BLIND SEPARATION OF SPARSE SOURCES USING VARIATIONAL EM

Ali Taylan Cemgil, Cédric Févotte and Simon J. Godsill

Engineering Dept, University of Cambridge
Trumpington St, CB2 1PZ, Cambridge, UK
phone: + 44 1223 765582, fax: + 44 1223 332662, email: {atc27,cf269,sjg}@eng.cam.ac.uk

ABSTRACT

In this paper, we tackle the general linear instantaneous model (possibly underdetermined and noisy) using the assumption of *sparsity* of the sources on a given dictionary. We model the sparsity of expansion coefficients with a Student t prior. The conjugate-exponential characterisation of the t distribution as an infinite mixture of scaled Gaussians enables us to derive an efficient variational expectation maximisation algorithm (V-EM). The resulting deterministic algorithm has superior properties in terms of computation time and achieves a separation performance comparable in quality to alternative methods based on Markov Chain Monte Carlo (MCMC).

1. INTRODUCTION

In Blind Source Separation (BSS), the task is to estimate n source signals from the sole observation of m linear mixtures. In this paper, we consider linear instantaneous mixtures.

The (over)determined case ($m \geq n$) for non-noisy linear instantaneous mixtures has been widely studied and many solutions now exist for this scenario. Conceptual and computational difficulties arise when dealing with the underdetermined case ($m < n$) and/or with noisy mixtures. The underdetermined case in particular is very challenging because contrary to (over)determined mixtures, estimating the mixing system is not sufficient for reconstruction of the sources, since for $m < n$ the mixing matrix is not invertible. It appears that separation of underdetermined mixtures requires important prior information on the sources to allow their reconstruction. Such prior information is also helpful for reconstructing the sources in noisy environments.

In this paper, we tackle the general linear instantaneous model (possibly underdetermined, possibly noisy) using the assumption of *sparsity* of the sources on a given dictionary. This assumption means that only a few coefficients of the decomposition of the sources on the basis are significantly non-zero. The use of sparsity to handle the general linear instantaneous model, has arisen in several papers in the areas of learning [1, 2, 3] and source separation [4, 5, 6, 7, 8, 9, 10]. In the latter case the aim of the methods becomes the estimation of the coefficients of the sources in the dictionary and not the time series themselves.

In particular, in [9, 10], audio sources are decomposed on a MDCT basis (a local cosine transform orthonormal basis) and sparsity of the coefficients is modelled with a Student t distribution. A Gibbs sampler (a standard Markov Chains Monte Carlo simulation method) is derived to sample from the posterior distribution of the parameters (the mixing matrix, the sources coefficients, the additive noise variance and the hyper-parameters of the Student t prior distributions). Minimum Mean Square Error estimates of the coefficients of the sources are then computed and time domain estimates of the sources are reconstructed by applying inverse-MDCT.

A. T. Cemgil and C. Févotte respectively acknowledge European Commission funded Research Training Networks MUSCLE and HASSIP (HPRN-CT-2002-00285) for support of this work.

The main advantage of the MCMC approach is its generality and attractive theoretical properties. In earlier work [9, 10], the MCMC based method for source separation is found to be robust, for example one can estimate mixing matrices with almost linearly dependent columns. Moreover, the results were reproducible, i.e., independent of initialisation, the method could compute a perceptually reasonable separation of the sources. However, the method comes at the price of heavy computational burden which for certain applications renders the method unpractical. To alleviate this, we investigate in this paper an alternative approach for computing the required integrals based on variational approximations [11, 12].

The paper is organised as follows: Section 2 introduces notations and assumptions, in Section 3 we present briefly the variational approach. Section 4 presents separation results with noisy underdetermined mixtures of audio signals decomposed on a MDCT basis. Our approach is compared with the MCMC approach of [9, 10]. Conclusions and perspectives are given in Section 5.

