Inference and Parameter Estimation in Gamma Chains

Onur Dikmen, A. Taylan Cemgil

CUED/F-INFENG/TR.596
February 2008

University of Cambridge
Department of Engineering
Trumpington Street
Cambridge CB2 1PZ
United Kingdom

Email: onuro@boun.edu.tr
atc27@cam.ac.uk

Inference and Parameter Estimation in Gamma Chains*

Onur Dikmen, A. Taylan Cemgil

Abstract

We investigate a class of prior models, called Gamma chains, for modelling depedicies in time-frequency representations of signals. We assume transform coefficients are drawn independently from Gaussians where the latent variances are coupled using Markov chains of inverse Gamma random variables. Exact inference is not feasible but this model class is conditionally conjugate, so standard approximate inference methods like Gibbs sampling, variational Bayes or sequential Monte Carlo can be applied effectively and efficiently. We show how hyperparameters, that determine the coupling between prior variances of transform coefficients, can be optimised. We discuss the pros and cons of various inference schemata (variational Bayes, Gibbs sampler and Sequential Monte Carlo) 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

Statistical description of complex phenomena encountered in many applications requires construction of non-stationary 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 these representations, a time series x_t for $t = 1, 2, \ldots, T$ is represented as a linear combination of basis functions, $\phi_{\alpha,t}$:

$$x_t = \sum_{\alpha} \phi_{\alpha,t} \tilde{x}_{\alpha}, \tag{1}$$

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 $\tilde{x}_{\alpha}$. In this paper, we assume an orthogonal transformation and we write in matrix-vector notation

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

where $\boldsymbol{x}$ is a $T \times 1$ vector denoting the signal of length T, $\tilde{\boldsymbol{x}}$ 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 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 $\tilde{\boldsymbol{x}}$.

*This research is funded by EPSRC (Engineering and Physical Sciences Research Council) under the grant EP/D03261X/1 entitled "Probabilistic Modelling of Musical Audio for Machine Listening" and Dikmen is supported by TUBITAK (Scientific and Technological Research Council of Turkey).

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.

$$\tilde{\boldsymbol{x}} = \sum_{i=1}^{N_s} \tilde{\boldsymbol{s}}_i \tag{6}$$

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 report, 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 the Bayesian framework by inferring the posterior distribution

$$p(\boldsymbol{s}|\boldsymbol{x}) = \frac{1}{Z_{\boldsymbol{x}}(\boldsymbol{\psi})} \int p(\boldsymbol{x}|\boldsymbol{s}) p(\boldsymbol{s}|\boldsymbol{\theta}_s) p(\boldsymbol{\theta}_s|\boldsymbol{\psi}_s)\, d\boldsymbol{\theta}_s \tag{7}$$

In this paper, we assume Eq.6, hence the observation model is degenerate $p(\boldsymbol{x}|\boldsymbol{s}) = \delta(\boldsymbol{x} - \sum_i \boldsymbol{s}_i)$. The model is completed by choosing a prior distribution for the sources, $p(\boldsymbol{s}|\boldsymbol{\theta}_s)$, with source parameters, $\boldsymbol{\theta}_s$ and prior over source parameters $p(\boldsymbol{\theta}_s|\boldsymbol{\psi}_s)$ with corresponding hyperparameters $\boldsymbol{\psi}_s$. The normalisation term $Z_{\boldsymbol{x}}(\boldsymbol{\psi})$ is the marginal likelihood (evidence) of the observed signals under the complete set of hyperparameters $\boldsymbol{\psi}_s$. 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, $Z_{\boldsymbol{x}}(\boldsymbol{\psi})$, is intractable but can be approximated by stochastic simulation or analytic lower bounding methods.

When assumed to be apriori independent, time-frequency domain coefficients of audio sources are shown to be better modelled with heavy-tailed distributions [1, 2, 3]. In source separation literature, specific proposals include mixture of Gaussians [4],[5], Laplace [6],[7] and Student-t distribution [3, 8]. 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); see the sequence of publications [9, 10, 11] for a systematic treatment of this model class.

However, typical audio sources have periodic components, with occasional, transient bursts in energy content. In a time-frequency representation, these properties reflect as harmonic continuity of tonal components and impulsive activation in a range of frequencies. More realistic statistical models, that capture such phenomena can be obtained by introducing dependencies by coupling the prior variances. In [12], a discrete Markov random field is proposed, where, the variances are drawn conditionally, given the discrete labels. More recently, a more flexible alternative model is proposed in [13], named as Gamma and inverse Gamma Markov random fields where the random field is directly defined on Gamma variables. One property of this model is that exact inference in chains is not analytically tractable. The main objective of this study is to illustrate and compare various inference strategies in terms of solution quality and computation time in Gamma chains to pave the way to fields.

In the next section we will define inverse Gamma Markov chains. Then in Section 3, we will review three inference methods, variational Bayes, Markov chain Monte Carlo and sequential Monte Carlo. In section 4, we will focus on EM based hyperparameter estimation that use the previous techniques as a subroutine. Simulation results on denoising and single channel source separation and comparisons on synthetic and real data as well as some toy problems will be presented in Section 5.

2 Inverse Gamma Markov Chains

An inverse Gamma Markov chain (IGMC), proposed first in [13], is a sequence of random variables which have inverse Gamma[1] priors conditional on only the preceding variable. It is defined as

$$z_1 \sim \mathcal{IG}(z_1; a_z, b/a_z) \tag{8}$$

$$z_t|v_{t-1} \sim \mathcal{IG}(z_t; a_z, v_{t-1}/a_z), \quad t > 1 \tag{9}$$

$$v_t|z_t \sim \mathcal{IG}(v_t; a_v, z_t/a_v) \tag{10}$$

where v_t and z_t are the variables of the chain and a_v, a_z, b are hyperparameters.

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

Full conditional distributions

Full conditional distributions of all the variables in the chain are inverse Gamma:

$$\begin{aligned} p(v_t|z_t, z_{t+1}, \boldsymbol{\theta}) &\propto p(z_{t+1}|v_t, \boldsymbol{\theta}) p(v_t|z_t, \boldsymbol{\theta}) \\ &= \mathcal{IG}(z_{t+1}; a_z, v_t/a_z)\mathcal{IG}(v_t; a_v, z_t/a_v) \\ &\propto \exp(-\frac{a_z}{z_{t+1}}\frac{1}{v_t} - a_z \log v_t) \exp(-(a_v+1)\log v_t - \frac{a_v}{z_t}\frac{1}{v_t}) \\ &\propto \mathcal{IG}(v_t; \alpha, \beta_t^v) \end{aligned}$$

where $\alpha = a_v + a_z$ and $\beta_t^v = 1/(a_v/z_t + a_z/z_{t+1})$. Similarly,

$$p(z_t|v_t, v_{t-1}, \boldsymbol{\theta}) \propto \mathcal{IG}(v_t; \alpha, \beta_t^z)$$

where $\alpha = a_v + a_z$ and $\beta_t^z = 1/(a_v/v_t + a_z/v_{t-1})$.

Transition Kernel

The probability of a variance variable conditional on the previous one (transition kernel of the inverse Gamma Markov chain) is found by integrating out the auxiliary variables z_t:

$$\begin{aligned} p(v_t|v_{t-1}) &= \int dz_t p(v_t|z_t) p(z_t|v_{t-1}) \\ &= \int dz_t \mathcal{IG}(v_t; a_v, z_t/a_v)\mathcal{IG}(z_t; a_z, v_{t-1}/a_z) \\ &= \frac{\Gamma(a_v+a_z)}{\Gamma(a_z)\Gamma(a_v)} \frac{(a_z v_{t-1}^{-1})^{a_z}(a_v v_t^{-1})^{a_v}}{(a_z v_{t-1}^{-1} + a_v v_t^{-1})^{(a_z+a_v)}} v_t^{-1} \end{aligned} \tag{11}$$

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.

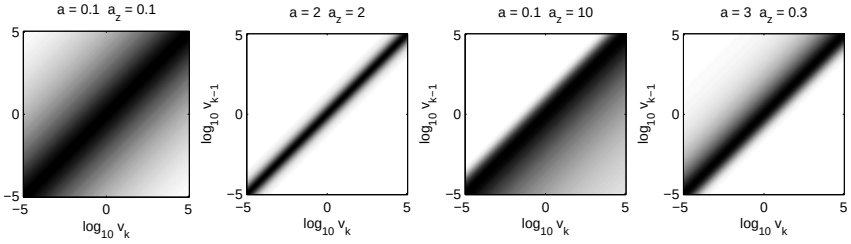


Figure 1: The first two figures show that when the parameters, a_v and a_z, are equal, there is no skewness and the value determines the strength of the coupling. The last two figures correspond to kernels where typical realisations have positive and negative drifts, respectively.

$$\begin{aligned} \langle v_t|v_{t-1}\rangle &= \int dv_t dz_t v_t p(v_t|z_t) p(z_t|v_{t-1}) \\ &= \int dv_t dz_t \exp(-a_v \log v_t - \frac{a_v}{z_t v_t} - \log\Gamma(a_v) - a_v \log z_t + a_v \log a_v) \\ &\quad \exp(-(a_v+1)\log z_t - \frac{a_v}{v_{t-1} z_t} - \log\Gamma(a_v) - a_v \log v_{t-1} + a_v \log a_v) \end{aligned}$$

$$\langle v_t|z_t\rangle = \int dv_t \exp(-(a+1)\log v_t - \frac{a+1}{z_t v_t} - \log\Gamma(a+1) - (a+1)\log z_t + a_v \log a_v)$$

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.

3 Inference

3.1 Variational Bayes

Variational Bayes (mean field) [14] 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.

$$\begin{aligned}
\mathrm{KL}(q||p) &= \int d\boldsymbol{x}\, q(\boldsymbol{x}) \log \frac{q(\boldsymbol{x})}{p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})} && (12)\\
&= \int d\boldsymbol{x}\, q(\boldsymbol{x}) \log \frac{q(\boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{\theta})}{p(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{\theta})} && (13)\\
&= \log p(\boldsymbol{y}|\boldsymbol{\theta}) + \int d\boldsymbol{x}\, q(\boldsymbol{x}) \log \frac{q(\boldsymbol{x})}{p(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{\theta})} && (14)\\
&= \log p(\boldsymbol{y}|\boldsymbol{\theta}) + \mathrm{KL}(q||p(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{\theta})) && (15)\\
&\equiv \log p(\boldsymbol{y}|\boldsymbol{\theta}) + \mathcal{E}(q,\boldsymbol{\theta}) && (16)
\end{aligned}$$

Since the evidence, $p(\boldsymbol{y}|\boldsymbol{\theta})$, is independent of the variational distribution, $q(\boldsymbol{x})$, minimising the Kullback-Leibler 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 [15], so Equation 16 defines a lower bound on the evidence:

$$\begin{aligned}
p(\boldsymbol{y}|\boldsymbol{\theta}) &\geq -\mathcal{E}(q,\boldsymbol{\theta}) && (17)\\
&= \langle \log p(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{\theta})\rangle_q - \langle \log q(\boldsymbol{x})\rangle_q && (18)
\end{aligned}$$

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

$$\log q(\boldsymbol{x}_i) =^{+} \langle \log p(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{\theta})\rangle_{q(\boldsymbol{x}_{-i})} \tag{19}$$

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:

$$E(f) = \langle f(\boldsymbol{x})\rangle_{p(\boldsymbol{x})} = \int f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} \approx \frac{1}{N}\sum_{i=1}^{N} f(\boldsymbol{x}^{(i)}) \equiv \hat{E}_N(f) \tag{20}$$

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:

$$\sigma_f^2 = \int (f(\boldsymbol{x}) - E(f))^2 p(\boldsymbol{x})\, d\boldsymbol{x}. \tag{21}$$

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 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. 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 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:

$$a(\boldsymbol{x}';\boldsymbol{x}^{(t)}) = \min\left\{\frac{p(\boldsymbol{x}')}{p(\boldsymbol{x}^{(t)})}\frac{q(\boldsymbol{x}^{(t)}|\boldsymbol{x}')}{q(\boldsymbol{x}'|\boldsymbol{x}^{(t)})}, 1\right\}. \tag{22}$$

MH algorithm has an irreducible and aperiodic transition kernel and its invariant distribution is $p(\boldsymbol{x})$. This algorithm allows us to draw samples from probability distributions, $p(\boldsymbol{x}) = \phi(\boldsymbol{x})/Z$ where the normalising constant Z is not known, because Z is independent of $\boldsymbol{x}$ and two normalising constants in Equation 22 are cancelled out.

