Acoustics'08 Paris
June 29-July 4, 2008
www.acoustics08-paris.org

Inference and learning in gamma chains for Bayesian audio processing

Ali Taylan Cemgil[a] and Onur Dikmen[b]

[a]University of Cambridge, Trumpington street, CB2 1PZ Cambridge, UK
[b]Bogazici University, Dept. of Computer Engineering, 80815 Istanbul, Turkey
atc27@cam.ac.uk

Statistical description of complex phenomena encountered in many applications requires construction of nonstationary models, where source statistics are varying over time. A first step in analysis of such nonstationary sources involves typically a traditional time-frequency analysis such as Gabor, Short time Fourier transform (STFT) or modified discrete cosine transform (MDCT). In all these techniques, the underlying assumption is that the process is piecewise stationarity, however dependencies across frequency bands or time frames are not explicitly characterised. We investigate a class of prior models, called Gamma chains, for modelling such statistical dependencies in the time-frequency representations of signals. In particular, we model the prior variance of transform coefficients using Markov chains of inverse Gamma random variables. This model class is Markovian and conditionally conjugate, so standard inference methods like Gibbs sampling, variational Bayes or sequential Monte Carlo can be applied effectively and efficiently. We also show how hyperparameters, that determine the coupling between prior variances of transform coefficients, can also be optimised. We discuss the pros and cons of various inference schemata (variational Bayes, Gibbs sampler and particle filtering) in terms of complexity and optimisation performance for this model class. We illustrate the effectiveness of our approach in audio denoising and single channel audio source separation applications.

1 Introduction

Time-frequency representation of a signal represents time varying spectral components of a signal, and such a representations are often more compact and easier to interpret than a time or frequency domain representations alone. Modified discrete cosine transform (MDCT), short time Fourier transform (STFT), Gabor transform and wavelet transform 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, $\phi_{\alpha,t}$:

$$y_t = \sum_{\alpha} \phi_{\alpha,t} \tilde{y}_{\alpha}, \tag{1}$$

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

$$\boldsymbol{y} = \boldsymbol{\Phi} \tilde{\boldsymbol{y}}, \tag{2}$$

where $\boldsymbol{y}$ is a $T \times 1$ vector denoting the signal of length T, $\tilde{\boldsymbol{y}}$ is a column vector of all the coefficients $(K \times 1)$ and $\boldsymbol{\Phi}$ is a basis matrix $(T \times K)$ formed by concatenating individual basis vectors ϕ's. Here, $K = WN$. Note that the matrix $\boldsymbol{\Phi}$ is the inverse transform matrix. When the transform basis is orthogonal (e.g. inverse MDCT, orthogonal wavelets), certain statistical properties are preserved under transformations. To illustrate this, we consider a denoising problem where the original signal $\boldsymbol{s}$ is observed in additive noise $\boldsymbol{\epsilon}$ to yield the observed signal $\boldsymbol{x}$. Now suppose we transform the observed signal $\boldsymbol{x}$ via an orthogonal transform

$$\begin{aligned}
\boldsymbol{x} &= \boldsymbol{s} + \boldsymbol{\epsilon} && (3)\\
\boldsymbol{\Phi}^{-1}\boldsymbol{x} &= \boldsymbol{\Phi}^{-1}(\boldsymbol{s} + \boldsymbol{\epsilon}) = \boldsymbol{\Phi}^{-1}\boldsymbol{s} + \boldsymbol{\Phi}^{-1}\boldsymbol{\epsilon} && (4)\\
\tilde{\boldsymbol{x}} &= \tilde{\boldsymbol{s}} + \tilde{\boldsymbol{\epsilon}} && (5)
\end{aligned}$$

The correlation structure between $\boldsymbol{s}$ and $\boldsymbol{\epsilon}$ is preserved since

$$\left\langle \tilde{\boldsymbol{\epsilon}}^{\top} \tilde{\boldsymbol{s}} \right\rangle = \left\langle (\boldsymbol{\Phi}^{-1}\boldsymbol{\epsilon})^{\top} \boldsymbol{\Phi}^{-1}\boldsymbol{s} \right\rangle = \left\langle \boldsymbol{\epsilon}^{\top} \boldsymbol{\Phi}\boldsymbol{\Phi}^{-1}\boldsymbol{s} \right\rangle = \left\langle \boldsymbol{\epsilon}^{\top} \boldsymbol{s} \right\rangle$$

