 17]. In this thesis, our interest is on audio source separation, which has important uses in hearing aids, cocktail party problem and denoising of recordings. In addition, it may be used as a preprocessing step for music transcription and enhancement.

1.1. The Source Separation Problem

In source separation, the goal is to extract the underlying sources from observations which are mixtures of these sources. When no information about the particular sources or mixing conditions is used, the problem is referred to as blind source separation. Mixing is generally assumed to be instantaneous, i.e., signals emanating from the sources instantly arrive at the sensors via a single path. In this case, observations are linear mixtures of the sources. When the number of sources is equal to the number of observations (even-determined case), the source separation problem boils down to estimating an unmixing matrix. The over-determined case, where there are more observations than the sources, can be solved using least-squares estimation. However, the under-determined case is ill-defined with more parameters than that can be uniquely estimated [2].

One of the earliest studies in the field is by Jutten and Herault [18]. They defined the concept of independent component analysis (ICA) and proposed an algorithm to solve the even-determined source separation problem with the assumption that the

sources are independent and non-Gaussian. The algorithm was based on maximising the non-Gaussianity of the sources. Comon [19] showed that minimising the mutual information was equivalent to maximising non-Gaussianity. Many fast and sophisticated methods were proposed based on the above approaches [20, 21, 22] and also information maximisation [23] and maximum likelihood estimation [24, 25], which were later shown to be equivalent [26, 27].

The above methods deal with the even-determined and over-determined cases, in which the problem is estimating the mixing matrix. Then, the sources can be reconstructed by a linear transformation using the (pseudo) inverse of this matrix. For the under-determined case, more information should be incorporated into the problem, such as the sparsity of the source coefficients in the transform domain. If only one source is assumed to be active at a particular time, the scatter plots of the observations contain lines that are determined by the columns of the mixing matrix. A thread of research has concentrated on clustering the observed coefficients in order to estimate the mixing matrix. Various methods are used to cluster the coefficients, such as fuzzy C-means [28], topographic maps [29, 30], modified k-means [31], and EM-based clustering [32, 33] to extract the line orientations. Once the mixing matrix is estimated, the source coefficients can be found by assigning the observations to the closest column of the mixing matrix [34, 35, 36] or solving a linear optimisation problem to find the most sparse sources that generate the observations [37, 38].

Another line of research is focused on source separation from one observation signal using non-negative matrix factorisation (NMF) [39]. NMF is a technique for decomposing a non-negative matrix into two non-negative components. In source separation domain, it is used to decompose the transform domain coefficients of a source into a compact set of bases and their excitations in time. It is successfully applied to the single-channel audio source separation problem with additional temporal continuity [15, 14, 16] and sparsity [14] constraints. It is also incorporated into the ICA [40] and sparse coding [41] frameworks.

The source separation problem can also be expressed in the Bayesian framework.

After defining prior distributions for the sources and other parameters and an observation model that describes the generation process of the observations, the problem becomes an inference problem. In the literature, various realistic source models were proposed. These models generally make use of the sparsity of the sources [42, 43, 44, 45, 12, 46, 47, 48, 49]. In audio source modelling, temporal and spectral continuity of the source coefficients are also incorporated into the models [48, 50, 51, 52, 53, 13, 54]. The Bayesian framework can also be used to define prior structures for ICA [55] and NMF [56, 57].

Below, we will give the generative model for the audio source separation problem and then we will explain some basic properties of audio signals. We will review ICA, NMF and the Bayesian paradigm, detailing how these properties can be made use of.

1.2. Generative Model

In audio source separation, what we have in hand is the observed signals recorded by m microphones, $x_i(t)$, where i denotes the microphone number and t is the time index in samples. If the duration of the signal is T samples, t takes values between 1 and T. The observations are thought to be a mixture of n source signals with additive noise, $\varepsilon_i(t)$:

$$x_i(t) = f_i(s_1(t), s_2(t), \ldots, s_n(t)) + \varepsilon_i(t) \tag{1.1}$$

where $f_i(\cdot)$ denotes the mixing function associated with the i^{th} microphone. The function mixes the source signals depending on their distances to the microphone, the reverberation conditions of the room and whether the sources or the microphone are in motion. $\varepsilon_i(t)$ can be seen as sensor noise which is added by the microphone. It is generally assumed to be white Gaussian. More structured, non-stationary noise may be modelled as another source. Equation 1.1 can be expressed in a more compact way

if we gather the row vectors representing the signals in matrices:

$$\mathbf{X} = f(\mathbf{S}) + \boldsymbol{\varepsilon} \tag{1.2}$$

where $\mathbf{X}$, $\mathbf{S}$ and $\boldsymbol{\varepsilon}$ are matrices containing the observed, source and noise signals, of sizes $m \times T$, $n \times T$ and $m \times T$, respectively. $f(\cdot)$ is the mixing function of the whole system, from domain $\mathbb{R}^{n \times T}$ to codomain $\mathbb{R}^{m \times T}$.

If we assume that the sources and the microphones are motionless and there is no reverberation in the room, it is possible to consider $f(\cdot)$ as a linear function. So, each observed signal becomes a linear combination of the source signals plus the noise

$$\mathbf{X} = \mathbf{AS} + \boldsymbol{\varepsilon} \tag{1.3}$$

where $\mathbf{A}$ is an $m \times n$ mixing matrix. This is a linear instantaneous model where observed samples at time t depend on the source samples at time t, but no other samples

$$x_i(t) = \sum_{j=1}^{n} A_{i,j} s_j(t) + \varepsilon_i(t) \tag{1.4}$$

1.3. Time Frequency Representations

It is also possible to define the source separation problem in linear time-frequency representations of signals. Such representations describe the behavior of the spectral content of a signal in time, thus provide a more efficient means to analyse real world signals with time varying spectra. Modified discrete cosine transform (MDCT) [58], short time Fourier transform (STFT) [59], Gabor transform [60] and wavelet transform [61] are popular examples of such linear time-frequency representations. In these representations, a time series $y(t)$ for $t = 1, 2, \ldots, T$ is represented as a linear combination of

basis functions, $\phi_{\alpha,t}$:

$$y(t) = \sum_{\alpha} \phi_{\alpha,t} \tilde{y}_{\alpha}, \tag{1.5}$$

where the time-frequency indices are denoted by α. In this notation, each time-frequency index is a tuple $\alpha = (\tau, \nu)$, where $\tau = 1 \dots N$ is a frame index and $\nu = 1 \dots W$ a frequency index. The expansion coefficients are denoted by $\tilde{y}_{\alpha}$.

Let us denote the dictionary of basis functions with $\mathbf{\Phi}$, a $W \times T$ matrix where W is the number of waveforms and T is the length of the signals. The source separation problem in Equation 1.3 can be written in the transform domain as

$$\mathbf{X}\mathbf{\Phi}^{\top} = \mathbf{A}\mathbf{S}\mathbf{\Phi}^{\top} + \boldsymbol{\varepsilon}\mathbf{\Phi}^{\top} \quad \equiv \quad \tilde{\mathbf{X}} = \mathbf{A}\tilde{\mathbf{S}} + \tilde{\boldsymbol{\varepsilon}} \tag{1.6}$$

The problems in time and transform domains are equivalent when the transformation is orthogonal, i.e. $\mathbf{\Phi}^{-1} = \mathbf{\Phi}^{\top}$. To see this, let us assume the original sources are $\hat{\mathbf{S}}$ and $\hat{\tilde{\mathbf{S}}}$ in time and transform domains, respectively. If the two problems are equivalent, $\hat{\tilde{\mathbf{S}}}\mathbf{\Phi}$ should also be equal to the sources in time domain. Similarly, $\hat{\mathbf{S}}\mathbf{\Phi}^{\top}$ should be equal to the time-frequency domain sources. This requires $\hat{\mathbf{S}} = \hat{\mathbf{S}}\mathbf{\Phi}^{\top}\mathbf{\Phi}$, which is true only when the transformation is orthogonal. Otherwise, when reconstructing signals from time-frequency estimates, an additional effect will be added. MDCT and orthogonal wavelets are examples of orthogonal transformations and we will be working on the source separation problem in MDCT domain.

1.4. Sparsity and Structure of Audio Sources

In source separation problems, the number of sources is generally greater than the number of observations. Even when they are equal, the system is underdetermined unless the mixing matrix and the noise parameters are known. That means many solutions that are compatible with the underlying model exist. In addition, some of these solutions may not be physically meaningful and lead to reconstructions with

artefacts. This problem may be overcome by imposing constraints onto the model such as the independence of the sources or incorporating prior knowledge about the variables in the model. Since time-frequency representations of audio signals include more structure than time domain coefficients, it is appropriate to define such constraints and prior knowledge upon them in this domain.

A typical property of time-frequency representations of natural audio signals is that the coefficients are sparse, i.e. only a small number of the coefficients have high magnitudes, the rest being close to zero [62, 63, 12]. In addition to this, the coefficients with high magnitudes are not independently distributed over the spectrogram but they form clusters. This means that adjacent coefficients in the time-frequency lattice have dependency among each other. More specifically, in audio signals with tonal components such as the recordings of musical instruments, there is high amount of dependency between the coefficients along the time axis at the harmonics of the fundamental frequency [64]. The onset of the musical notes may include transients or an adjustment period due to the properties of the instrument. Percussive sounds are composed mainly of transients and fast decaying components. The time-frequency representations of signals containing such transients have strong dependency along the frequency axis because of the simultaneous activation of a range of frequencies. Time-frequency coefficients of speech signals tend to have continuity along both axes, having clusters of high magnitude around formant frequencies [65]. Figure 1.1 presents spectrograms of some audio signals.

In order to ensure the sparsity of the source coefficients, solutions may be penalised according to their magnitudes. Similar constraints can be incorporated into the objective function to enforce continuity along time and frequency axes, e.g. by penalising high differences between adjacent coefficients. Defining the problem as a generative model with an appropriate prior distribution for the source coefficients is another way of solving this problem. However, these constraints and prior distributions should be adaptive because the dependencies between the coefficients are not of the same strength in both directions.

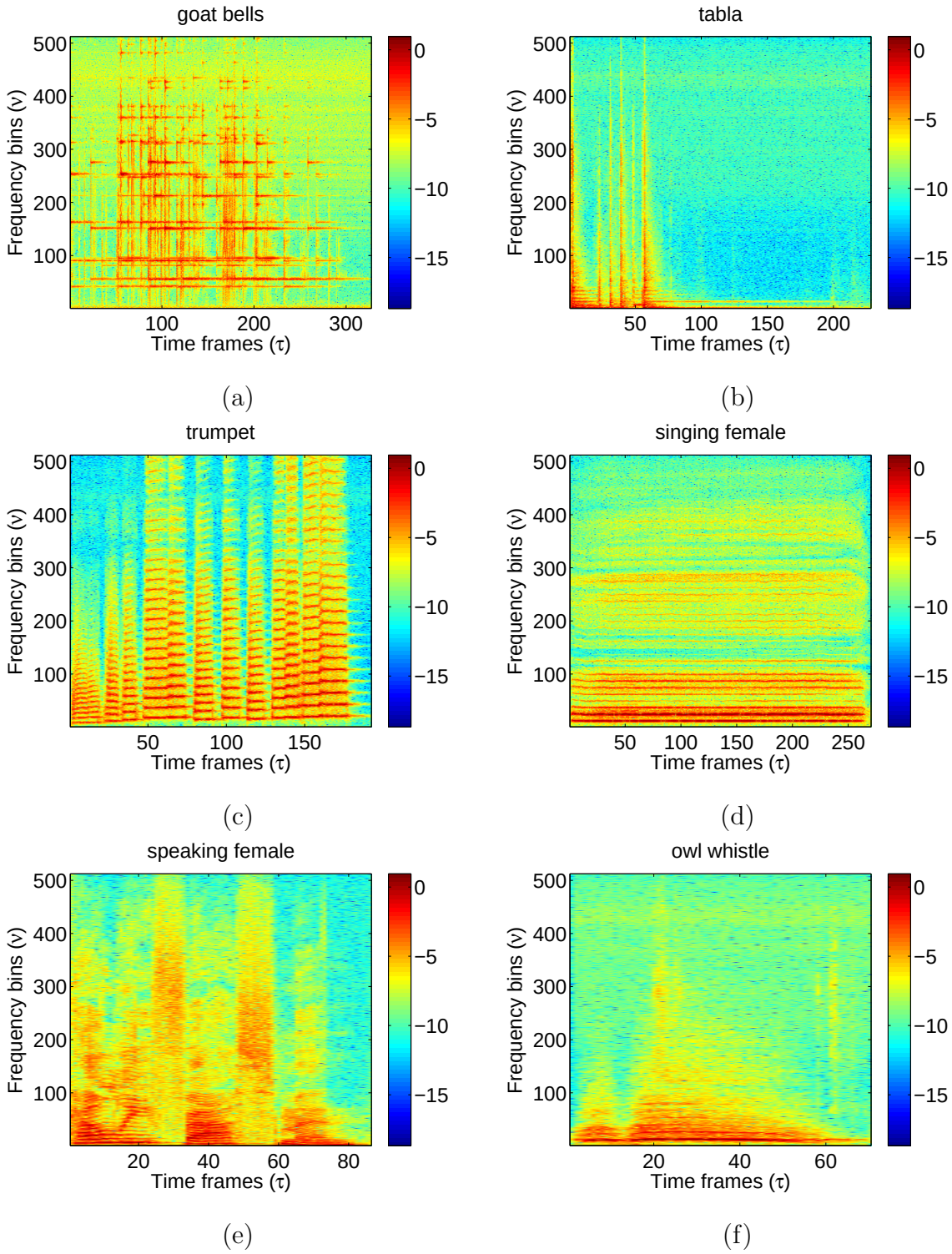


Figure 1.1. Spectrograms of some audio signals. (a) and (b) are examples of percussive sounds. Note that goat bells also have harmonic components. (c) and (d) are tonal signals. The trumpet plays several notes, while the singing contains only one note. (e) is an example of a speech signal. The owl whistle in (e) shares some characteristics with the speech signal, which also has harmonic components.

1.5. Independent Component Analysis

ICA is a widely used source separation approach. The research on ICA started in 1990s with very restricted models and built up to cover more realistic scenarios. Moreover, state-of-the-art ICA methods incorporate time and time-frequency structures in their models.

The generative model of the basic ICA [18, 19] is given by

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{1.7}$$

where $\mathbf{x} = [x_1\, x_2 \ldots x_n]^\top$ and $\mathbf{s} = [s_1\, s_2 \ldots s_n]^\top$ are the observation and source vectors, respectively. Here, each mixture, x_i, and source, s_i, are scalar random variables, that is, the values of signals $x_i(t)$ and $s_i(t)$ constitute samples of these random variables.

The main idea of ICA is to estimate the mixing matrix with the assumption that all sources are statistically independent. This assumption leads to the following factorisation of the joint probability density

$$p(\mathbf{s}) = \prod_{i=1}^{n} p(s_i). \tag{1.8}$$

In addition, all sources are required to be non-Gaussian. When the sources are Gaussian ICA boils down to decorrelation (whitening) [2]. A decorrelated source vector, $\mathbf{s}$, is not unique because any orthogonal transformation is also decorrelated. Because decorrelated Gaussians are also independent, the sources estimated will not be unique.

The basic ICA, in which the number of sources is equal to the number of observations and there is no noise, is a well-studied problem. Methods based on maximum likelihood estimation [25, 23], maximising non-Gaussianity [20, 21, 22] or minimising mutual information [19] were proposed for its solution.

The ICA problem becomes much more complicated in more realistic scenarios,

such as in the presence of noise or where the number of sources is greater than the number of observations (the overcomplete case). Methods tackling these problems should make more assumptions about the nature of the sources. For example, in [66, 67] time dependencies of the signals are made use of. These methods find the independent source signals for which one-time-lagged covariances are zero [2].

An important extension of ICA is independent subspace analysis (ISA) [68] in which multi-component subspaces (sets of basis vectors) of an input vector are separated. Each source is assigned a subset of basis vectors by minimising their cross entropies. ISA has been successfully used in single-channel audio source separation [69, 70].

1.6. Non-Negative Matrix Factorisation

NMF, proposed for decomposition of non-negative data [39], is a method for multivariate data analysis. The goal is to approximate an $W \times K$ non-negative matrix, $\mathbf{X}$, as the product of two non-negative matrices, $\mathbf{T}$ and $\mathbf{V}$, of sizes $W \times I$ and $I \times K$, respectively. This is done via minimising the dissimilarity between $\mathbf{X}$ and $\mathbf{TV}$

$$\mathbf{T}^*, \mathbf{V}^* = \arg\min_{\mathbf{T},\mathbf{V}} D(\mathbf{X} \| \mathbf{TV}) \tag{1.9}$$

where the dissimilarity can be defined as the Kullback-Leibler (KL) divergence

$$D(\mathbf{A} \| \mathbf{B}) = -\sum_{\nu=1}^{W} \sum_{\tau=1}^{K} \left(A_{\nu,\tau} \log \frac{A_{\nu,\tau}}{B_{\nu,\tau}} + A_{\nu,\tau} - B_{\nu,\tau} \right) \tag{1.10}$$

KL divergence is always non-negative and is equal to zero when $\mathbf{X} = \mathbf{T}V$. The minimisation problem is effectively solved using variational bound optimisation in [39].

NMF can be seen as summarising the rows of $\mathbf{X}$ in the rows of $\mathbf{V}$ and columns in the columns of $\mathbf{T}$ [71]. For example, non-negative factorisation of the magnitude spectrogram of an audio signal provides a compact form of the spectrogram with re-

dundancies removed. The rows of $\mathbf{T}$ again corresponds to the frequency bins, the columns show the dominant spectral structures of the spectrogram. These columns can be thought of as codebooks of spectra or basis vectors. The matrix $\mathbf{V}$ contains the excitations of these basis vectors along the time frames.

Spectrogram decomposition using NMF is successfully applied to single-channel audio source separation [72, 14]. In addition to the KL divergence between $\mathbf{X}$ and $\mathbf{TV}$, the objective function also contains terms such that temporal continuity and sparseness of the excitations in $\mathbf{V}$ are satisfied. In [56, 57], the NMF model is defined in the Bayesian framework and the temporal continuity is incorporated through Gamma Markov chains [13]. The Bayesian extension was shown to be more successful than the previous NMF methods [56].

We will describe the Bayesian approach to the source separation problem, next. This methodology not only provides a consistent way to incorporate prior knowledge about the solutions into the problem, but also enables us to generalise our approach to other applications.

1.7. Bayesian Paradigm and the Approach of this Thesis

Bayesian paradigm provides a natural way to incorporate our prior beliefs into the solution. In crude terms, we start with our prior belief and update our belief into a posterior with the arrival of data. More formally, we infer the posterior distribution of the sources $p(\mathbf{s}|\mathbf{x})$, which is, by Bayes theorem, given by

$$p(\mathbf{s}|\mathbf{x}) = \frac{1}{Z_{\mathbf{x}}(\psi)} \int p(\mathbf{x}|\mathbf{s},\theta_m)p(\mathbf{s}|\theta_s)p(\theta_m|\psi_m)p(\theta_s|\psi_s)\, d\theta_m\, d\theta_s$$

The observation model, $p(\mathbf{x}|\mathbf{s},\theta_m)$, describes how the observed data is generated given the sources, $\mathbf{s}$, and the model parameters, θ_m. The observation model also defines the likelihood of the source coefficients and the model parameters. By maximising the marginal likelihood ($\mathbf{s}^* = \arg\max_{\mathbf{s}} p(\mathbf{x}|\mathbf{s},\theta_m)$) we find the most likely values of source coefficients that might have generated the observed signals. But, as we discussed in

the previous paragraph, the system is underdetermined and it is not possible to obtain a unique solution by maximum likelihood since the observation model gives the same likelihood value [52] for various coefficient values. Thus, we have to add what we know or believe about the sources and parameters into the model: A prior distribution for the sources, $p(\mathbf{s}|\theta_s)$, with source parameters, θ_s and prior distributions for the model and source parameters ($p(\theta_m|\psi_m)$ and $p(\theta_s|\psi_s)$) with hyperparameters ψ_m and ψ_s. The normalisation term $Z_{\mathbf{x}}(\psi)$ is the marginal likelihood (evidence) of the observed signals given the complete set of hyperparameters, $\psi \equiv [\psi_m\, \psi_s]$. Although evaluation of the marginal likelihood can be avoided during the inference, it needs to be evaluated or approximated when the optimisation of the hyperparameters will be accomplished through maximum likelihood. In the audio source models we mentioned, the marginal likelihood, $Z_{\mathbf{x}}(\psi)$, is intractable but can be approximated by stochastic simulation or analytic lower bounding methods.

In this thesis, we model the variances of audio source coefficients using Gamma Markov chains (GMCs) and Gamma Markov random fields (GMRFs). These two models ensure positive correlation between the variance variables, so the energy in the time-frequency domain is slowly-changing. In addition, all the variables in the model have full conditional conjugacy, i.e. their full conditional distributions belong to the same probability distribution class as their priors. So, inference on the variables in the model can be efficiently fulfilled using the Gibbs sampler [73] or variational Bayes (VB) [74].

One problem with the source model based on GMRFs is that the hyperparameters of the model, which determine the degree of coupling between the variables, cannot be optimised using the standard maximum likelihood approach because the marginal likelihood term contains an unknown normalising constant. This fact originates from the unknown normalising constant of the source model which includes a Markov random field. There are, however, optimisation methods designed for learning in models where the normalising constant is not known: Namely, pseudolikelihood [75] and contrastive divergence [76] both of which can be seen as approximate likelihood methods and score matching [77]. These methods were all proposed for learning in fully-observed models.

However the idea of contrastive divergence can be extended to cover latent variable models. In our audio model, the variances of time-frequency coefficients are modelled with GMRFs. We have no way of observing the variances in the source separation applications we focus on, so the hyperparameter optimisation will be accomplished through contrastive divergence. Pseudolikelihood and score matching are not directly applicable to models with latent variables including our audio model, but they can be effectively used in fully observed or marginalised GMRF models as we will show in the experiments section. The hyperparameter optimisation in GMCs is less problematic and can be carried out using the Monte Carlo EM method. However, variational EM algorithm performs poorly in this problem because it ignores the correlation between variables.

The organisation of this thesis is as follows: In Chapter 2, theoretical backround on the inference and learning methods used throughout this thesis is given. We explain the inference methods that can be used in the proposed methods in Section 2.1. These are the Gibbs sampler, VB and sequential Monte Carlo (SMC) methods. Then, we explain the optimisation methods which enable hyperparameter learning in normalised and non-normalised models in Section 2.2. In Chapter 3, we review some audio source models that are found in the literature. Then, we give the formal definitions and some properties of GMCs and GMRFs. We present audio denoising and single-channel source separation results in Section 4 along with experiments on some simple models to compare the performances of different inference and learning methods. The accomplishments of this work are summarised and possible future work is explained in Section 5.

2. THEORETICAL BACKGROUND

2.1. Inference

In this chapter, we explain the inference methods that are applicable to the inference of the variables of the audio models based on GMCs and GMRFs. In both models, full conditional distributions of the variables are standard distributions and inference can be carried out using the Gibbs sampler and variational Bayes. GMCs also have a sequential structure that enables us to use particle filter as well.

2.1.1. Variational Bayes

Variational Bayes (mean field) [74] methods make use of tractable distributions to effectively approximate intractable integrals in Bayesian inference problems. They also provide a lower bound on the marginal likelihood (evidence) which can be used in model selection and hyperparameter optimization tasks.

The idea is to approximate the posterior distribution of the latent variables, $p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})$, with a variational distribution, $q(\boldsymbol{x})$, that minimises the dissimilarity (KL divergence) between the two distributions.

$$\begin{aligned}
\mathrm{KL}(q||p) &= \int d\boldsymbol{x}\, q(\boldsymbol{x}) \log \frac{q(\boldsymbol{x})}{p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})} && (2.1)\\
&= \int d\boldsymbol{x}\, q(\boldsymbol{x}) \log \frac{q(\boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{\theta})}{p(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{\theta})} && (2.2)\\
&= \log p(\boldsymbol{y}|\boldsymbol{\theta}) + \int d\boldsymbol{x}\, q(\boldsymbol{x}) \log \frac{q(\boldsymbol{x})}{p(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{\theta})} && (2.3)\\
&= \log p(\boldsymbol{y}|\boldsymbol{\theta}) + \mathrm{KL}(q||p(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{\theta})) && (2.4)\\
&\equiv \log p(\boldsymbol{y}|\boldsymbol{\theta}) + \mathcal{E}(q,\boldsymbol{\theta}) && (2.5)
\end{aligned}$$

Since the evidence, $p(\boldsymbol{y}|\boldsymbol{\theta})$, is independent of the variational distribution, $q(\boldsymbol{x})$, minimising the KL divergence between the posterior and the variational distributions is

equal to minimising the variational free energy $\mathcal{E}(q, \boldsymbol{\theta})$. KL divergence is always non-negative due to Gibbs' inequality [78], so Equation 2.5 defines a lower bound on the evidence:

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{\theta}) &\geq -\mathcal{E}(q, \boldsymbol{\theta}) && (2.6) \\ &= \langle \log p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}) \rangle_q - \langle \log q(\boldsymbol{x}) \rangle_q && (2.7) \end{aligned}$$

where $\langle . \rangle_{\pi(\mathcal{X})}$ denotes expectation under probability distribution $\pi(\mathcal{X})$.

Having reduced the inference problem to the minimisation of the variational free energy (or equally, maximisation of the lower bound), we can compute each independent distribution $q(\boldsymbol{x}_i)$ using the fixed point equation

$$\log q(\boldsymbol{x}_i) =^{+} \langle \log p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}) \rangle_{q(\boldsymbol{x}_{-i})} \tag{2.8}$$

where $\boldsymbol{x}_{-i}$ refers to all variables $\boldsymbol{x}_j$ except for $\boldsymbol{x}_i$ itself.

2.1.2. Markov Chain Monte Carlo Methods

Monte Carlo methods are computational methods to approximate expectations in which the integration (or summation) is not analytically tractable and classical grid-based integration techniques perform poorly, e.g. due to high dimensionality. The expectation of a test function, $f(\mathbf{x})$, under a target distribution, $p(\mathbf{x})$, is estimated using a set of i.i.d. samples, $\{\mathbf{x}^{(i)}\}_{i=1}^{N}$, drawn from this distribution:

$$\begin{aligned} E(f) &= \langle f(\mathbf{x}) \rangle_{p(\mathbf{x})} = \int f(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} && (2.9) \\ &\approx \frac{1}{N} \sum_{i=1}^{N} f(\mathbf{x}^{(i)}) \equiv \hat{E}_N(f) && (2.10) \end{aligned}$$

This estimator is unbiased and almost surely converges to the true expectation $E(f)$ as a result of the strong law of large numbers. The variance of $\hat{E}_N(f)$ is equal to σ_f^2/N,

where σ_f^2 is the variance of the function f:

$$\sigma_f^2 = \int (f(\mathbf{x}) - E(f))^2 p(\mathbf{x})\, d\mathbf{x}. \tag{2.11}$$

Intuitively, Monte Carlo methods perform better than numerical integration techniques in high dimensions because they make a finer representation of the areas with high probability. However, drawing independent samples from a multidimensional probability distribution is often not straightforward.

Markov chain Monte Carlo (MCMC) approaches are used in cases where it is very difficult to draw independent samples from the target distribution, $p(\mathbf{x})$, but it can be evaluated up to a normalising constant. If an ergodic (irreducible and aperiodic) transition kernel is constructed, the Markov chain will converge to the target density as its invariant distribution.