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 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 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:

$$p(\boldsymbol{x}_{1:t+1}|\boldsymbol{y}_{1:t+1}) = p(\boldsymbol{x}_{1:t}|\boldsymbol{y}_{1:t}) \frac{p(\boldsymbol{y}_{t+1}|\boldsymbol{x}_{t+1}) p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)}{p(\boldsymbol{y}_{t+1}|\boldsymbol{y}_{1:t})}. \tag{23}$$

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).

Importance sampling 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:

$$\begin{aligned}
\langle f(\boldsymbol{x}) \rangle_{p(\boldsymbol{x})} &= \int f(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} && (24)\\
&= \frac{1}{Z} \int f(\boldsymbol{x}) \phi(\boldsymbol{x}) d\boldsymbol{x} && (25)\\
&= \frac{\int f(\boldsymbol{x}) \phi(\boldsymbol{x}) d\boldsymbol{x}}{\int \phi(\boldsymbol{x}) d\boldsymbol{x}} && (26)\\
&= \frac{\int f(\boldsymbol{x}) W(\boldsymbol{x}) q(\boldsymbol{x}) d\boldsymbol{x}}{\int W(\boldsymbol{x}) q(\boldsymbol{x}) d\boldsymbol{x}} && (27)\\
&\approx \frac{\sum_{i=1}^N f(\boldsymbol{x}^{(i)}) W^{(i)}}{\sum_{i=1}^N W^{(i)}} && (28)\\
&= \sum_{i=1}^N f(\boldsymbol{x}^{(i)}) w^{(i)} && (29)
\end{aligned}$$

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 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:

$$W_{t+1}^{(i)} = W_t^{(i)} \frac{p(\boldsymbol{y}_{t+1}|\boldsymbol{x}_{t+1}^{(i)}) p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t^{(i)})}{q(\boldsymbol{x}_{t+1})} \tag{30}$$

$$w_{t+1}^{(i)} = \frac{W_{t+1}^{(i)}}{\sum_{j=1}^{N} W_{t+1}^{(j)}} \tag{31}$$

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 [16]. 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.

4 Hyperparameter Estimation in Inverse Gamma Chains

The expectation-maximisation (EM) algorithm[17] is a classic algorithm for maximum likelihood (or maximum a posteriori) 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, $p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}_t)$:

$$Q(\boldsymbol{\theta}) = \int \log p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}) p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}_t)\, d\boldsymbol{x} \tag{32}$$

  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:

$$\boldsymbol{\theta}_{t+1} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}) \tag{33}$$

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

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) = \log p(\boldsymbol{y}|\boldsymbol{\theta}) &= \log \int p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})\, d\boldsymbol{x} && (34)\\
&= \log \int q(\boldsymbol{x}) \frac{p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})}{q(\boldsymbol{x})}\, d\boldsymbol{x} && (35)\\
&\geq \int q(\boldsymbol{x}) \log \frac{p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})}{q(\boldsymbol{x})}\, d\boldsymbol{x} && (36)\\
&= \int q(\boldsymbol{x}) \log p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})\, d\boldsymbol{x} - \int q(\boldsymbol{x}) \log q(\boldsymbol{x})\, d\boldsymbol{x} && (37)\\
&= \mathcal{F}(q, \boldsymbol{\theta}) && (38)
\end{aligned}$$

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 36. 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 32 is one constituent of the lower bound that is a function of the hyperparameters.

The variational inference algorithm explained in Section 3.1 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:

$$\hat{Q}_{VB}(\boldsymbol{\theta}) = \int \log p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}) q(\boldsymbol{x})\, d\boldsymbol{x} \tag{39}$$

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:

$$\hat{Q}_{MC}(\boldsymbol{\theta}) = \frac{1}{N_i} \sum_{j=1}^{N_i} \log p(\boldsymbol{x}^{(j)}, \boldsymbol{y}|\boldsymbol{\theta}) \tag{40}$$

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.

5 Simulations

Our main goal in this report is to optimise the hyperparameters of dynamic systems offline (given a fixed sequence of observations, $y_{1:T}$), particularly the inverse Gamma chains. The variants of the EM algorithm explained in Section 4 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[18, 19]. This model is very similar to the audio source model we use throughout this report. The state transitions of these models are Gaussian and inverse Gamma respectively. However, 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 3 with the ground truth.

Our 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 inverse Gamma Markov chains. 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.

5.1 Linear Gaussian State Space Model

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

$$x_1 \sim \mathcal{N}(x_1; 0, P) \tag{41}$$
$$x_k \sim \mathcal{N}(x_k; Ax_{k-1}, Q),\ k > 2 \tag{42}$$
$$y_k \sim \mathcal{N}(y_k; Cx_{k-1}, R) \tag{43}$$

where P, Q and R are the variances of Gaussian perturbations, A and C are linear operators. Optimal filtering can be done on this model with the Kalman filter [18, 19], so this model provides a platform to compare the algorithms in terms of the accuracy of their likelihood estimates and time complexity.

Figure 2 presents log-likelihood attained by the algorithms (Kalman filter, bootstrap filter, SIS with optimal proposal distribution, Gibbs sampler and variational Bayes) 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 [20], 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 3 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

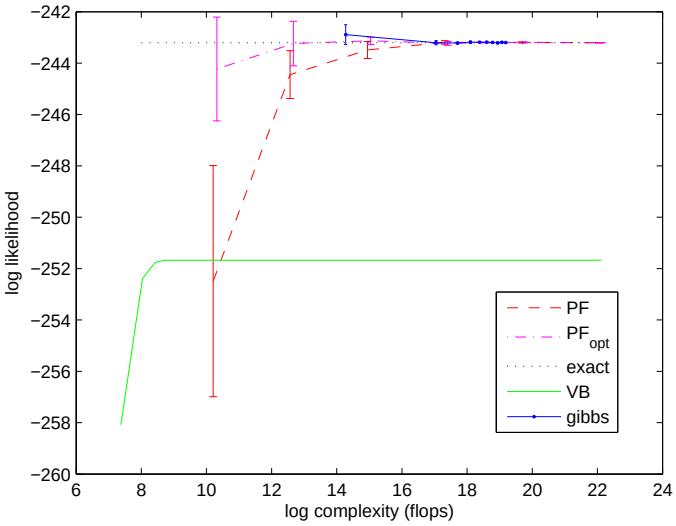


Figure 2: Log-likelihood approximations on a linear Gaussian state-space model of length 100 by different methods (Kalman filter (exact), bootstrap filter (PF), SIS with optimal proposal distribution ($\mathrm{PF}_{\mathrm{opt}}$), Gibbs sampler (gibbs) and variational Bayes (VB)). More complex particle filters and Gibbs samplers are obtained by increasing the number of samples while in VB total number of iterations is increased.

mean estimates (minimum mean square error, MMSE, estimates) of the two methods overlap as depicted in Figure 4.

In Figure 5, the likelihood surfaces estimated by the Kalman filter, SIS 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 with optimal proposal distribution and Gibbs sampler converge to the exact likelihood, while VB suffers from loose lower bound and estimates an incorrect surface. The EM variants using the posterior distributions approximated by these algorithms converge to their respective maximum shown in this figure.

5.2 Inverse Gamma Markov Chains

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

$$
\begin{aligned}
z_1 &\sim \mathcal{IG}(z_1; a_z, b/a_z) && (44)\\
v_t|z_t &\sim \mathcal{IG}(v_t; a_v, z_t/a_v) && (45)\\
z_{t+1}|v_t &\sim \mathcal{IG}(z_{t+1}; a_z, v_t/a_z) && (46)\\
s_t &\sim \mathcal{N}(s_t; 0, v_t) && (47)
\end{aligned}
$$

which is a simplified version of the audio source model explained in Section 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 6). 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.

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 7, the Gibbs samples representing the distributions of consecutive variables in a chain have high correlations between them, whereas VB estimates these variables independently.

Figure 8 shows the hyperparameter (a_v and a_z) values that maximise the likelihood estimates of the SIS 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 and the same hyperparameter set maximises the likelihood.

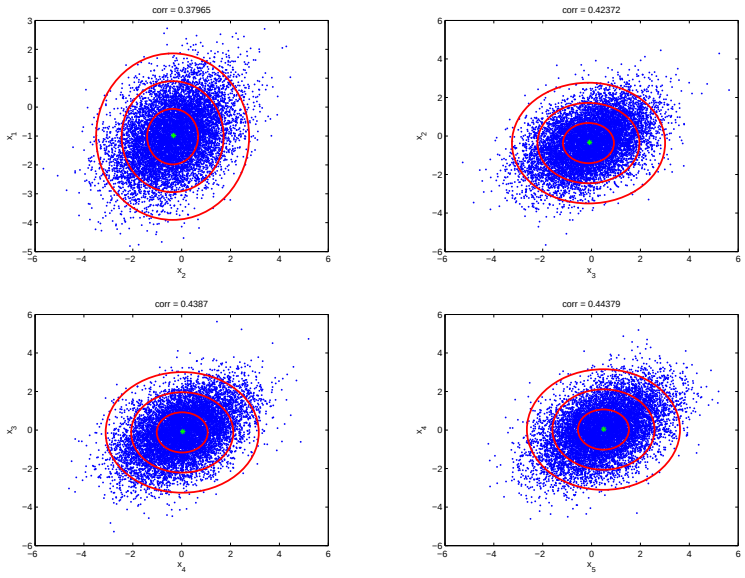


Figure 3: Samples drawn by a Gibbs sampler and variational estimates for consecutive variables in a chain of length 5. Variational distributions are shown in red circles or horizontal or vertical ellipses up to three standard deviations whereas samples are blue dots.

5.3 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 atoms of sources obtained by MDCT with inverse gamma Markov chains. As mentioned in [13], 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).

The horizontal model can be summarised as:

$$
\begin{aligned}
z_{\nu,1} &\sim \mathcal{IG}(z_{\nu,1}; a_z, b/a_z) && (48)\\
z_{\nu,\tau}|v_{\nu,\tau-1} &\sim \mathcal{IG}(z_{\nu,\tau}; a_z, v_{\nu,\tau-1}/a_z),\ \tau > 1 && (49)\\
v_{\nu,\tau}|z_{\nu,\tau} &\sim \mathcal{IG}(v_{\nu,\tau}; a_v, z_{\nu,\tau}/a_v) && (50)\\
s_{\nu,\tau}|v_{\nu,\tau} &\sim \mathcal{N}(s_{\nu,\tau}; 0, v_{\nu,\tau}) && (51)\\
x_{\nu,\tau}|s_{\nu,\tau}, r &\sim \mathcal{N}(x_{\nu,\tau}; s_{\nu,\tau}, r) && (52)\\
r &\sim \mathcal{IG}(r; a_r, b_r) && (53)
\end{aligned}
$$

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 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 9 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.

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 10). 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 11 spectrograms and SNRs of the estimated sources by the three

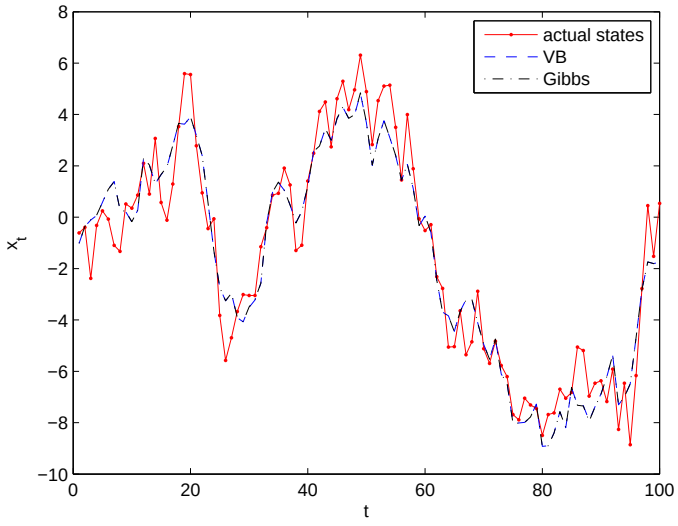


Figure 4: Actual state sequence and state estimates by VB and Gibbs sampler. Note that the estimates by the two methods coincide.

methods are presented. This audio signal is a piano recording and its MDCT coefficients are modelled with horizontal IGMCs.

5.4 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: $x_{\nu,\tau} = \sum_{j=1}^{N} s_{\nu,\tau}^{j}$.