Here, $\langle \cdot \rangle$ denotes the expectation. For example, if $\boldsymbol{s}$ and $\boldsymbol{\epsilon}$ are a priori uncorrelated, i.e. $\left\langle \boldsymbol{\epsilon}^{\top} \boldsymbol{s} \right\rangle = 0$, so are $\tilde{\boldsymbol{s}}$ and $\tilde{\boldsymbol{\epsilon}}$. In denoising, where our task is to estimate $\boldsymbol{s}$ given $\boldsymbol{x}$, this observation motivates the fact that we can do modelling equivalently in the transform domain and aim at recovering the transform coefficients $\tilde{\boldsymbol{s}}$ given $\tilde{\boldsymbol{x}}$.

A closely related problem to denoising is single channel source separation problem. Here, our goal is to extract N_s source signals from a single observation signal which is expressed the sum of the sources.

$$\boldsymbol{x} = \sum_{i=1}^{N_s} \boldsymbol{s}_i \qquad \tilde{\boldsymbol{x}} = \sum_{i=1}^{N_s} \tilde{\boldsymbol{s}}_i$$

In fact, this problem is a simple generalisation of denoising: we can view denoising as a single channel source separation problem with $N_s = 2$ where one source is the noise component. Hence, single channel source separation problem can be modelled in the time-frequency domain as in the same manner as above.

In this paper, we will concentrate on the denoising and the single channel source separation problems to demonstrate the advantages of modelling the dependencies in the time-frequency representations of audio signals. Source separation (and denoising as a special case) can be solved in Bayesian framework by inferring the posterior distribution of the sources $p(\boldsymbol{s}|\boldsymbol{x})$.

Time-frequency domain coefficients of audio sources are shown to be better modelled with heavy-tailed distributions [1, 2, 3]. In source separation literature source coefficients are modelled with mixture of Gaussians [4],[5], Laplace [6],[7] and Student-*t* distribution [3, 8]. These models make use of mutually independent and identical (having the same distribution with the same hyperparameters) prior distributions of the source coefficients. This approach misses the inherent dependency in the time-frequency representation of audio signals, such as the harmonic continuity of tonal components of a signal over a period of time or impulsive activation of a range of frequencies by transients of a signal. The independent models can be extended to include such dependencies by coupling the variances or other parameters that control the sources. Inverse Gamma Markov chains (IGMC) and Markov random fields (IGMRF) are introduced to correlate the variances [9]. In [10], a Markov random field that controls the activation of the source coefficients is proposed to model the dependency.

The posterior distribution, $p(\boldsymbol{s}|\boldsymbol{x})$, contains an intractable marginal likelihood (evidence). 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 is intractable but can be approximated by the lower bounds which are evaluated using the inferred sufficient statistics of latent variables. One of the main objectives of this study is to compare the inference methods in terms of the accuracy of the approximate likelihoods (tightness of the lower bounds, which leads to the success of the optimisation) and time complexity. In audio source models using IGMCs, the hyperparameters determine the magnitude of the coupling between the variance variables, which is directly related to the strength of the model.

In the next section we will explain inverse Gamma Markov chains which are simple and efficient dependency models for time-frequency representation of audio signals. Then in Section 3 we will review the relevant inference methods (variational Bayes, Markov chain Monte Carlo and sequential Monte Carlo methods). These methods are used on the inference of the sources in denoising and single channel source separation applications. The optimisation of the hyperparameters is done using EM variants based on the estimates of these methods. The results will be presented in Section 4.