The Metropolis-Hastings (MH) algorithm uses a proposal density, $q(\mathbf{x}'|\mathbf{x}^{(t)})$, to generate a new sample that depends on the current state of the Markov chain. The proposed sample is accepted with probability:

$$a(\mathbf{x}'; \mathbf{x}^{(t)}) = \min \left\{ \frac{p(\mathbf{x}')}{p(\mathbf{x}^{(t)})} \frac{q(\mathbf{x}^{(t)}|\mathbf{x}')}{q(\mathbf{x}'|\mathbf{x}^{(t)})}, 1 \right\}. \tag{2.12}$$

MH algorithm has an irreducible and aperiodic transition kernel and its invariant distribution is $p(\mathbf{x})$. This algorithm allows us to draw samples from probability distributions, $p(\mathbf{x}) = \phi(\mathbf{x})/Z$ where the normalising constant Z is not known, because Z is independent of $\mathbf{x}$ and two normalising constants in Equation 2.12 are cancelled out.

The Gibbs sampler [73] can be seen as a special case of the MH algorithm where the proposal distribution for the variables are their full conditionals, $p(\mathbf{x}_i|\mathbf{x}_{-i})$. First a variable ($\mathbf{x}_i$, i^{th} dimension of $\mathbf{x}$) is chosen uniformly, and then a sample for that dimension is drawn from its full conditional density. This way we obtain a sample that differs from the previous one, only in one dimension. In this case the acceptance probability

of a newly generated sample becomes one. When the full conditional distributions of the model are distributions from which efficient methods exist for sampling, it is highly convenient to use the Gibbs sampler.

2.1.3. Particle Filtering

Sequential Monte Carlo (SMC) methods are point-mass approximations to time evolving target distributions in dynamic systems, such as state-space models. A state-space model is represented by a state transition equation, $\boldsymbol{x}_t \sim f(.|\boldsymbol{x}_{t-1}, \boldsymbol{\theta}_{\boldsymbol{x}})$, i.e. prior of the hidden Markov process, and a observation equation $\boldsymbol{y}_t \sim g(.|\boldsymbol{x}_t, \boldsymbol{\theta}_{\boldsymbol{y}})$, i.e. the likelihood of the observed data. At time t, the target distribution for inference is the posterior $p(\boldsymbol{x}_{1:t}|\boldsymbol{y}_{1:t}) = p(\boldsymbol{x}_1, ..., \boldsymbol{x}_t|\boldsymbol{y}_1, ..., \boldsymbol{y}_t)$ or particularly the marginal posterior $p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t})$ (also called the filtering distribution).

It is impossible to evaluate these posterior distributions analytically except in hidden Markov models with finite states and linear Gaussian state-space models (Kalman filters). Monte Carlo methods can be employed to infer about the hidden variables in the general case. However, MCMC methods are not completely suitable for on-line update of a dynamic system because of their "batch" nature. When the system moves into a new time slice, $t+1$, an MCMC algorithm has to repeat the iterations to approximate $p(\boldsymbol{x}_{1:t+1}|\boldsymbol{y}_{1:t+1})$ because the previous samples are discarded.

SMC methods enable a way to reuse the previous samples, $\{\boldsymbol{x}_t^{(i)}\}_{i=1}^N$, in drawing the new generation of samples over the next time slice, $t+1$. Our target distribution in the state-space models, i.e. the posterior distribution, can be defined recursively as:

$$p(\boldsymbol{x}_{1:t+1}|\boldsymbol{y}_{1:t+1}) = p(\boldsymbol{x}_{1:t}|\boldsymbol{y}_{1:t}) \frac{p(\boldsymbol{y}_{t+1}|\boldsymbol{x}_{t+1}) p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)}{p(\boldsymbol{y}_{t+1}|\boldsymbol{y}_{1:t})}. \tag{2.13}$$

At time $t+1$, if we assume we already have an approximation for $p(\boldsymbol{x}_{1:t}|\boldsymbol{y}_{1:t})$ and samples $\{\boldsymbol{x}_t^{(i)}\}_{i=1}^N$, we can draw new samples from $p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)$ depending on the previous ones and evaluate $p(\boldsymbol{y}_{t+1}|\boldsymbol{x}_{t+1})$ and $p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)$ on these new samples. But, the denominator $p(\boldsymbol{y}_{t+1}|\boldsymbol{y}_{1:t})$ is not easy to evaluate analytically. This issue can be resolved making use

of importance sampling (IS).

IS lets us draw samples from a proposal (sampling) distribution and assign importance to these samples, indicating how likely it was for these samples to have been drawn from the actual target distribution. Then, the expectations under the target distribution, $p(\boldsymbol{x}) = \phi(\boldsymbol{x})/Z$ where Z is the generally unknown normalising constant, can be estimated as:

$$\begin{aligned}
\langle f(\boldsymbol{x})\rangle_{p(\boldsymbol{x})} &= \int f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} && (2.14)\\
&= \frac{1}{Z}\int f(\boldsymbol{x})\phi(\boldsymbol{x})d\boldsymbol{x} && (2.15)\\
&= \frac{\int f(\boldsymbol{x})\phi(\boldsymbol{x})d\boldsymbol{x}}{\int \phi(\boldsymbol{x})d\boldsymbol{x}} && (2.16)\\
&= \frac{\int f(\boldsymbol{x})W(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x}}{\int W(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x}} && (2.17)\\
&\approx \frac{\sum_{i=1}^{N} f(\boldsymbol{x}^{(i)})W^{(i)}}{\sum_{i=1}^{N} W^{(i)}} && (2.18)\\
&= \sum_{i=1}^{N} f(\boldsymbol{x}^{(i)})w^{(i)} && (2.19)
\end{aligned}$$

where $W^{(i)}$ and $w^{(i)} = W^{(i)}/\sum_{i=1}^{N} W^{(i)}$ are the unnormalised and normalised importance weights of the i^{th} sample, respectively.

Performing the IS method recursively on the arrival of new observations, we obtain the sequential importance sampling (SIS) algorithm. At each step we draw N samples from the proposal distribution $q(\boldsymbol{x}_{t+1})$ and update and normalise the importance weights:

$$W_{t+1}^{(i)} = W_t^{(i)}\frac{p(\boldsymbol{y}_{t+1}|\boldsymbol{x}_{t+1}^{(i)})p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t^{(i)})}{q(\boldsymbol{x}_{t+1})} \quad (2.20)$$

$$w_{t+1}^{(i)} = \frac{W_{t+1}^{(i)}}{\sum_{j=1}^{N} W_{t+1}^{(j)}} \quad (2.21)$$

One of the most important design choices in the SIS algorithm is the proposal

distribution, $q(\boldsymbol{x})$. A poor choice of the proposal degrades the performance of the algorithm, but at the same time the proposal should be easy to sample from. The best possible proposal would be the posterior itself, if was possible to draw samples from it. Using the prior, $p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)$, may be the simplest choice (Bootstrap filter, condensation algorithm), but it causes to explore the state space without any information about the observations.

A problem of the SIS algorithm in general is degeneracy, i.e. the unconditional variance of the importance weights increases over time [79]. This is because all but one of the weights tend to go to zero after a few steps and their contribution thereafter becomes negligible. Using the optimal proposal distribution, which minimises the variance of the importance weights conditional on $\boldsymbol{x}_{1:t}^{(i)}$ and $\boldsymbol{y}_{1:t}$, can be shown to be $p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t^{(i)}, \boldsymbol{y}_{t+1})$. But it may not be possible to draw samples from this distribution or evaluate $p(\boldsymbol{y}_{t+1}|\boldsymbol{x}_t^{(i)})$, which is used to update the weights. Besides, optimal proposal distribution decreases the degeneracy, but cannot solve the problem completely.

Another method to reduce degeneracy is to perform resampling whenever needed. Resampling is to sample current set with replacement from $p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t}) = \sum_{i=1}^{N} w^{(i)}\delta(\boldsymbol{x}_t - \boldsymbol{x}_t^{(i)})$ to generate a new set of samples in which unimportant samples of the original set are discarded and the important ones are stressed. The new set induces the same probability $p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t}) = \frac{1}{N}\sum_{i=1}^{N}\delta(\boldsymbol{x}_t - \boldsymbol{x}_t^{(i)})$ with equal weights. The resulting algorithm is called sequential importance sampling with resampling (SIS/R).

2.2. Hyperparameter Optimisation

GMCs and GMRFs are constructed to model the dependencies between source coefficients through coupling their variances. They are flexible so that different degrees of dependency can be modelled. Finding the best model is accomplished via optimising the values of the coupling hyperparameters. Hyperparameter optimisation in GMCs can be carried out through the EM algorithm. In GMRFs, maximum likelihood estimation is problematic because of the intractable normalising constant. In this chapter, we will review three methods (contrastive divergence, pseudolikelihood and score matching) that deal with such cases.

2.2.1. Expectation-Maximisation (EM) Algorithm

The EM algorithm [80] is a classic algorithm for maximum likelihood (ML) (or maximum a posteriori (MAP)) estimation of parameters in the presence of latent variables. It consists of two iteratively applied steps to find a local maximum of the likelihood $p(\boldsymbol{y}|\boldsymbol{\theta})$:

- Expectation (E) step: Compute the expectation of the complete log likelihood under the posterior distribution of the latent variables, $p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}_t)$:

$$Q(\boldsymbol{\theta}) = \int \log p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}) p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}_t)\, d\boldsymbol{x} \tag{2.22}$$

  where $\boldsymbol{\theta}_t$ represents the current values of the parameters.
- Maximisation (M) step: Find the values of the parameters that maximise the above expectation:

$$\boldsymbol{\theta}_{t+1} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}) \tag{2.23}$$

We can define a lower bound on the likelihood, $\mathcal{L}(\boldsymbol{\theta})$, using any distribution $q(\boldsymbol{x})$:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) = \log p(\boldsymbol{y}|\boldsymbol{\theta}) &= \log \int p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})\, d\boldsymbol{x} && (2.24)\\
&= \log \int q(\boldsymbol{x}) \frac{p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})}{q(\boldsymbol{x})}\, d\boldsymbol{x} && (2.25)\\
&\geq \int q(\boldsymbol{x}) \log \frac{p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})}{q(\boldsymbol{x})}\, d\boldsymbol{x} && (2.26)\\
&= \int q(\boldsymbol{x}) \log p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})\, d\boldsymbol{x} - \int q(\boldsymbol{x}) \log q(\boldsymbol{x})\, d\boldsymbol{x} && (2.27)\\
&= \mathcal{F}(q, \boldsymbol{\theta}) && (2.28)
\end{aligned}$$

making use of Jensen's inequality, which says that the value of a concave function of a weighted sum is greater than or equal to the weighted summation of the function values, in Equation 2.26. This lower bound is equal to the likelihood, $\mathcal{L}(\boldsymbol{\theta})$, when $q(\boldsymbol{x})$ is selected to be the posterior, $p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})$. Since we do not have the exact posterior distribution in most of the problems, we may use estimates of the posterior to evaluate the lower bound. The expectation in Equation 2.22 is one constituent of the lower bound that is a function of the hyperparameters.

The variational inference algorithm explained in Section 2.1.1 can be seen as the approximate E-step, in which the expected complete log likelihood is estimated using the tractable distribution, of a variational EM algorithm:

$$\hat{Q}_{VB}(\boldsymbol{\theta}) = \int \log p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}) q(\boldsymbol{x})\, d\boldsymbol{x} \tag{2.29}$$

The M-step of this EM-algorithm is the same as that of the exact EM-algorithm, except that it performs the parameter optimisation on an approximate expectation.

Likewise, in Monte Carlo EM the lower bound is evaluated using Monte Carlo estimate of the posterior of the latent variables:

$$\hat{Q}_{MC}(\boldsymbol{\theta}) = \frac{1}{N_i} \sum_{j=1}^{N_i} \log p(\boldsymbol{x}^{(j)}, \boldsymbol{y}|\boldsymbol{\theta}) \tag{2.30}$$

where N_i denotes the number of samples.

At each iteration, one (stochastic EM) or more (Monte Carlo EM) samples can be used to approximate the expectation.

In Markov random fields or any non-normalised model in general, the EM algorithm or its variants cannot be directly applied for hyperparameter optimisation because of the intractable normalising constant of the full joint distribution. Several methods such as the maximum pseudolikelihood [75], contrastive divergence [76] and score matching [77] were proposed for the estimation and optimisation in non-normalised densities. These methods are mainly proposed for fully observed models. In the presence of latent variables they may be incorporated within the EM fixed point iterations. Below, we will review these methods in detail.

2.2.2. Contrastive Divergence

Contrastive divergence [76] is an approximate maximum likelihood method based on estimating the gradient of the logarithm of the normalising constant, Z_θ, using MCMC samples. The gradient of the log likelihood, $\mathcal{L}(\theta; \mathbf{x})$, is

$$\frac{\partial \mathcal{L}(\theta; \mathbf{x})}{\partial \theta} = \frac{1}{T} \sum_{t=1}^{T} \frac{\partial \log \pi(\mathbf{x}(t); \theta)}{\partial \theta} - \frac{\partial \log Z_\theta}{\partial \theta} \tag{2.31}$$

where T is the number of observation vectors. The gradient of Z_θ can be expressed as:

$$\frac{\partial \log Z_\theta}{\partial \theta} = \frac{1}{Z_\theta}\frac{\partial Z_\theta}{\partial \theta} \tag{2.32}$$

$$= \frac{1}{Z_\theta}\frac{\partial}{\partial \theta}\int \pi(\mathbf{x};\theta)\, d\mathbf{x} \tag{2.33}$$

$$= \frac{1}{Z_\theta}\int \frac{\partial \pi(\mathbf{x};\theta)}{\partial \theta}\, d\mathbf{x} \tag{2.34}$$

$$= \frac{1}{Z_\theta}\int \pi(\mathbf{x};\theta)\frac{\partial \log \pi(\mathbf{x};\theta)}{\partial \theta}\, d\mathbf{x} \tag{2.35}$$

$$= \int \frac{\pi(\mathbf{x};\theta)}{Z_\theta}\frac{\partial \log \pi(\mathbf{x};\theta)}{\partial \theta}\, d\mathbf{x} \tag{2.36}$$

$$= \int p(\mathbf{x}|\theta)\frac{\partial \log \pi(\mathbf{x};\theta)}{\partial \theta}\, d\mathbf{x} \tag{2.37}$$

$$= \left\langle \frac{\partial \log \pi(\mathbf{x};\theta)}{\partial \theta} \right\rangle_{p(\mathbf{x}|\theta)} \tag{2.38}$$

Although still intractable, this gradient can be estimated using samples drawn from $p(\mathbf{x}|\theta_k)$ as it would be done in the expectation step of an EM algorithm, θ_k denoting the value of θ at k^{th} iteration:

$$\frac{\partial \log Z_\theta}{\partial \theta} \approx \frac{1}{N_i}\sum_{j=1}^{N_i}\frac{\partial \log \pi(\mathbf{x}^{(j)}(\theta_k);\theta)}{\partial \theta} \tag{2.39}$$

and the maximum likelihood solution can be sought iteratively. Here, the samples are written as $\mathbf{x}^{(j)}(\theta_k)$ to show that they depend on the current value θ_k. The computational complexity of estimating the gradient this way is too high because the evaluation of each gradient requires an MCMC, of which stationary distribution is $p(\mathbf{x}|\theta$, to reach equilibrium.

Empirical study of Hinton [76] showed that only a few MCMC steps (even one) is sufficient to estimate the direction of the gradient if the Markov chain is initialised at the observed values. This approach corresponds to minimising the following objective function

$$\mathrm{CD}_n = \mathrm{KL}(p_{\mathbf{x}}(\mathbf{x})\|p(\mathbf{x}|\theta)) - \mathrm{KL}(p_{\mathbf{x}^n}(\mathbf{x})\|p(\mathbf{x}|\theta)) \tag{2.40}$$

which is called the contrastive divergence. In this expression, $p_{\mathbf{x}^n}(\mathbf{x})$ is the reconstructed data distribution after running the MCMC for n steps. Note that, ML minimises $\mathrm{KL}(p_{\mathbf{x}}(\mathbf{x})\|p(\mathbf{x}|\theta))$ so that the data distribution, $p_{\mathbf{x}}(\mathbf{x})$, and the model distribution, $p(\mathbf{x}|\theta)$, get as similar as possible. The idea behind the contrastive divergence is to move towards the stationary distribution without getting far away from the data distribution, so that the variance is kept small. This leads to an approximate ML estimator, being equal to it if the MCMC is run for infinite steps. The terms containing the normalising constant, Z_θ, in the objective function cancel out:

$$\begin{aligned}\mathrm{CD}_n &= -\frac{1}{T}\sum_{t=1}^{T}\log p(\mathbf{x}(t)|\theta) + \frac{1}{T}\sum_{t=1}^{T}\log p(\mathbf{x}^n(t)|\theta) \\ &= -\frac{1}{T}\sum_{t=1}^{T}\log \pi(\mathbf{x}(t);\theta) + \frac{1}{T}\sum_{t=1}^{T}\log \pi(\mathbf{x}^n(t);\theta)\end{aligned}$$

There are mixed results about the convergence of CD. It has been shown to be unbiased in the optimisation of bivariate Gaussian distributions [81] and Gaussian Boltzmann machines [82]. In [81], empirical and theoretical evidences were given to show that CD has a small bias in learning Boltzmann machines. However, the work by Hyvärinen [83] showed that pseudolikelihood and CD were equivalent for Boltzmann machines and pseudolikelihood was a consistent estimator, which implies the consistency of CD in these models. This may be due to the difference between the definitions of bias and consistency they used. Yuille [84] expressed CD as a stochastic approximation algorithm in which convergence is guaranteed under certain conditions.

Contrastive divergence is also applicable to latent variable models. In a model $p(\mathbf{x},\mathbf{y}|\theta) = \pi(\mathbf{x},\mathbf{y};\theta)/Z_\theta$ with observed variables $\mathbf{y}$ and latent variables $\mathbf{x}$, we can define

the contrastive divergence as

$$\begin{aligned}\mathrm{CD}_n &= -\int d\mathbf{x}\, d\mathbf{y}\ \log \pi(\mathbf{y},\mathbf{x};\theta) p_{\mathbf{y}}(\mathbf{y}) p(\mathbf{x}|\mathbf{y},\theta_k) \\ &\quad + \int d\mathbf{x}\, d\mathbf{y}\ \log \pi(\mathbf{y},\mathbf{x};\theta) p_{\mathbf{y}^n}(\mathbf{y}) p_{\mathbf{x}^n}(\mathbf{x}) \\ &= -\frac{1}{T}\sum_{t=1}^{T} \langle \log \pi(\mathbf{y}(t),\mathbf{x};\theta)\rangle_{p(\mathbf{x}|\mathbf{y}(t),\theta_k)} \\ &\quad + \frac{1}{T}\sum_{t=1}^{T} \log \pi(\mathbf{y}^n(t),\mathbf{x}^n(t);\theta)\end{aligned}$$

where $\mathbf{y}^n(t)$ and $\mathbf{x}^n(t)$ are the n-step reconstructions of the visible and latent variables, respectively. The expectation can be approximated using samples drawn from $p(\mathbf{x}|\mathbf{y}(t),\theta_k)$.

2.2.3. Pseudolikelihood

Pseudolikelihood [75] is an approximate likelihood function based on a pseudo joint density function of the random variables. The pseudo joint density is defined as the product of the full conditional distributions of each variable in the model:

$$\tilde{\mathcal{L}}(\theta;\mathbf{x}) = \log \tilde{p}(\mathbf{x}|\theta) = \sum_i \log p(x_i|\mathbf{x}_{-i},\theta)\ . \tag{2.41}$$

Pseudolikelihood makes the learning problem in Markov random fields tractable, because the full conditionals do not contain the unknown normalising constant, Z_θ. Still, these conditional densities should be normalised:

$$p(x_i|\mathbf{x}_{-i},\theta) = \frac{p(\mathbf{x}|\theta)}{\int\ dx_i p(\mathbf{x}|\theta)}\ . \tag{2.42}$$

But this normalisation is over one variable and numerical integration techniques, as well as MCMC methods, can be applied to calculate this one-dimensional integral. Besides, in many Markov random fields the integral has an analytical solution and the variables have standard full conditional densities.

Although it is an approximation to the likelihood, pseudolikelihood preserves the dependencies between variables, to some degree, through the full conditional distributions. For some special Markov random fields [85, 86, 83], pseudolikelihood was shown to be consistent, i.e. maximised by the true parameter values when the size of the sample is infinite. But, a general consistency proof is not yet available.

2.2.4. Score Matching

Another method to estimate non-normalised densities is score matching [77], which is based on the idea that the score functions of the data and the model densities should be equal. The score function here is the gradient of the log-density with respect to the input variable:

$$\psi(\mathbf{x};\theta) = \nabla_{\mathbf{x}} \log p(\mathbf{x}|\theta) = \nabla_{\mathbf{x}} \log \pi(\mathbf{x};\theta) \tag{2.43}$$

This score function is analogous to the negative of the Fisher score function with respect to a location parameter, μ, evaluated at $\mu = \mathbf{0}$.

The objective of the method is accomplished through minimising the expected squared distance between the score functions of the data and the model densities:

$$J(\theta) = \frac{1}{2} \int d\mathbf{x}\, p_{\mathbf{x}}(\mathbf{x}) \|\psi(\mathbf{x};\theta) - \psi_{\mathbf{x}}(\mathbf{x})\|^2 \tag{2.44}$$

where $\psi_{\mathbf{x}}(\mathbf{x}) = \nabla_{\mathbf{x}} \log p_{\mathbf{x}}(\mathbf{x})$ is the score function of the observed data distribution $p_{\mathbf{x}}(\mathbf{x})$. This quantity does not contain the normalising constant Z_{a} but still requires a non-parametric estimator for the observed data. However, when the integrand in Equation 2.44 is expanded, it can be seen that the score function of the data is not actually needed.

$$\begin{aligned} J(\theta) \;=\; & \int d\mathbf{x}\, p_{\mathbf{x}}(\mathbf{x}) \left[\frac{1}{2} \|\psi_{\mathbf{x}}(\mathbf{x})\|^2 + \frac{1}{2} \|\psi(\mathbf{x};\theta)\|^2 \right. \\ & \left. - \psi_{\mathbf{x}}(\mathbf{x})^T \psi(\mathbf{x};\theta) \right] \end{aligned} \tag{2.45}$$

The first term in the brackets is independent of θ, so it can be ignored during the minimisation of $J(\theta)$ w.r.t. θ. The second term does not contain the data score function, $\psi_{\mathbf{x}}(\mathbf{x})$, and is equal to

$$\int d\mathbf{x}\, p_{\mathbf{x}}(\mathbf{x}) \frac{1}{2} \|\psi(\mathbf{x};\theta)\|^2 = \int d\mathbf{x}\, p_{\mathbf{x}}(\mathbf{x}) \sum_{i=1}^{N} \frac{1}{2} \psi_i(\mathbf{x}(t);\theta)^2$$

And finally, the following set of equations show that $\psi_{\mathbf{x}}(\mathbf{x})$ does not need to be known in order to calculate the third term

$$\begin{aligned}
&- \int d\mathbf{x}\, p_{\mathbf{x}}(\mathbf{x})\ \psi_{\mathbf{x}}(\mathbf{x})^T \psi(\mathbf{x};\theta) \\
&\qquad = \ -\int d\mathbf{x}\, p_{\mathbf{x}}(\mathbf{x}) \sum_{i=1}^{N} \psi_{\mathbf{x},i}(\mathbf{x}) \psi_i(\mathbf{x};\theta) \\
&\qquad = \ -\sum_{i=1}^{N} \int d\mathbf{x}\, p_{\mathbf{x}}(\mathbf{x}) \psi_{\mathbf{x},i}(\mathbf{x}) \psi_i(\mathbf{x};\theta) \\
&\qquad = \ -\sum_{i=1}^{N} \int d\mathbf{x}\, p_{\mathbf{x}}(\mathbf{x}) \frac{\partial \log p_{\mathbf{x}}(\mathbf{x})}{\partial x_i} \psi_i(\mathbf{x};\theta) \\
&\qquad = \ -\sum_{i=1}^{N} \int d\mathbf{x}\, \frac{p_{\mathbf{x}}(\mathbf{x})}{p_{\mathbf{x}}(\mathbf{x})} \frac{\partial p_{\mathbf{x}}(\mathbf{x})}{\partial x_i} \psi_i(\mathbf{x};\theta) \\
&\qquad = \ -\sum_{i=1}^{N} \int d\mathbf{x}\, \frac{\partial p_{\mathbf{x}}(\mathbf{x})}{\partial x_i} \psi_i(\mathbf{x};\theta) \\
&\qquad = \ \sum_{i=1}^{N} \int d\mathbf{x}\, p_{\mathbf{x}}(\mathbf{x}) \frac{\partial \psi_i(\mathbf{x};\theta)}{\partial x_i}
\end{aligned}$$

The last equation is derived using the partial integration rule with the assumption that $p_{\mathbf{x}}(\mathbf{x})\psi(\mathbf{x};\theta)$ goes to zero as $\|\mathbf{x}\| \to \infty$. Thus, $J(\theta)$ can be expressed in the following form

$$\begin{aligned}
J(\theta) \ &= \ \int d\mathbf{x}\, p_{\mathbf{x}}(\mathbf{x}) \sum_{i=1}^{N} \left[\partial_i \psi_i(\mathbf{x}(t);\theta) + \frac{1}{2} \psi_i(\mathbf{x}(t);\theta)^2 \right] \\
&\quad + \text{const.}
\end{aligned} \tag{2.46}$$

This expression is an expectation with respect to the data density, $p_{\mathbf{x}}(\mathbf{x})$, and can be

approximated by Monte Carlo integration using the observed samples from $p_{\mathbf{x}}(\mathbf{x})$:

$$\tilde{J}(\theta) = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N}\left[\partial_i\psi_i(\mathbf{x}(t);\theta) + \frac{1}{2}\psi_i(\mathbf{x}(t);\theta)^2\right] + \text{const.} \tag{2.47}$$

where $\mathbf{x}(t)$ denotes t^{th} observation vector, T is the number of observations and N is the size of $\mathbf{x}$. As T goes to infinity, $\tilde{J}(\theta)$ converges to $J(\theta)$.

In basic score matching, the probability density functions are assumed to be differentiable in $(-\infty, \infty)$. For distributions from other domains, the above derivations should be adapted. The expected squared distance for non-negative distributions (e.g. GMRFs) is approximated as [87]:

$$\begin{aligned}\tilde{J}_{\text{NN}}(\theta) \;=\; & \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N}\Big[x_i(t)^2\partial_i\psi_i(\mathbf{x}(t);\theta) \\ & + \frac{1}{2}\psi_i(\mathbf{x}(t);\theta)^2 x_i(t)^2 + 2\psi_i(\mathbf{x}(t);\theta)x_i(t)\Big] \\ & + \text{const.}\end{aligned} \tag{2.48}$$

When the data are assumed to come from the model $p(\mathbf{x}|\theta_{\text{true}})$ (i.e. this is the optimal model for the data) and there is no degeneracy in the model, score matching estimator, $\theta^* = \arg\min_\theta J(\theta)$, is equal to θ_{true}. As the sample size goes to infinity, $\tilde{J}(\theta)$ converges to $J(\theta)$, so the estimator that minimises $\tilde{J}(\theta)$ is consistent. However, this consistency requires an optimisation method that can find the global minimum. If it is liable to get stuck in local minima, the overall estimator becomes locally consistent, i.e. consistent for the subspace around the global minimum.