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 evaluated in closed form:

$$\Sigma_{i,k} = v_{i,k}\,(1 - \kappa_{i,k}) \tag{54}$$

$$m_{i,k} = \kappa_{i,k} x_k \tag{55}$$

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 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. An ad hoc propagation scheme in which each step in the frequency axis is followed by a step in the time axis and vice versa can be adapted but it is beyond the scope of this report and omitted. Tables 1 and 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 artifacts ratio (SAR), defined as follows [21]

$$\text{SDR} \equiv 10 \log_{10} \frac{\|s_{\text{target}}\|^2}{\|e_{\text{interf}} + e_{\text{artif}}\|^2} \tag{56}$$

$$\text{SIR} \equiv 10 \log_{10} \frac{\|s_{\text{target}}\|^2}{\|e_{\text{interf}}\|^2} \tag{57}$$

$$\text{SAR} \equiv 10 \log_{10} \frac{\|s_{\text{target}} + e_{\text{interf}}\|^2}{\|e_{\text{artif}}\|^2} \tag{58}$$

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 artifacts due to the separation algorithm, e_{artif}.

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

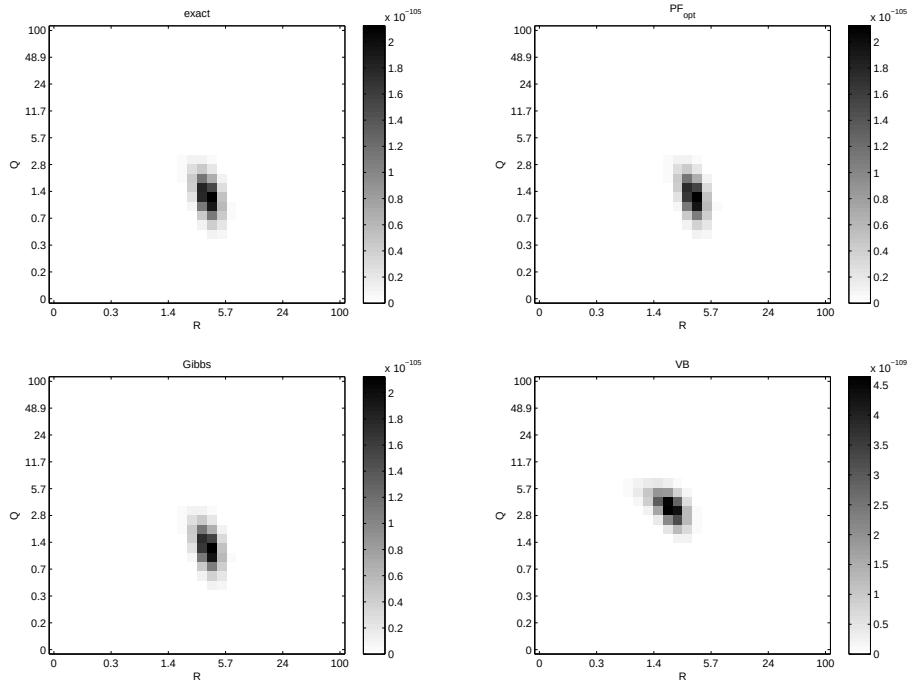


Figure 5: Likelihood versus model parameters (Q and R) for an observation sequence of length 100. The surfaces are almost the same for the Kalman filter (exact), SIS with optimal proposal distribution ($\mathrm{PF}_{\mathrm{opt}}$) and the Gibbs sampler (Gibbs). In particular, the Q and R pair that maximises 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).

	$\hat{\boldsymbol{s}}_1$			$\hat{\boldsymbol{s}}_2$		
	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
$\mathrm{Gibbs}_{\mathrm{EM}}$	-4.23	-2.42	4.82	1.34	13.13	1.85

Table 1: Single channel source separation results on a mixture of guitar ("Matte Kudasai") and drums ("Territory")

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 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 12. 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.

6 Conclusion and Discussions

In this report 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. In tonal audio signals there is a high correlation between the variances along the same frequency, whereas in percussive signals the correlation is higher along the same time frame. It is suitable to model these signals with horizontal and vertical IGMCs, respectively.

IGMCs are conditionally conjugate for all the variables in the chain. Moreover, when they are used to model the variances of an audio signal along with a Gaussian generative model for the source coefficients, the conditional conjugacy is preserved. As a result, inference of the latent variables in this model is very easy using variational Bayes and the Gibbs sampler. It is also possible to regard the model as a dynamic system and apply Sequential Monte Carlo methods for the inference. This model can be effectively used in source separation,

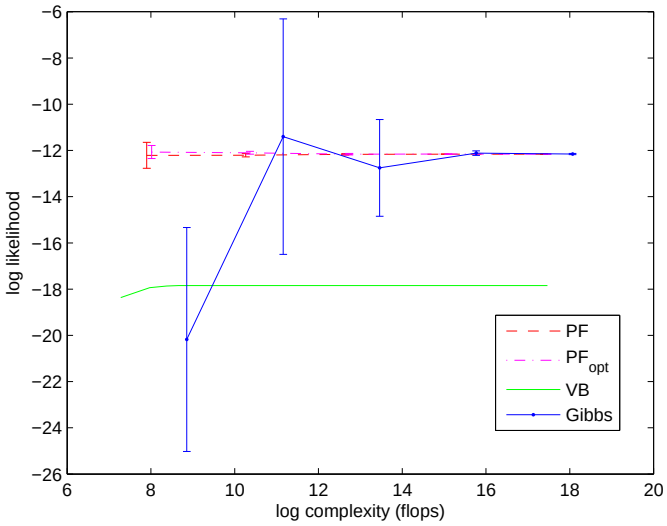


Figure 6: Log-likelihood approximations on an inverse Gamma state-space model of length 100 by different methods (bootstrap filter (PF), SIS with optimal proposal distribution ($\mathrm{PF}_{\mathrm{opt}}$), Gibbs sampler (Gibbs) and variational Bayes(VB)). More complex particle filters and Gibbs samplers are obtained by increasing the number of samples while in VB total number of iterations is increased.

	$\hat{\boldsymbol{s}}_1$			$\hat{\boldsymbol{s}}_2$		
	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
$\mathrm{Gibbs}_{\mathrm{EM}}$	-7.74	-6.19	4.62	-1.14	16.62	-0.97

Table 2: Single channel source separation results on a mixture of flute ("Vandringar I Vilsenhet") and drums ("Moby Dick")

transcription and interpolation problems. In this report we worked on the source separation problem.

Although inference is convenient in this model, 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. In order to make a better representation of the dependencies of source coefficients, inverse Gamma Markov random fields were proposed in [13]. Hyperparameter optimisation in these structures is difficult due to the unknown normalising constant. We will investigate this problem as a future work.

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.

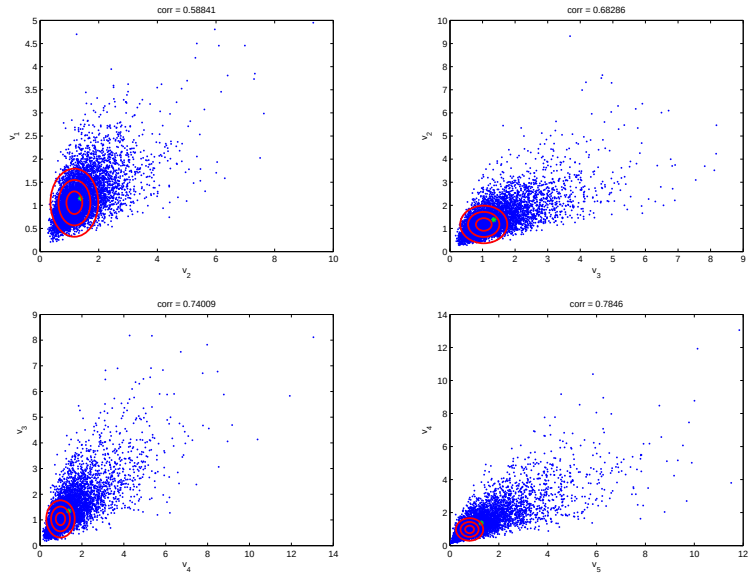


Figure 7: Samples drawn by a Gibbs sampler and variational estimates for consecutive variables in a chain of length 5. Variational distributions are shown in red circles or horizontal or vertical ellipses up to three standard deviations whereas samples are blue dots.

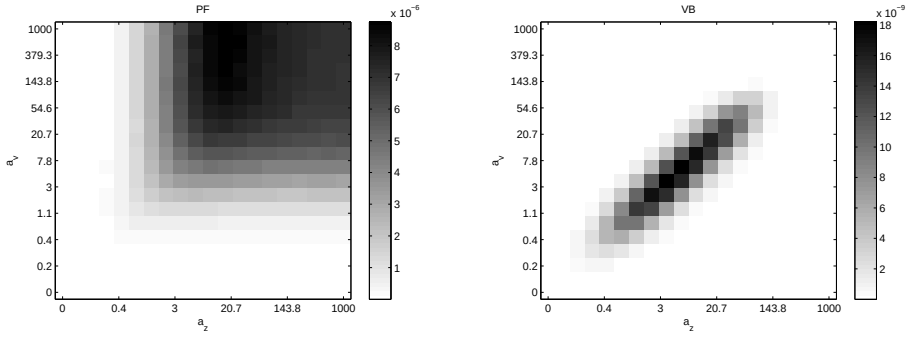


Figure 8: Likelihood versus hyperparameters (a_v and a_z) for an observation sequence of length 10. The likelihood functions attained by the SIS algorithm with optimal proposal distribution (on the left) and the VB lower bound (on the right) have very different characteristics and maxima.

[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. 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 *Advances in Neural Information Processing Systems*, 2000, pp. 841–847.

[5] M. Davies and N. Mitianoudis, "A Simple Mixture Model for Sparse Overcomplete ICA," in *IEEE proceedings in Vision, Image and Signal Processing*, vol. 151, no. 1, 2004, 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.

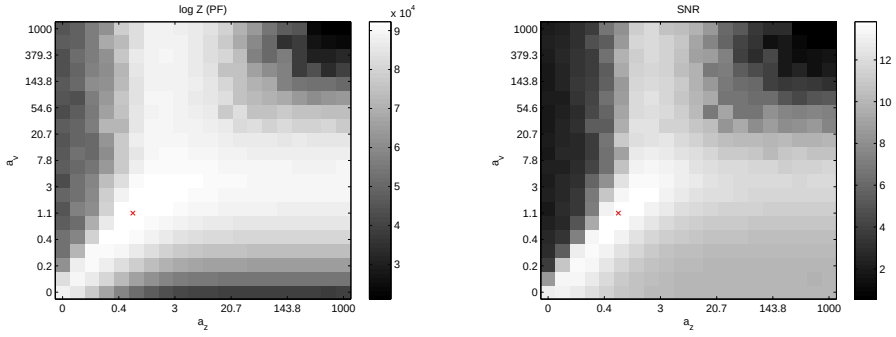


Figure 9: Log likelihood and reconstruction SNR values obtained by the SIS/R algorithm using the optimal proposal distribution. The surfaces are evaluated using a fixed value of b ($b = 10^{-4}$).

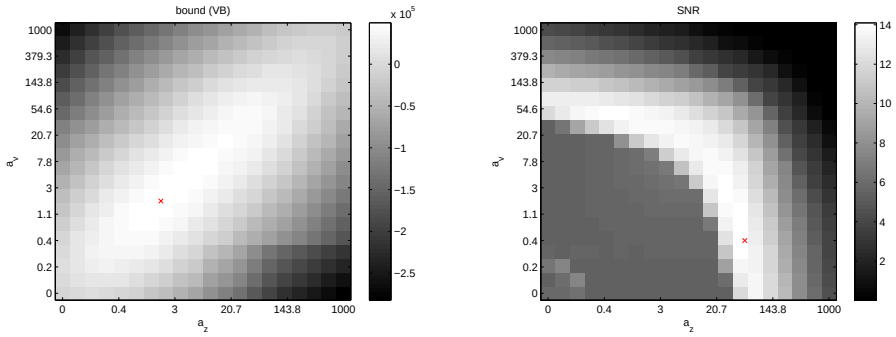


Figure 10: Lower bound and SNR values obtained by the variational Bayes method. The surfaces are evaluated using a fixed value of b ($b = 10^{-4}$).

[9] J. A. Palmer, K. Kreutz-Delgado, B. D. Rao, and S. Makeig, "Modeling and estimation of dependent subspaces with non-radially symmetric and skewed densities," in *Proceedings of the 7th International Symposium on Independent Component Analysis*, 2007.

[10] J. A. Palmer, "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. dissertation, University of California San Diego, 2006.

[11] J. A. Palmer, K. Kreutz-Delgado, D. P.Wipf, and B. D. Rao, "Variational EM algorithms for non-gaussian latent variable models," in *NIPS 2005*, 2005.

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

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

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

[15] D. MacKay, *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[16] A. Doucet, "On sequential Monte Carlo methods for Bayesian filtering," University Of Cambridge, UK, Department Of Engineering, Tech. Rep., 1998.

[17] A. P. Dempster, 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.