2 Inverse Gamma Markov Chains

An inverse Gamma Markov chain (IGMC), proposed in [9], is a sequence of random variables which have inverse Gamma[1] priors conditional on only the preceding variable. It is defined as

$$\begin{aligned} z_1 &\sim \mathcal{IG}(z_1; a_z, b/a_z) && (6)\\ v_t|z_t &\sim \mathcal{IG}(v_t; a_v, z_t/a_v) && (7)\\ z_t|v_{t-1} &\sim \mathcal{IG}(z_t; a_z, v_{t-1}/a_z), \quad t>1 && (8)\end{aligned}$$

where v_t and z_t are the variables of the chain and a_v, a_z, b are hyperparameters. There are efficient algorithms to perform inference on this model, because the prior distributions are conditionally conjugate for the variables in the model. In a model with variables y and $\boldsymbol{x}$, if the prior distribution of a variable y, $p(y)$, is in the same class with the conditional posterior distribution $p(y|\boldsymbol{x})$, that distribution family is said to be conditionally conjugate for y[11]. That means it is as easy to draw samples from the conditional distribution, $p(y|\boldsymbol{x})$, with a Gibbs sampler as from the prior, $p(y)$. The same fact enables a variational distribution of this family to be updated very easily in the mean field algorithm, as will be seen in Section 3.1.

Full conditional distributions of all the variables in the chain are inverse Gamma. For example, full conditional of the variable v_t is expressed as

$$\begin{aligned} p(v_t|z_t, z_{t+1}, \boldsymbol{\theta}) &= \frac{p(z_{t+1}|v_t, \boldsymbol{\theta})p(v_t|z_t, \boldsymbol{\theta})}{\int p(z_{t+1}|v_t, \boldsymbol{\theta})p(v_t|z_t, \boldsymbol{\theta})\, dv_t}\\ &= \frac{\mathcal{IG}(v_t; \alpha, \beta)}{\int \mathcal{IG}(v_t; \alpha, \beta)\, dv_t} = \mathcal{IG}(v_t; \alpha, \beta)\end{aligned}$$

where $\alpha = a_v + a_z$ and $\beta = 1/(a_v/z_t + a_z/z_{t+1})$.

An IGMC is a chain of strictly positive variables with positive correlation between v_t's (and separately, between z_t's). These v_t's can be used to model the slowly varying variances of a nonstationary audio signal. When the sources are assumed zero mean Gaussian, $\mathcal{N}(s_t; 0, v_t)$, this source model becomes another instantiation of the scale mixture of Gaussians family and reduces to the Student-t model when z_t's are known. This source prior distribution is still conditionally conjugate.

The conditional distribution of variance variables are given by

$$p(v_t|v_{t-1}) = \frac{\Gamma(a_v + a_z)}{\Gamma(a_z)\Gamma(a_v)} \frac{(a_z v_{t-1}^{-1})^{a_z} (a_v v_t^{-1})^{a_v}}{(a_z v_{t-1}^{-1} + a_v v_t^{-1})^{(a_z+a_v)}} v_t^{-1}$$

As it can be seen in Figure 1, there is positive correlation between the variances for various values of a_v and a_z. The larger these parameters are, the higher coupling between the variables exists. The ratio a_z/a_v is a measure of the skewness of correlation and it can lead to positive and negative drifts.

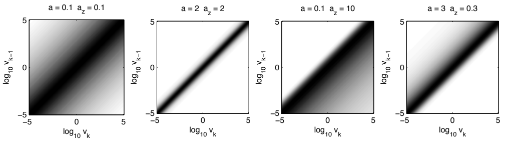

Figure 1: The first two figures show that when the parameters, a_v and a_z, are nearly equal, there is no skewness and the value determines the strength of the coupling. The last two figures are examples of positive and negative drifts, respectively.