3. GAMMA MARKOV CHAINS AND RANDOM FIELDS

In this chapter, GMCs and GMRFs are explained. These models are used to define dependency structures on the variances of audio source coefficients. With GMCs, positive correlation between the variance variables along one axis (time or frequency) is ensured. GMRFs ensure correlation on both axes, thus provide more general means to model dependency of source coefficients. First, we will review some audio source models that incorporate basic properties of audio signals.

3.1. Audio Source Modelling in Bayesian Framework

In audio processing tasks, such as source separation and denoising, it is important to construct realistic and flexible models to capture the physical properties of audio signals. In a Bayesian framework, this requires development of appropriate prior distributions. Once a probabilistic model is constructed, many audio processing tasks can be solved, at least in principle, as inference problems. However, constructing physically realistic and flexible models is not very easy. In this section, we will review the models that reflect two physical properties of audio signals in the time-frequency representation: sparseness and dependency along time and frequency axes.

3.1.1. Sparse Priors

When assumed to be a priori independent, time-frequency domain coefficients of audio sources are shown to be better modelled with heavy-tailed distributions [62, 63, 12]. In source separation literature, specific proposals include mixture of Gaussians [42, 43], Laplace [44, 45], alpha-stable distributions [88] and Student-t distribution [12, 46]. These distributions can be defined in a hierarchical manner as a scale mixture of Gaussians (SMoG): $p(s_t) = \int p(s_t|v_t)p(v_t)\,dv_t$. $p(s_t|v_t)$ has a fairly simple form: a zero mean Gaussian with variance v_t which has its own prior distribution, $p(v_t)$. SMoG distributions are a large family of heavy tailed distributions, with every prior distribution $p(v_t)$ leading to a different instantiation: e.g. inverse gamma (Student-t),

exponential (Laplace) [89, 90, 47].

The prior distribution of time-frequency coefficients of a source signal is factorised as

$$p(\mathbf{s}|\theta_{\mathbf{s}}) = \prod_{\nu=1}^{W}\prod_{\tau=1}^{N} p(s_{\nu,\tau}|\theta_{\mathbf{s}}) \tag{3.1}$$

when the coefficients are assumed to be independent and identically distributed. A heavy-tailed prior can be defined for each $p(s_{\nu,\tau}|\theta_{\mathbf{s}})$ using mixture of Gaussians as

$$p(s_{\nu,\tau}) = \sum_{i=1}^{I} p(s_{\nu,\tau}|\gamma_{\nu,\tau}=i)p(\gamma_{\nu,\tau}=i) \tag{3.2}$$

where $\gamma_{\nu,\tau}$ is an indicator variable to select from I Gaussians

$$p(s_{\nu,\tau}|\gamma_{\nu,\tau}=i) = \mathcal{N}(s_{\nu,\tau};\mu_i,\sigma_i^2) \tag{3.3}$$

A heavy tailed distribution can simply be constructed with two Gaussians, one with a low variance around zero and another with high variance. Other mixtures can also be defined with different means and variances and different state priors.

The probability density function (pdf) of the prior of a source coefficient, $s_{\nu,\tau}$, when modelled with a Student-t distribution, is given by

$$p(s_{\nu,\tau}|k) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\,\Gamma(\frac{k}{2})}\left(1+\frac{s_{\nu,\tau}^2}{k}\right)^{-(\frac{k+1}{2})} \tag{3.4}$$

where k is the degrees of freedom and $\Gamma(\cdot)$ is the Gamma (generalised factorial) function. Student-t is a heavy-tailed distribution and is equal to the Gaussian distribution as k goes to infinity. Figure 3.1 presents pdf's of Student-t distributions with different degrees of freedom (k). As k gets bigger, the distribution gets closer to the standard Gaussian distribution ($\mathcal{N}(0,1)$).

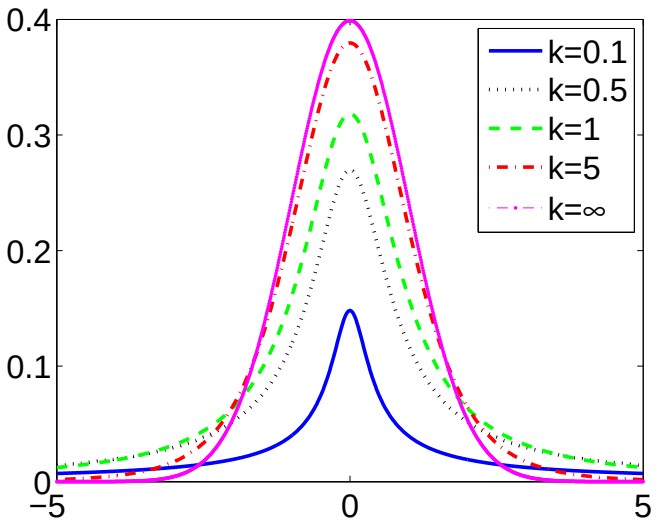


Figure 3.1. Probability density functions of Student-t distributions with different degrees of freedom (k). Note that when $k = \infty$, the distribution is equal to standard Gaussian distribution.

The Student-t source prior can also be defined in an hierarchical manner:

$$p(s_{\nu,\tau}|v_{\nu,\tau}) = \mathcal{N}(s_{\nu,\tau}; 0, v_{\nu,\tau})$$
$$p(v_{\nu,\tau}|k) = \mathcal{IG}(v_{\nu,\tau}; k/2, k/2)$$

where $\mathcal{IG}$ represents the inverse Gamma distribution of which details are given in

Appendix A. The marginal source prior in this model is derived by

$$
\begin{aligned}
p(s_{\nu,\tau}|k) &= \int_0^\infty p(s_{\nu,\tau}|v_{\nu,\tau})p(v_{\nu,\tau}|k)\,dv_{\nu,\tau} \\
&= \int_0^\infty \exp\left(-\frac{1}{2}\log(2\pi v_{\nu,\tau}) - \frac{s_{\nu,\tau}^2}{2v_{\nu,\tau}}\right. \\
&\left.-(k/2+1)\log v_{\nu,\tau} - \frac{k}{2v_{\nu,\tau}} + (k/2)\log(k/2) - \log\Gamma(k/2)\right)\,dv_{\nu,\tau} \\
&= \exp\left(-\frac{1}{2}\log(2\pi) + (k/2)\log(k/2) - \log\Gamma(k/2)\right) \\
&\int_0^\infty \exp\left(-\left(\frac{k+3}{2}\right)\log v_{\nu,\tau} - \frac{s_{\nu,\tau}^2+k}{2v_{\nu,\tau}}\right)\,dv_{\nu,\tau} \\
&= \exp\left(-\frac{1}{2}\log(2\pi) + (k/2)\log(k/2) - \log\Gamma(k/2)\right. \\
&\left.-\left(\frac{k+1}{2}\right)\log\left(\frac{s_{\nu,\tau}^2+k}{2}\right) + \log\Gamma\left(\frac{k+1}{2}\right)\right) \\
&= \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\,\Gamma(\frac{k}{2})}\left(1+\frac{s_{\nu,\tau}^2}{k}\right)^{-(\frac{k+1}{2})}
\end{aligned}
$$

This representation is the scale mixture of Gaussians (SMoG), where the prior is a weighted sum of infinite number of Gaussians. Here, the weights are inverse Gamma densities associated with the variances, $v_{\nu,\tau}$. Laplace distribution can be defined similarly, using an exponential distribution for $v_{\nu,\tau}$. Student-t and Laplace distributions show similar characteristics in terms of shape and scale, but the hierarchical definition of Student-t provides one more advantage: inverse Gamma distribution is the conjugate prior for the variances, $v_{\nu,\tau}$, in this model, i.e. in the presence of a Gaussian observation model, $\mathcal{N}(s_{\nu,\tau}; 0, v_{\nu,\tau})$, the prior and posterior distributions of $v_{\nu,\tau}$ are from the same probability family, namely inverse Gamma. This fact can be taken advantage of during the inference of the variables, as we have seen in Section 2.1.

Another hierarchical model which introduces the flexibility to have more sparse values was proposed in [48]:

$$
\begin{aligned}
p(s_{\nu,\tau}|v_{\nu,\tau},\gamma_{\nu,\tau}) &= (1-\gamma_{\nu,\tau})\delta_0(s_{\nu,\tau}) + \gamma_{\nu,\tau}\mathcal{N}(s_{\nu,\tau}; 0, v_{\nu,\tau}) \qquad (3.5)\\
p(v_{\nu,\tau}|\alpha,\beta) &= \mathcal{IG}(v_{\nu,\tau};\alpha,\beta)
\end{aligned}
$$

where $\delta_0(\cdot)$ is the Dirac delta function and $\gamma_{\nu,\tau}$ is a Bernoulli distributed indicator variable. When $\gamma_{\nu,\tau} = 0$, $s_{\nu,\tau}$ is equal to zero, otherwise, it is distributed as Student-t. By increasing $p(\gamma_{\nu,\tau} = 0)$, a more sparse model for $s_{\nu,\tau}$ is obtained.

3.1.2. Priors with Dependency Structures

It is possible to define more realistic prior distributions by introducing dependencies between source coefficients. The magnitudes of source coefficients in a time-frequency representation is slowly-changing. The majority of the coefficients are close to zero and there are some local clusters of coefficients that have high values. The dependency structure placed upon the source coefficients should prevent sudden changes in the magnitudes.

In [48, 50, 51, 52], discrete Markov chains and Markov random fields are used to put a dependency structure on indicator variables which are used for the selection of source variables. Selected variables have a Student-t distribution, while the rest are set to zero as explained in the model associated with Equation 3.5. The distribution of the indicators are defined conditional on the other indicators of the model.

A Markov chain structure to provide continuity between the indicators along the time axis is defined as [48, 50]

$$p(\gamma_{\nu,\tau}|\gamma_{\nu,\tau-1}, \theta^{\nu}) = \theta^{\nu}(\gamma_{\nu,\tau-1}, \gamma_{\nu,\tau}) \tag{3.6}$$

where θ^{ν} is a transition matrix for the frequency bin ν. θ^{ν} is a 2×2 matrix where the entries on the diagonal, $\theta^{\nu}(0,0)$ and $\theta^{\nu}(1,1)$, have beta prior distributions and the remaining values are set to $\theta^{\nu}(0,1) = 1 - \theta^{\nu}(0,0)$ and $\theta^{\nu}(1,0) = 1 - \theta^{\nu}(1,1)$. In this model, dependency between the coefficients along the time axis is satisfied by using the same transition matrix for the indicators of those coefficients.

In order to capture the dependencies on both axes, an Ising model, which is a type of Markov random field, with fixed parameters is used in [48]. An Ising model is

originally defined for the spins (of atoms) that are coupled to each other. Each variable takes a binary value, $x_i \in \{-1, +1\}$ and the energy of a state, $\mathbf{x}$, is defined as

$$E(\mathbf{x}; \mathbf{J}, \mathbf{H}) = -\left[\sum_{i,j} J_{i,j} x_i x_j + \sum_i H_i x_i\right]$$

where $J_{i,j}$ is the coupling between variables x_i and x_j; H_i is the applied field for each variable x_i. The probability of a state, $\mathbf{x}$, is then

$$p(\mathbf{x}|\boldsymbol{\theta}) = \frac{\exp[-\beta E(\mathbf{x}; \mathbf{J}, \mathbf{H})]}{Z(\boldsymbol{\theta})}$$
$$Z(\boldsymbol{\theta}) = \sum_{\mathbf{x}} \exp[-\beta E(\mathbf{x}; \mathbf{J}, \mathbf{H})]$$

where β is a hyperparameter inversely proportional to the Boltzmann's constant and $Z(\boldsymbol{\theta})$ is the normalising constant of the distribution.

In an Ising model, states, in which coupled variables have the same value, have high probability. By coupling the adjacent indicator variables, it is possible to define a dependency structure on the source coefficients. In [48], the hyperparameters of the model were fixed. Normally, the calculation of the normalising constant, $Z(\boldsymbol{\theta})$ is intractable for large Ising models, so the maximum likelihood estimation of the hyperparameters is not possible. Inference of the indicator variables can be carried out using the Gibbs sampling, which is explained in Section 2.1.

Another way to construct the dependency structure is by coupling the prior variances. In this thesis, variances of the source coefficients are directly coupled through GMCs and GMRFs. These models are explained in Chapter 3.

3.2. Gamma Markov Chains

A first-order Markov chain is a sequence of random variables, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$, where the distribution of each variable, $\mathbf{x}_t$, conditional on the previous variable, $\mathbf{x}_{t-1}$ is independent from all the other preceding variables [91, 92]

$$p(\mathbf{x}_t|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{t-1}) = p(\mathbf{x}_t|\mathbf{x}_{t-1}).$$

A GMC is a Markov chain in which the variables have alternatingly Gamma and inverse Gamma[1] prior distributions conditional on the preceding variable. That is, if a variable is Gamma distributed conditional on the previous variable, the previous variable is inverse Gamma distributed and vice versa. The formal definition is given as

$$\begin{aligned} p(v_1|\mathrm{a}) &= \mathcal{IG}(v_1; a_w, a_w b) \\ p(v_t|z_{t-1}, \mathrm{a}) &= \mathcal{IG}(v_t; a_w, a_w z_{t-1}), \quad t = 2..N \\ p(z_t|v_t, \mathrm{a}) &= \mathcal{G}(z_t; a_e, v_t/a_e), \quad t = 1..N-1. \end{aligned}$$

where $\mathrm{a} = [a_w\, a_e\, b]$ is the hyperparameter vector. b is used to define the distribution of the first variable, v_1. It may be thought of as z_0 without a prior distribution, thus a hyperparameter. a_w and a_e are the coupling hyperparameters that determine the strength of the dependency between the variables. The graphical model associated with this GMC distribution is presented in Figure 3.2.

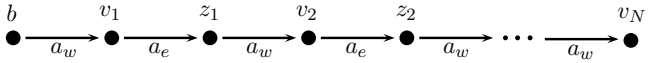


Figure 3.2. A Gamma Markov chain.

A GMC constitutes a chain of strictly positive variables with positive correlation between consecutive v_t's and consecutive z_t's. This naming convention is chosen for the variables such that the ones with Gamma and inverse Gamma prior distributions are separated. In addition, throughout this thesis, we will be modelling the variances of the source coefficients and v variables will denote these variance variables. GMCs

[1]The details of these distributions can be found in Appendix A

can be used to model the slowly-varying variances of a non-stationary audio signal as follows: The variances of source signal values, s_t, are associated with v_t's of the chain. The sources are then modelled with zero-mean Gaussians, $\mathcal{N}(s_t; 0, v_t)$. An example of a source signal generated from this model is presented in Figure 3.3. In this model, z_t's are auxiliary variables that are not associated with physical entities. In this thesis, we used this model in time-frequency representations of audio signals, to include the spectral dependencies across time frames or the dependencies between frequency bins. GMCs can also be used to model the intensities of a Poisson observation model [13, 56]. In this case, z_t variables, of which prior distributions are Gamma, are the intensity variables and v_t are auxiliary variables. Sources are Poisson distributed conditional on the intensities modelled by the GMC.

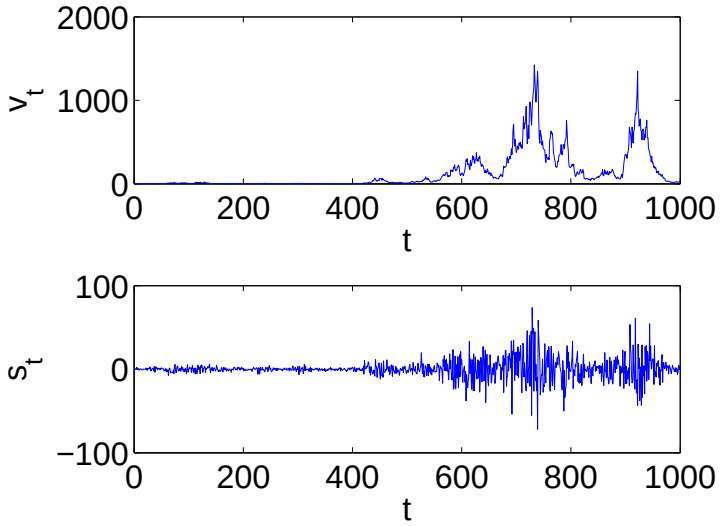


Figure 3.3. A signal generated using a GMC and Gaussian observation model. The hyperparameters used for the generation of the signal are $a_w = 50$, $a_e = 50$ and $b = 1$.

There are two important properties of GMCs: positive correlation between consecutive v variables, which also holds for z variables, and conditional conjugacy of the variables which is advantageous during the inference of the variables. Figure 3.4 shows the distribution of v_t conditional on v_{t-1} for various values of a_w and a_e. The positive correlation between consecutive variance variables can be seen in the figure. The ratio a_e/a_w is a measure of the skewness of correlation and it can lead to positive and negative drifts. When $a_e = a_w$, correlation is not skewed and larger values result in higher coupling between the variables. The probability of a variance variable condi-

tional on the previous one (transition kernel of the GMC) is found by integrating out the auxiliary variables z_{t-1}:

$$\begin{aligned} p(v_t|v_{t-1}) &= \int dz_{t-1} p(v_t|z_{t-1}) p(z_{t-1}|v_{t-1}) \\ &= \int dz_{t-1} \mathcal{IG}(v_t; a_w, a_w z_{t-1}) \mathcal{G}(z_{t-1}; a_e, v_{t-1}/a_e) \\ &= \frac{\Gamma(a_w + a_e)}{\Gamma(a_e)\Gamma(a_w)} \frac{(a_e v_{t-1}^{-1})^{a_e} (a_w v_t^{-1})^{a_e}}{(a_e v_{t-1}^{-1} + a_w v_t^{-1})^{(a_w + a_e)}} v_t^{-1} \end{aligned} \tag{3.7}$$

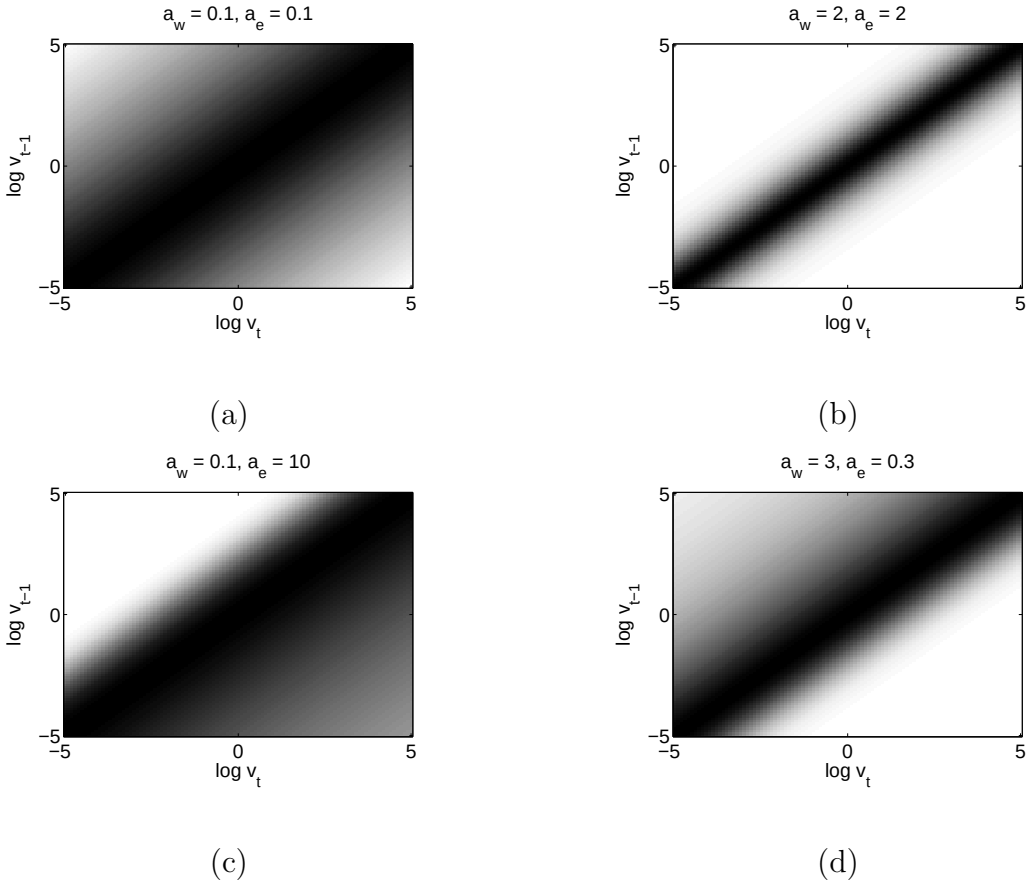


Figure 3.4. Correlation between v_t and v_{t-1} for various values of a_w and a_e. Figures (a) and (b) show that when the parameters, a_w and a_e, are equal, there is no skewness and the value determines the strength of the coupling. Figures (c) and (d) correspond to transition kernels where typical realisations have positive and negative drifts, respectively.

The other important feature of GMCs is that the prior distributions are conditionally conjugate. A model with variables $\mathbf{y}$ (observed) and $\mathbf{x}$ (latent) exhibits conditional

conjugacy [93], if the full conditional prior distribution of each variable x_i, $p(x_i|\mathbf{x}_{-i})$, is in the same class with its full conditional posterior distribution $p(x_i|\mathbf{x}_{-i}, \mathbf{y})$. That means it is as easy to draw samples from the posterior distribution, $p(\mathbf{x}|\mathbf{y})$, with the Gibbs sampler if the full conditionals of the prior, $p(\mathbf{x})$, are in the form of standard distributions. Moreover, such models are suitable for hierarchical expansion, e.g. by making the current observation model a part of the prior and defining a new observation model which preserves the conditional conjugacy. In GMCs, the full conditional distributions of the variables are Gamma and inverse Gamma, as their priors

$$\begin{aligned}
p(v_1|z_1, \mathrm{a}) &= \mathcal{IG}(v_1; a_w + a_e, a_w b + a_e z_1) \\
p(v_i|z_{i-1}, z_i, \mathrm{a}) &= \mathcal{IG}(v_i; a_w + a_e, a_w z_{i-1} + a_e z_i), \quad i = 2..N-1 \\
p(v_N|z_{N-1}\mathrm{a}) &= \mathcal{IG}(v_N; a_w, a_w z_{N-1}), \\
p(z_i|v_i, v_{i+1}, \mathrm{a}) &= \mathcal{G}(z_i; a_w + a_e, 1/(a_e/v_i + a_w/v_{i+1})), \quad i = 1..N-1.
\end{aligned}$$

3.3. Gamma Markov Random Fields

A Markov random field defines the joint probability distribution of a set of random variables, $\mathbf{x}$, based on the dependencies encoded in an undirected graph [94]. The joint distribution is written in terms of a set of potential functions, $\psi_C(\mathbf{x}_C)$, which map the possible assignments of the variables in a clique, $\mathbf{x}_C$, to a non-negative real value:

$$p(\mathbf{x}|\theta) = \frac{1}{Z_\theta} \prod_C \psi_C(\mathbf{x}_C; \theta). \tag{3.8}$$

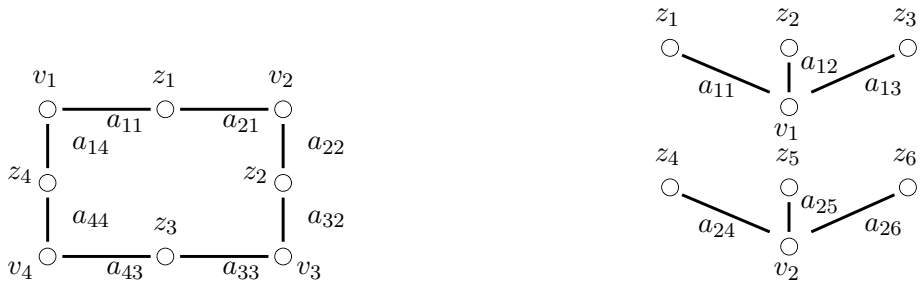
Here θ denotes the hyperparameters of the model and Z_θ is the normalisation constant to ensure $p(\mathbf{x}|\theta)$ is a pdf:

$$Z_\theta = \int d\mathbf{x} \prod_C \psi_C(\mathbf{x}_C; \theta). \tag{3.9}$$

Z_θ is generally analytically unavailable because the integration over the whole set of $\mathbf{x}$ is intractable. Even in the discrete case, the summation is over a number of values

which is exponential in the size of $\mathbf{x}$. In the evaluation of conditional probabilities, the normalising constant is cancelled out, so it does not have an effect on the inference of the variables, $\mathbf{x}$. However, in the optimisation of the hyperparameters, θ, of the model, we have to evaluate Z_θ because it depends on θ.

A GMRF models a joint probability distribution, $p(\mathbf{v}, \mathbf{z})$, using a bipartite undirected graph which consists of a vertex set $\mathcal{V} = \mathcal{V}_\mathbf{v} \cup \mathcal{V}_\mathbf{z}$, where partitions $\mathcal{V}_\mathbf{v}$ and $\mathcal{V}_\mathbf{z}$ denotes the collection of variables $\mathbf{v}$ and $\mathbf{z}$ that are conditionally distributed Gamma and inverse Gamma respectively. The edge set, $\mathcal{E}$, consists of pairs (i, j) representing the connection between the variables v_i and z_j which is associated with the coupling parameter a_{ij}. Examples of two GMRFs are given in Figure 3.5.



(a) (b)

Figure 3.5. Two examples of GMRFs. In the GMRF on the left, all variables are dependent on the other variables, whereas, the one on the right has partitions independent from each other.

Since a GMRF is associated with a bipartite undirected graph, it contains maximal cliques of size two. The joint probability, $p(\mathbf{v}, \mathbf{z})$, is defined in terms of pairwise and singleton potentials:

$$p(\mathbf{v}, \mathbf{z}|\mathrm{a}) = \frac{1}{Z_\mathrm{a}} \prod_{i\,:v_i \in \mathcal{V}_\mathbf{v}} \phi_v \left(v_i, \sum_{(v_i, z_j) \in \mathcal{E}} a_{ij} \right) \prod_{j\,:z_j \in \mathcal{V}_\mathbf{z}} \phi_z \left(z_j, \sum_{(v_i, z_j) \in \mathcal{E}} a_{ij} \right) \prod_{i,j:\,(v_i, z_j) \in \mathcal{E}} \phi_e \left(v_i^{-1}, a_{ij} z_j \right)$$

where the pairwise and singleton potentials are defined as follows:

$$\phi_v(\xi;\alpha) = \exp(-(\alpha+1)\log\xi) \tag{3.10}$$

$$\phi_z(\xi;\alpha) = \exp((\alpha-1)\log\xi) \tag{3.11}$$

$$\phi_e(\xi,\eta) = \exp((-\xi\eta)) \tag{3.12}$$

It is easy to see that the GMC in Figure 3.2 is equivalent to a GMRF with the joint density

$$p(\mathbf{v},\mathbf{z}|\mathrm{a}) = \frac{1}{Z_\mathrm{a}}\left(\prod_{i=1}^{N}\phi_v\left(v_i, a_w+a_e\right)\right)\left(\prod_{i=1}^{N}\phi_z(z_i, a_w+a_e)\right)\phi_e(v_1^{-1}, a_w z_0)$$
$$\left(\prod_{i=1}^{N-1}\phi_e(v_i^{-1}, a_e z_i)\phi_e(z_i, a_w v_{i+1}^{-1})\right)$$

where the normalising constant Z_a can be evaluated analytically.