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

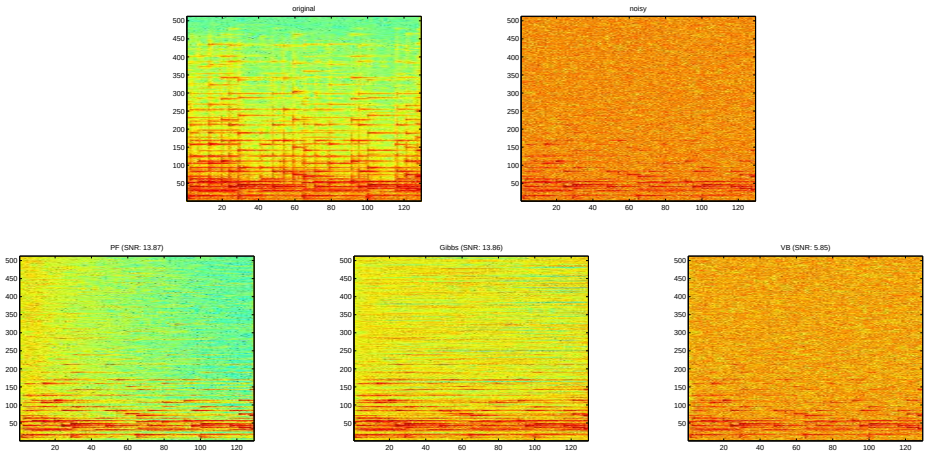


Figure 11: The figures on top are the spectrograms of the original and the noisy signals. The others are the reconstructed signals output by the three inference methods. In this example, results obtained by VB-EM are poor because the hyperparameters optimised by this method did not lead to better results. There are hyperparameter values that result in better reconstructions, but these parameters do not correspond to a local maxima of the lower bound.

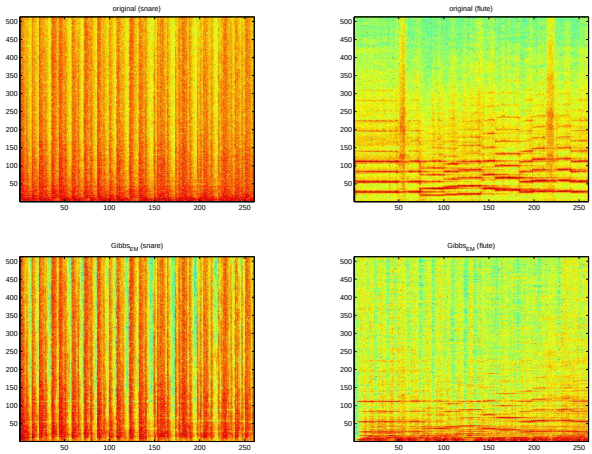


Figure 12: The spectrograms of the original sources (top) and the sources estimated by the Gibbs-EM algorithm (bottom) in the second experiment.

[19] R. E. Kalman 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.

[20] S. Chib, "Marginal likelihood from the gibbs output," *Journal of the American Statistical Association*, vol. 90, no. 432, pp. 1313–1321, 1995.

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

A Linear Gaussian State Space Model

A.1 Generative Model

x_t: state variable at time t

$$\begin{aligned} x_1 &\sim \mathcal{N}(x_1; 0, P) \\ x_t &\sim \mathcal{N}(x_t; Ax_{t-1}, Q), \quad t = 2..T \end{aligned}$$

y_t: observation at time t

$$y_t \sim \mathcal{N}(y_t; Cx_t, R), \quad t = 1..T$$

A.2 Expression of the Joint Posterior

$$\phi \equiv p(x_1)p(y_1|x_1)\prod_{t=2}^{T} p(x_t|x_{t-1})p(y_t|x_t)$$

$$\begin{aligned}
\log\phi &= -\frac{1}{2}\log 2\pi P - \frac{1}{2}\frac{x_1^2}{P} \\
&+ \sum_{t=2}^{T} -\frac{1}{2}\log 2\pi Q - \frac{1}{2}\frac{A^2 x_{t-1}^2}{Q} + \frac{A x_{t-1} x_t}{Q} - \frac{1}{2}\frac{x_t^2}{Q} \\
&+ \sum_{t=1}^{T} -\frac{1}{2}\log 2\pi R - \frac{1}{2}\frac{C^2 x_t^2}{R} + \frac{C x_t y_t}{R} - \frac{1}{2}\frac{y_t^2}{R}
\end{aligned}$$

A.3 Structure of the Q Distribution

$$Q = \prod_{t=1}^{T} Q(x_t)$$

A.4 Expression of the Variational Lower Bound

$$\log p(\boldsymbol{y}) \geq \langle \log\phi \rangle_{Q(\boldsymbol{x})} - \langle \log Q(\boldsymbol{x}) \rangle_{Q(\boldsymbol{x})}$$

$$\begin{aligned}
\langle \log\phi \rangle_{Q(\boldsymbol{x})} &= -\frac{1}{2}\log 2\pi P - \frac{1}{2}\frac{\langle x_1^2 \rangle}{P} \\
&+ \sum_{t=2}^{T} -\frac{1}{2}\log 2\pi Q - \frac{1}{2}\frac{A^2 \langle x_{t-1}^2 \rangle}{Q} + \frac{A\langle x_{t-1}\rangle\langle x_t\rangle}{Q} - \frac{1}{2}\frac{\langle x_t^2 \rangle}{Q} \\
&+ \sum_{t=1}^{T} -\frac{1}{2}\log 2\pi R - \frac{1}{2}\frac{C^2 \langle x_t^2 \rangle}{R} + \frac{C\langle x_t\rangle y_t}{R} - \frac{1}{2}\frac{y_t^2}{R}
\end{aligned}$$

A.5 Expressions of Q Distributions

x_t: state variable at time t

For $t = 1$

$$\begin{aligned}
\log Q(x_1) &=^{+} -\frac{1}{2}\left(\frac{1}{P} + \frac{A^2}{Q} + \frac{C^2}{R}\right) x_1^2 + \left(\frac{A\langle x_2\rangle}{Q} + \frac{C y_1}{R}\right) x_1 \\
Q(x_1) &= \mathcal{IG}(x_1; \mu_1, \Sigma_1) \\
\Sigma_1 &= \left(\frac{1}{P} + \frac{A^2}{Q} + \frac{C^2}{R}\right)^{-1} \\
\mu_1 &= \Sigma_1\left(\frac{A\langle x_2\rangle}{Q} + \frac{C y_1}{R}\right)
\end{aligned}$$

For $t = 2..T-1$

$$\begin{aligned}
\log Q(x_t) &=^{+} -\frac{1}{2}\left(\frac{A^2+1}{Q} + \frac{C^2}{R}\right) x_t^2 + \left(\frac{A(\langle x_{t-1}\rangle + \langle x_{t+1}\rangle)}{Q} + \frac{C y_t}{R}\right) x_t \\
Q(x_t) &= \mathcal{IG}(x_t; \mu_t, \Sigma_t) \\
\Sigma_t &= \left(\frac{A^2+1}{Q} + \frac{C^2}{R}\right)^{-1} \\
\mu_t &= \Sigma_t\left(\frac{A(\langle x_{t-1}\rangle + \langle x_{t+1}\rangle)}{Q} + \frac{C y_t}{R}\right)
\end{aligned}$$

For $t = T$

$$\begin{aligned}
\log Q(x_T) &=^{+} -\frac{1}{2}\left(\frac{1}{Q}+\frac{C^2}{R}\right)x_T^2 + \left(\frac{A\langle x_{T-1}\rangle}{Q}+\frac{Cy_T}{R}\right)x_T \\
Q(x_T) &= \mathcal{IG}(x_T;\mu_T,\Sigma_T) \\
\Sigma_T &= \left(\frac{1}{Q}+\frac{C^2}{R}\right)^{-1} \\
\mu_T &= \Sigma_T\left(\frac{A\langle x_{T-1}\rangle}{Q}+\frac{Cy_T}{R}\right)
\end{aligned}$$

A.6 Expressions of Full Conditionals

x_t: state variable at time t
For $t = 1$

$$\begin{aligned}
\log \phi(x_1; x_2^{(i)}, y_1) &=^{+} -\frac{1}{2}\left(\frac{1}{P}+\frac{A^2}{Q}+\frac{C^2}{R}\right)x_1^2 + \left(\frac{Ax_2^{(i)}}{Q}+\frac{Cy_1}{R}\right)x_1 \\
p(x_1|x_2^{(i)}, y_1) &= \mathcal{IG}(x_1;\mu_1,\Sigma_1) \\
\Sigma_1 &= \left(\frac{1}{P}+\frac{A^2}{Q}+\frac{C^2}{R}\right)^{-1} \\
\mu_1 &= \Sigma_1\left(\frac{Ax_2^{(i)}}{Q}+\frac{Cy_1}{R}\right)
\end{aligned}$$

For $t = 2..T-1$

$$\begin{aligned}
\log \phi(x_t; x_{t-1}^{(i)}, x_{t+1}^{(i)}, y_t) &=^{+} -\frac{1}{2}\left(\frac{A^2+1}{Q}+\frac{C^2}{R}\right)x_t^2 + \left(\frac{A(x_{t-1}^{(i)}+x_{t+1}^{(i)})}{Q}+\frac{Cy_t}{R}\right)x_t \\
p(x_t|x_{t-1}^{(i)}, x_{t+1}^{(i)}, y_t) &= \mathcal{IG}(x_t;\mu_t,\Sigma_t) \\
\Sigma_t &= \left(\frac{A^2+1}{Q}+\frac{C^2}{R}\right)^{-1} \\
\mu_t &= \Sigma_t\left(\frac{A(x_{t-1}^{(i)}+x_{t+1}^{(i)})}{Q}+\frac{Cy_t}{R}\right)
\end{aligned}$$

For $t = T$

$$\begin{aligned}
\log \phi(x_T; x_{T-1}^{(i)}, y_T) &=^{+} -\frac{1}{2}\left(\frac{1}{Q}+\frac{C^2}{R}\right)x_T^2 + \left(\frac{Ax_{T-1}^{(i)}}{Q}+\frac{Cy_T}{R}\right)x_T \\
p(x_T|x_{T-1}^{(i)}, y_T) &= \mathcal{IG}(x_T;\mu_T,\Sigma_T) \\
\Sigma_T &= \left(\frac{1}{Q}+\frac{C^2}{R}\right)^{-1} \\
\mu_T &= \Sigma_T\left(\frac{Ax_{T-1}^{(i)}}{Q}+\frac{Cy_T}{R}\right)
\end{aligned}$$

A.7 Optimal Proposal Distribution for SIS/R

For $t = 1$

$$\begin{aligned} p(x_1|y_1) &= \frac{p(y_1|x_1)p(x_1)}{p(y_1)} \\ &= \frac{p(y_1|x_1)p(x_1)}{\int p(y_1|x_1)p(x_1)\,dx_1} \\ &= \frac{\mathcal{N}(y_1; Cx_1, R)\mathcal{N}(x_1; 0, P)}{\int \mathcal{N}(y_1; Cx_1, R)\mathcal{N}(x_1; 0, P)\,dx_1} \\ &= \mathcal{N}(x_1; \mu_1, \Sigma_1) \\ \Sigma_t &= \frac{RP}{R + PC^2} \\ \mu_t &= \frac{Cy_1 P}{R + P} \end{aligned}$$

For $t = 2..T$

$$\begin{aligned} p(x_t|x_{t-1}^{(i)}, y_t) &= \frac{p(y_t|x_t)p(x_t|x_{t-1}^{(i)})}{p(y_t|x_{t-1}^{(i)})} \\ &= \frac{p(y_t|x_t)p(x_t|x_{t-1}^{(i)})}{\int p(y_t|x_t)p(x_t|x_{t-1}^{(i)})\,dx_t} \\ &= \frac{\mathcal{N}(y_t; Cx_t, R)\mathcal{N}(x_t; Ax_{t-1}^{(i)}, Q)}{\int \mathcal{N}(y_t; Cx_t, R)\mathcal{N}(x_t; Ax_{t-1}^{(i)}, Q)\,dx_t} \\ &= \mathcal{N}(x_t; \mu_t, \Sigma_t) \\ \Sigma_t &= \frac{RQ}{R + QC^2} \\ \mu_t &= \left(\frac{Cy_t}{R} + \frac{Ax_{t-1}^{(i)}}{Q}\right)\Sigma_t \end{aligned}$$

A.8 Marginal Likelihood Using Chib's Method

$$p(\boldsymbol{y}) = \frac{p(x_1^*)p(y_1|x_1^*)\prod_{t=2}^T p(x_t^*|x_{t-1}^*)p(y_t|x_t^*)}{p(x_1^*|\boldsymbol{y})\prod_{t=2}^T p(x_t^*|x_{t-1}^*, \boldsymbol{y})}$$

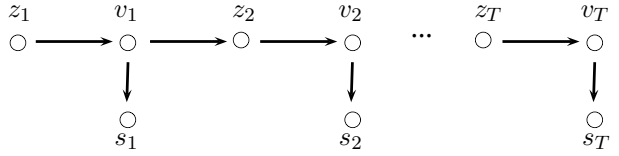
with

$$\begin{aligned} p(x_1^*|\boldsymbol{y}) &= \int p(x_1^*|x_2, \boldsymbol{y})p(x_2|\boldsymbol{y})\,dx_2 = \frac{1}{N}\sum_{i=1}^N p(x_1^*|x_2^{(i)}, \boldsymbol{y}) \\ p(x_t^*|x_{t-1}^*, \boldsymbol{y}) &= \int p(x_t^*|x_{t+1}, x_{t-1}^*, \boldsymbol{y})p(x_{t+1}|x_{t-1}^*, \boldsymbol{y})\,dx_{t+1} = \frac{1}{N}\sum_{i=1}^N p(x_t^*|x_{t+1}^{(i)}, x_{t-1}^*, \boldsymbol{y}) \end{aligned}$$

where $\left\{x_{t+1}^{(i)}\right\}_{i=1}^N$ for $t = 2..T-1$ are drawn from $p(x_{t+1}|x_{t-1}^*, \boldsymbol{y})$. $p(x_T^*|x_{T-1}^*, \boldsymbol{y})$ is available in closed form and there is no need to draw new samples.