3 Inference

3.1 Variational Bayes

Variational Bayes (mean field) [12] 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 (Kullback-Leibler divergence) between the two distributions.

$$\begin{aligned} \mathrm{KL}(q||p) &= \log p(\boldsymbol{y}|\boldsymbol{\theta}) + \mathrm{KL}(q||p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})) && (9)\\ &\equiv \log p(\boldsymbol{y}|\boldsymbol{\theta}) + \mathcal{E}(q, \boldsymbol{\theta}) && (10)\end{aligned}$$

Since the evidence, $p(\boldsymbol{y}|\boldsymbol{\theta})$, is independent of the variational distribution, $q(\boldsymbol{x})$, minimising the Kullback-Leibler

[1] Inverse Gamma distribution is defined as:
$\mathcal{IG}(x; \alpha, \beta) \equiv \exp\left((\alpha+1)\log x^{-1} - \beta^{-1}x^{-1} + \alpha \log \beta^{-1} - \log \Gamma(\alpha)\right)$

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, so Equation 10 defines a lower bound on the evidence:

$$p(\boldsymbol{y}|\boldsymbol{\theta}) \geq -\mathcal{E}(q, \boldsymbol{\theta}) = \langle \log p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}) \rangle_q - \langle \log q(\boldsymbol{x}) \rangle_q \quad (11)$$

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

$$\log q(\boldsymbol{x}_i) =^+ \langle \log p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}) \rangle_{q(\boldsymbol{x}_{-i})} \quad (12)$$

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

3.2 Markov Chain Monte Carlo Methods

Monte Carlo methods are used to approximate expectations in which the integration (or summation) is not analytically tractable and numerical integration techniques perform poorly, e.g. due to high dimensionality. Expectations of functions under a target distribution, $p(\boldsymbol{x})$, are estimated using a set of i.i.d. samples, $\{\boldsymbol{x}^{(i)}\}_{i=1}^N$, drawn from this distribution:

$$\langle f(\boldsymbol{x}) \rangle_{p(\boldsymbol{x})} = \int f(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} \approx \frac{1}{N} \sum_{i=1}^{N} f(\boldsymbol{x}^{(i)}) \quad (13)$$

Markov chain Monte Carlo approaches are used in cases where it is very difficult to draw independent samples from the target distribution, $p(\boldsymbol{x})$, but it can be evaluated up to a normalising constant.

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

$$a(\boldsymbol{x}'; \boldsymbol{x}^{(t)}) = \min \left\{ \frac{p(\boldsymbol{x}')}{p(\boldsymbol{x}^{(t)})} \frac{q(\boldsymbol{x}^{(t)}|\boldsymbol{x}')}{q(\boldsymbol{x}'|\boldsymbol{x}^{(t)})}, 1 \right\}. \quad (14)$$

The Gibbs sampler can be seen as a special case of the Metropolis-Hastings algorithm where the proposal distribution for the variables are their full conditionals, $p(\boldsymbol{x}_i|\boldsymbol{x}_{-i})$. First a variable ($\boldsymbol{x}_i$, i^{th} dimension of $\boldsymbol{x}$) 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.

3.3 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 possible to evaluate these posterior distributions analytically in hidden Markov models with finite states and linear Gaussian state-space models (Kalman filters). In the general case, Monte Carlo methods can be employed to infer about the hidden variables. However, MCMC methods are not completely suitable for online 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.

Sequential Monte Carlo 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:

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

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). Performing the importance sampling 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:

$$\begin{aligned} W_{t+1}^{(i)} &= W_t^{(i)} \frac{p(\boldsymbol{y}_{t+1}|\boldsymbol{x}_{t+1}^{(i)}) p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t^{(i)})}{q(\boldsymbol{x}_{t+1})} \\ w_{t+1}^{(i)} &= \frac{W_{t+1}^{(i)}}{\sum_{j=1}^N W_{t+1}^{(j)}} \end{aligned}$$

4 Simulations

4.1 Denoising

We modelled dependencies of the time-frequency atoms of sources obtained by MDCT with inverse gamma Markov chains. As mentioned in [9], this can be done in two ways: either tying atoms of each frequency bin across time frames (horizontal) or tying frequency atoms in each frame (vertical).

In this problem, the observed signal, $\boldsymbol{x}$, is the sum of the source signal, $\boldsymbol{s}$, and independent white Gaussian noise with variance r. Each source coefficient, $s_{\nu,\tau}$ is a zero mean Gaussian with variance $v_{\nu,\tau}$ and the variance variables an IGMC prior. ν and τ stand for the indices of frequency bins and time frames, respectively:

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

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 SNR between the original signal and the reconstructed signal.