In GMRFs, the full conditional distribution of each v_i variable in the field is inverse Gamma:

$$\begin{aligned} p(v_i|\mathbf{M}(v_i)) &= \frac{p(v_i,\mathbf{M}(v_i)|\mathbf{v}_{-i})}{p(\mathbf{M}(v_i)|\mathbf{v}_{-i})} \\ &= \phi_v\left(v_i, \sum_{j:\, z_j\in\mathbf{M}(v_i)} a_{ij}\right)\frac{\prod_{j:\, z_j\in\mathbf{M}(v_i)}\phi_e(v_i^{-1}, a_{ij}z_j)}{Z'} \\ &= \mathcal{IG}\left(v_i; \sum_{j:\, z_j\in\mathbf{M}(v_i)} a_{ij}, \sum_{j:\, z_j\in\mathbf{M}(v_i)} a_{ij}z_j\right) \end{aligned} \tag{3.13}$$

where $\mathbf{M}(v_i)$ is the set of z_j variables in the Markov blanket of v_i, the summations are over these variables and $\mathbf{v}_{-i}$ represents all the variables in $\mathcal{V}_\mathbf{v}$ except v_i. Z' is the normalising constant of the density function in the numerator. Similarly, z_j's are

conditionally Gamma distributed

$$p(z_j|\mathbf{M}(z_j)) = \mathcal{G}\left(z_j; \sum_{i:\, v_i \in \mathbf{M}(z_j)} a_{ij}, \left(\sum_{i:\, v_i \in \mathbf{M}(z_j)} a_{ij}/v_i\right)^{-1}\right) \tag{3.14}$$

GMRFs are proposed to model the variances of the time-frequency coefficients of audio signals so that the dependency between adjoining coefficients is captured. Although we know that there can be correlation between any pair or set of coefficients, it is difficult to find a generic model for the full joint density. Rather, we make use of the fact that the coefficients show a degree of persistence to change, i.e. they change slowly. GMRFs ensure a positive correlation between the coefficient variances to satisfy this property.

To see the conditional conjugacy of the audio source model based on GMRFs, let us consider the denoising scenario. Here, the aim is to extract the original source from the observed signal which is thought of as the original source plus white Gaussian noise

$$x_{\nu,\tau} \sim \mathcal{N}(s_{\nu,\tau}, r), \;\; \tau = 1..T, \nu = 1..N \tag{3.15}$$

Our audio model assumes a GMRF prior distribution for the variances of the time-frequency coefficients, $s_{\nu,\tau}$, as shown in Figure 3.5a. Conditioned on the variances, $v_{\nu,\tau}$, each coefficient is a zero-mean Gaussian

$$s_{\nu,\tau}|v_{\nu,\tau} \sim \mathcal{N}(s_{\nu,\tau}; 0, v_{\nu,\tau}), \;\; \tau = 1..T, \nu = 1..N$$

Then, full conditional distribution of a source coefficient becomes

$$p(s_{\nu,\tau}|v_{\nu,\tau}, x_{\nu,\tau}, r) = \mathcal{N}(s_{\nu,\tau}; \mu_{\nu,\tau}, \Sigma_{\nu,\tau})$$
$$\Sigma_{\nu,\tau} = \left(\frac{1}{r} + \frac{1}{v_{\nu,\tau}}\right)^{-1}, \qquad \mu_{\nu,\tau} = \frac{x_{\nu,\tau}\Sigma_{\nu,\tau}}{r}$$

This observation model is conditionally conjugate for the GMRF. So, the full conditional distribution of each variance variable, $v_{\nu,\tau}$, is again inverse Gamma

$$p(v_{\nu,\tau}|\mathbf{M}(v_{\nu,\tau})) = \mathcal{IG}(v_{\nu,\tau}; \alpha_{\nu,\tau}, \beta_{\nu,\tau}) \tag{3.16}$$

$$\alpha_{\nu,\tau} = \frac{1}{2} + \sum_{j:\, z_j \in \mathbf{M}(v_{\nu,\tau})} a_{ij}, \qquad \beta_{\nu,\tau} = \frac{s_{\nu,\tau}^2}{2} + \sum_{j:\, z_j \in \mathbf{M}(v_{\nu,\tau})} a_{ij} z_j \tag{3.17}$$

The $\mathbf{z}$ variables, which do not correspond to physical entities but are just auxiliary variables to construct the GMRF, are not coupled to any new variables, so they preserve Gamma full conditional distributions.

In this denoising scenario, we also assume an inverse Gamma prior distribution for the variance of observation noise, r

$$r \sim \mathcal{IG}(r; a_r, b_r)$$

The full conditional distribution of r is also inverse Gamma

$$p(r|s_{1:N,1:T}, x_{1:N,1:T}) = \mathcal{IG}(r; \alpha_r, \beta_r)$$

$$\alpha_r = a_r + \frac{NT}{2}, \qquad \beta_r = \left(\sum_{\nu=1}^{N} \sum_{\tau=1}^{T} \left(\frac{1}{2} x_{\nu,\tau}^2 - x_{\nu,\tau} s_{\nu,\tau} + \frac{1}{2} {s_{\nu,\tau}}^2 \right) + \frac{1}{b_r} \right)^{-1}$$

The Gibbs sampler and variational Bayes, which are suitable methods to perform inference on GMRFs, are explained in Section 2.1.

3.4. NMF Using GMCs

The statistical interpretation of NMF can be derived by seeking a maximum likelihood solution to the following model

$$s_{\nu,i,\tau} \sim \mathcal{PO}(s_{\nu,i,\tau}; t_{\nu,i} v_{i,\tau}) \tag{3.18}$$

$$x_{\nu,\tau} = \sum_{i}^{I} s_{\nu,i,\tau} \tag{3.19}$$

where $\mathbf{S}_i = \{s_{\nu,i,\tau}\}$ are latent sources and $\mathcal{PO}(\cdot)$ denotes the Poisson distribution. In the presence of these latent variables, the solution can be obtained using the EM algorithm. This approach leads to the same update rules as the original NMF minimising the information divergence between $\mathbf{X}$ and $\mathbf{TV}$ [57].

In order to obtain template and excitation matrices satisfying some properties, we can define prior distributions on $\mathbf{T}$ and $\mathbf{V}$, such as

$$\mathbf{T} \sim p(\mathbf{T}|\mathbf{\Theta}^t)$$
$$\mathbf{V} \sim p(\mathbf{V}|\mathbf{\Theta}^v)$$

where $\mathbf{\Theta}^t$ and $\mathbf{\Theta}^v$ are the parameters of these distributions. Then, $\mathbf{T}$ and $\mathbf{V}$ can be estimated by the maximum a posteriori solution or Bayesian inference.

The topology of our model is designed to separate the underlying tonal and percussive sources from an audio signal. This is accomplished through assigning different prior structures to different parts of the template and excitation matrices, $\mathbf{T}$ and $\mathbf{V}$. Spectral templates of tonal signals have high values for the fundamental frequency and the harmonics of the notes that are being played. The other values are close to zero. These templates are excited for the duration that the notes are audible. However, a percussive hit excites a band of frequencies at the same time. These excitations are generally for short time intervals, except for the bass drum hits.

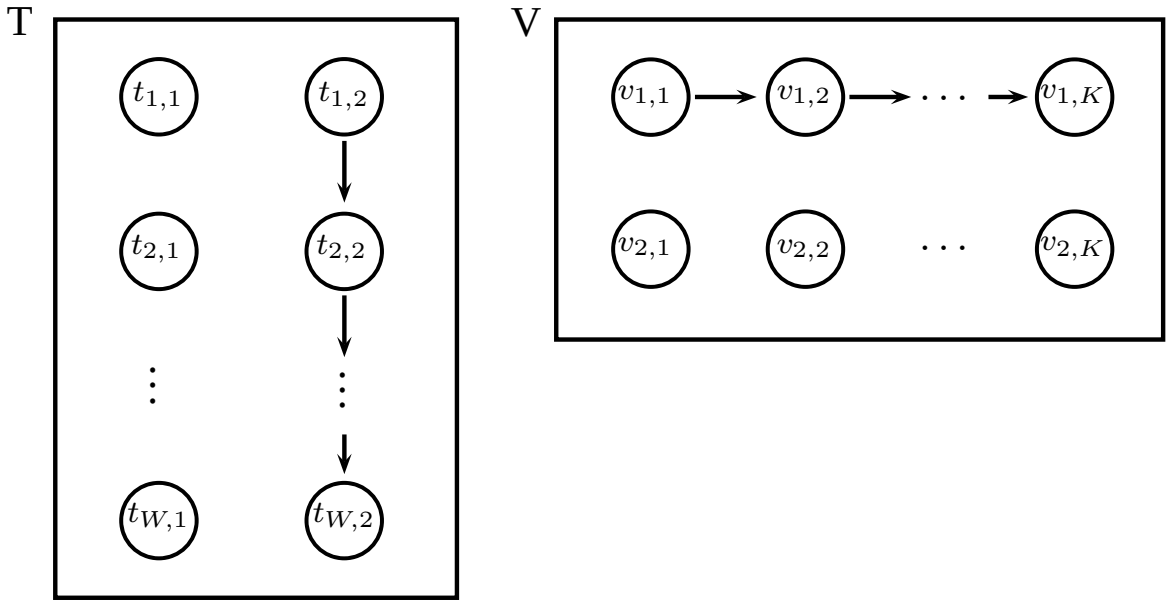


Figure 3.6. **T** and **V** matrices for one tonal and one percussive components.

Our model makes use of GMC priors for columns of **T** or rows of **V** to enable continuity along those vectors and independent Gamma priors to have sparse values with occasional peaks. So, the tonal vectors of **T** are modelled with independent Gamma distributions for sparsity, whereas the vectors for percussions are modelled with GMCs. In contrast, excitation vectors for tonal components are modelled with GMCs to enforce continuity in time. Excitation vectors of percussive sources have independent Gamma priors which are suitable for short-time excitations. **T** and **V** matrices for one tonal and one percussive component are presented in Figure 3.6.

The choice of Gamma and Gamma Markov chains as priors is mainly for the sake of simplicity. Gamma distribution is the conjugate prior for the Poisson observation model and this enables us to use faster and more convenient inference methods such as the Gibbs sampler or variational Bayes. In addition, we can incorporate the above mentioned requirements of tonal and percussive sources into the model using these prior distributions.

The density of a Gamma distributed random variable, $x \in \Re_+$, with shape and scale parameters, a and b, is given by $\mathcal{G}(x; a, b) = \exp((a-1)\log x - x/b - \log\Gamma(a) - a\log b)$. The mean of this distribution is ab and the variance is ab^2. With small ab and

a larger b, the distribution will be sparse, i.e. mainly close to zero but with a heavy tail.

A Gamma Markov chain (Section 3.2) is a prior structure for a chain of positive variables, where the correlation between consecutive variables is positive. In addition, each variable is conditionally conjugate, i.e. their prior and full conditional distributions are Gamma. In the Poisson observation model, this conjugacy is preserved. A GMC of $v_{1:K}$ can be defined as

$$\begin{aligned} v_1 &\sim \mathcal{G}(v_1; a_v, b/a_v) \\ z_i|v_i &\sim \mathcal{IG}(z_i; a_z, a_z v_i), \quad i = 1..K-1 \\ v_{i+1}|z_i &\sim \mathcal{G}(v_{i+1}; a_v, z_i/a_v), \quad i = 1..K-1 \end{aligned}$$

where a_v, a_z, and b are the hyperparameters of the chain and $z_{1:K-1}$ are auxiliary variables introduced to have positive correlation and conjugacy properties simultaneously. a_v and a_z are the coupling hyperparameters and they determine the degree of correlation between variables. Prior and full conditional distributions of $z_{1:K-1}$ are inverse Gamma and consecutive z variables have positive correlation between them.

Denoting the number of tonal components with I_{ton} and percussive components with $I_{perc} = I - I_{ton}$, the overall NMF model can be written as

$$\begin{aligned} t_{\nu,i} &\sim \mathcal{G}(t_{\nu,i}; a_t^i, b_t^i/a_t^i), \quad i = 1..I_{ton}, \nu = 1..W \\ t_{1:W,i} &\sim \text{GMC}(t_{1:K,i}; a_{tv}^i, a_{tz}^i, b_t^i), \quad i = I_{ton}+1..I \\ v_{i,1:K} &\sim \text{GMC}(v_{i,1:K}; a_{vv}^i, a_{vz}^i, b_v^i), \quad i = 1..I_{ton} \\ v_{i,\tau} &\sim \mathcal{G}(v_{i,\tau}; a_v^i, b_v^i/a_v^i), \quad i = I_{ton}+1..I, \tau = 1..K \end{aligned}$$

The observation model is again given as in Equations 3.18 and 3.19.

Because of the conditional conjugacy, the full conditional distribution of each variable in the model is a standard distribution: Gamma for $t_{\nu,i}$ and $v_{i,\tau}$, multinomial

for the latent sources $s_{\nu,i,\tau}$ and inverse Gamma for the auxiliary variables of the GMCs. This makes it feasible to use the Gibbs sampler or variational Bayes to infer about the variables.

The optimisation of the hyperparameters of the model can be performed using an EM algorithm, which makes use of the posterior distribution estimated during the inference: samples drawn by the Gibbs sampler or the sufficient statistics estimated by variational Bayes. In this thesis, we assume a uniform distribution for the hyperparameters and estimate them by sampling from their full conditional distributions using the Metropolis algorithm.

3.4.1. An Extension To The Model

As mentioned before, bass drums have a hybrid behaviour: they excite a band of frequencies as the other percussive sources but the duration is longer. This causes the bass drum and tonal instrument components to get mixed. As a remedy, we added another partition of size I_{bass} to the $\mathbf{T}$ and $\mathbf{V}$ matrices. A template vector in this partition has high values until a change point λ_i and very low values afterwards.

$$t_{1:W,i} \sim \prod_{\nu=1}^{\lambda_i} \mathcal{G}(t_{\nu,i}; a_B^i, b_B^i/a_B^i) \prod_{\nu=\lambda_i+1}^{W} \mathcal{G}(t_{\nu,i}; a_b^i, b_b^i/a_b^i)$$

$$v_{i,1:K} \sim \mathrm{GMC}(v_{i,1:K}; a_{vv}^i, a_{vz}^i, b_v^i), \quad i = I_{ton} + I_{perc} + 1..I$$

where a_B^i and b_B^i are selected such that the mean of distribution is high and variance low, in contrast, a_b^i and b_b^i ensure that the distribution is highly sparse. Here, $t_{\nu,i}$ and $v_{i,\tau}$ variables again have Gamma full conditional distributions. λ_i is discrete and its full conditional distribution can be evaluated at each W. So, this extended model can again be inferred using the Gibbs sampler. The pseudocode of the overall method is given in Appendix D.

4. EXPERIMENTS & DISCUSSIONS

In Chapter 3, we defined two models, GMC and GMRF, which are appropriate to model the variances of audio source coefficients in time-frequency representations such that the temporal and spectral dependencies of the coefficients are captured. The GMC model couples the variances either across the time axis or the frequency axis, while the GMRF model can capture dependencies along both directions. The strengths of the couplings are determined by the hyperparameters of the models which results in flexible models able to model different dependency structures. This makes the optimisation of the hyperparameters of the models important during the inference.

GMCs and GMRFs share some basic ideas such as conditional conjugacy which enables the inference of the variables to be carried out through the Gibbs sampler and variational Bayes. However, the hyperparameter optimisation schemes of the two methods are highly different. GMCs can be optimised by using an appropriate EM method during the inference, while this cannot be done in GMRFs due to the intractable normalising constant and approximate optimisation methods are needed. This chapter is divided into two sections dedicated to these two methods. Each section contains experiments with synthetic data to assess the most suitable inference-optimisation scheme and then presents results of denoising and single-channel source separation results using this scheme.

4.1. Gamma Markov Chains

The main goal of this section is to optimise the hyperparameters of dynamic systems offline (given a fixed sequence of observations, $y_{1:T}$), particularly the GMCs. The variants of the EM algorithm explained in Section 2.2.1 maximise approximate likelihoods and accuracy of these approximations are crucial to the maximum likelihood optimisation. In the remainder of this section we will demonstrate how accurate and efficient the methods are in estimating likelihoods on different problems.

We start with the linear Gaussian state-space model, in which the likelihood, $p(\boldsymbol{y}|\boldsymbol{\theta})$, and the posterior filtering distributions, $p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t}, \boldsymbol{\theta})$, can be calculated exactly by the Kalman filter[95, 96]. This model is very similar to the GMC based audio source model. In both of the models observations are Gaussian and hyperparameters determine the coupling between the state variables. Since we can calculate the likelihood of the linear Gaussian state-space model exactly, we will have the chance to compare the inference methods explained in Section 2.1 with the ground truth.

Our GMC-based audio source model is a similar state-space model where the states are the variances and the observations are the source coefficients. The state transitions are modelled with GMCs. There is no exact analytical solution in this problem, so we will compare the methods among themselves. Then we will use these source priors in denoising and single channel source separation problems. We will demonstrate the relation between the objective source separation evaluation criteria and the approximate likelihoods and how the results are affected by hyperparameter optimisation.

4.1.1. Linear Gaussian State Space Model

Linear Gaussian state-space model is given by the following state transition and observation models

$$x_1 \sim \mathcal{N}(x_1; 0, P) \tag{4.1}$$

$$x_k \sim \mathcal{N}(x_k; Ax_{k-1}, Q),\ k \geq 2 \tag{4.2}$$

$$y_k \sim \mathcal{N}(y_k; Cx_{k-1}, R) \tag{4.3}$$

where P, Q and R are the variances of Gaussian perturbations, A and C are linear operators. Optimal filtering can be performed on this model with the Kalman filter [95, 96], so this model provides a platform to compare the algorithms in terms of the accuracy of their likelihood estimates and time complexity.

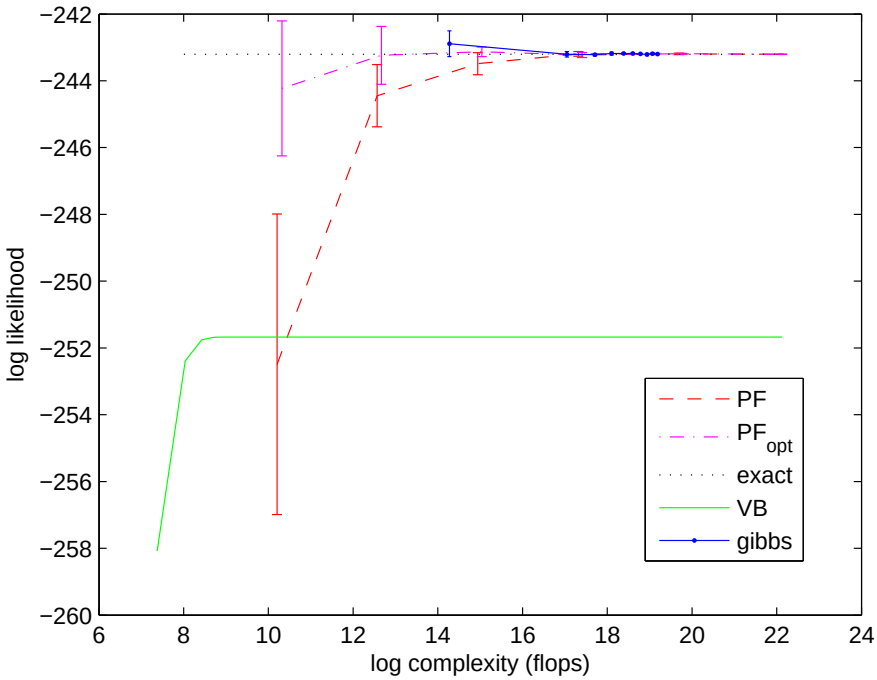


Figure 4.1. Log-likelihood approximations on a linear Gaussian state-space model of length 100 by different methods (Kalman filter (exact), bootstrap filter (PF), SIS/R with optimal proposal distribution ($\mathrm{PF}_{\mathrm{opt}}$), Gibbs sampler (gibbs) and VB). More complex particle filters and Gibbs samplers are obtained by increasing the number of samples while in VB total number of iterations is increased.

Figure 4.1 presents log-likelihood attained by the algorithms (Kalman filter, bootstrap filter, SIS/R with optimal proposal distribution, Gibbs sampler and VB) in a certain amount of CPU cycles (flops). In this model Kalman filter finds the exact likelihood $p(\boldsymbol{y}|\boldsymbol{\theta})$ making use of the fact that the convolution of two Gaussians is an unnormalised Gaussian. The likelihood estimates of the particle filters and the Gibbs samplers converge to the exact likelihood as the number of samples is increased. We have to note that Gibbs sampler does not output the likelihood as a by-product. We estimated the Gibbs likelihood using Chib's method [97], which needs extra sampling for this model. VB is quick to converge but the lower bound of the likelihood estimated by VB cannot reach the exact likelihood due to the factorised approximation of the model. Figure 4.2 shows how the Gibbs sampler and VB estimates consecutive variables, x_t and x_{t+1}. It is certain that the correlations between the variables are lost in

the variational estimation which in turn results in a loose lower bound. However, the mean estimates (minimum mean square error (MMSE) estimates) of the two methods overlap as depicted in Figure 4.3.

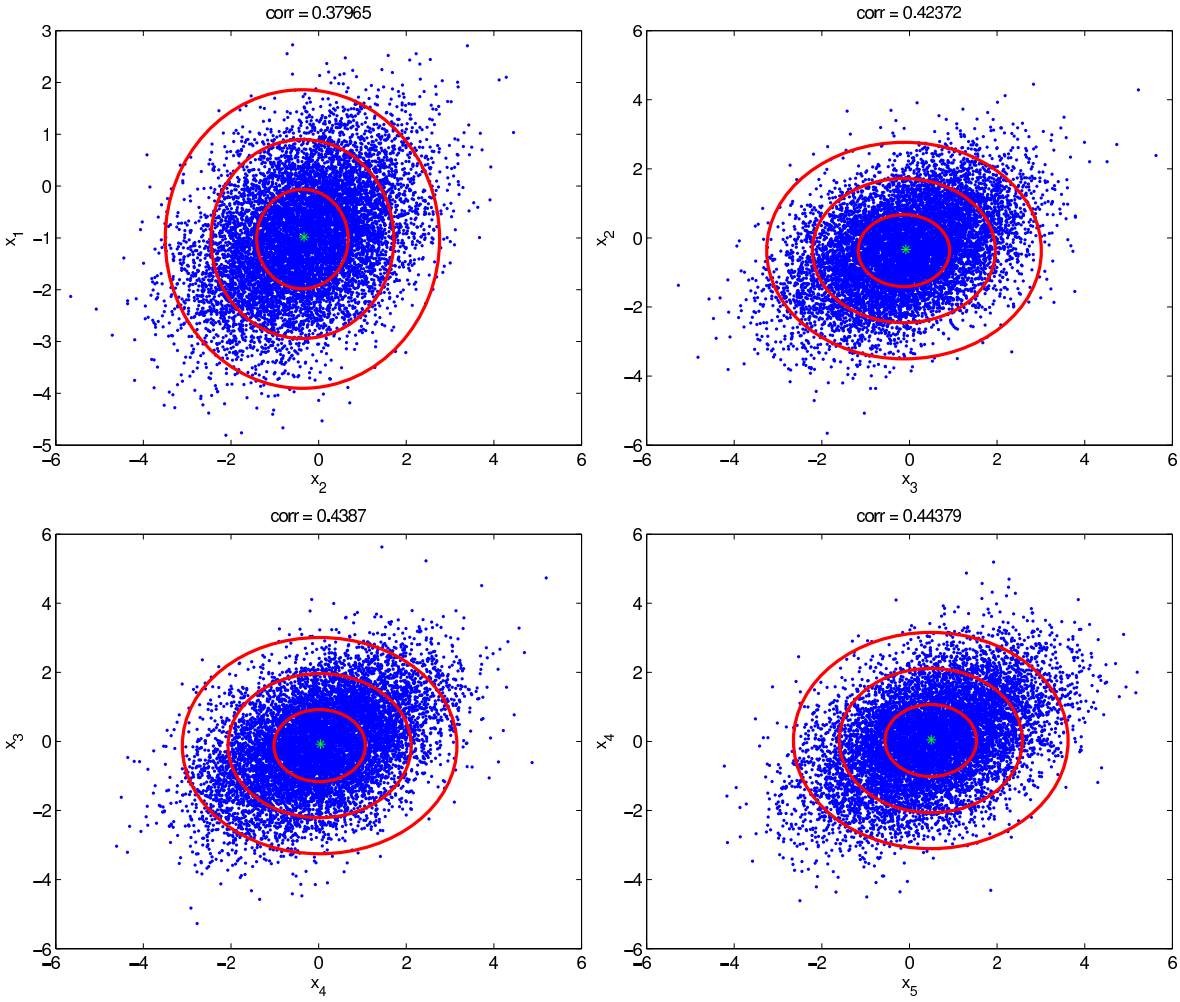


Figure 4.2. Samples drawn by a Gibbs sampler and variational estimates for consecutive variables in a chain of length 5. Variational distributions are represented with horizontal or vertical ellipses up to three standard deviations whereas samples are shown as dots.

In Figure 4.4, the likelihood surfaces estimated by the Kalman filter, SIS/R with optimal proposal distribution, Gibbs sampler and VB are presented. The methods are run with a grid of different values of hyperparameters Q and R while the others are fixed. Likelihood estimates of SIS/R with optimal proposal distribution and Gibbs sampler converge to the exact likelihood, while VB suffers from loose lower bound and

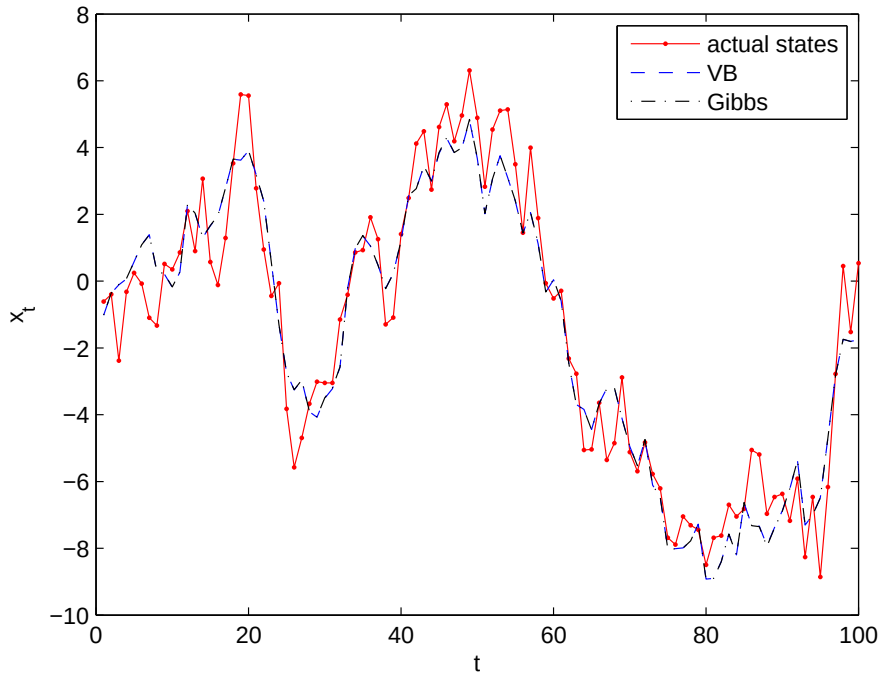


Figure 4.3. Actual state sequence and state estimates by VB and Gibbs sampler. The estimates by the two methods overlap.

estimates an incorrect surface. The EM variants, which use the posterior distributions approximated by these algorithms, converge to their respective maximum shown in this figure.