B Inverse Gamma Markov Chains

B.1 Generative Model



z_t: auxiliary variable for the inverse Gamma chain

$$\begin{aligned} z_1 &\sim \mathcal{IG}(z_1; a_z, b/a_z) \\ z_t|v_{t-1} &\sim \mathcal{IG}(z_t; a_z, v_{t-1}/a_z), \;\; t = 2..T \end{aligned}$$

v_t: main variable of the chain at time t

$$v_t|z_t \sim \mathcal{IG}(v_t; a_v, z_t/a_v), \;\; t = 1..T$$

s_t: observation at time t

$$s_t|v_t \sim \mathcal{N}(s_t; 0, v_t), \;\; t = 1..T$$

B.2 Expression of the Joint Posterior

$$\phi \equiv \left(p(z_1)p(v_1|z_1) \prod_{t=2}^{T} p(v_t|z_t)p(z_t|v_{t-1}) \right) \left(\prod_{t=1}^{T} p(s_t|v_t) \right)$$

$$\begin{aligned} \log\phi &= \sum_{t=1}^{T} \left(-\frac{1}{2}\log 2\pi v_t - \frac{1}{2}\frac{s_t^2}{v_t} \right) \\ &+ \sum_{t=1}^{T} \left(-(a_v+1)\log v_t - \frac{a_v}{z_t v_t} - \log\Gamma(a_v) - a_v \log z_t + a_v \log a_v \right) \\ &+ \sum_{t=2}^{T} \left(-(a_z+1)\log z_t - \frac{a_z}{z_t v_{t-1}} - \log\Gamma(a_z) - a_z \log v_{t-1} + a_z \log a_z \right) \\ &\quad -(a_z+1)\log z_1 - \frac{a_z}{b z_1} - \log\Gamma(a_z) - a_z \log b + a_z \log a_z \end{aligned}$$

B.3 Structure of the Q Distribution

$$Q = \prod_{t=1}^{T} Q(z_t)Q(v_t)$$

B.4 Expression of the Variational Lower Bound

$$\log p(\boldsymbol{s}) \geq \langle \log\phi \rangle_{Q(\boldsymbol{v},\boldsymbol{z})} - \langle \log Q(\boldsymbol{v},\boldsymbol{z}) \rangle_{Q(\boldsymbol{v},\boldsymbol{z})}$$

$$\begin{aligned} \langle \log\phi \rangle_{Q(\boldsymbol{v},\boldsymbol{z})} &= \sum_{t=1}^{T} \left(-\frac{1}{2}\log 2\pi - \frac{1}{2}\langle \log v_t \rangle - \frac{1}{2}\langle s_t^2 \rangle \left\langle \frac{1}{v_t} \right\rangle \right) \\ &+ \sum_{t=1}^{T} \left(-(a_v+1)\langle \log v_t \rangle - a_v \left\langle \frac{1}{z_t} \right\rangle \left\langle \frac{1}{v_t} \right\rangle - \log\Gamma(a_v) - a_v \langle \log z_t \rangle + a_v \log a_v \right) \\ &+ \sum_{t=2}^{T} \left(-(a_z+1)\langle \log z_t \rangle - a_z \left\langle \frac{1}{z_t} \right\rangle \left\langle \frac{1}{v_{t-1}} \right\rangle - \log\Gamma(a_z) - a_z \langle \log v_{t-1} \rangle + a_z \log a_z \right) \\ &\quad -(a_z+1)\langle \log z_1 \rangle - \frac{a_z}{b z_1} - \log\Gamma(a_z) - a_z \log b + a_z \log a_z \end{aligned}$$

B.5 Expressions of Q Distributions

z_t: auxiliary variable for the inverse Gamma chain

For $t = 1$

$$\begin{aligned} \log Q(z_1) &=^{+} -(a_z + a_v + 1)\log z_1 - \left(a_v \left\langle \frac{1}{v_1} \right\rangle + \frac{a_z}{b} \right) \frac{1}{z_1} \\ Q(z_1) &= \mathcal{IG}(z_1; \alpha_{z,1}, \beta_{z,1}) \\ \alpha_{z,1} &= a_z + a_v \\ \beta_{z,1} &= \left(a_v \left\langle \frac{1}{v_1} \right\rangle + \frac{a_z}{b} \right)^{-1} \end{aligned}$$

For $t = 2..T$

$$\begin{aligned}
\log Q(z_t) &=^+ -(a_z + a_v + 1)\log z_t - \left(a_v \left\langle \frac{1}{v_t} \right\rangle + a_z \left\langle \frac{1}{v_{t-1}} \right\rangle\right) \frac{1}{z_t} \\
Q(z_t) &= \mathcal{IG}(z_t; \alpha_{z,t}, \beta_{z,t}) \\
\alpha_{z,t} &= a_z + a_v \\
\beta_{z,t} &= \left(a_v \left\langle \frac{1}{v_t} \right\rangle + a_z \left\langle \frac{1}{v_{t-1}} \right\rangle\right)^{-1}
\end{aligned}$$

v_t: main variable of the chain at time t
For $t = 1..T-1$

$$\begin{aligned}
\log Q(v_t) &=^+ -\left(a_v + a_z + \frac{1}{2} + 1\right) \log v_t - \left(\frac{1}{2}\langle s_t^2 \rangle + a_v \left\langle \frac{1}{z_t} \right\rangle + a_z \left\langle \frac{1}{z_{t+1}} \right\rangle\right) \frac{1}{v_t} \\
Q(v_t) &= \mathcal{IG}(v_t; \alpha_{v,t}, \beta_{v,t}) \\
\alpha_{v,t} &= a_v + a_z + \frac{1}{2} \\
\beta_{v,t} &= \left(\frac{1}{2}\langle s_t^2 \rangle + a_v \left\langle \frac{1}{z_t} \right\rangle + a_z \left\langle \frac{1}{z_{t+1}} \right\rangle\right)^{-1}
\end{aligned}$$

For $t = T$

$$\begin{aligned}
\log Q(v_T) &=^+ -\left(a_v + \frac{1}{2} + 1\right) \log v_T - \left(\frac{1}{2}\langle s_T^2 \rangle + a_v \left\langle \frac{1}{z_T} \right\rangle\right) \frac{1}{v_T} \\
Q(v_T) &= \mathcal{IG}(v_T; \alpha_{v,T}, \beta_{v,T}) \\
\alpha_{v,T} &= a_v + \frac{1}{2} \\
\beta_{v,T} &= \left(\frac{1}{2}\langle s_T^2 \rangle + a_v \left\langle \frac{1}{z_T} \right\rangle\right)^{-1}
\end{aligned}$$

B.6 Expressions of Full Conditionals

z_t: auxiliary variable for the inverse Gamma chain
For $t = 1$

$$\begin{aligned}
\log \phi(z_1; v_1^{(i)}) &=^+ -(a_z + a_v + 1)\log z_1 - \left(\frac{a_v}{v_1^{(i)}} + \frac{a_z}{b}\right) \frac{1}{z_1} \\
p(z_1 | v_1^{(i)}) &= \mathcal{IG}(z_1; \alpha_{z,1}, \beta_{z,1}) \\
\alpha_{z,1} &= a_z + a_v \\
\beta_{z,1} &= \left(\frac{a_v}{v_1^{(i)}} + \frac{a_z}{b}\right)^{-1}
\end{aligned}$$

For $t = 2..T$

$$\begin{aligned}
\log \phi(z_t; v_{t-1}^{(i)}, v_t^{(i)}) &=^+ -(a_z + a_v + 1)\log z_t - \left(\frac{a_v}{v_t^{(i)}} + \frac{a_z}{v_{t-1}^{(i)}}\right) \frac{1}{z_t} \\
p(z_t | v_{t-1}^{(i)}, v_t^{(i)}) &= \mathcal{IG}(z_t; \alpha_{z,t}, \beta_{z,t}) \\
\alpha_{z,t} &= a_z + a_v \\
\beta_{z,t} &= \left(\frac{a_v}{v_t^{(i)}} + \frac{a_z}{v_{t-1}^{(i)}}\right)^{-1}
\end{aligned}$$

v_t: main variable of the chain at time t
For $t = 1..T-1$

$$\begin{aligned}
\log \phi(v_t; z_t^{(i)}, z_{t+1}^{(i)}, s_t) &=^+ -\left(a_v + a_z + \frac{1}{2} + 1\right) \log v_t - \left(\frac{1}{2} {s_t^{(i)}}^2 + \frac{a_v}{z_t^{(i)}} + \frac{a_z}{z_{t+1}^{(i)}}\right) \frac{1}{v_t} \\
p(v_t | z_t^{(i)}, z_{t+1}^{(i)}, s_t) &= \mathcal{IG}(v_t; \alpha_{v,t}, \beta_{v,t}) \\
\alpha_{v,t} &= a_v + a_z + \frac{1}{2} \\
\beta_{v,t} &= \left(\frac{1}{2} {s_t^{(i)}}^2 + \frac{a_v}{z_t^{(i)}} + \frac{a_z}{z_{t+1}^{(i)}}\right)^{-1}
\end{aligned}$$

For $t = T$

$$\begin{aligned}
\log \phi(v_T; z_T^{(i)}, s_T) &=^+ -\left(a_v + \frac{1}{2} + 1\right) \log v_T - \left(\frac{1}{2} {s_T^{(i)}}^2 + \frac{a_v}{z_T^{(i)}}\right) \frac{1}{v_T} \\
p(v_T | z_T^{(i)}, s_T) &= \mathcal{IG}(v_T; \alpha_{v,T}, \beta_{v,T}) \\
\alpha_{v,T} &= a_v + \frac{1}{2} \\
\beta_{v,T} &= \left(\frac{1}{2} {s_T^{(i)}}^2 + \frac{a_v}{z_T^{(i)}}\right)^{-1}
\end{aligned}$$

B.7 Optimal Proposal Distribution for SIS/R

$$\begin{aligned}
p(v_t | z_t^{(i)}, s_t) &= \frac{p(s_t|v_t) p(v_t | z_t^{(i)})}{p(s_t | z_t^{(i)})} \\
&= \frac{p(s_t|v_t) p(v_t | z_t^{(i)})}{\int p(s_t|v_t) p(v_t | z_t^{(i)}) \, dv_t} \\
&= \frac{\mathcal{N}(s_t; 0, v_t) \mathcal{IG}(v_t; a_v, z_t^{(i)}/a_v)}{\int \mathcal{N}(s_t; 0, v_t) \mathcal{IG}(v_t; a_v, z_t^{(i)}/a_v) \, dv_t} \\
&= \mathcal{IG}(v_t; \alpha_t, \beta_t) \\
\alpha_t &= a_v + \frac{1}{2} \\
\beta_t &= \left(\frac{a_v}{z_t^{(i)}} + \frac{s_t^2}{2}\right)^{-1}
\end{aligned}$$