The top two plots in Figure 2 present 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.

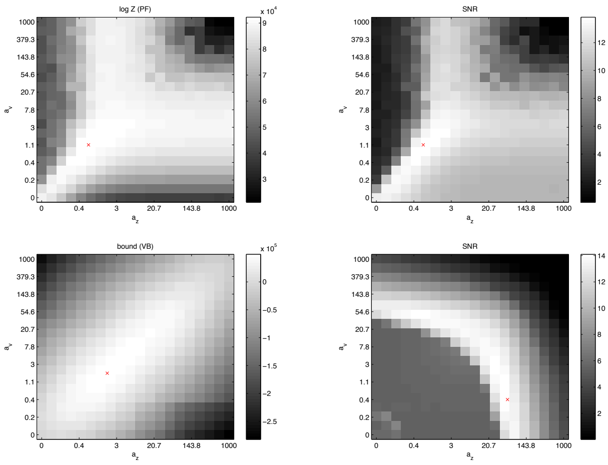

Figure 2: Log likelihood and reconstruction SNR values obtained by the SIS/R algorithm using the optimal proposal distribution (top) and variational Bayes (bottom). The surfaces are evaluated using a fixed value of b ($b = 10^{-4}$).

On the other hand, in the case of variational Bayes, there is no correlation between the lower bound of the log likelihood and the SNR (Figure 2). 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 transfer domain, after transforming the signals using MDCT with 512 frequency bins. In Figure 3 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 IGMCs. Results obtained by VB-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 parameters do not correspond to a local maxima of the lower bound.

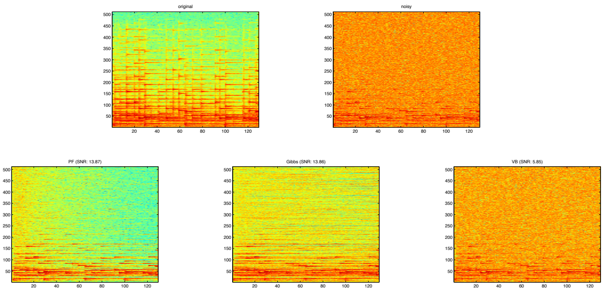

Figure 3: The figures on top are the spectrograms of the original and the noisy signals. The others are the reconstructed signals output by the three inference methods.

4.2 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 IGMCs to ensure dependency along time or frequency axis. The source coefficients are then Gaussian distributed with zero mean: $s_{\nu,\tau} \sim \mathcal{N}(0, v_{\nu,\tau})$. The observed signal is the sum of N sources: $x_{\nu,\tau} = \sum_{j=1}^{N} s^j_{\nu,\tau}$. 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 evaluated in closed form:

$$\Sigma_{i,k} = v_{i,k}\,(1 - \kappa_{i,k}) \qquad m_{i,k} = \kappa_{i,k} x_k$$

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 [9] and also known as Wiener filter factors.

Modelling the variances of a source using horizontal IGMCs and another with vertical IGMCs, 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 variational Bayes and Gibbs sampler. Since we have two directions of propagation in this model, we cannot apply classical particle filter methods directly. Table 1 shows 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 artifacts ratio (SAR) [13].

In the experiments, we applied variational Bayes (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-EM algorithm of which initial hyperparameter values are the

	$\hat{s}_1$			$\hat{s}_2$		
	SDR	SIR	SAR	SDR	SIR	SAR
VB	-4.74	-3.28	5.67	-1.58	15.46	-1.37
Gibbs	-4.50	-2.62	4.57	1.05	12.46	1.61
Gibbs$_{\mathrm{EM}}$	-4.23	-2.42	4.82	1.34	13.13	1.85
VB	-7.80	-6.22	4.53	-2.35	18.40	-2.25
Gibbs	-8.46	-7.53	6.93	-4.04	14.59	-3.83
Gibbs$_{\mathrm{EM}}$	-7.74	-6.19	4.62	-1.14	16.62	-0.97

Table 1: Source separation results (top) on a mixture of guitar ("Matte Kudasai") and drums ("Territory") and (bottom) on a mixture of flute ("Vandringar I Vilsenhet") and drums ("Moby Dick")

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 spectromrams of the sources estimated by the Gibbs-EM in Figure 4. As expected, the variational EM algorithm converges to a set of hyperparameters that lead to a worse performance, so those results are omitted. The results of these source separation and denoising experiments can be found at `http://www.cmpe.boun.edu.tr/~dikmen/igmc_report/` as audio files.