4.1.2. Gamma Markov Chains

We define a state-space model with transitions modelled with an GMC and observations with Gaussians:

$$z_1 \sim \mathcal{G}(z_1; a_z, b/a_z) \tag{4.4}$$

$$v_t|z_t \sim \mathcal{IG}(v_t; a_v, a_v z_t) \tag{4.5}$$

$$z_{t+1}|v_t \sim \mathcal{G}(z_{t+1}; a_z, v_t/a_z) \tag{4.6}$$

$$s_t \sim \mathcal{N}(s_t; 0, v_t) \tag{4.7}$$

which is a simplified version of the audio source model explained in Section 3.2.

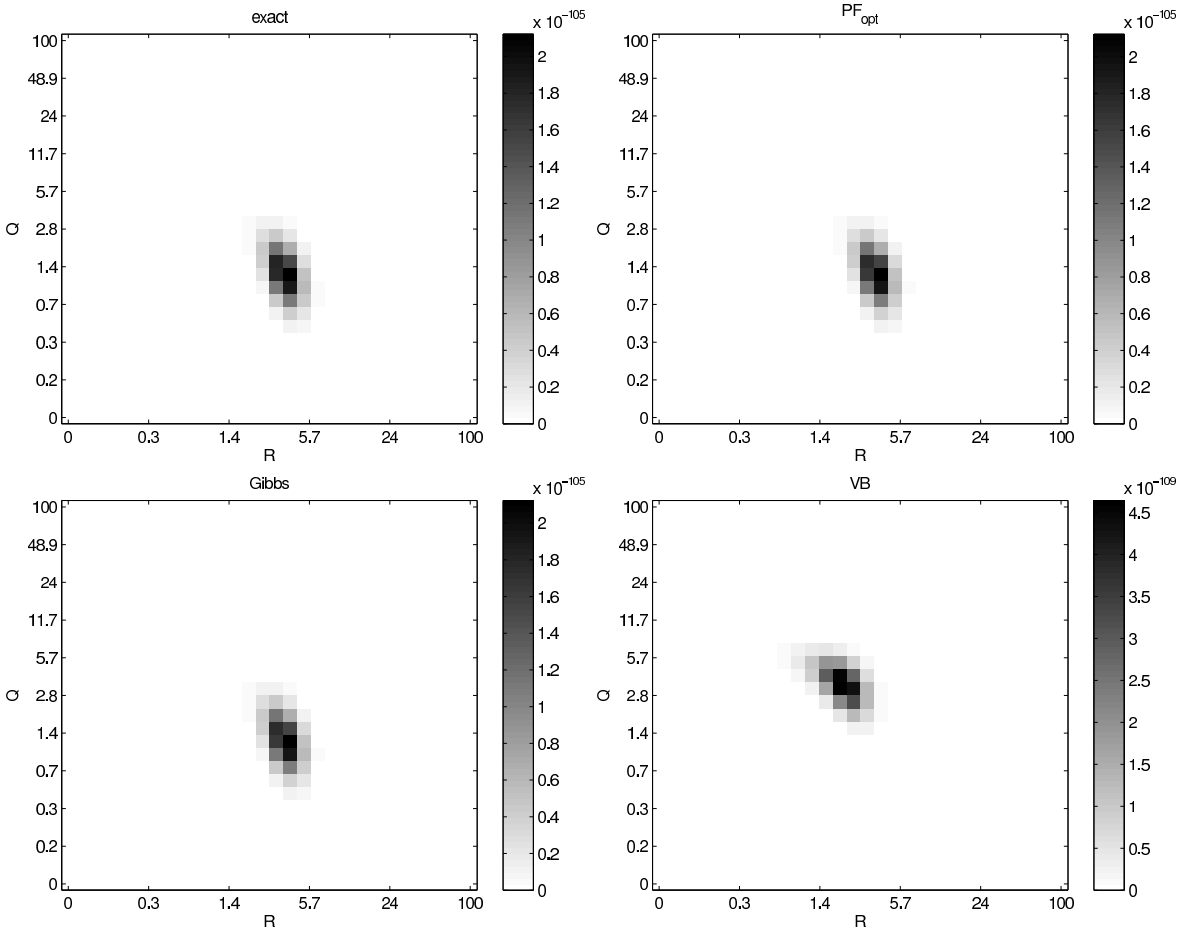


Figure 4.4. Likelihood versus model parameters (Q and R) for an observation sequence of length 100. The surfaces are almost the same for the Kalman filter (exact), SIS/R with optimal proposal distribution ($\mathrm{PF}_{\mathrm{opt}}$) and the Gibbs sampler (Gibbs). In particular, the Q and R pair that maximises the likelihood and the maximum value are the same.The VB lower bound has completely different characteristics. In the experiment, the hyperparameters A, C and P are fixed ($A = C = 1$, P=2).

As in the linear Gaussian case, the lower bound on the log-likelihood estimated by the VB algorithm is not a tight bound (Figure 4.5). Likelihood estimates of all the sampling based methods converge to a fixed value. While we do not know the relation between this value and the exact likelihood as there is no known analytical solution for it, the experiments in the previous section proved these estimates to be consistent with the exact likelihood.

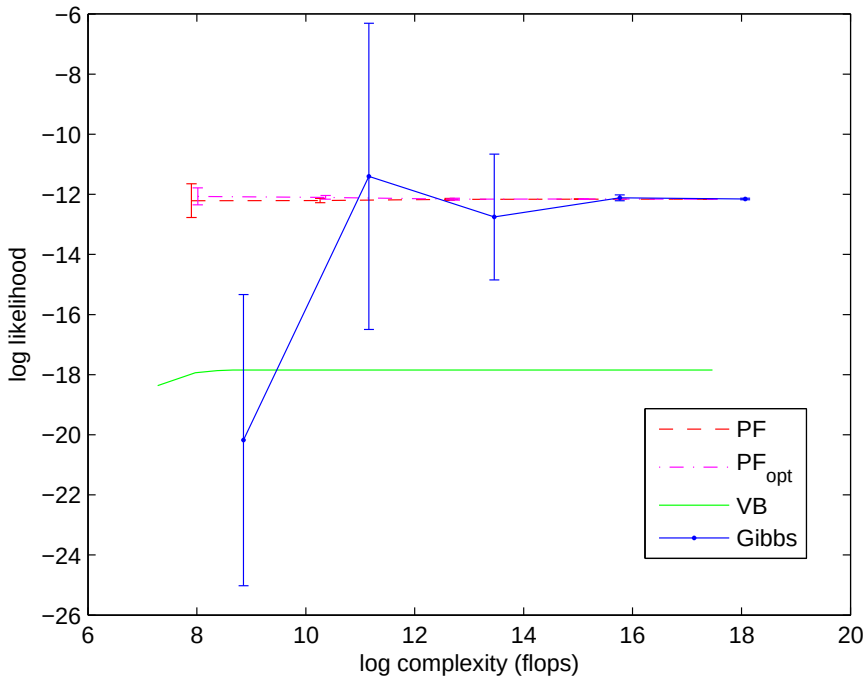


Figure 4.5. Log-likelihood approximations on an Gamma state-space model of length 100 by different methods (bootstrap filter (PF), SIS/R with optimal proposal distribution ($\mathrm{PF}_{\mathrm{opt}}$), Gibbs sampler (Gibbs) and VB). More complex particle filters and Gibbs samplers are obtained by increasing the number of samples while in VB total number of iterations is increased.

The most important reason for the difference between the likelihood estimated by the sampling methods and the variational lower bound is that variational Bayes method discards the correlations between the variables. As we can see in Figure 4.6, the Gibbs samples representing the distributions of consecutive variables in a chain have high correlations between them, whereas VB estimates these variables independently.

Figure 4.7 shows the hyperparameter (a_v and a_z) values that maximise the likelihood estimates of the SIS/R algorithm with optimal proposal distribution and VB. Indeed, the EM methods based on estimates from these methods converge to these maxima, respectively. Although not presented here, the likelihood surface for the Gibbs sampler is similar to that of the optimal SIS/R and the same hyperparameter set maximises the likelihood.

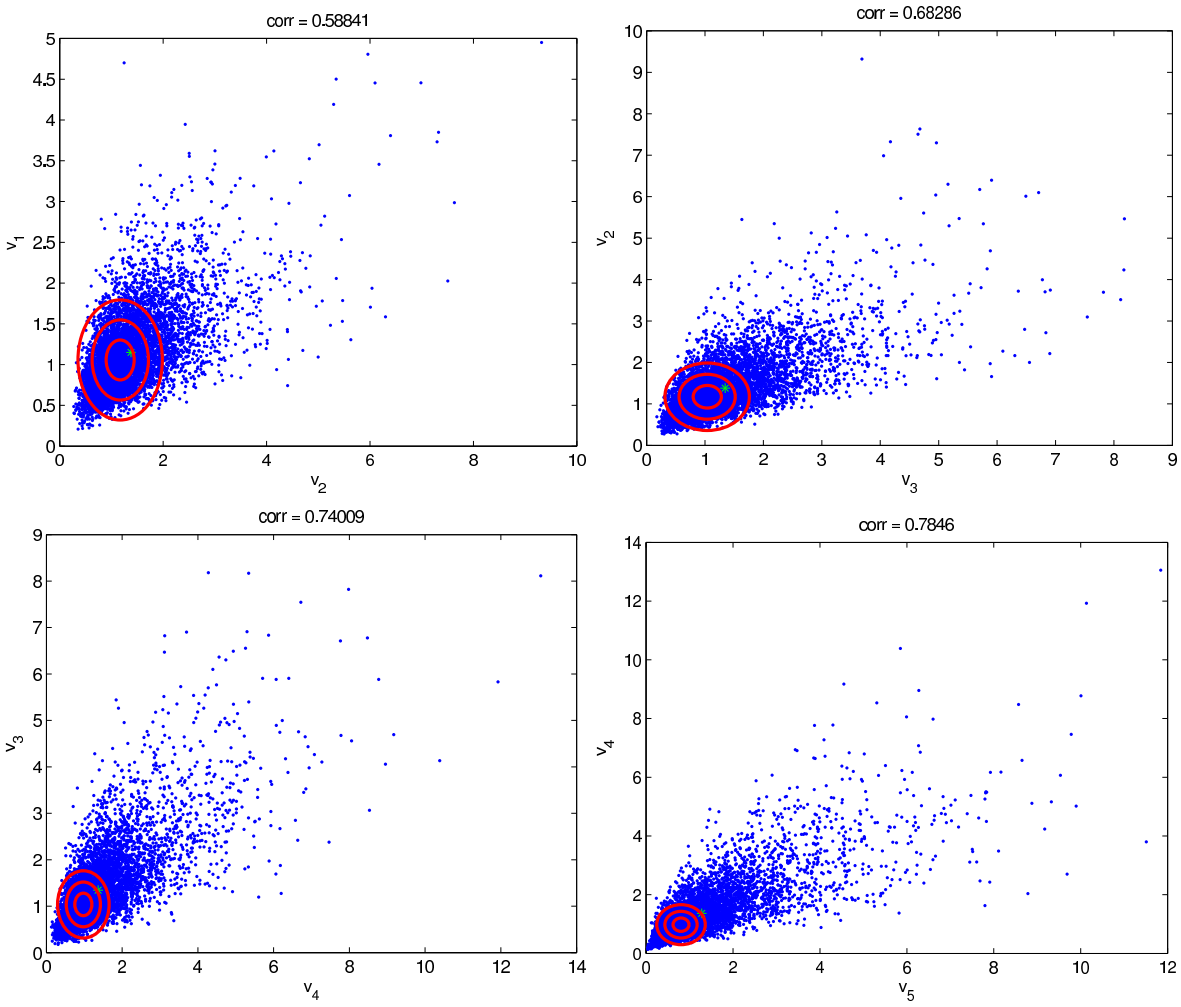


Figure 4.6. Samples drawn by a Gibbs sampler and variational estimates for consecutive variables in a chain of length 5. Variational distributions are represented with horizontal or vertical ellipses up to three standard deviations whereas samples are shown as dots.

4.1.3. Audio Experiments

4.1.3.1. Denoising. Denoising is a special case of source separation with one source and one observation ($M = N = 1$). Estimating the source signal is equivalent to denoising the observation.

We modelled dependencies of the time-frequency coefficients of sources obtained by MDCT with GMCs. This can be done in two ways: either tying coefficients of each frequency bin across time frames (horizontal) or tying frequency coefficients in each

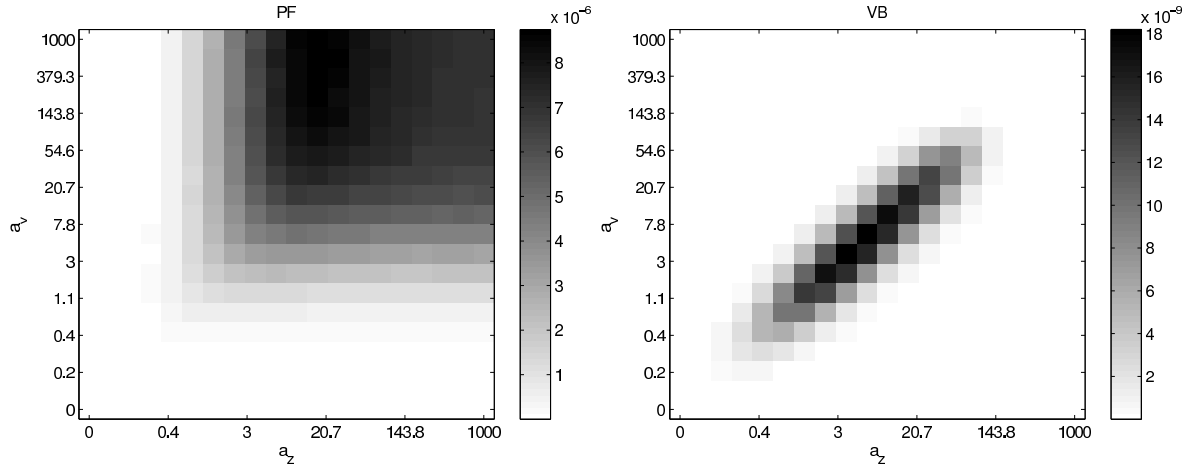


Figure 4.7. Likelihood versus hyperparameters (a_v and a_z) for an observation sequence of length 10. The likelihood functions attained by the SIS/R algorithm with optimal proposal distribution (on the left) and the VB lower bound (on the right) have very different characteristics and maxima.

frame (vertical).

The horizontal model can be summarised as:

$$
\begin{aligned}
z_{\nu,1} &\sim \mathcal{G}(z_{\nu,1}; a_z, b/a_z) && (4.8)\\
z_{\nu,\tau}|v_{\nu,\tau-1} &\sim \mathcal{G}(z_{\nu,\tau}; a_z, v_{\nu,\tau-1}/a_z),\ \tau > 1 && (4.9)\\
v_{\nu,\tau}|z_{\nu,\tau} &\sim \mathcal{IG}(v_{\nu,\tau}; a_v, a_v z_{\nu,\tau}) && (4.10)\\
s_{\nu,\tau}|v_{\nu,\tau} &\sim \mathcal{N}(s_{\nu,\tau}; 0, v_{\nu,\tau}) && (4.11)\\
x_{\nu,\tau}|s_{\nu,\tau}, r &\sim \mathcal{N}(x_{\nu,\tau}; s_{\nu,\tau}, r) && (4.12)\\
r &\sim \mathcal{IG}(r; a_r, b_r) && (4.13)
\end{aligned}
$$

where the indices ν and τ are for the frequency bins and time frames, respectively. The observed signal, $\boldsymbol{x}$, is the sum of the source signal, $\boldsymbol{s}$, and independent white Gaussian noise with variance r.

In order to be able to have an objective measure of success we added noise to the original signals and obtained noisy observation signals. To assess the quality of the

reconstructions, we used the signal-to-noise ratio (SNR) between the original signal and the reconstructed signal:

$$\mathrm{SNR}(\boldsymbol{s}_{\mathrm{org}}, \boldsymbol{s}_{\mathrm{rec}}) = 10 \log_{10} \left(\frac{\|\boldsymbol{s}_{\mathrm{org}}\|^2}{\|\boldsymbol{s}_{\mathrm{org}} - \boldsymbol{s}_{\mathrm{rec}}\|^2} \right)$$

Figure 4.8 presents the log likelihoods and reconstruction SNRs attained by the SIS/R with the optimal proposal distribution using different values for hyperparameters a_v and a_z. The two surfaces are very similar and they have their peaks at the same point. This correlation between the log likelihood and the SNR encourages hyperparameter optimisation using maximum likelihood.

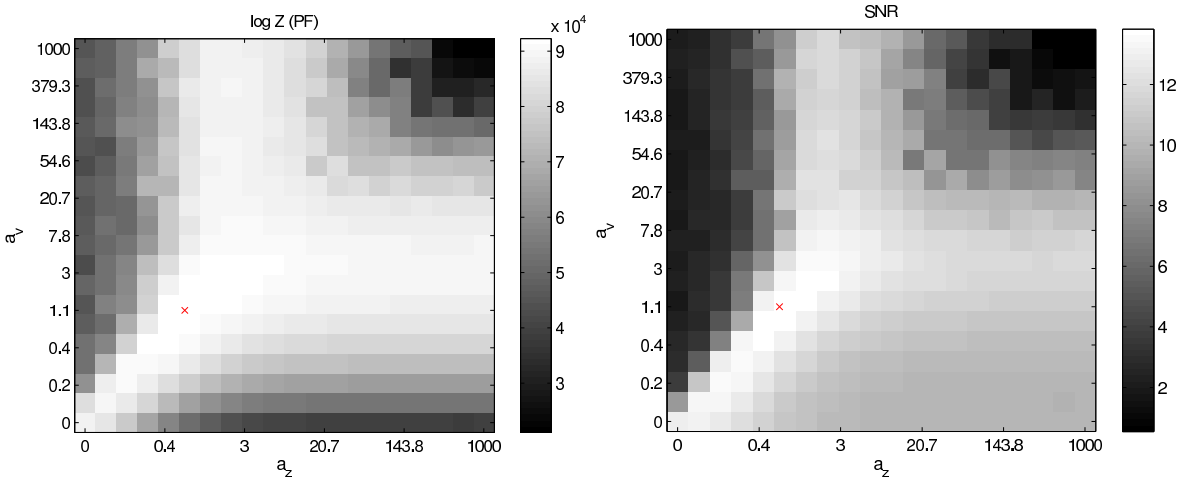


Figure 4.8. Log likelihood and reconstruction SNR values obtained by the SIS/R algorithm using the optimal proposal distribution. The surfaces are evaluated using a fixed value of b ($b = 10^{-4}$).

On the other hand, in the case of VB, there is no correlation between the lower bound of the log likelihood and the SNR (Figure 4.9). Although this method can obtain higher SNR values than the SIS/R algorithm, the SNR surface is neither like the bound surface nor the surfaces obtained by the SIS/R. So, the values of hyperparameters that maximise the SNR cannot be found by optimising an available function.

In these denoising simulations we obtained the noisy signal by adding around 0 dB white noise to a noise-free audio clip. We modelled the source coefficients in the

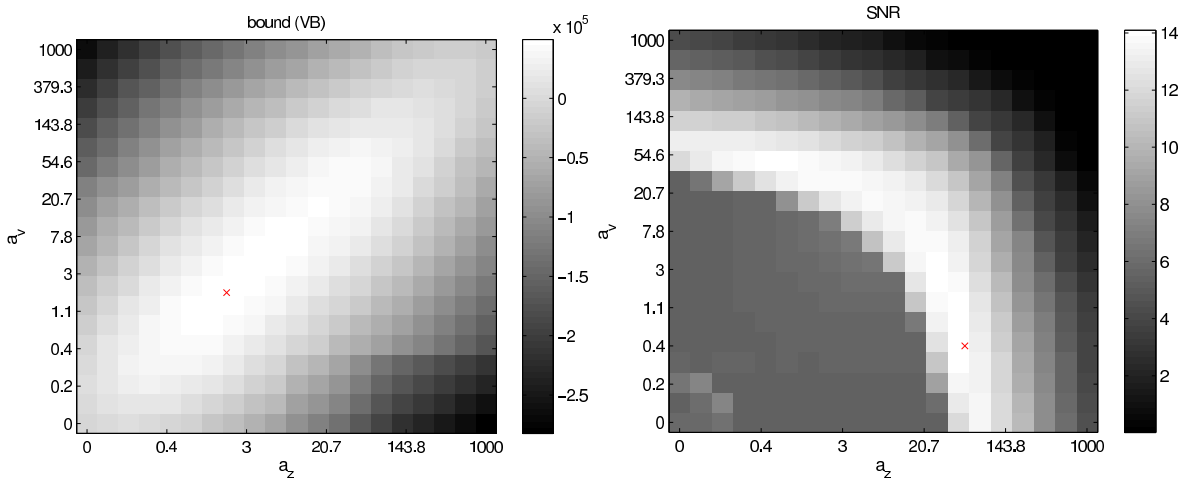


Figure 4.9. Lower bound and SNR values obtained by the variational Bayes method. The surfaces are evaluated using a fixed value of b ($b = 10^{-4}$).

transfer domain, after transforming the signals using MDCT with 512 frequency bins. In Figure 4.10 spectrograms and SNRs of the estimated sources by the three methods are presented. This audio signal is a piano recording and its MDCT coefficients are modelled with horizontal GMCs. In this example, results obtained by variational EM are poor because the hyperparameters optimised by this method did not lead to better results. There are hyperparameter values that result in better reconstructions, but these do not correspond to a local maxima of the lower bound.

4.1.3.2. Single-Channel Source Separation. In single-channel source separation, we try to estimate the N sources that comprise a single observation signal. We again approach the problem in the time-frequency representation and model the variances of the sources with GMCs to ensure dependency along time or frequency axis. The source coefficients are then Gaussian distributed with zero mean: $s_{\nu,\tau} \sim \mathcal{N}(s_{\nu,\tau}; 0, v_{\nu,\tau})$. The observed signal is the sum of N sources: $x_{\nu,\tau} = \sum_{j=1}^{N} s_{\nu,\tau}^{j}$.

In this problem, full conditional distributions of the source coefficients, $p(s_{i,k}|x_k, v_{i,k})$ (of i^{th} source and k^{th} index), are in Gaussian form and their sufficient statistics can be

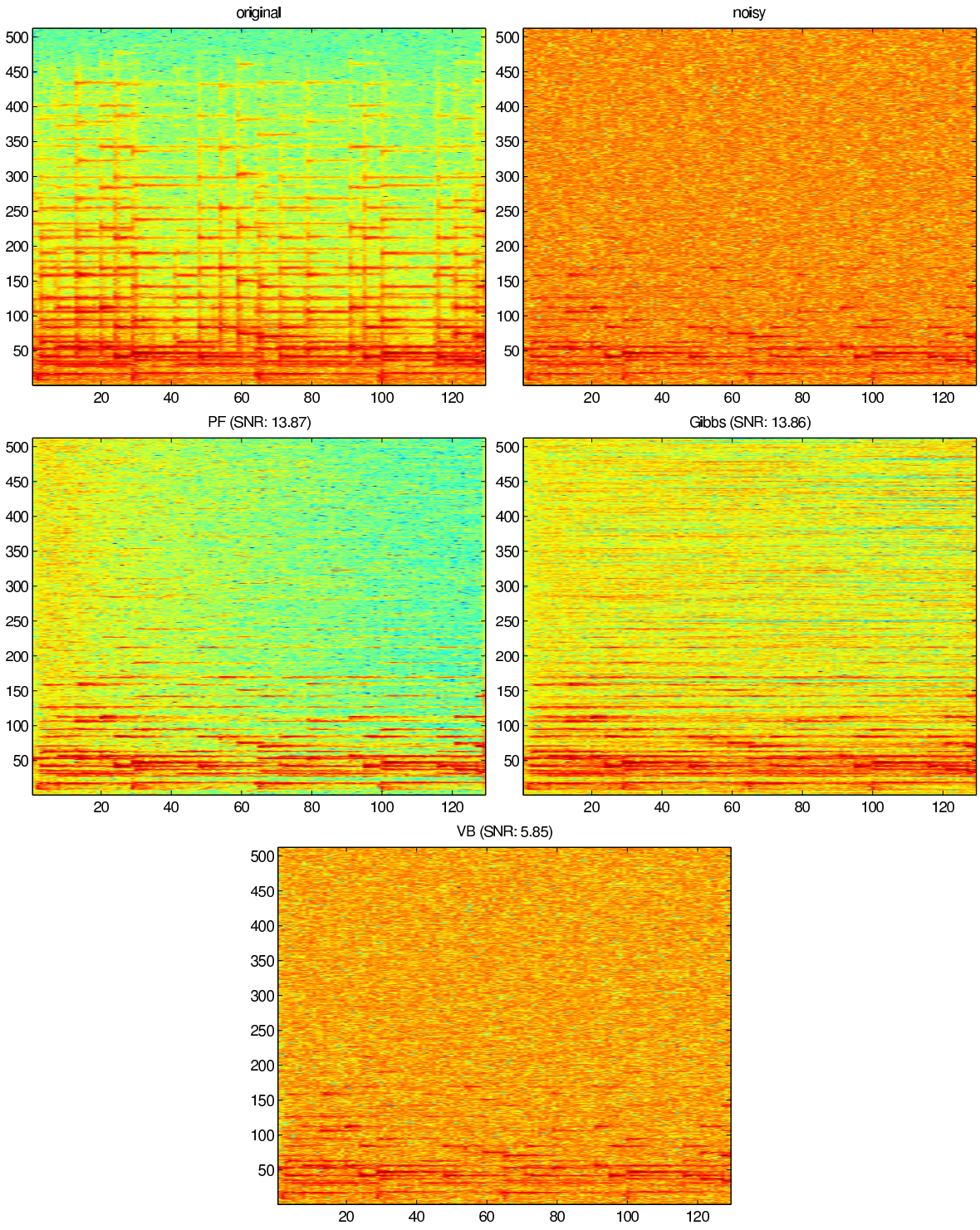


Figure 4.10. Denoising results of a piano recording of which coefficients are modelled with a GMC. The figures on top are the spectrograms of the original and the noisy signals. The others are the estimations of the three inference methods.

evaluated in closed form:

$$\Sigma_{i,k} = v_{i,k}\,(1-\kappa_{i,k}) \tag{4.14}$$

$$m_{i,k} = \kappa_{i,k} x_k \tag{4.15}$$

where $\kappa_{i,k} = v_{i,k} / \sum_j^N v_{j,k}$ represents what portion of the observation can be attributed to the i^{th} source. κ's are called responsibilities in [13] and also known as Wiener filter factors.