B.8 Marginal Likelihood Using Chib's Method

$$p(\boldsymbol{s}) = \frac{p(z_1^*) p(v_1^* | z_1^*) p(s_1 | v_1^*) \prod_{t=2}^T p(z_t^* | v_{t-1}^*) p(v_t^* | z_t^*) p(s_t | v_t^*)}{p(z_1^* | \boldsymbol{s}) p(v_1^* | z_1^*, \boldsymbol{s}) \prod_{t=2}^T p(z_t^* | v_{t-1}^*, \boldsymbol{s}) p(v_t^* | z_t^*, \boldsymbol{s})}$$

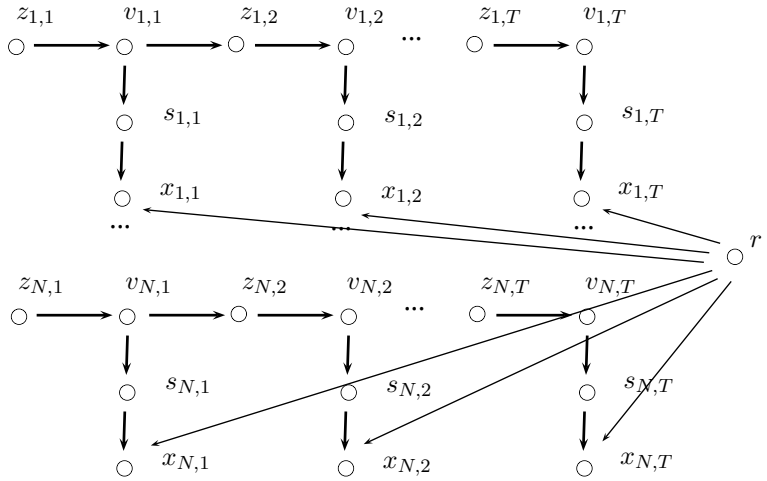
with

$$\begin{aligned}
p(z_1^* | \boldsymbol{s}) &= \int p(z_1^* | v_1, \boldsymbol{s}) p(v_1 | \boldsymbol{s}) \, dv_1 = \frac{1}{N} \sum_{i=1}^N p(z_1^* | v_1^{(i)}, \boldsymbol{s}) \\
p(z_t^* | v_{t-1}^*, \boldsymbol{s}) &= \int p(z_t^* | v_t, v_{t-1}^*, \boldsymbol{s}) p(v_t | v_{t-1}^*, \boldsymbol{s}) \, dv_t = \frac{1}{N} \sum_{i=1}^N p(z_t^* | v_t^{(i)}, v_{t-1}^*, \boldsymbol{s}), \quad t = 2..T \\
p(v_t^* | z_t^*, \boldsymbol{s}) &= \int p(v_t^* | z_{t+1}, z_t^*, \boldsymbol{s}) p(z_{t+1} | z_t^*, \boldsymbol{s}) \, dz_{t+1} = \frac{1}{N} \sum_{i=1}^N p(v_t^* | z_{t+1}^{(i)}, z_t^*, \boldsymbol{s}), \quad t = 1..T-1
\end{aligned}$$

where $\left\{v_t^{(i)}\right\}_{i=1}^N$ for $t = 2..T$ are drawn from $p(v_t | v_{t-1}^*, \boldsymbol{s})$ and $\left\{z_{t+1}^{(i)}\right\}_{i=1}^N$ for $t = 1..T-1$ are drawn from $p(z_{t+1} | z_t^*, \boldsymbol{s})$. $p(v_T^* | z_T^*, \boldsymbol{y})$ is available in closed form and there is no need to draw new samples.

C Denoising using Inverse Gamma Markov Chains

C.1 Generative Model



$z_{\nu,\tau}$: auxiliary variable for the inverse gamma chain

$$\begin{aligned} z_{\nu,1} &\sim \mathcal{IG}(z_{\nu,1}; a_z, b/a_z), \ \ \nu = 1..N \\ z_{\nu,\tau}|v_{\nu,\tau-1} &\sim \mathcal{IG}(z_{\nu,\tau}; a_z, v_{\nu,\tau-1}/a_z), \ \ \tau = 2..T, \nu = 1..N \end{aligned}$$

$v_{\nu,\tau}$: variance of the source coefficient with frequency index ν at time τ

$$v_{\nu,\tau}|z_{\nu,\tau} \sim \mathcal{IG}(v_{\nu,\tau}; a_v, z_{\nu,\tau}/a_v), \ \ \tau = 1..T, \nu = 1..N$$

$s_{\nu,\tau}$: source coefficient with frequency index ν at time τ

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

r: variance of observation noise

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

$x_{\nu,\tau}$: observation coefficient with frequency index ν at time τ

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

Vertical source models can be obtained simply by interchanging ν and τ indices of the z and v variables.

C.2 Expression of the Joint Posterior

$$\phi \equiv \left(\prod_{\nu=1}^{N} \left(p(z_{\nu,1}) p(v_{\nu,1}|z_{\nu,1}) \prod_{\tau=2}^{T} p(v_{\nu,\tau}|z_{\nu,\tau}) p(z_{\nu,\tau}|v_{\nu,\tau-1}) \right) \left(\prod_{\tau=1}^{T} p(x_{\nu,\tau}|s_{\nu,\tau}, r) p(s_{\nu,\tau}|v_{\nu,\tau}) \right) \right) p(r)$$

$$\begin{aligned} \log \phi &= \sum_{\nu=1}^{N} \sum_{\tau=1}^{T} \left(-\frac{1}{2} \log 2\pi r - \frac{1}{2} \frac{x_{\nu,\tau}^2}{r} + \frac{x_{\nu,\tau} s_{\nu,\tau}}{r} - \frac{1}{2} \frac{s_{\nu,\tau}^2}{r} \right) \\ &+ \sum_{\nu=1}^{N} \sum_{\tau=1}^{T} \left(-\frac{1}{2} \log 2\pi v_{\nu,\tau} - \frac{1}{2} \frac{s_{\nu,\tau}^2}{v_{\nu,\tau}} \right) \\ &+ \sum_{\nu=1}^{N} \sum_{\tau=1}^{T} \left(-(a_v + 1) \log v_{\nu,\tau} - \frac{a_v}{z_{\nu,\tau} v_{\nu,\tau}} - \log \Gamma(a_v) - a_v \log z_{\nu,\tau} + a_v \log a_v \right) \\ &+ \sum_{\nu=1}^{N} \sum_{\tau=2}^{T} \left(-(a_z + 1) \log z_{\nu,\tau} - \frac{a_z}{z_{\nu,\tau} v_{\nu,\tau-1}} - \log \Gamma(a_z) - a_z \log v_{\nu,\tau-1} + a_z \log a_z \right) \\ &+ \sum_{\nu=1}^{N} \left(-(a_z + 1) \log z_{\nu,1} - \frac{a_z}{b z_{\nu,1}} - \log \Gamma(a_z) - a_z \log b + a_z \log a_z \right) \\ &+ \left(-(a_r + 1) \log r - \frac{1}{b_r r} - \log \Gamma(a_r) - a_r \log b_r \right) \end{aligned}$$

C.3 Structure of the Q Distribution

$$Q = Q(r) \prod_{\nu=1}^{N} \prod_{\tau=1}^{T} Q(z_{\nu,\tau}) Q(v_{\nu,\tau}) Q(s_{\nu,\tau})$$

C.4 Expressions of Q Distributions

$z_{\nu,\tau}$: auxiliary variable for the gamma chain
For $\tau = 1$

$$\begin{aligned}
\log Q(z_{\nu,1}) &=^{+} -(a_z + a_v + 1)\log z_{\nu,1} - \left(a_v \left\langle \frac{1}{v_{\nu,1}} \right\rangle + \frac{a_z}{b} \right) \frac{1}{z_{\nu,1}} \\
Q(z_{\nu,1}) &= \mathcal{IG}(z_{\nu,1}; \alpha_{z,\nu,1}, \beta_{z,\nu,1}) \\
\alpha_{z,\nu,1} &= a_z + a_v \\
\beta_{z,\nu,1} &= \left(a_v \left\langle \frac{1}{v_{\nu,1}} \right\rangle + \frac{a_z}{b} \right)^{-1}
\end{aligned}$$

For $\tau = 2..T$

$$\begin{aligned}
\log Q(z_{\nu,\tau}) &=^{+} -(a_z + a_v + 1)\log z_{\nu,\tau} - \left(a_v \left\langle \frac{1}{v_{\nu,\tau}} \right\rangle + a_z \left\langle \frac{1}{v_{\nu,\tau-1}} \right\rangle \right) \frac{1}{z_{\nu,\tau}} \\
Q(z_{\nu,\tau}) &= \mathcal{IG}(z_{\nu,\tau}; \alpha_{z,\nu,\tau}, \beta_{z,\nu,\tau}) \\
\alpha_{z,\nu,\tau} &= a_z + a_v \\
\beta_{z,\nu,\tau} &= \left(a_v \left\langle \frac{1}{v_{\nu,\tau}} \right\rangle + a_z \left\langle \frac{1}{v_{\nu,\tau-1}} \right\rangle \right)^{-1}
\end{aligned}$$

$v_{\nu,\tau}$: variance of the source coefficient with frequency index ν at time τ
For $\tau = 1..T-1$

$$\begin{aligned}
\log Q(v_{\nu,\tau}) &=^{+} -\left(a_v + a_z + \frac{1}{2} + 1 \right) \log v_{\nu,\tau} - \left(\frac{1}{2}\langle s_{\nu,\tau}^2 \rangle + a_v \left\langle \frac{1}{z_{\nu,\tau}} \right\rangle + a_z \left\langle \frac{1}{z_{\nu,\tau+1}} \right\rangle \right) \frac{1}{v_{\nu,\tau}} \\
Q(v_{\nu,\tau}) &= \mathcal{IG}(v_{\nu,\tau}; \alpha_{v,\nu,\tau}, \beta_{v,\nu,\tau}) \\
\alpha_{v,\nu,\tau} &= a_v + a_z + \frac{1}{2} \\
\beta_{v,\nu,\tau} &= \left(\frac{1}{2}\langle s_{\nu,\tau}^2 \rangle + a_v \left\langle \frac{1}{z_{\nu,\tau}} \right\rangle + a_z \left\langle \frac{1}{z_{\nu,\tau+1}} \right\rangle \right)^{-1}
\end{aligned}$$

For $\tau = T$

$$\begin{aligned}
\log Q(v_{\nu,T}) &=^{+} -\left(a_v + \frac{1}{2} + 1 \right) \log v_{\nu,T} - \left(\frac{1}{2}\langle s_{\nu,T}^2 \rangle + a_v \left\langle \frac{1}{z_{\nu,T}} \right\rangle \right) \frac{1}{v_{\nu,T}} \\
Q(v_{\nu,T}) &= \mathcal{IG}(v_{\nu,T}; \alpha_{v,\nu,T}, \beta_{v,\nu,T}) \\
\alpha_{v,\nu,T} &= a_v + \frac{1}{2} \\
\beta_{v,\nu,T} &= \left(\frac{1}{2}\langle s_{\nu,T}^2 \rangle + a_v \left\langle \frac{1}{z_{\nu,T}} \right\rangle \right)^{-1}
\end{aligned}$$

$s_{\nu,\tau}$: source coefficient with frequency index ν at time τ

$$\begin{aligned}
\log Q(s_{\nu,\tau}) &=^{+} -\frac{1}{2} \left(\left\langle \frac{1}{r} \right\rangle + \left\langle \frac{1}{v_{\nu,\tau}} \right\rangle \right) s_{\nu,\tau}^2 + x_{\nu,\tau} \left\langle \frac{1}{r} \right\rangle s_{\nu,\tau} \\
Q(s_{\nu,\tau}) &= \mathcal{N}(s_{\nu,\tau}; \mu_{\nu,\tau}, \Sigma_{\nu,\tau}) \\
\Sigma_{\nu,\tau} &= \left(\left\langle \frac{1}{r} \right\rangle + \left\langle \frac{1}{v_{\nu,\tau}} \right\rangle \right)^{-1} \\
\mu_{\nu,\tau} &= x_{\nu,\tau} \left\langle \frac{1}{r} \right\rangle \Sigma_{\nu,\tau}
\end{aligned}$$

r: variance of observation noise

$$\begin{aligned}
\log Q(r) &=^{+} -\left(a_r + \frac{NT}{2} + 1\right)\log r - \left(\sum_{\nu}\sum_{\tau}\left(\frac{1}{2}x_{\nu,\tau}^2 - x_{\nu,\tau}\langle s_{\nu,\tau}\rangle + \frac{1}{2}\langle s_{\nu,\tau}^2\rangle\right) + \frac{1}{b_r}\right)\frac{1}{r} \\
Q(r) &= \mathcal{IG}(r;\alpha_r,\beta_r) \\
\alpha_r &= a_r + \frac{NT}{2} \\
\beta_r &= \left(\sum_{\nu}\sum_{\tau}\left(\frac{1}{2}x_{\nu,\tau}^2 - x_{\nu,\tau}\langle s_{\nu,\tau}\rangle + \frac{1}{2}\langle s_{\nu,\tau}^2\rangle\right) + \frac{1}{b_r}\right)^{-1}
\end{aligned}$$