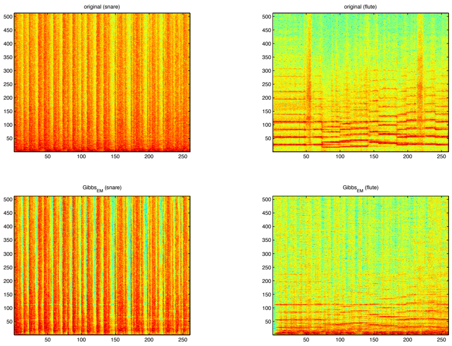

Figure 4: The spectrograms of the original sources (top) and the sources estimated by the Gibbs-EM algorithm (bottom) in the second experiment.

5 Conclusion

We modelled the variances of the time-frequency representation coefficients of non-stationary audio signals with inverse Gamma Markov chains to include the positive correlation among the variances at consecutive indices. It is suitable to model these signals with horizontal and vertical IGMCs, respectively.

Although inference is convenient in this model due to conditional conjugacy, optimisation of the hyperparameters that determine the coupling between the chain variables is essential. We performed extensive simulations to deduce facts about the model and various inference methods. Despite the fact that the Gibbs sampler needs a high number of samples for the estimation, the best model can be obtained using the Gibbs-EM algorithm. The run time of the algorithm is generally several hours. Sequential Monte Carlo performs as well as the Gibbs sampler, but with less number of 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.

The reconstructions we get with this model still have some artifacts even when all the hyperparameters are optimised. This is because, with IGMCs we can capture the dependencies in one dimension. In most of the audio signals, there is a correlation between the coefficients in both directions, although the correlation in one direction may be more prominent. Moreover, the model with IGMCs needs prior knowledge about the nature of the signal, such as whether it is tonal or percussive, for better performance.

References

[1] R. Martin, "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.

[2] M. Crouse, 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.

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

[4] B. A. Olshausen and K. J. Millman, "Learning sparse codes with a mixture-of-Gaussians prior," in *NIPS*, 2000, pp. 841–847.

[5] M.E. Davies and N. Mitianoudis, "A Simple Mixture Model for Sparse Overcomplete ICA," in *IEEE proceedings in Vision, Image and Signal Processing*, 2004, vol. 151, pp. 35–43.

[6] M. S. Lewicki and T. J. Sejnowski, "Learning Overcomplete Representations," *Neural Computation*, vol. 12, no. 2, pp. 337–365, 2000.

[7] M. Girolami, "A Variational Method for Learning Sparse and Overcomplete Representations," *Neural Computation*, vol. 13, no. 11, pp. 2517–2532, 2001.

[8] A. T. Cemgil, C. Fevotte, and S. J. Godsill, "Variational and Stochastic Inference for Bayesian Source Separation," *Digital Signal Processing*, vol. in Print, 2007.

[9] A. T. Cemgil and O. Dikmen, "Conjugate gamma Markov random fields for modelling nonstationary sources," in *ICA*, 2007, pp. 697–705.

[10] S.J. Godsill, A.T. Cemgil, C. Fevotte, and P.J. Wolfe, "Bayesian computational methods for sparse audio and music processing," in *EURASIP*, 2007.

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

[12] H. Attias, "A variational bayesian framework for graphical models," in *NIPS*, 2000.

[13] C. Févotte, R. Gribonval, and E. Vincent, "BSS_EVAL Toolbox User Guide," Tech. Rep. 1706, IRISA, Rennes, France, 2005.