Modelling the variances of a source using horizontal GMCs and another with vertical GMCs, we can separate the harmonic components and transients of an observed signal. We mixed tonal audio signals with percussive ones and performed single-channel source separation using VB and Gibbs sampler. Since we have two directions of propagation in this model, we cannot apply classical particle filter methods directly. Tables 4.1 and 4.2 show the results of two single-channel source separation experiments. Here, the performance criteria are the source to distortion ratio (SDR), the source to interference ratio (SIR) and the source to artefacts ratio (SAR), defined as follows [98]

$$\text{SDR} \equiv 10\log_{10}\frac{\|s_{\text{target}}\|^2}{\|e_{\text{interf}}+e_{\text{artif}}\|^2} \tag{4.16}$$

$$\text{SIR} \equiv 10\log_{10}\frac{\|s_{\text{target}}\|^2}{\|e_{\text{interf}}\|^2} \tag{4.17}$$

$$\text{SAR} \equiv 10\log_{10}\frac{\|s_{\text{target}}+e_{\text{interf}}\|^2}{\|e_{\text{artif}}\|^2} \tag{4.18}$$

where an estimate of a source is decomposed into an allowed deformation of the target source, s_{target}, interferences from the other sources, e_{interf} and the artefacts due to the separation algorithm, e_{artif}.

Table 4.1. Single-channel source separation results on a mixture of guitar ("Matte Kudasai") and drums ("Territory")

	$\hat{\boldsymbol{s}}_1$			$\hat{\boldsymbol{s}}_2$		
	SDR	SIR	SAR	SDR	SIR	SAR
VB	-4.74	-3.28	5.67	-1.58	15.46	-1.37
Gibbs	-4.5	-2.62	4.57	1.05	12.46	1.61
Gibbs+MCEM	-4.23	-2.42	4.82	1.34	13.13	1.85

Table 4.2. Single-channel source separation results on a mixture of flute ("Vandringar I Vilsenhet") and drums ("Moby Dick")

	$\hat{\boldsymbol{s}}_1$			$\hat{\boldsymbol{s}}_2$		
	SDR	SIR	SAR	SDR	SIR	SAR
VB	-7.8	-6.22	4.53	-2.35	18.4	-2.25
Gibbs	-8.46	-7.53	6.93	-4.04	14.59	-3.83
Gibbs+MCEM	-7.74	-6.19	4.62	-1.14	16.62	-0.97

In the experiments, we applied VB (with 3000 iterations) and Gibbs sampler (with 5000 samples) using the same set of parameters ($a_v = 3$, $a_z = 3$ and $b = 10^{-4}$). This random choice of the hyperparameters seems suitable due to the good quality of the results. We obtained slightly better results using a Gibbs sampler of which hyperparameters are optimised with MCEM. The initial hyperparameter values are the same as the above. The values converge within 150 iterations of the EM algorithm which makes use of 5000 samples for the E-step. We present the spectrograms of the sources estimated by the Gibbs+MCEM algorithm in Figure 4.11. As expected, the variational EM algorithm converges to a set of hyperparameters that lead to a worse performance, so those results are omitted.

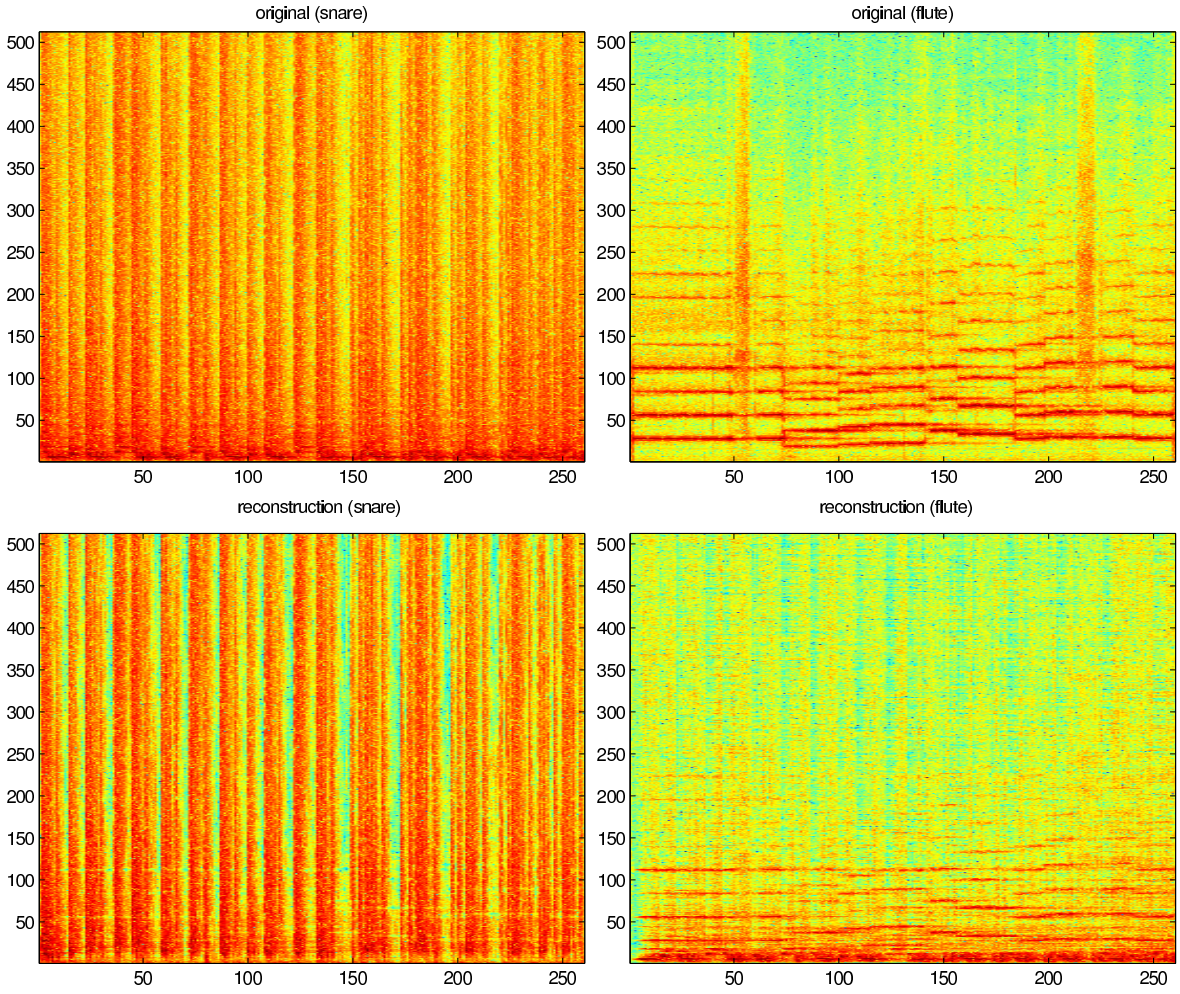


Figure 4.11. The spectrograms of the original sources (top) and the sources estimated by the Gibbs+MCEM algorithm (bottom) in the second single-channel source separation experiment.

4.2. Gamma Markov Random Fields

This section focuses on the optimisation of hyperparameters in GMRFs. As explained before, the probability distribution encoded by a GMRF contains an intractable normalising constant which makes ML estimation difficult. In this section, we will compare three methods that deal with optimisation in non-normalised models: CD, PL, and SM. It turns out that only CD is directly applicable to models containing latent variables, whereas PL and SM are originally proposed for fully-observed models. This is important because our ultimate aim is to optimise the hyperparameters of a (latent) GMRF model directly from data. In an audio processing application where the vari-

ances of source coefficients are modelled with a GMRF, the variables that constitute the GMRF are not observable. In this section, first, we will conduct some experiments on synthetic data in order to investigate the consistency of CD in simpler models and compare its performance to ML, PL, and SM, where applicable. Then, we present results with real data.

The setup of our synthetic data experiments is summarised in Table 4.3, where we test each algorithm for the particular scenarios. We include GMCs, as those are special cases of GMRFs, whose normalising constants can be calculated analytically. The experiments featuring GMCs enable us to compare the estimates of the approximate optimisation methods with the maximum likelihood estimates. In the general case where the maximum likelihood is intractable, we will compare the results to the true hyperparameters that were used to generate the data. We will consider three cases of these models: Fully-observed (FO) models can be optimised using all three approximate optimisation methods so we can compare the performances of these methods. The partially-observed (PO) case is suitable to judge on the performance of contrastive divergence in case of latent variables. Partially-observed GMCs and GMRFs can be marginalised over the observed variables and the resulting model can be optimised using score matching (and maximum likelihood in case of GMCs). But, contrastive divergence is not applicable on this marginalised model. This enables us to observe whether the performance of CD degrades in the presence of latent variables or not.

Table 4.3. Setup for synthetic data experiments. In the experiments two models (Gamma Markov chains (GMC) and Gamma Markov random fields (GMRF)) are considered in three scenarios: fully-observed (FO), partially-observed (PO) and fully-latent (FL). The entries in the table show the optimisation methods applicable for the particular cases.

	GMC	**GMRF**
FO	ML, PL, SM, CD	PL, SM, CD
PO	ML, SM, CD	SM, CD
FL	ML, CD	CD

The case which we call fully-latent (FL) is a more realistic scenario where all the variables of the GMRF are latent and there are observations conditional on the **v** variables. At present, only contrastive divergence can be applied to this case, we compare its estimates to the maximum likelihood estimates in GMCs and to the true parameters in GMRFs. Then, we present results of denoising and single-channel source separation experiments where we performed the inference using the Gibbs sampler and optimised the hyperparameters using CD.

4.2.1. Synthetic Data Results

4.2.1.1. Fully-Observed Models. The model in Figure 4.12 is a GMC with $2N-1$ variables. The hyperparameters $\mathrm{a} = [a_w \; a_e]$ determine the coupling between the variables and b is just a constant.

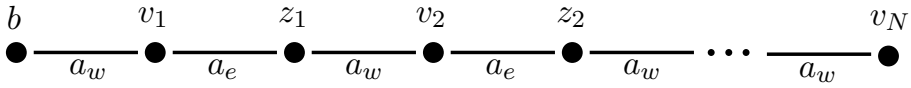
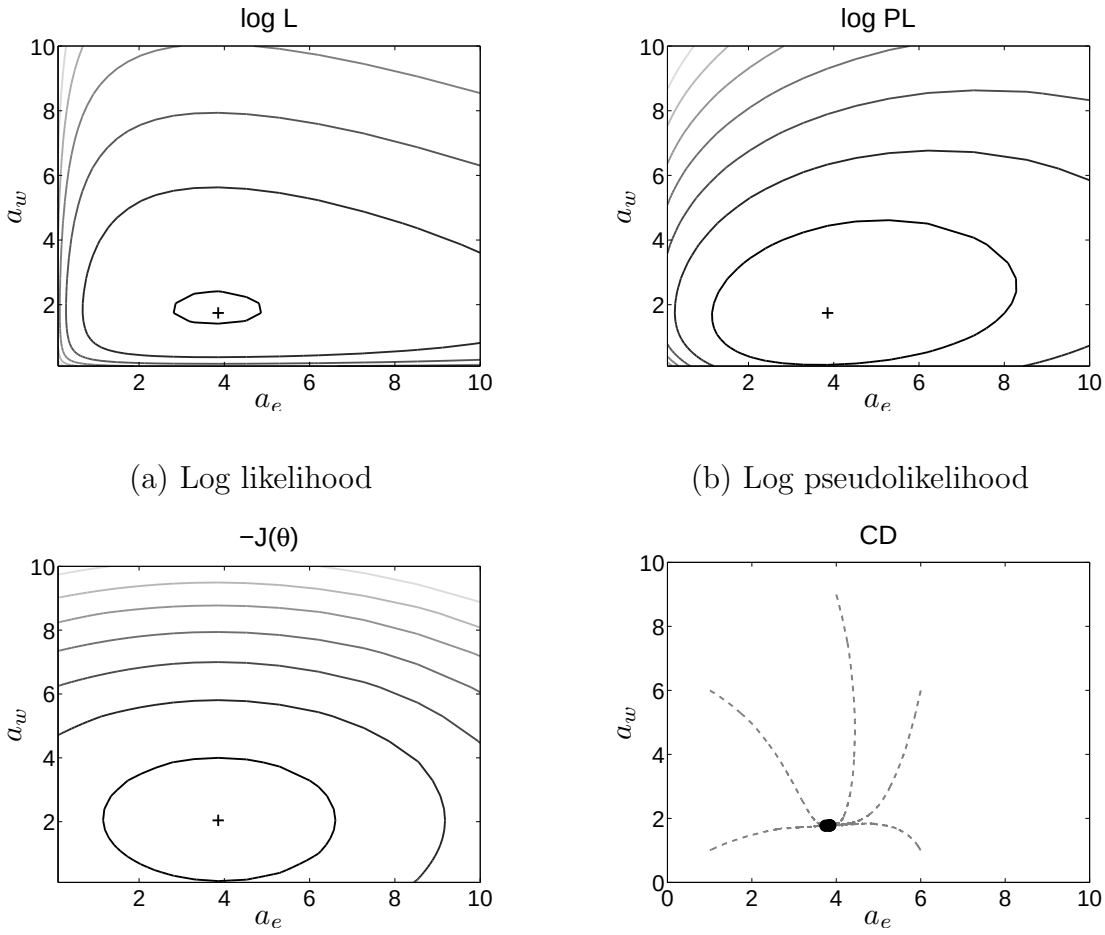


Figure 4.12. A fully observed Gamma chain.

This model can also be interpreted as a chain-structured GMRF with $2N-1$ variables, $\mathcal{V}_{\mathbf{v}} = \{v_1,\ v_2,\ \ldots\ , v_N\}$ and $\mathcal{V}_{\mathbf{z}} = \{z_1,\ z_2,\ \ldots\ , z_{N-1}\}$. It is a special GMRF yet the normalising constant, Z_{a}, can be analytically calculated. However, throughout these experiments Z_{a} will be treated as unknown in order to investigate the accuracy of the approximation methods explained in Section 2.2.

ML estimation of the hyperparameters of a Gamma chain is straightforward. The likelihood surface of a particular chain as a function of the hyperparameters, a_w and a_e, is given in Figure 4.13a. The maximum of this function, depicted by a plus sign (+), can be easily found using Newton's method. Figure 4.13b and 4.13c present the PL and negative score distance[2] surfaces, respectively. The hyperparameters that maximise the pseudolikelihood of this model can again be found using Newton's method, whereas the gradient of the negative score distance has a more complicated form because the

[2]Since this model constitutes a non-negative distribution, the score distance is calculated as explained in [87].



(a) Log likelihood

(b) Log pseudolikelihood

(c) Score values

(d) CD trajectories

Figure 4.13. Optimal hyperparameter values for a fully observed Gamma chain of length 1000. The original hyperparameter values used to generate the chain are $a_w = 2$ and $a_e = 4$. b is fixed at 1.

distance already contains gradients with respect to the observed variables. In this simple model, the hyperparameters can be optimised using gradient ascent, but in more complicated models, surrogate-based optimisation such as response surface methods (RSM) [99] can be used. In Figure 4.13d, we present the optimal hyperparameter values attained by the CD algorithm initialised with different values (paths are presented as dashed lines and optimal values are dots). We used one-step reconstructions of the data and a slowly-decaying gain parameter, $\eta_t = 1/t$, where t is the iteration number. In this experiment, we see that all three approximation methods converge to the true optimum. CD has the simplest optimisation criterion but it needs to generate data and evaluate the gradient for each hyperparameter setting. Moreover, the stopping

criterion and the gain parameter affects the convergence rate.

In Figure 4.14, we present a more general GMRF. Here, we do not know Z_{a}; that means the ML estimates are not available. The algorithmic details of hyperparameter optimisation is similar to the previous case. Figure 4.15 shows that the estimates of the approximation methods are consistent with each other and the hyperparameter values used to generate the data.

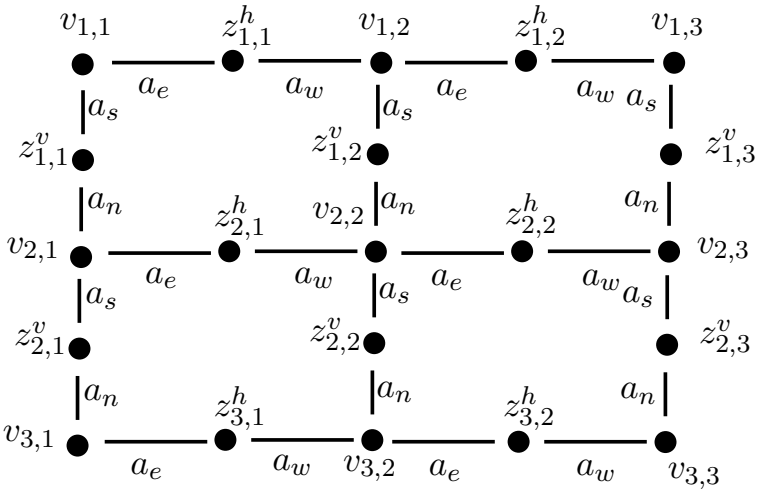
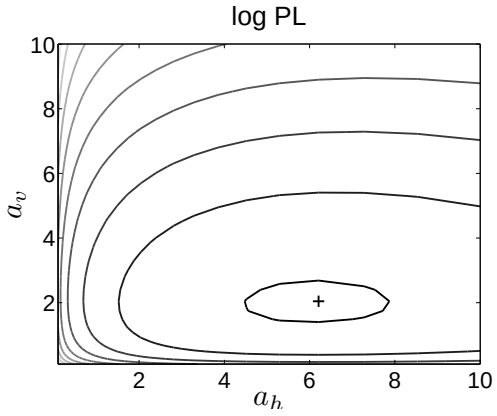


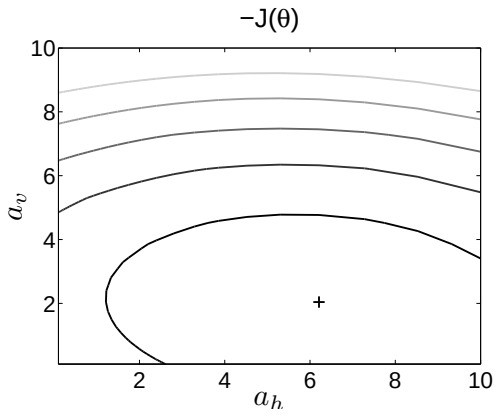
Figure 4.14. A fully observed GMRF. a = $[a_w\ a_e\ a_n\ a_s]$ are the hyperparameters.

We repeated these experiments with data generated with random hyperparameter values and compared the estimates with the true values. In GMCs, maximum likelihood estimates are used as the true values and in general GMRFs where Z_{a} is not available, the values used to generate the data are treated as the true parameters. The average of mean squared errors (MSE) of the estimates in 10 replications of 10 different experiments are presented in Table 4.4.

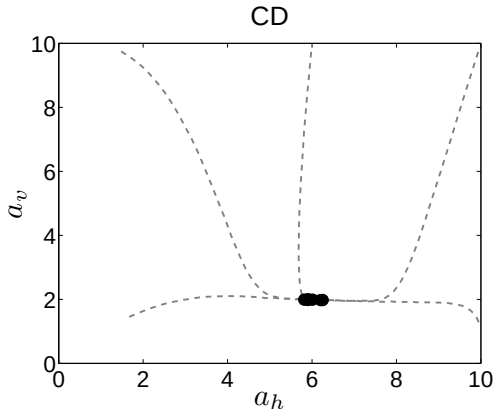
In the fully-observed models, all three approximation methods are successful. However, CD has an important disadvantage: it needs its gain parameter and stopping criterion to be adjusted. A general setting for these criteria may lead to early stopping in some of the experiments, which is the cause of the high standard deviations of CD in Table 4.4.



(a) Log pseudolikelihood



(b) Score values



(c) CD trajectories

Figure 4.15. Optimal hyperparameter values for a fully-observed GMRF of size 50×50. Here, we decrease the number of hyperparameters to two by setting $a_w = a_e \equiv a_h$ and $a_n = a_s \equiv a_v$. The original hyperparameter values used to generate data are $a_v = 2$ and $a_h = 6$.

4.2.1.2. Partially-Observed Models. Now, we consider the case where $\mathbf{z}$ are latent variables. The joint distribution, $p(\mathbf{v}, \mathbf{z}|\mathrm{a})$, is the same as in the previous case and again we treat it as if we do not know the normalising constant.

Two of the approximate optimisation techniques, namely, maximum PL and SM are not designed for models with latent variables. However, this is not a problem for this particular model because the latent variables can be integrated out. The marginal model is not suitable for sampling; so, CD is not applied on the marginal model but the original one. This model constitutes a suitable platform to compare the performance

Table 4.4. MSEs of the estimates averaged over 10 experiments. In each experiment data are generated using random hyperparameter values and each method is run for 10 times with different initial values. Variances of the estimates within the experiment are higher in contrastive divergence (CD) because of its stochastic nature. In pseudolikelihood (PL) and score matching (SM), this is not the case because they are deterministic and the surfaces are smooth. The table summarises averages of 10 such experiments.

	GMC	**GMRF**
PL	0.16 ± 0.09	0.20 ± 0.07
SM	0.24 ± 0.13	0.21 ± 0.10
CD	0.18 ± 0.15	0.26 ± 0.16

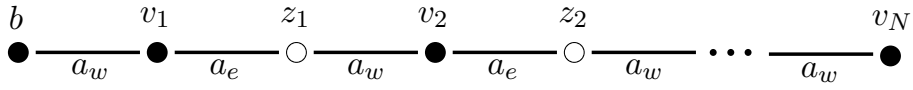


Figure 4.16. A partially observed Gamma chain. Black and white circles denote observed and latent variables, respectively.

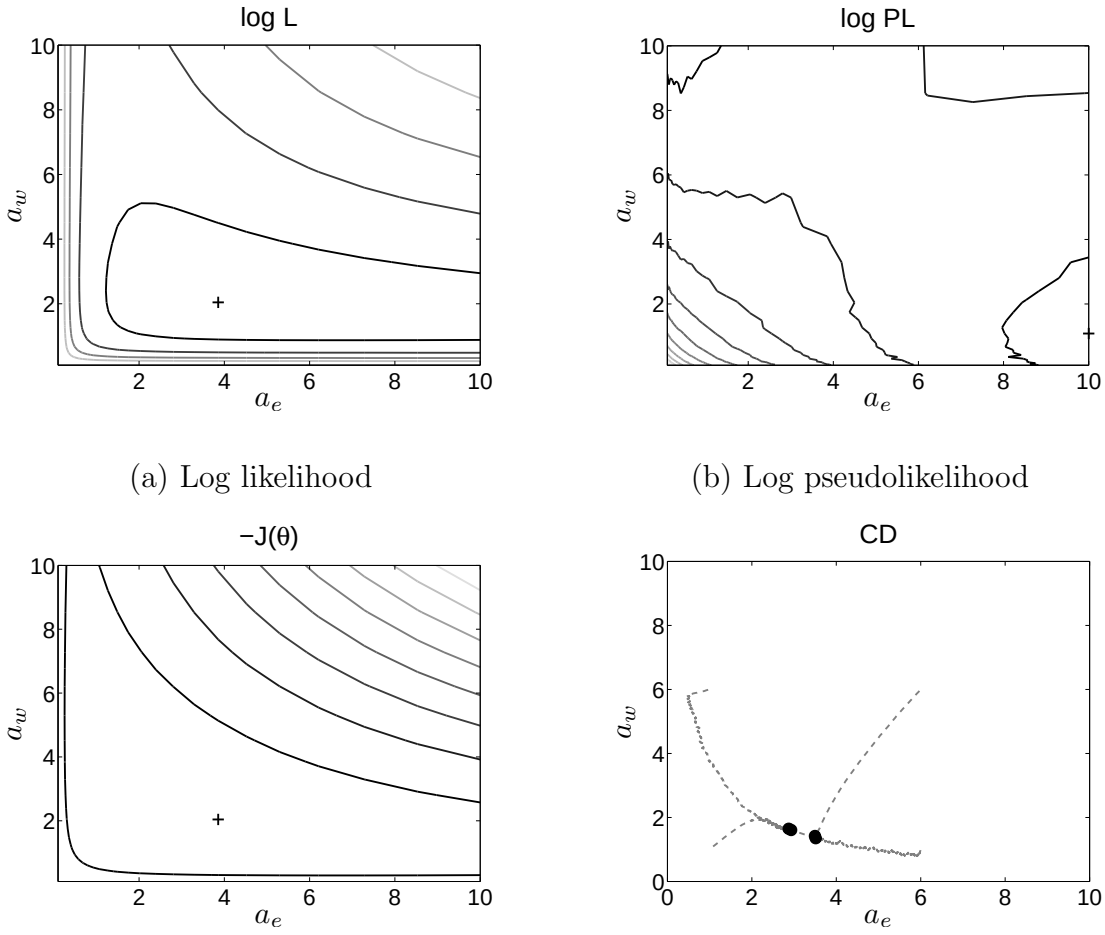
of the latent variable version of CD with the other methods which run on a marginal (so, fully-observed) model.

In Figure 4.17a, the marginal likelihood surface is depicted to set the ground truth for the case in which we suppose the normalising constant is not known and optimise the hyperparameters with approximation techniques. The PL of this partially visible model is difficult to calculate. The full conditionals of the marginal model cannot be evaluated analytically

$$p(v_i|\mathbf{v}_{-i}, \mathrm{a}) = \frac{p(\mathbf{v}|\mathrm{a})}{p(\mathbf{v}_{-i}|\mathrm{a})} = \frac{p(\mathbf{v}|\mathrm{a})}{\int dv_i\, p(\mathbf{v}|\mathrm{a})} = \frac{\pi(\mathbf{v}|\mathrm{a})}{\int dv_i\, \pi(\mathbf{v}|\mathrm{a})} \tag{4.19}$$

It is possible to approximate the integral in the denominator using numerical quadrature methods. However, the optimal hyperparameter values found this way diverge from the true values as can be seen in Figure 4.17b. This may be because of inconsistency of PL for this model. A less likely cause may be the approximation error. Apart

from having slightly complicated terms, score matching remains consistent for this model. The surface of negative of the distance values versus the hyperparameter values are depicted in Figure 4.17c. Contrastive divergence suffers from early convergence in some instantiations as can be seen in Figure 4.17d, which shows the optimisation trajectories obtained from different initialisations.



(a) Log likelihood

(b) Log pseudolikelihood

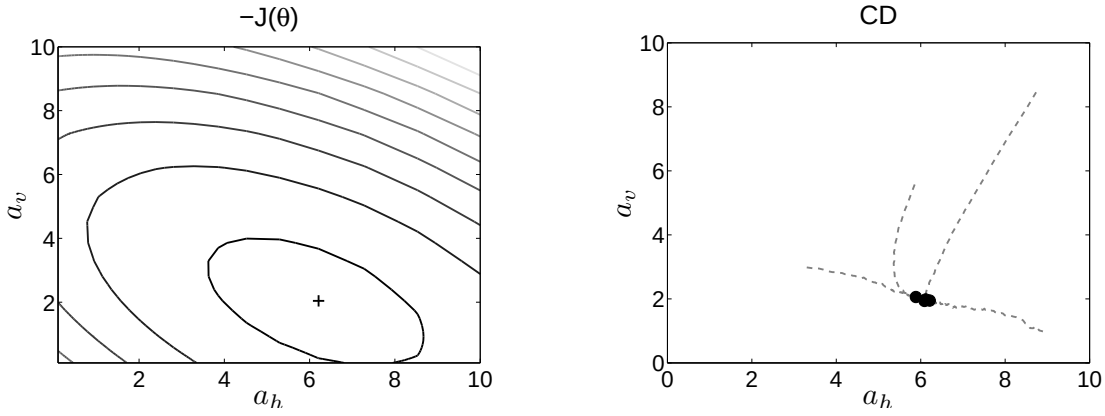
(c) Score values

(d) CD trajectories

Figure 4.17. Optimal hyperparameter values for a partially-observed Gamma chain of length 1000. The original hyperparameter values used to generate the chain are $a_w = 2$ and $a_e = 4$. b is fixed at 1.