C.5 Expressions of Full Conditionals

$z_{\nu,\tau}$: auxiliary variable for the gamma chain
For $\tau = 1$

$$\begin{aligned}
\log\phi(z_{\nu,1};v_{\nu,1}^{(i)}) &=^{+} -(a_z + a_v + 1)\log z_{\nu,1} - \left(\frac{a_v}{v_{\nu,1}^{(i)}} + \frac{a_z}{b}\right)\frac{1}{z_{\nu,1}} \\
p(z_{\nu,1}|v_{\nu,1}^{(i)}) &= \mathcal{IG}(z_{\nu,1};\alpha_{z,\nu,1},\beta_{z,\nu,1}) \\
\alpha_{z,\nu,1} &= a_z + a_v \\
\beta_{z,\nu,1} &= \left(\frac{a_v}{v_{\nu,1}^{(i)}} + \frac{a_z}{b}\right)^{-1}
\end{aligned}$$

For $\tau = 2..T$

$$\begin{aligned}
\log\phi(z_{\nu,\tau};v_{\nu,\tau-1}^{(i)},v_{\nu,\tau}^{(i)}) &=^{+} -(a_z + a_v + 1)\log z_{\nu,\tau} - \left(\frac{a_v}{v_{\nu,\tau}^{(i)}} + \frac{a_z}{v_{\nu,\tau-1}^{(i)}}\right)\frac{1}{z_{\nu,\tau}} \\
p(z_{\nu,\tau}|v_{\nu,\tau-1}^{(i)},v_{\nu,\tau}^{(i)}) &= \mathcal{IG}(z_{\nu,\tau};\alpha_{z,\nu,\tau},\beta_{z,\nu,\tau}) \\
\alpha_{z,\nu,\tau} &= a_z + a_v \\
\beta_{z,\nu,\tau} &= \left(\frac{a_v}{v_{\nu,\tau}^{(i)}} + \frac{a_z}{v_{\nu,\tau-1}^{(i)}}\right)^{-1}
\end{aligned}$$

$v_{\nu,\tau}$: variance of the source coefficient with frequency index ν at time τ
For $\tau = 1..T-1$

$$\begin{aligned}
\log\phi(v_{\nu,\tau};z_{\nu,\tau}^{(i)},z_{\nu,\tau+1}^{(i)},s_{\nu,\tau}^{(i)}) &=^{+} -\left(a_v + a_z + \frac{1}{2} + 1\right)\log v_{\nu,\tau} - \left(\frac{1}{2}{s_{\nu,\tau}^{(i)}}^2 + \frac{a_v}{z_{\nu,\tau}^{(i)}} + \frac{a_z}{z_{\nu,\tau+1}^{(i)}}\right)\frac{1}{v_{\nu,\tau}} \\
p(v_{\nu,\tau}|z_{\nu,\tau}^{(i)},z_{\nu,\tau+1}^{(i)},s_{\nu,\tau}^{(i)}) &= \mathcal{IG}(v_{\nu,\tau};\alpha_{v,\nu,\tau},\beta_{v,\nu,\tau}) \\
\alpha_{v,\nu,\tau} &= a_v + a_z + \frac{1}{2} \\
\beta_{v,\nu,\tau} &= \left(\frac{1}{2}{s_{\nu,\tau}^{(i)}}^2 + \frac{a_v}{z_{\nu,\tau}^{(i)}} + \frac{a_z}{z_{\nu,\tau+1}^{(i)}}\right)^{-1}
\end{aligned}$$

For $\tau = T$

$$\begin{aligned}
\log\phi(v_{\nu,T};z_{\nu,T}^{(i)},s_{\nu,T}^{(i)}) &=^{+} -\left(a_v + \frac{1}{2} + 1\right)\log v_{\nu,T} - \left(\frac{1}{2}{s_{\nu,T}^{(i)}}^2 + \frac{a_v}{z_{\nu,T}^{(i)}}\right)\frac{1}{v_{\nu,T}} \\
p(v_{\nu,T}|z_{\nu,T}^{(i)},s_{\nu,T}^{(i)}) &= \mathcal{IG}(v_{\nu,T};\alpha_{v,\nu,T},\beta_{v,\nu,T}) \\
\alpha_{v,\nu,T} &= a_v + \frac{1}{2} \\
\beta_{v,\nu,T} &= \left(\frac{1}{2}{s_{\nu,T}^{(i)}}^2 + \frac{a_v}{z_{\nu,T}^{(i)}}\right)^{-1}
\end{aligned}$$

$s_{\nu,\tau}$: source coefficient with frequency index ν at time τ

$$\begin{aligned}
\log \phi(s_{\nu,\tau}; v^{(i)}_{\nu,\tau}, x_{\nu,\tau}, r^{(i)}) &=^{+} -\frac{1}{2}\left(\frac{1}{r^{(i)}} + \frac{1}{v^{(i)}_{\nu,\tau}}\right) s^2_{\nu,\tau} + x_{\nu,\tau}\frac{1}{r^{(i)}} s_{\nu,\tau} \\
p(s_{\nu,\tau}|v^{(i)}_{\nu,\tau}, x_{\nu,\tau}, r^{(i)}) &= \mathcal{N}(s_{\nu,\tau}; \mu_{\nu,\tau}, \Sigma_{\nu,\tau}) \\
\Sigma_{\nu,\tau} &= \left(\frac{1}{r^{(i)}} + \frac{1}{v^{(i)}_{\nu,\tau}}\right)^{-1} \\
\mu_{\nu,\tau} &= \frac{x_{\nu,\tau}\Sigma_{\nu,\tau}}{r^{(i)}}
\end{aligned}$$

r: variance of observation noise

$$\begin{aligned}
\log \phi(r; s^{(i)}_{1:N,1:T}, x_{1:N,1:T}) &=^{+} -\left(a_r + \frac{NT}{2} + 1\right)\log r \\
&\quad - \left(\sum_{\nu=1}^{N}\sum_{\tau=1}^{T}\left(\frac{1}{2}x^2_{\nu,\tau} - x_{\nu,\tau}s^{(i)}_{\nu,\tau} + \frac{1}{2}{s^{(i)}_{\nu,\tau}}^2\right) + \frac{1}{b_r}\right)\frac{1}{r} \\
p(r|s^{(i)}_{1:N,1:T}, x_{1:N,1:T}) &= \mathcal{IG}(r; \alpha_r, \beta_r) \\
\alpha_r &= a_r + \frac{NT}{2} \\
\beta_r &= \left(\sum_{\nu=1}^{N}\sum_{\tau=1}^{T}\left(\frac{1}{2}x^2_{\nu,\tau} - x_{\nu,\tau}s^{(i)}_{\nu,\tau} + \frac{1}{2}{s^{(i)}_{\nu,\tau}}^2\right) + \frac{1}{b_r}\right)^{-1}
\end{aligned}$$

C.6 Optimal Proposal Distribution for SIS/R

$$\begin{aligned}
p(s_{\nu,\tau}|v^{(i)}_{\nu,\tau}, r^{(i)}, x_{\nu,\tau}) &= \frac{p(x_{\nu,\tau}|s_{\nu,\tau}, r^{(i)})p(s_{\nu,\tau}|v^{(i)}_{\nu,\tau})}{p(x_{\nu,\tau}|v^{(i)}_{\nu,\tau}, r^{(i)})} \\
&= \frac{p(x_{\nu,\tau}|s_{\nu,\tau}, r^{(i)})p(s_{\nu,\tau}|v^{(i)}_{\nu,\tau})}{\int p(x_{\nu,\tau}|s_{\nu,\tau}, r^{(i)})p(s_{\nu,\tau}|v^{(i)}_{\nu,\tau})\, ds_{\nu,\tau}} \\
&= \frac{\mathcal{N}(x_{\nu,\tau}; s_{\nu,\tau}, r^{(i)})\mathcal{N}(s_{\nu,\tau}; 0, v^{(i)}_{\nu,\tau})}{\int \mathcal{N}(x_{\nu,\tau}; s_{\nu,\tau}, r^{(i)})\mathcal{N}(s_{\nu,\tau}; 0, v^{(i)}_{\nu,\tau})\, ds_{\nu,\tau}} \\
&= \mathcal{N}(s_{\nu,\tau}; \mu_{\nu,\tau}, \Sigma_{\nu,\tau}) \\
\Sigma_{\nu,\tau} &= \frac{v^{(i)}_{\nu,\tau} r^{(i)}}{v^{(i)}_{\nu,\tau} + r^{(i)}} \\
\mu_{\nu,\tau} &= \frac{x_{\nu,\tau}\Sigma_{\nu,\tau}}{r^{(i)}}
\end{aligned}$$

D Single Channel Source Separation using Inverse Gamma Markov Chains

D.1 Generative Model

$z^j_{\nu,\tau}$: auxiliary variables for the inverse gamma chains of the j^{th} source

$$\begin{aligned}
z^j_{\nu,1} &\sim \mathcal{IG}(z^j_{\nu,1}; a^j_z, b^j/a^j_z), \;\; \nu = 1..N, j = 1..J \\
z^j_{\nu,\tau}|v^j_{\nu,\tau-1} &\sim \mathcal{IG}(z^j_{\nu,\tau}; a^j_z, v^j_{\nu,\tau-1}/a^j_z), \;\; \nu = 1..N, \tau = 2..T, j = 1..J
\end{aligned}$$

$v^j_{\nu,\tau}$: variance of the coefficient of the j^{th} source with frequency index ν at time τ

$$v^j_{\nu,\tau}|z^j_{\nu,\tau} \sim \mathcal{IG}(v^j_{\nu,\tau}; a^j_v, z^j_{\nu,\tau}/a^j_v), \;\; \nu = 1..N, \tau = 1..T, j = 1..J$$

$s^j_{\nu,\tau}$: coefficient of the j^{th} source with frequency index ν at time τ

$$s^j_{\nu,\tau}|v^j_{\nu,\tau} \sim \mathcal{N}(s^j_{\nu,\tau}; 0, v^j_{\nu,\tau}), \;\; \nu = 1..N, \tau = 1..T, j = 1..J$$

$x_{\nu,\tau}$: observation coefficient with frequency index ν at time τ

$$x_{\nu,\tau} = \sum_{j=1}^{J} s_{\nu,\tau}^j \sim \mathcal{N}\left(x_{\nu,\tau}; \sum_{j=1}^{J-1} s_{\nu,\tau}^j, v_{\nu,\tau}^J\right), \;\; \nu = 1..N, \tau = 1..T$$

Vertical source models can be obtained simply by interchanging ν and τ indices of the z and v variables.

D.2 Expression of the Joint Posterior

$$\begin{aligned}
\phi &\equiv \prod_{j=1}^{J}\left(\prod_{\nu=1}^{N}\left(p(z_{\nu,1}^j)p(v_{\nu,1}^j|z_{\nu,1}^j)\prod_{\tau=2}^{T}p(v_{\nu,\tau}^j|z_{\nu,\tau}^j)p(z_{\nu,\tau}^j|v_{\nu,\tau-1}^j)\right)\left(\prod_{\tau=1}^{T}p(s_{\nu,\tau}^j|v_{\nu,\tau}^j)\right)\right) \\
&\quad \prod_{\nu=1}^{N}\prod_{\tau=1}^{T}p(x_{\nu,\tau}|s_{\nu,\tau}^{1:J-1}, v_{\nu,\tau}^J)
\end{aligned}$$

$$\begin{aligned}
\log\phi &= \sum_{j=1}^{J}\sum_{\nu=1}^{N}\sum_{\tau=1}^{T}\left(-(a_v^j+1)\log v_{\nu,\tau}^j - \frac{a_v^j}{z_{\nu,\tau}^j v_{\nu,\tau}^j} - \log\Gamma(a_v^j) - a_v^j\log z_{\nu,\tau}^j + a_v^j\log a_v^j\right) \\
&+ \sum_{j=1}^{J}\sum_{\nu=1}^{N}\sum_{\tau=2}^{T}\left(-(a_z^j+1)\log z_{\nu,\tau}^j - \frac{a_z^j}{z_{\nu,\tau}^j v_{\nu,\tau-1}^j} - \log\Gamma(a_z^j) - a_z^j\log v_{\nu,\tau-1}^j + a_z^j\log a_z^j\right) \\
&+ \sum_{j=1}^{J}\sum_{\nu=1}^{N}\left(-(a_z^j+1)\log z_{\nu,1}^j - \frac{a_z^j}{b z_{\nu,1}^j} - \log\Gamma(a_z^j) - a_z^j\log b^j + a_z^j\log a_z^j\right) \\
&+ \sum_{j=1}^{J-1}\sum_{\nu=1}^{N}\sum_{\tau=1}^{T}\left(-\frac{1}{2}\log 2\pi v_{\nu,\tau}^j - \frac{1}{2}\frac{{s_{\nu,\tau}^j}^2}{v_{\nu,\tau}^j}\right) \\
&+ \sum_{\nu=1}^{N}\sum_{\tau=1}^{T}\left(-\frac{1}{2}\log 2\pi v_{\nu,\tau}^J - \frac{1}{2}\frac{{x_{\nu,\tau}}^2}{v_{\nu,\tau}^J} + \frac{x_{\nu,\tau}\sum_{j=1}^{J-1}s_{\nu,\tau}^j}{v_{\nu,\tau}^J} - \frac{1}{2}\frac{(\sum_{j=1}^{J-1}s_{\nu,\tau}^j)^2}{v_{\nu,\tau}^J}\right)
\end{aligned}$$

D.3 Structure of the Q Distribution

$$Q = \prod_{\nu=1}^{N}\prod_{\tau=1}^{T}Q(s_{\nu,\tau}^{1:J-1})\prod_{j=1}^{J}Q(z_{\nu,\tau}^j)Q(v_{\nu,\tau}^j)$$

D.4 Expressions of Q Distributions

$z_{\nu,\tau}^j$: auxiliary variables for the inverse gamma chains of the j^{th} source

For $\tau = 1$

$$\begin{aligned}
\log Q(z_{\nu,1}^j) &=^+ -(a_z^j + a_v^j + 1)\log z_{\nu,1}^j - \left(a_v^j\left\langle\frac{1}{v_{\nu,1}^j}\right\rangle + \frac{a_z^j}{b^j}\right)\frac{1}{z_{\nu,1}^j} \\
Q(z_{\nu,1}^j) &= \mathcal{IG}(z_{\nu,1}^j; \alpha_{z,\nu,1}^j, \beta_{z,\nu,1}^j) \\
\alpha_{z,\nu,1}^j &= a_z^j + a_v^j \\
\beta_{z,\nu,1}^j &= \left(a_v^j\left\langle\frac{1}{v_{\nu,1}^j}\right\rangle + \frac{a_z^j}{b^j}\right)^{-1}
\end{aligned}$$

For $\tau = 2..T$

$$\begin{aligned}
\log Q(z_{\nu,\tau}^j) &=^+ -(a_z^j + a_v^j + 1)\log z_{\nu,\tau}^j - \left(a_v^j\left\langle\frac{1}{v_{\nu,\tau}^j}\right\rangle + a_z^j\left\langle\frac{1}{v_{\nu,\tau-1}^j}\right\rangle\right)\frac{1}{z_{\nu,\tau}^j} \\
Q(z_{\nu,\tau}^j) &= \mathcal{IG}(z_{\nu,\tau}^j; \alpha_{z,\nu,\tau}^j, \beta_{z,\nu,\tau}^j) \\
\alpha_{z,\nu,\tau}^j &= a_z^j + a_v^j \\
\beta_{z,\nu,\tau}^j &= \left(a_v^j\left\langle\frac{1}{v_{\nu,\tau}^j}\right\rangle + a_z^j\left\langle\frac{1}{v_{\nu,\tau-1}^j}\right\rangle\right)^{-1}
\end{aligned}$$

$v^j_{\nu,\tau}$: variance of the coefficient of the j^{th} source with frequency index ν at time τ
For $\tau = 1..T-1$

$$\begin{aligned}
\log Q(v^j_{\nu,\tau}) &=^+ -\left(a^j_v + a^j_z + \frac{1}{2} + 1\right)\log v^j_{\nu,\tau} - \left(\frac{1}{2}\langle {s^j_{\nu,\tau}}^2\rangle + a^j_v\left\langle\frac{1}{z^j_{\nu,\tau}}\right\rangle + a^j_z\left\langle\frac{1}{z^j_{\nu,\tau+1}}\right\rangle\right)\frac{1}{v^j_{\nu,\tau}} \\
Q(v^j_{\nu,\tau}) &= \mathcal{IG}(v^j_{\nu,\tau}; \alpha^j_{v,\nu,\tau}, \beta^j_{v,\nu,\tau}) \\
\alpha^j_{v,\nu,\tau} &= a^j_v + a^j_z + \frac{1}{2} \\
\beta^j_{v,\nu,\tau} &= \left(\frac{1}{2}\langle {s^j_{\nu,\tau}}^2\rangle + a^j_v\left\langle\frac{1}{z^j_{\nu,\tau}}\right\rangle + a^j_z\left\langle\frac{1}{z^j_{\nu,\tau+1}}\right\rangle\right)^{-1}
\end{aligned}$$

For $\tau = T$

$$\begin{aligned}
\log Q(v^j_{\nu,T}) &=^+ -\left(a^j_v + \frac{1}{2} + 1\right)\log v^j_{\nu,T} - \left(\frac{1}{2}\langle {s^j_{\nu,T}}^2\rangle + a^j_v\left\langle\frac{1}{z^j_{\nu,T}}\right\rangle\right)\frac{1}{v^j_{\nu,T}} \\
Q(v^j_{\nu,T}) &= \mathcal{IG}(v^j_{\nu,T}; \alpha^j_{v,\nu,T}, \beta^j_{v,\nu,T}) \\
\alpha^j_{v,\nu,T} &= a^j_v + \frac{1}{2} \\
\beta^j_{v,\nu,T} &= \left(\frac{1}{2}\langle {s^j_{\nu,T}}^2\rangle + a^j_v\left\langle\frac{1}{z^j_{\nu,T}}\right\rangle\right)^{-1}
\end{aligned}$$

$s^j_{\nu,\tau}$: coefficient of the j^{th} source with frequency index ν at time τ

$$\begin{aligned}
Q(s^{1:J-1}_{\nu,\tau}) &= \mathcal{N}(s^{1:J-1}_{\nu,\tau}; \mu^{1:J-1}_{\nu,\tau}, \Sigma^{1:J-1}_{\nu,\tau}) \\
\kappa^j_{\nu,\tau} &= \frac{\left\langle\frac{1}{v^j_{\nu,\tau}}\right\rangle^{-1}}{\sum_{j=1}^J\left\langle\frac{1}{v^j_{\nu,\tau}}\right\rangle^{-1}} \\
\Sigma^j_{\nu,\tau} &= \left\langle\frac{1}{v^j_{\nu,\tau}}\right\rangle^{-1}(1-\kappa^j_{\nu,\tau}) \\
\mu^j_{\nu,\tau} &= \kappa^j_{\nu,\tau}x_{\nu,\tau}
\end{aligned}$$

D.5 Expressions of Full Conditionals

$z^j_{\nu,\tau}$: auxiliary variables for the inverse gamma chains of the j^{th} source
For $\tau = 1$

$$\begin{aligned}
\log\phi(z^j_{\nu,1}; v^{j,(i)}_{\nu,1}) &=^+ -(a^j_z + a^j_v + 1)\log z^j_{\nu,1} - \left(\frac{a^j_v}{v^{j,(i)}_{\nu,1}} + \frac{a^j_z}{b^j}\right)\frac{1}{z^j_{\nu,1}} \\
p(z^j_{\nu,1}|v^{j,(i)}_{\nu,1}) &= \mathcal{IG}(z^j_{\nu,1}; \alpha^j_{z,\nu,1}, \beta^j_{z,\nu,1}) \\
\alpha^j_{z,\nu,1} &= a^j_z + a^j_v \\
\beta^j_{z,\nu,1} &= \left(\frac{a^j_v}{v^{j,(i)}_{\nu,1}} + \frac{a^j_z}{b^j}\right)^{-1}
\end{aligned}$$

For $\tau = 2..T$

$$\begin{aligned}
\log\phi(z^j_{\nu,\tau}; v^{j,(i)}_{\nu,\tau-1}, v^{j,(i)}_{\nu,\tau}) &=^+ -(a^j_z + a^j_v + 1)\log z^j_{\nu,\tau} - \left(\frac{a^j_v}{v^{j,(i)}_{\nu,\tau}} + \frac{a^j_z}{v^{j,(i)}_{\nu,\tau-1}}\right)\frac{1}{z^j_{\nu,\tau}} \\
p(z^j_{\nu,\tau}|v^{j,(i)}_{\nu,\tau-1}, v^{j,(i)}_{\nu,\tau}) &= \mathcal{IG}(z^j_{\nu,\tau}; \alpha^j_{z,\nu,\tau}, \beta^j_{z,\nu,\tau}) \\
\alpha^j_{z,\nu,\tau} &= a^j_z + a^j_v \\
\beta^j_{z,\nu,\tau} &= \left(\frac{a^j_v}{v^{j,(i)}_{\nu,\tau}} + \frac{a^j_z}{v^{j,(i)}_{\nu,\tau-1}}\right)^{-1}
\end{aligned}$$

$v^j_{\nu,\tau}$: variance of the coefficient of the j^{th} source with frequency index ν at time τ
For $\tau = 1..T-1$

$$\begin{aligned}
\log \phi(v^j_{\nu,\tau}; z^{j,(i)}_{\nu,\tau}, z^{j,(i)}_{\nu,\tau+1}, s^{j,(i)}_{\nu,\tau}) &=^+ -\left(a^j_v + a^j_z + \frac{1}{2} + 1\right) \log v^j_{\nu,\tau} - \left(\frac{1}{2} {s^{j,(i)}_{\nu,\tau}}^2 + \frac{a^j_v}{z^{j,(i)}_{\nu,\tau}} + \frac{a^j_z}{z^{j,(i)}_{\nu,\tau+1}}\right) \frac{1}{v^j_{\nu,\tau}} \\
p(v^j_{\nu,\tau} | z^{j,(i)}_{\nu,\tau}, z^{j,(i)}_{\nu,\tau+1}, s^{j,(i)}_{\nu,\tau}) &= \mathcal{IG}(v^j_{\nu,\tau}; \alpha^j_{v,\nu,\tau}, \beta^j_{v,\nu,\tau}) \\
\alpha^j_{v,\nu,\tau} &= a^j_v + a^j_z + \frac{1}{2} \\
\beta^j_{v,\nu,\tau} &= \left(\frac{1}{2} {s^{j,(i)}_{\nu,\tau}}^2 + \frac{a^j_v}{z^{j,(i)}_{\nu,\tau}} + \frac{a^j_z}{z^{j,(i)}_{\nu,\tau+1}}\right)^{-1}
\end{aligned}$$

For $\tau = T$

$$\begin{aligned}
\log \phi(v^j_{\nu,T}; z^{j,(i)}_{\nu,T}, s^{j,(i)}_{\nu,T}) &=^+ -\left(a^j_v + \frac{1}{2} + 1\right) \log v^j_{\nu,T} - \left(\frac{1}{2} {s^{j,(i)}_{\nu,T}}^2 + \frac{a^j_v}{z^{j,(i)}_{\nu,T}}\right) \frac{1}{v^j_{\nu,T}} \\
p(v^j_{\nu,T} | z^{j,(i)}_{\nu,T}, s^{j,(i)}_{\nu,T}) &= \mathcal{IG}(v^j_{\nu,T}; \alpha^j_{v,\nu,T}, \beta^j_{v,\nu,T}) \\
\alpha^j_{v,\nu,T} &= a^j_v + \frac{1}{2} \\
\beta^j_{v,\nu,T} &= \left(\frac{1}{2} {s^{j,(i)}_{\nu,T}}^2 + \frac{a^j_v}{z^{j,(i)}_{\nu,T}}\right)^{-1}
\end{aligned}$$

$s^j_{\nu,\tau}$: coefficient of the j^{th} source with frequency index ν at time τ

$$\begin{aligned}
p(s^{1:J-1}_{\nu,\tau} | v^{1:J,(i)}_{\nu,\tau}, x_{\nu,\tau}) &= \mathcal{N}(s^{1:J-1}_{\nu,\tau}; \mu^{1:J-1}_{\nu,\tau}, \Sigma^{1:J-1}_{\nu,\tau}) \\
\kappa^j_{\nu,\tau} &= \frac{v^{j,(i)}_{\nu,\tau}}{\sum_{j=1}^J v^{j,(i)}_{\nu,\tau}} \\
\Sigma^j_{\nu,\tau} &= v^{j,(i)}_{\nu,\tau} \left(1 - \kappa^j_{\nu,\tau}\right) \\
\mu^j_{\nu,\tau} &= \kappa^j_{\nu,\tau} x_{\nu,\tau}
\end{aligned}$$