We performed the same experiment on the partially-observed GMRF, which is based on the model in Figure 4.14 with the difference that $\mathbf{z}$ variables are latent. As we can see in Figure 4.18, score matching and contrastive divergence estimates are consistent with each other and with the true hyperparameter values that are used to generate data. However, in general, contrastive divergence may stop before converging

to the true values. Therefore, its parameters should be adjusted well. The average mean squared errors obtained from batch simulations are presented in Table 4.5. Note that PL was not applicable to this scenario.



(a) Score values (b) CD trajectories

Figure 4.18. Optimal hyperparameter values for a partially-observed GMRF of size 50×50. Here we decrease the number of hyperparameters to two by setting $a_w = a_e \equiv a_h$ and $a_n = a_s \equiv a_v$. The original hyperparameter values used to generate data are $a_v = 2$ and $a_h = 6$.

Table 4.5. MSEs of the estimates averaged over 10 experiments. Each experiment is repeated 10 times with different initial values and MSEs are calculated. The entries in the table are averages of these MSEs.

	GMC	**GMRF**
SM	0.92 ± 0.40	1.39 ± 0.68
CD	0.80 ± 0.62	2.31 ± 1.63

4.2.1.3. Fully-Latent Models With Gaussian Observations. The model in Figure 4.19 introduces new variables s_i to the previous model. Each s_i is conditionally Gaussian with zero mean and variance, v_i. We only observe s_i's and all the variables that constitute the GMC are latent.

In this model, we cannot marginalise $p(\mathbf{s})$, so, do not have a way to optimise the hyperparameters, a, using score matching or maximum pseudolikelihood. However,

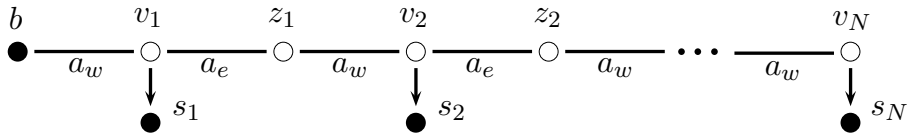


Figure 4.19. A Gamma chain with Gaussian observations.

contrastive divergence is applicable to this case as explained in Section 2.2.2. The expression for the gradient is given as

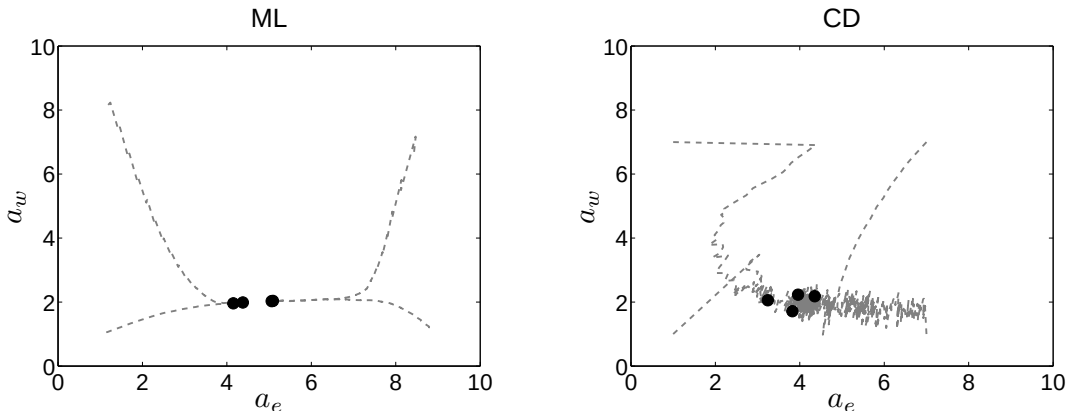
$$\mathrm{CD}_n = -\left\langle \log \pi(\mathbf{s}, \mathbf{v}, \mathbf{z}; \mathrm{a}) \right\rangle_{p(\mathbf{v},\mathbf{z}|\mathbf{s},\mathrm{a}_k)} + \left\langle \log \pi(\mathbf{s}, \mathbf{v}, \mathbf{z}; \mathrm{a}) \right\rangle_{p^n(\mathbf{s},\mathbf{v},\mathbf{z}|\mathrm{a}_k)}$$

where a_k denotes the current value of the hyperparameters, $p^n(\mathbf{v}, \mathbf{z}|\mathbf{s}, \mathrm{a}_k)$ is the distribution of n-step reconstructions. The expectation in the first term is approximated using samples drawn from $p(\mathbf{v}, \mathbf{z}|\mathbf{s}, \mathrm{a}_k)$.

In Figure 4.20, we compare ML and CD estimations of hyperparameters of a chain with $N = 1000$. We used MCEM with 500 samples to obtain the ML solutions. Similarly, CD uses 500 samples and 1-step reconstructions. Although both methods are stochastic, we observe that CD follows more distorted paths; but, the results are close to those of ML. Again, the gain parameter and the stopping criterion should be adjusted. Table 4.6 shows the average of mean squared errors obtained from 10 simulations. The fully-latent GMRF here is again based on the model in Figure 4.14. All the variables in the GMRF are latent and we have observations, $s_{\nu,\tau}$, which are conditionally Gaussian, $\mathcal{N}(s_{\nu,\tau}; 0, v_{\nu,\tau})$. Only CD can be applied to this case.

Table 4.6. MSEs of the estimates averaged over 10 experiments. In each experiment, data are generated from the model with random hyperparameter values. Experiments are repeated 10 times with different initial values and MSEs are calculated. the average of these MSEs are presented in the table.

	GMC	**GMRF**
CD	0.86 ± 0.81	2.11 ± 1.26



(a) ML trajectories (b) CD trajectories

Figure 4.20. ML and CD estimates for a Gamma chain of length 1000. The chain was generated using the hyperparameter values $a_w = 2$ and $a_e = 4$. b is fixed at 1.

4.2.2. Audio Experiments

We performed two types of audio experiments as in Section 4.1: denoising (with stationary and non-stationary noise) and single-channel source separation. In these experiments, we modelled the prior distributions of time-frequency coefficients of audio sources with GMRFs. Inference is carried out using the Gibbs sampler and the hyperparameters of the model are optimised using contrastive divergence. The audio model does not make any assumptions about the structure of the dependencies (e.g., vertical or horizontal) and is fully adaptive.

Throughout these audio experiments, we used audio recordings of 6-10 seconds duration and transferred them into MDCT domain with 512 frequency bins. MDCT is an orthogonal transformation, so, denoising and source separation problems are equivalent in time and transfer domains.

The first experiment is denoising where the noise is white Gaussian. The variances of source coefficients are modelled with a GMRF, the coefficients are zero mean Gaussians conditional on these variances and we have the noisy observations of these

coefficients

$$s_{\nu,\tau} \sim \mathcal{N}(s_{\nu,\tau}; 0, v_{\nu,\tau}) \tag{4.20}$$

$$x_{\nu,\tau} \sim \mathcal{N}(x_{\nu,\tau}; s_{\nu,\tau}, r) \tag{4.21}$$

where r is the noise variance, ν and τ denotes frequency and time indices, respectively.

During the inference using the Gibbs sampler, we run an optimisation step at every N_i iterations. The objective function that CD minimises becomes

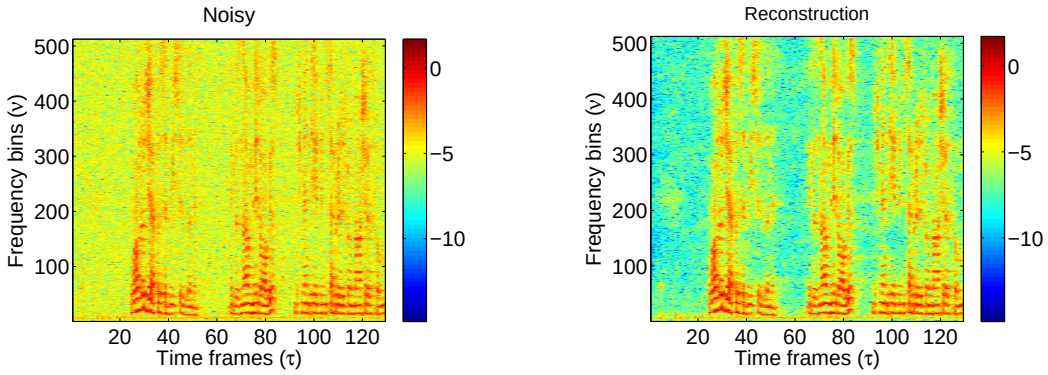
$$\mathrm{CD}_n = -\left\langle \log \pi(\mathbf{x}, \mathbf{s}, \mathbf{v}, \mathbf{z}, r; \mathrm{a}) \right\rangle_{p(\mathbf{s},\mathbf{v},\mathbf{z},r|\mathbf{x},\mathrm{a}_k)} \tag{4.22}$$

$$+ \left\langle \log \pi(\mathbf{x}, \mathbf{s}, \mathbf{v}, \mathbf{z}, r; \mathrm{a}) \right\rangle_{p^n(\mathbf{x},\mathbf{s},\mathbf{v},\mathbf{z},r|\mathrm{a}_k)} \tag{4.23}$$

in this problem. The N_i samples drawn by the Gibbs sampler are used to approximate the expectation in the CD gradient. Then, 1-step reconstructions for both observed and latent variables are generated to evaluate the second term. The pseudocode of the overall algorithm is given in Appendix C.

In Figure 4.21, we present the result of a denoising experiment. On the left, we see the spectrogram of the observed signal, which is obtained by adding noise onto a clean speech recording in order to be able to judge the denoising performance. The spectrogram of the reconstruction, i.e., the mean of $p(\mathbf{s}|\mathbf{x}, \mathrm{a})$ that we infer, is depicted on the right. The reconstructed signal (by inverse MDCT) is clean and sounds natural, i.e., no artefacts exist. Another thing to mention is that, the optimal hyperparameters that were found by CD have high vertical and low horizontal values as one would expect looking at the spectrogram.

In Table 4.7, we present the results of three denoising experiments including the one above. Here, we compare the performance of two models based on GMRFs and GMCs according to the reconstruction SNRs, the ratio of the power of the original signal to that of the estimation error, of their reconstructions. Two audio models can



(a) Noisy audio clip

(b) Reconstructed audio clip

Figure 4.21. Result of denoising a speech recording with white Gaussian noise. The original source is the first 8 seconds of "The Gift" by "The Velvet Underground" sampled at 8kHz.

be obtained using GMCs: a horizontal model to couple the variance variables along the time frames and a vertical model to tie them along the frequency axis. Table 4.7 shows that slightly better results are obtained by GMCs in terms of SNR. But we are indecisive about the significance of these numbers, especially after listening to the results. On the contrary, the reconstructions by GMCs contain some artefacts, at higher frequencies in general, that are not found in the reconstructions by GMRFs. Considering that GMC results are the best of two models (vertical (V) and horizontal (H)) and hyperparameter optimisation in these methods can be successfully performed using MCEM, we can say that the optimisation of hyperparameters of a GMRF corresponds to model selection between horizontal and vertical models. Of course, the audio model with GMRFs is more general than these two models; it can define a loose dependency along one axis while stressing the dependencies along the other.

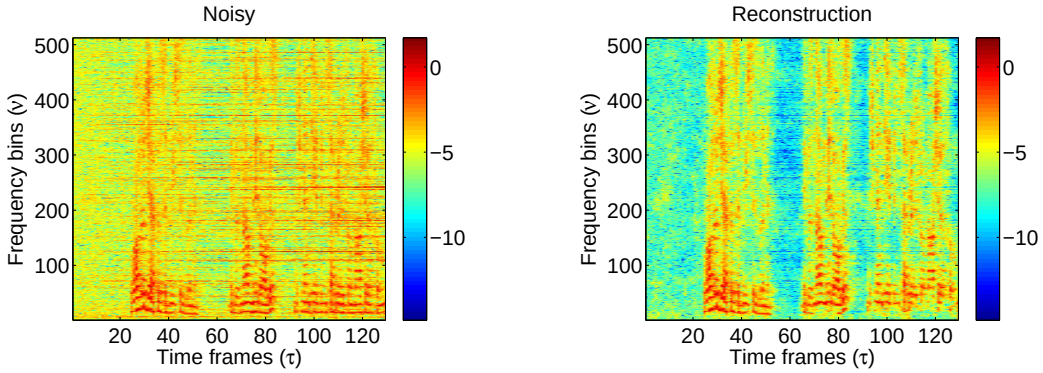
We also performed denoising in the existence of non-stationary noise. We modelled the noise parameters with a horizontal GMC. We generated noise using this model and added onto the original signal:

$$x_{\nu,\tau} \sim \mathcal{N}(x_{\nu,\tau}; s_{\nu,\tau}, r_{\nu,\tau}) \tag{4.24}$$

Table 4.7. Reconstruction SNRs assessed in the denoising of three artificially noised audio signals. The second column shows the amount of noise added. Third and fourth columns contain the reconstruction SNRs of the GMC and GMRF models, respectively. The letters in the parentheses denote the type of the GMC used: v for vertical, h for horizontal.

	SNR	GMC SNR	GMRF SNR
Speech	17.5	20.79 (V)	20.77
Piano	3	8.92 (H)	8.79
Guitar	3.9	8.66 (H)	8.53

The result of this experiment is presented in Figure 4.22. Here, the hyperparameters of the source model are optimised using CD. The Gamma chain that constitutes the noise model is a normalised model and its optimisation is carried out using MCEM as was done in Section 4.1.

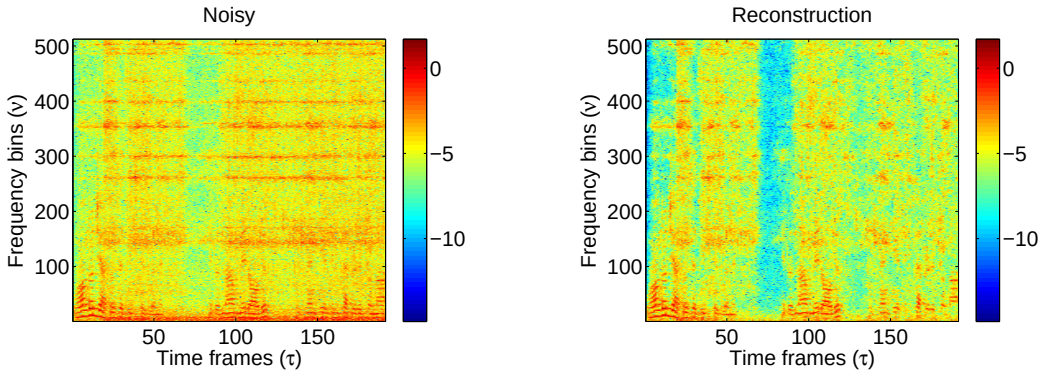


(a) Noisy audio clip

(b) Reconstructed audio clip

Figure 4.22. Denoising a speech recording with non-stationary noise (1.4dB SNR). The reconstruction SNR is 6.4dB.

With the same source and noise models, we separated drill noise from speech. The two signals are artificially mixed, and the spectrogram of this mixed signal is given in Figure 4.23a. The horizontal noise model is not sufficient for the drill noise, because it has both horizontal and vertical dependency in its spectrogram. The reconstruction in Figure 4.23b shows that at least the horizontal components of the noise are removed.



(a) Noisy audio clip (b) Reconstructed audio clip

Figure 4.23. Denoising a speech recording with drill noise (0dB SNR). The reconstruction SNR is 3.2dB.

We performed some single-channel source separation experiments on music signals to separate tonal and percussive components. We used a GMRF based audio model for the tonal source and a vertical GMC model for the percussion. The overall model is similar to the one above. The hyperparameters of the GMRF are estimated using CD. Starting from random values, the GMRF ended up with prominent horizontal hyperparameters, which is expected in a tonal source. GMC optimisation is carried out using MCEM. In Tables 4.8[3] and 4.9[4] , we compared the results with those in Section 4.1.3.2 where the sources are modelled with one horizontal and one vertical GMCs.

Table 4.8. Single-channel source separation results on a mixture of guitar and drums

	$\hat{\mathbf{s}}_1$			$\hat{\mathbf{s}}_2$		
	SDR	SIR	SAR	SDR	SIR	SAR
GMC	-4.23	-2.42	4.82	1.34	13.13	1.85
GMRF	-0.85	3.5	2.74	7.67	10.61	11.11

The reconstructions of the tonal sources (denoted as $\hat{\mathbf{s}}_1$) with GMRFs are higher

[3]A mixture of 6-second excerpts from “Matte Kudasai” by King Crimson and “Territory” by Sepultura sampled at 16kHz.

[4]A mixture of 8-second excerpts from “Vandringar I Vilsenhet” by Änglagård and “Moby Dick” by Led Zeppelin sampled at 16kHz.

Table 4.9. Single-channel source separation results on a mixture of flute and drums

	$\hat{\mathbf{s}}_1$			$\hat{\mathbf{s}}_2$		
	SDR	SIR	SAR	SDR	SIR	SAR
GMC	-7.74	-6.19	4.62	-1.14	16.62	-0.97
GMRF	-4.27	-1.61	3.0	5.59	19.82	5.8

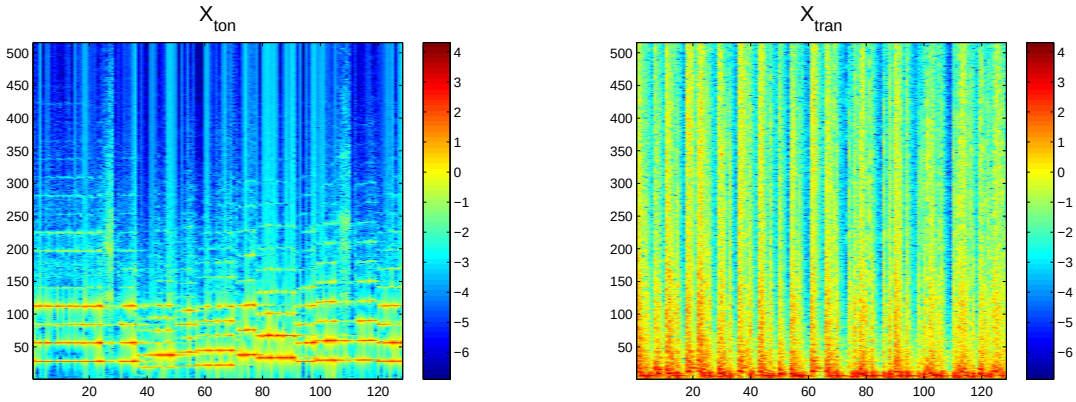
than the GMCs in terms of SDR and SIR. Actually, the reconstructions are not very different from each other perceptually, but the GMRF results sound more natural and without artefacts.

4.3. NMF Using GMCs

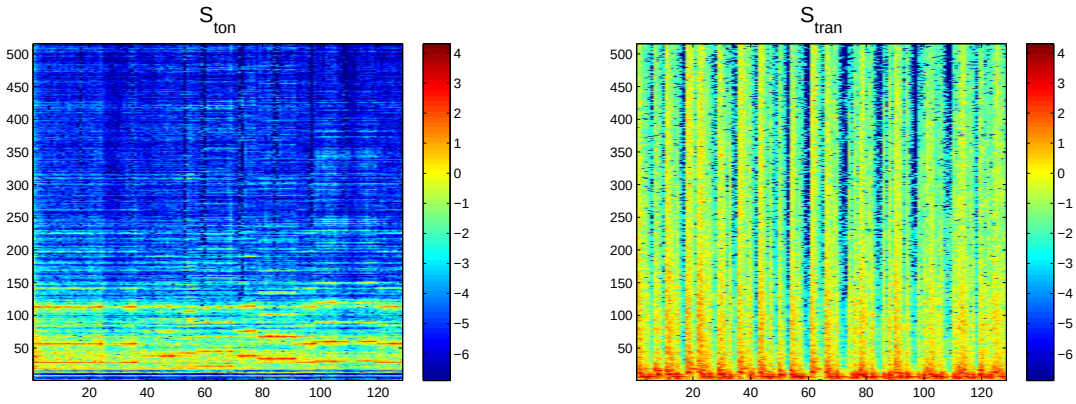
In our experiments regarding NMF, we used the recordings from the previous sections. Magnitude spectrograms are obtained using STFT, with non-overlapping windows of length 1024. Consequently, we work on spectrograms with 513 frequency bins and roughly 120-140 time frames. Phases of the original signal are stored and added to each estimated source before reconstruction.

The unsupervised NMF method infers the posterior distributions of the $\mathbf{T}$, $\mathbf{V}$ and $\mathbf{S}_i$, $i = 1 : I$ matrices using the Gibbs sampler. The hyperparameters of the model are also estimated during the inference, using the Metropolis algorithm with Gaussian proposal distributions. The only input to the model, apart from $\mathbf{X}$, are the number of components for each source: I_{ton}, I_{perc} and I_{bass}. The model is based on Poisson observations and needs integer-valued $\mathbf{X}$ matrices. Magnitude spectrograms of audio signals have a large number of elements between zero and one. In order to decrease the effect of round-off error, we multiply the $\mathbf{X}$ matrix with a constant C and round. Estimated components are divided to C accordingly.

We made use of both manually mixed percussive and tonal signals and original recordings where we do not have the individual sources. First type of examples enables us to assess the performance using objective criteria such as signal to distortion ratio



(a) Tonal Signal (b) Percussive Signal



(c) Tonal Estimate (d) Percussive Estimate

Figure 4.24. Sources estimated from a mixture of flute and drums recording.

(SDR), signal to interference ratio (SIR) or signal to artefacts ratio (SAR) [98]. For the latter type, we judged on the performance perceptually.

In Figure 4.24, we present the spectrograms of the separated sources from a mixed signal of flute and drums. On the top row, spectrograms of the original sources are given. Below them are the corresponding estimation obtained by our extended NMF model. In this experiment we used ten components for the tonal source ($I_{ton} = 10$), six components for the percussive sources ($I_{bass} = I_{perc} = 3$).

We compared the performances of our two models with the previous two models in Tables 4.10 and 4.11. The results show that our extended model (UNMF-e) performs better separation than the other models. GMRF results are also successful and has the

highest SAR values in one of the experiments. According to the objective performance criteria, our simpler model (UNMF) performs very poorly. However, by listening to the reconstructed signals, we see that the problem mainly lies in assigning the bass drum to the wrong source.

Table 4.10. Single-channel source separation results on a mixture of guitar and drums.

	$\hat{\mathbf{s}}_{ton}$			$\hat{\mathbf{s}}_{tran}$		
	SDR	SIR	SAR	SDR	SIR	SAR
GMC	-4.23	-2.42	4.82	1.34	13.13	1.85
GMRF	-0.85	3.5	2.74	7.67	10.61	11.11
UNMF	-5.02	-4.40	9.44	-1.58	13.67	-1.26
UNMF-e	-0.32	5.84	1.88	7.46	13.53	8.89

Table 4.11. Single-channel source separation results on a mixture of flute and drums.

	$\hat{\mathbf{s}}_{ton}$			$\hat{\mathbf{s}}_{tran}$		
	SDR	SIR	SAR	SDR	SIR	SAR
GMC	-7.74	-6.19	4.62	-1.14	16.62	-0.97
GMRF	-4.27	-1.61	3.0	5.59	19.82	5.8
UNMF	-13.82	-13.48	11.11	-7.26	-2.69	-0.84
UNMF-e	6.03	15.50	6.67	15.72	24.15	16.41

5. CONCLUSIONS

In this thesis, we have introduced GMCs and GMRFs, which are generic and flexible models to define prior structures that capture general properties of audio signals. We modelled the variances of time-frequency coefficients of audio signals with these models to ensure positive correlation between consecutive coefficients. By this way, temporal and/or spectral smoothness of time-frequency representations is conserved. We also used GMCs to model the dependencies in excitation and template vectors in source separation using NMF.

GMCs couple the variances in one direction of the spectrogram. In tonal audio signals there is a high correlation between the variances along the time axis, whereas in percussive signals the correlation is higher along the frequency axis. It is suitable to model these signals with horizontal and vertical GMCs, respectively. But, in general, adjacent coefficients are correlated across both time and frequency axes. Since GMRFs couple the variance variables in both directions, they are suitable to model general audio sources.

Both GMCs and GMRFs are conditionally conjugate when they are used to model the variances of an audio signal along with a Gaussian generative model for the source coefficients. As a result, inference of the latent variables in this model is very easy using variational Bayes and the Gibbs sampler. For GMCs, it is also possible to regard the model as a dynamic system and apply Sequential Monte Carlo methods for the inference.

While the dependency is existent along both axes, the strength of this dependency is not a priori known. This fact makes the optimisation of the coupling parameters of the models crucial during the inference.

In GMCs, the optimal model can be obtained using the Monte Carlo EM algorithm. The run time of the algorithm is generally several hours. Sequential Monte

Carlo performs as well as the Gibbs sampler, but with fewer samples. One problem with SMC methods is to adapt a propagation scheme due to the offline nature of the problem as we handle. Optimisation with variational Bayes is not consistent. Although VB works very well and fast when it runs on the "correct" parameters, the optimised hyperparameters are not guaranteed to increase the performance, because the optimisation of the variational lower bound does not correspond to the optimisation of the true likelihood.

GMRFs encode non-normalised probability distributions; so, optimisation of their hyperparameters cannot be directly carried out using a straightforward maximum likelihood estimation. We have reviewed three methods that deal with such cases: pseudolikelihood, score matching and contrastive divergence. Except for contrastive divergence, these former methods are not designed for models containing latent variables.

We performed several source separation experiments where the variances of source coefficients are modelled with GMRFs and sources are assumed conditionally Gaussian. During the inference, the hyperparameters are optimised using CD. The results show that CD estimates are consistent with our expectations and lead to successful reconstructions. However, there are some disadvantages of CD, as well. For example in the case of latent variables, at each gradient evaluation we have to run full MCMC iterations to approximate an expectation. This makes the overall method too costly. In addition, selection of the gain parameter and the stopping criterion is important in the success of the optimisation. It would be better to find current optimal values at the end of each iteration rather than evaluating the gradient. For this reason, we will further investigate pseudolikelihood and score matching methods to adapt them to latent variable models.

Our audio model based on GMRFs captures the spectral dependency of the source coefficients and is flexible such that it can emphasise the prominent dependency structures of different audio sources. Without any information about the nature of the audio source, we can find the appropriate dependency model for that source. With the audio models that define more specific dependency structures, such as GMCs with

vertical and horizontal structure, a model selection step is needed. However, GMRFs are flexible so that model selection is not needed.

We also proposed a model based on GMCs to separate percussive and tonal sources from single-channel audio signals via partitioning the spectrogram using NMF. The model makes use of some basic properties of the spectral behaviour of musical instruments. The separation process is totally unsupervised, i.e. there is no need to learn the template vectors from training data or manual assignment of each component to sources. The only parameters that should be set are the number of components each source will have. However, this is not a critical decision. Setting a higher number of components to a source than that is actually needed does not change the performance very much.

The inference of the parameters of the model is carried out using the Gibbs sampler. Good results can be obtained even using 100 MCMC steps. Since the hyperparameters are estimated using Metropolis algorithm, it is better to use more steps if the rejection rate is high. Our method does not include model selection, e.g., for the number of components. Estimation of marginal likelihood can be costly with the Gibbs sampler. Variational Bayes can be used for faster inference. In that case, model selection can be carried out using the variational lower bound of the marginal likelihood.

To summarise, in this thesis, we introduced two models, GMCs and GMRFs, in order to define prior distributions for non-negative and locally dependent variables. We built two audio source models in which the dependencies among the time-frequency coefficients are captured at the variance level, by modelling the variances using GMCs and GMRFs. We obtained successful results in denoising and single-channel source separation problems. With GMRFs, it is possible to get slightly better results according to both objective and subjective criteria. For single-channel source separation, we proposed another model which makes use of Gamma and GMC prior distributions in an NMF setting. With this model, we got even better quality reconstructions. In addition, the computational complexity of inference on this model is much less than

the other two models. But, still, the first two models are general audio source models and their applicability is not limited to source separation.

The optimisation in GMRFs is an important problem. In this thesis, it is carried out using CD, since it is the only method that can be used in models containing latent variables. It would be very profitable to extend SM for such cases, because it is a locally consistent estimator. For CD, consistency cannot be taken for granted.

GMRFs are very general models. The topology we used in this thesis captures local dependencies. Other topologies which introduce other dependencies, e.g., between the harmonics, can also be considered for audio modelling. GMRFs can also be used to model interdependent template vectors in source separation using NMF.

APPENDIX A: Standard Distributions

The probability density functions, sufficient statistics, and entropies of the standard distributions used in this thesis are given below.

Gaussian.

$$\begin{aligned}
\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) &\equiv \exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}+\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}-\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\frac{1}{2}\log|2\pi\boldsymbol{\Sigma}|\right)\\
\langle\boldsymbol{x}\rangle_{\mathcal{N}} &= \boldsymbol{\mu}\\
\langle\boldsymbol{x}\boldsymbol{x}^T\rangle_{\mathcal{N}} &= \boldsymbol{\Sigma}+\boldsymbol{\mu}\boldsymbol{\mu}^T\\
H[\mathcal{N}] &\equiv -\langle\log\mathcal{N}\rangle_{\mathcal{N}}=\frac{1}{2}\log|2\pi e\boldsymbol{\Sigma}|
\end{aligned}$$

Gamma.

$$\begin{aligned}
\mathcal{G}(x;a,b) &\equiv \exp\left((a-1)\log x-\frac{1}{b}x-\log\Gamma(a)-a\log b\right)\\
\langle x\rangle_{\mathcal{G}} &= ab\\
\langle\log x\rangle_{\mathcal{G}} &= \Psi(a)+\log b\\
H[\mathcal{G}] &\equiv -\langle\log\mathcal{G}\rangle_{\mathcal{G}}=-(a-1)\Psi(a)+\log b+a+\log\Gamma(a)
\end{aligned}$$

Inverse Gamma.

$$\begin{aligned}
\mathcal{IG}(x;a,b) &\equiv \exp\left(-(a+1)\log x-\frac{b}{x}-\log\Gamma(a)+a\log b\right)\\
\langle 1/x\rangle_{\mathcal{IG}} &= \frac{a}{b}\\
\langle\log x\rangle_{\mathcal{IG}} &= -(\Psi(a)-\log b)\\
H[\mathcal{IG}] &\equiv -\langle\log\mathcal{IG}\rangle_{\mathcal{IG}}=-(a+1)\Psi(a)+\log b+a+\log\Gamma(a)
\end{aligned}$$

APPENDIX B: Denoising with GMCs

Algorithm 1 GMC_DENOISE_GIBBS_EM($\mathbf{x}$, $N_{samples}$)

Initialise a, $\mathbf{v}^{(0)}$, $r^{(0)}$

Set N_{burn_in}

repeat

for $i = 1 : N_{samples}$ **do**

Draw $\mathbf{z}^{(i)}$ from $p(\mathbf{z}|\mathbf{v}^{(i-1)}, \mathrm{a})$

Draw $\mathbf{s}^{(i)}$ from $p(\mathbf{s}|\mathbf{x}, r^{(i-1)}, \mathbf{v}^{(i-1)})$

Draw $\mathbf{v}^{(i)}$ from $p(\mathbf{v}|\mathbf{z}^{(i)}, \mathbf{s}^{(i)}, \mathrm{a})$

Draw $r^{(i)}$ from $p(r|\mathbf{x}, \mathbf{s}^{(i)})$

end for

repeat

Set $\partial\mathrm{logL} = \sum_{i=N_{burn\_in}}^{N_{samples}} \frac{\partial \log \pi(\mathbf{x}, \mathbf{s}^{(i)}, \mathbf{v}^{(i)}, \mathbf{z}^{(i)}, r^{(i)}; \mathrm{a})}{\partial \mathrm{a}}$

Set $\partial^2\mathrm{logL} = \sum_{i=N_{burn\_in}}^{N_{samples}} \frac{\partial^2 \log \pi(\mathbf{x}, \mathbf{s}^{(i)}, \mathbf{v}^{(i)}, \mathbf{z}^{(i)}, r^{(i)}; \mathrm{a})}{\partial \mathrm{a}^2}$

$\mathrm{a} = \mathrm{a} - \partial\mathrm{logL}/\partial^2\mathrm{logL}$

until Change in a is small

until Overall change in a is small

APPENDIX C: Denoising with GMRFs

Algorithm 2 GMRF_DENOISE_GIBBS_CD($\mathbf{x}$, $N_{samples}$)

Initialise η, a, $\mathbf{v}^{(0)}$, $r^{(0)}$

Set N_{burn_in}

repeat

for $i = 1 : N_{samples}$ **do**

Draw $\mathbf{z}^{(i)}$ from $p(\mathbf{z}|\mathbf{v}^{(i-1)}, \mathrm{a})$

Draw $\mathbf{s}^{(i)}$ from $p(\mathbf{s}|\mathbf{x}, r^{(i-1)}, \mathbf{v}^{(i-1)})$

Draw $\mathbf{v}^{(i)}$ from $p(\mathbf{v}|\mathbf{z}^{(i)}, \mathbf{s}^{(i)}, \mathrm{a})$

Draw $r^{(i)}$ from $p(r|\mathbf{x}, \mathbf{s}^{(i)})$

end for

Set $\partial\mathrm{CD} = -\sum_{i=N_{burn\_in}}^{N_{samples}} \frac{\partial \log \pi(\mathbf{x}, \mathbf{s}^{(i)}, \mathbf{v}^{(i)}, \mathbf{z}^{(i)}, r^{(i)}; \mathrm{a})}{\partial \mathrm{a}}$

Draw $\mathbf{x}'$ from $p(\mathbf{x}|\mathbf{s}, r)$

Set $\mathbf{v}^{(0)} = \mathbf{v}^{(N_{samples})}$ and $r^{(0)} = r^{(N_{samples})}$

for $i = 1 : N_{samples}$ **do**

Draw $\mathbf{z}^{(i)}$ from $p(\mathbf{z}|\mathbf{v}^{(i-1)}, \mathrm{a})$

Draw $\mathbf{s}^{(i)}$ from $p(\mathbf{s}|\mathbf{x}', r^{(i-1)}, \mathbf{v}^{(i-1)})$

Draw $\mathbf{v}^{(i)}$ from $p(\mathbf{v}|\mathbf{z}^{(i)}, \mathbf{s}^{(i)}, \mathrm{a})$

Draw $r^{(i)}$ from $p(r|\mathbf{x}', \mathbf{s}^{(i)})$

end for

Set $\partial\mathrm{CD} = \partial\mathrm{CD} + \frac{\partial \log \pi(\mathbf{x}', \mathbf{s}^{(N_{samples})}, \mathbf{v}^{(N_{samples})}, \mathbf{z}^{(N_{samples})}, r^{(N_{samples})}; \mathrm{a})}{\partial \mathrm{a}}$

$\mathrm{a} = \mathrm{a} - \eta\, \partial\mathrm{CD}$

Update η

until Change in a is small

APPENDIX D: NMF Using GMCs

Below we give the pseudocode of single-channel source separation method which uses GMCs as prior distributions in NMF. $\mathbf{\Theta}$ denotes the vector of all hyperparameters of the model, $\mathbf{T}_Z$ and $\mathbf{V}_Z$ represents the auxiliary variables of the GMCs in the template and excitation models. $\mathbf{S}_{ton}$ and $\mathbf{S}_{tran}$ are the estimated magnitude spectrograms of the tonal and percussive sources.

Algorithm 3 UNMF-e ($\mathbf{X}$, I_{ton}, I_{bass}, I_{perc}, N_{samples})

$I = I_{ton} + I_{bass} + I_{perc}$

Set $N_{\text{burn\_in}}$

Initialise $\mathbf{T}$, $\mathbf{T}_Z$, $\mathbf{V}$, $\mathbf{V}_Z$ and hyperparameters, $\mathbf{\Theta}$

for n = 1:N_{samples} **do**

  Draw $\mathbf{S}^n$, $\mathbf{T}^n$, $\mathbf{T}_Z^n$, $\mathbf{V}^n$ and $\mathbf{V}_Z^n$ from full conditionals

  for each hyperparameter Θ **do**

    Propose Θ', calculate acceptance probability a_Θ

    Accept Θ' with probability a_Θ

  end for

end for

for i = 1:I **do**

  $\hat{\mathbf{S}}_i = \sum_{i=N_{\text{burn\_in}}+1}^{N_{\text{samples}}} \mathbf{S}_i^n / (N_{\text{samples}} - N_{\text{burn\_in}})$

end for

$\mathbf{S}_{ton} = \sum_{i=1}^{I_{ton}} \hat{\mathbf{S}}_i$

$\mathbf{S}_{tran} = \sum_{i=I_{ton}+1}^{I} \hat{\mathbf{S}}_i$

REFERENCES

1. Rowe, D. B., *Multivariate Bayesian statistics: models for source separation and signal unmixing*, Chapman and Hall/CRC, Boca Raton, FL, USA, 2003.

2. Hyvärinen, A., J. Karhunen, and E. Oja, *Independent Component Analysis*, Wiley, New York, 2001.

3. Attias, H., "Independent Factor Analysis", *Neural Computation*, Vol. 11, No. 4, pp. 803–851, 1999.

4. Jung, T.-P., S. Makeig, M. Westerfield, J. Townsend, E. Courchesne, and T. J. Sejnowski, "Analyzing and visualizing single-trial event-related potentials", *Advances in neural information processing systems 11*, pp. 118–124, 1999.

5. Jung, T.-P., C. Humphries, T.-W. Lee, M. J. McKeown, V. Iragui, S. Makeig, and T. J. Sejnowski, "Removing electroencephalographic artifacts by blind source separation", *Psychophysiology*, Vol. 37, No. 2, pp. 163–178, 2000.

6. Tang, A. C., B. A. Pearlmutter, M. Zibulevsky, and S. A. Carter, "Blind source separation of multichannel neuromagnetic responses", *Neurocomputing*, Vol. 32-33, pp. 1115–1120, 2000.

7. Vigário, R., J. Särelä, V. Jousmäki, M. Hämäläinen, and E. Oja, "Independent component approach to the analysis of EEG and MEG recordings", *IEEE Transactions on Biomedical Engineering*, Vol. 47, No. 5, pp. 589–593, 2000.

8. Wübbeler, G., A. Ziehe, B.-M. Mackert, K.-R. Müller, L. Trahms, and G. Curio, "Independent component analysis of non-invasively recorded cortical magnetic DC-fields in humans", *IEEE Transactions on Biomedical Engineering*, Vol. 47, No. 5, pp. 594–599, 2000.

9. Ziehe, A., K.-R. Müller, G. Nolte, B.-M. Mackert, and G. Curio, "Artifact reduction in magnetoneurography based on time-delayed second-order correlations", *IEEE Transactions on Biomedical Engineering*, Vol. 47, No. 1, pp. 75–87, 2000.

10. Pearlmutter, B. A. and S. Jaramillo, "Progress in blind separation of magnetoencephalographic data", *Independent Component Analyses, Wavelets, and Neural Networks. Proceedings of the SPIE*, pp. 129–134, 2003.

11. Mckeown, M. J., S. Makeig, G. G. Brown, T.-P. Jung, R. S. Kindermann, A. J. Bell, and T. J. Sejnowski, "Analysis of fMRI data by blind separation into independent spatial components", *Human Brain Mapping*, Vol. 6, pp. 160–188, 1998.

12. Févotte, C. and S. Godsill, "A Bayesian approach for blind separation of sparse sources", *IEEE Trans. on Speech and Audio Processing*, 2007, (to appear).

13. Cemgil, A. T. and O. Dikmen, "Conjugate gamma Markov random fields for modelling nonstationary sources", *ICA 2007, 7th International Conference on Independent Component Analysis and Signal Separation*, pp. 697–705, 2007.

14. Virtanen, T., "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15, No. 3, pp. 1066–1074, 2007.

15. Virtanen, T., "Sound source separation using sparse coding with temporal continuity objective", *International Computer Music Conference (ICMC 2003)*, 2003.

16. Smaragdis, P., "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs", *Fifth International Conference on Independent Component Analysis*, pp. 494–499, 2004.

17. Ozerov, A., P. Philippe, R. Gribonval, and F. Bimbot, "One microphone singing voice separation using source-adapted models", *2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2005.

18. Jutten, C. and J. Herault, "Blind separation of sources, part 1: an adaptive algorithm based on neuromimetic architecture", *Signal Process.*, Vol. 24, No. 1, pp. 1–10, 1991.

19. Comon, P., "Independent component analysis, a new concept?", *Signal Process.*, Vol. 36, No. 3, pp. 287–314, 1994.

20. Delfosse, N. and P. Loubaton, "Adaptive blind separation of independent sources: a deflation approach", *Signal Processing*, Vol. 45, pp. 59–83, 1995.

21. Hyvärinen, A. and E. Oja, "A fast fixed-point algorithm for independent component analysis", *Neural Computation*, Vol. 9, No. 7, pp. 1483–1492, 1997.

22. Hyvärinen, A. and E. Oja, "Independent component analysis by general nonlinear Hebbian-like learning rules", *Signal Processing*, Vol. 64, No. 3, pp. 301–313, Feb 1998.

23. Bell, A. J. and T. J. Sejnowski, "An information maximization approach to blind separation and blind deconvolution", *Neural Computation*, Vol. 7, pp. 1129–1159, 1995.

24. Gaeta, M. and J.-L. Lacoume, "Source separation without prior knowledge: The maximum likelihood solution", *EUSIPCO'90*, pp. 621–624, 1990.

25. Pham, D.-T., P. Garrat, and C. Jutten, "Separation of a mixture of independent sources through a maximum likelihood approach", *Proceedings of EUSIPCO*, pp. 771–774, 1992.

26. Pearlmutter, B. A. and L. C. Parra, "A context-sensitive generalization of ICA", *Int Conf on Neural Information Processing*, pp. 151–157, 1996.

27. Cardoso, J. F., "Infomax and maximum likelihood for blind source separation", *IEEE Signal Processing Letters*, Vol. 4, No. 4, pp. 112–114, April 1997.

28. Zibulevsky, M., P. Kisilev, Y. Y. Zeevi, and B. A. Pearlmutter, "Blind source separation via multinode sparse representation", *Advances in Neural Information Processing Systems 14*, pp. 1049–1056, 2002.

29. van Hulle, M. M., "Kernel-based equiprobabilistic topographic map formation", *Neural Computation*, Vol. 10, No. 7, pp. 1847–1871, 1998.

30. van Hulle, M. M., "Clustering approach to square and non-square blind source separation", *IEEE Signal Processing Society Workshop*, pp. 315–323, 1999.

31. O'Grady, P. D. and B. A. Pearlmutter, "Hard-LOST: Modified k-means for oriented lines", *Irish Signals and Systems Conference*, pp. 247–252, 2004.

32. O'Grady, P. D. and B. A. Pearlmutter, "Soft-lost: EM on a mixture of oriented lines", *Fifth International Conference on Independent Component Analysis*, pp. 430–436, 2004.

33. Theis, F. J., "A geometric algorithm for overcomplete linear ICA", *Neurocomputing*, Vol. 56, pp. 381–398, 2003.

34. Lin, J. K., D. G. Grier, and J. D. Cowan, "Feature extraction approach to blind source separation", *IEEE Workshop on Neural Networks for Signal Processing (NNSP)*, pp. 398–405, 1997.

35. Vielva, L., D. Erdogmus, and J. Principe, "Underdetermined blind source separation using a probabilistic source sparsity model", *2nd International Workshop on Independent Component Analysis and Blind Signal Separation*, pp. 675–679, 2000.

36. Vielva, L., D. Erdogmus, C. Pantaleon, I. Santamaria, J. Pereda, and J. Principe, "Underdetermined blind source separation in a time-varying environment", *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3049–3052, 2002.

37. Chen, S. S., D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis

pursuit", *SIAM Journal on Scientific Computing*, Vol. 20, pp. 33–61, 1998.

38. Bofill, P. and M. Zibulevsky, "Blind separation of more sources than mixtures using the sparsity of the short-time Fourier transform", *2nd International Workshop on Independent Component Analysis and Blind Signal Separation*, pp. 87–92, 2000.

39. Lee, D. D. and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization", *Nature*, Vol. 401, pp. 788–791, 1999.

40. Plumbley, M. D., "Algorithms for non-negative independent component analysis", *IEEE Transactions on Neural Networks*, Vol. 14, No. 3, pp. 534–543, 2003.

41. Hoyer, P. O., "Non-negative sparse coding", *12th IEEE Workshop on Neural Networks for Signal Processing*, pp. 557–565, 2002.

42. Olshausen, B. A. and K. J. Millman, "Learning sparse codes with a mixture-of-Gaussians prior", *Advances in Neural Information Processing Systems*, pp. 841–847, 2000.

43. Davies, M. and N. Mitianoudis, "A Simple Mixture Model for Sparse Overcomplete ICA", *IEE proceedings in Vision, Image and Signal Processing*, Vol. 151, No. 1, pp. 35–43, 2004.

44. Lewicki, M. S. and T. J. Sejnowski, "Learning Overcomplete Representations", *Neural Computation*, Vol. 12, No. 2, pp. 337–365, 2000.

45. Girolami, M., "A Variational Method for Learning Sparse and Overcomplete Representations", *Neural Computation*, Vol. 13, No. 11, pp. 2517–2532, 2001.

46. Cemgil, A. T., C. Févotte, and S. J. Godsill, "Variational and Stochastic Inference for Bayesian Source Separation", *Digital Signal Processing*, Vol. 17, No. 5, pp. 891–913, 2007.

47. Palmer, J. A., K. Kreutz-Delgado, B. D. Rao, and S. Makeig, "Modeling and

estimation of dependent subspaces with non-radially symmetric and skewed densities", *Proceedings of the 7th International Symposium on Independent Component Analysis*, 2007.

48. Wolfe, P. J., S. J. Godsill, and W. J. Ng, "Bayesian variable selection and regularisation for time-frequency surface estimation", *Journal of the Royal Statistical Society, Series B*, Vol. 66, No. 3, pp. 575–589, 2004.

49. Rowe, D. B., "A bayesian approach to blind source separation", *Journal of Interdisciplinary Mathematics*, Vol. 5, No. 1, pp. 49–76, 2002.

50. Wolfe, P. J. and S. J. Godsill, "Interpolation of missing data values for audio signal restoration using a gabor regression model", *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2005.

51. Reyes-Gomez, M., N. Jojic, and D. Ellis, "Deformable spectrograms", *AI and Statistics Conference*, Barbados, 2005.

52. Godsill, S., A. Cemgil, C. Fevotte, and P. Wolfe, "Bayesian computational methods for sparse audio and music processing", *15th European Signal Processing Conference*, EURASIP, 2007.

53. Cemgil, A. T., C. Févotte, and S. J. Godsill, "Blind Separation of Sparse Sources using Variational EM", *13th European Signal Processing Conference*, EURASIP, Antalya/Turkey, 2005.

54. Dikmen, O. and A. T. Cemgil, "Inference and parameter estimation in Gamma chains", Technical Report CUED/F-INFENG/TR.596, University of Cambridge, February 2008.

55. Knuth, K. H., "A Bayesian approach to source separation", *ICA'99, International Conference on Independent Component Analysis*, pp. 283–288, 1999.

56. Virtanen, T., A. T. Cemgil, and S. J. Godsill, "Bayesian extensions to nonnegative

matrix factorisation for audio signal modelling", *Proc. of IEEE ICASSP 08*, Las Vegas, 2008.

57. Cemgil, A. T., "Bayesian inference in non-negative matrix factorisation models", Technical Report CUED/F-INFENG/TR.609, University of Cambridge, July 2008.

58. Princen, J. P., A. W. Johnson, and A. B. Bradley, "Subband/transform coding using filter bank designs based on time domain aliasing cancellation", *IEEE Proc. Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP) 12*, pp. 2161–2164, 1987.

59. Allen, J. B., "Short term spectral analysis, synthesis, and modification by discrete Fourier transform", *IEEE Transactions on Acoustics, Speech, Signal Processing*, Vol. ASSP-25, pp. 235–238, 1977.

60. Qian, S. and D. Chen, "Discrete Gabor transform", *IEEE Transactions on Signal Processing*, Vol. 41, No. 7, pp. 2429–2438, 1993.

61. Chui, C. K., *An Introduction to Wavelets*, Academic Press, San Diego, 1992.

62. Martin, R., "Speech enhancement based on minimum mean square error estimation and supergaussian priors", *IEEE Trans. on Speech and Audio Processing*, Vol. 13, No. 5, pp. 845–856, 2005.

63. Crouse, M., R. Nowak, and R. Baraniuk, "Wavelet-based statistical signal processing using hidden Markov models", *IEEE Transactions on Signal Processing*, Vol. 46, No. 4, pp. 886–902, 1998.

64. Cohen, L., *Time-Frequency Analysis*, Prentice-Hall, 1995.

65. Rabiner, L. R. and R. W. Schafer, *Digital processing of speech signals*, Prentice-Hall, 1978.

66. Tong, L., R.-W. Liu, V. C. Soon, and Y.-F. Huang, "Indeterminacy and identifia-

bility of blind identification", *IEEE Transactions on Circuits and Systems*, Vol. 38, No. 5, pp. 499–509, 1991.

67. Molgedey, L. and G. Schuster, "Separation of a mixture of independent signals using time delayed correlations", *Physical Review Letters*, Vol. 72, No. 23, pp. 3634–3637, 1994.

68. Hyvärinen, A. and P. Hoyer, "Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces", *Neural Computation*, Vol. 12, No. 7, pp. 1705–1720, 2000.

69. Casey, M. A. and A.Westner, "Separation of mixed audio sources by independent subspace analysis", *Proc. Int. Comp. Music Conf*, pp. 154–161, Berlin, Germany, 2000.

70. Brown, J. C. and P. Smaragdis, "Independent component analysis for automatic note extraction from musical trills", *Journal of the Acoustical Society of America*, Vol. 115, No. 5, pp. 2295–2306, 2004.

71. Smaragdis, P. and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription", *In IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 177–180, 2003.

72. Helén, M. and T. Virtanen, "Separation of drums from polyphonic music using nonnegtive matrix factorization and support vector machine", *European Signal Processing Conference*, Istanbul, Turkey, 2005.

73. Geman, S. and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 6, No. 6, pp. 721–741, 1984.

74. Attias, H., "A variational bayesian framework for graphical models", *Advances in Neural Information Processing Systems*, 2000.

75. Besag, J., "Statistical Analysis of Non-lattice Data", *Statistician*, Vol. 24, No. 3, pp. 179–195, 1975.

76. Hinton, G. E., "Training products of experts by minimizing contrastive divergence", *Neural Computation*, Vol. 14, No. 8, pp. 1771–1800, 2002.

77. Hyvärinen, A., "Estimation of non-normalized statistical models using score matching", *Journal of Machine Learning Research*, Vol. 6, pp. 695–709, 2005.

78. MacKay, D., *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, 2003.

79. Doucet, A., "On sequential Monte Carlo methods for Bayesian filtering", Technical report, University Of Cambridge, UK, Department Of Engineering, 1998.

80. Dempster, A. P., N. M. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society*, Vol. 1, No. 39, pp. 1–38, 1977.

81. Carreira-Perpiñán, M. A. and G. E. Hinton, "On contrastive divergence learning", *10th Int. Workshop on Artificial Intelligence and Statistics (AISTATS 2005)*, pp. 59–66, 2005.

82. Williams, C. K. I. and F. V. Agakov., "An analysis of contrastive divergence learning in Gaussian Boltzmann machines", Technical Report EDI-INF-RR-0120, Division of Informatics, University of Edinburgh, 2002.

83. Hyvärinen, A., "Consistency of pseudolikelihood estimation of fully visible Boltzmann machines", *Neural Computation*, Vol. 18, No. 10, pp. 2283–2292, 2006.

84. Yuille, A., "The convergence of contrastive divergences", *Advances in Neural Information Processing Systems 17*, pp. 1593–1600, 2005.

85. Mase, S., "Consistency of the maximum pseudo-likelihood estimator of continuous

state space Gibbsian processes", *The Annals of Applied Probability*, Vol. 5, No. 3, pp. 603–612, 1995.

86. Gidas, B., "Consistency of maximum likelihood and pseudo-likelihood estimators for Gibbsian distributions", Fleming, W. and P.-L. Lions, editors, *Stochastic differential systems, stochastic control theory and applications*, New York: Springer, 1988.

87. Hyvärinen, A., "Some extensions of score matching", *Comput. Stat. Data Anal.*, Vol. 51, No. 5, pp. 2499–2512, 2007.

88. Kidmose, P., "Alpha-stable distributions in signal processing of audio signals", *41st Conference on Simulation and Modelling, Scandinavian Simulation Society, SIMS2000*, pp. 87–94, 2000.

89. Palmer, J. A., K. Kreutz-Delgado, D. P.Wipf, and B. D. Rao, "Variational EM algorithms for non-Gaussian latent variable models", *NIPS 2005*, 2005.

90. Palmer, J. A., *Variational and Scale Mixture Representations of Non-Gaussian Densities for Estimation in the Bayesian Linear Model: Sparse Coding, Independent Component Analysis, and Minimum Entropy Segmentation*, Ph.D. thesis, University of California San Diego, 2006.

91. Howard, R., *Dynamic Probabilistic Systems, volume 1: Markov Chains*, John Wiley and Sons, 1971.

92. Bishop, C. M., *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer, August 2006.

93. Box, G. E. P. and G. C. Tiao, *Bayesian Inference in Statistical Analysis*, Addison-Wesley, Reading, MA, 1973.

94. Kindermann, R. and J. L. Snell, *Markov Random Fields and Their Applications*, American Mathematical Society, 1980.

95. Kalman, R. E., "A new approach to linear filtering and prediction problems", *Transactions of the ASME - Journal of Basic Engineering*, Vol. 82, pp. 35–45, 1960.

96. Kalman, R. E. and R. S. Bucy, "New results in linear filtering and prediction theory", *Transactions of the ASME - Journal of Basic Engineering*, Vol. 83, pp. 95–107, 1961.

97. Chib, S., "Marginal likelihood from the Gibbs output", *Journal of the American Statistical Association*, Vol. 90, No. 432, pp. 1313–1321, 1995.

98. Févotte, C., R. Gribonval, and E. Vincent, "BSS_EVAL Toolbox User Guide", Technical Report 1706, IRISA, Rennes, France, 2005.

99. Box, G. E. P. and K. Wilson, "On the experimental attainment of optimum conditions (with discussion)", *Journal of the Royal Statistical Society Series B*, Vol. 13, No. 1, pp. 1–45, 1951.