2. MODEL AND ASSUMPTIONS

2.1 Model and aim

We consider the following standard linear instantaneous model, $\forall t = 0, \ldots, N-1$:

$$\mathbf{x}_t \quad = \quad \mathbf{A}\,\mathbf{s}_t + \mathbf{n}_t \tag{1}$$

where $\mathbf{x}_t = [x_{1,t}, \ldots, x_{j,t}, \ldots, x_{m,t}]^T$ is a vector of size m containing the observations at each channel at time t. Similarly, $\mathbf{s}_t = [s_{1,t}, \ldots, s_{n,t}]^T$ is a vector of size n containing the sources, and $\mathbf{n}_t$ is a noise term. Variables without time index t denote entire sequences of samples, *e.g*, $\mathbf{x} = [\mathbf{x}_0, \ldots, \mathbf{x}_{N-1}]$.

In BSS, the aim is to estimate the sources $\mathbf{s}$ and possibly the mixing matrix $\mathbf{A}$ up to the standard BSS indetermination on gain and order, that is, compute $\hat{\mathbf{s}}$ and $\hat{\mathbf{A}}$ such that

$$\hat{\mathbf{A}} = \mathbf{A}\,\mathbf{D}\,\mathbf{P} \quad \hat{\mathbf{s}} = \mathbf{P}^T\,\mathbf{D}^{-1}\,\mathbf{s} \tag{2}$$

where $\mathbf{D}$ is a diagonal matrix and $\mathbf{P}$ is a permutation matrix.

2.2 Assumptions

2.2.1 Time domain / Transform domain

We assume that we are given a basis on which the sources adopt a sparse representation. Again, this means that only a low proportion of coefficients of the decompositions are significantly different from zero. Let $\mathbf{\Phi}$ be a $N \times N$ invertible matrix defining such a basis. We denote $\tilde{y} = y\,\mathbf{\Phi}$ the decomposition of a time series y in $\mathbf{\Phi}$. The decomposition of the observations is written:

$$\tilde{\mathbf{x}} = \mathbf{x}\,\mathbf{\Phi} = \mathbf{A}\,\tilde{\mathbf{s}} + \tilde{\mathbf{n}} \tag{3}$$

or, equivalently, $\forall k = 1, \ldots, N$:

$$\tilde{\mathbf{x}}_k = \mathbf{A}\,\tilde{\mathbf{s}}_k + \tilde{\mathbf{n}}_k \tag{4}$$

Because $\mathbf{\Phi}$ is a basis, Eq. (4) is strictly equivalent to Eq. (1), which means that separation can be performed equivalently either in the time domain or in the transform domain. In the following, we work in the transform domain.

2.2.2 Mixing Matrix

We assume that the mixing matrix has a Gaussian prior distribution.

$$p(\mathbf{A}) \sim \mathcal{N}(\mathbf{vec}\,\mathbf{A}; \mathbf{vec}\,\Omega_A, \Sigma_A) \tag{5}$$

The operator **vec** "reshapes" a matrix as a column vector by concatenating its columns. This distribution gives the expected orientations and scale of the mixing matrix.

2.2.3 Model of sparsity

We assume that the sequences of coefficients $\tilde{s}_i$ are independently and identically distributed (i.i.d) with Student t distribution $t(\alpha_i, \lambda_i)$ defined as

$$p(\tilde{s}_{i,k}|\alpha_i, \lambda_i) = \frac{\Gamma(\frac{\alpha_i+1}{2})}{\lambda_i \sqrt{\alpha_i\,\pi}\,\Gamma(\frac{\alpha_i}{2})} \left(1 + \frac{1}{\alpha_i}\left(\frac{\tilde{s}_{i,k}}{\lambda_i}\right)^2\right)^{-\frac{\alpha_i+1}{2}} \tag{6}$$

where index $i = 1 \dots n$ runs over sources, $k = 0 \dots N-1$ runs over transform coefficients, α_i is the "degrees of freedom" and λ_i is a scale parameter, tied across time slices. For small α_i, the Student t density gathers most of its probability mass around zero and exhibits "fatter tails" than the normal distribution. The Student t distribution has proved to be a relevant model for sparsity in [13, 9, 10].

An important feature of the Student t distribution is that it can be expressed as a Scale Mixture of Gaussians (SMoG) [14], such that

$$p(\tilde{s}_{i,k}|\alpha_i, \lambda_i) = \int_0^{+\infty} \mathcal{N}(\tilde{s}_{i,k}; 0, v_{i,k}^{-1})\, \mathcal{G}\left(v_{i,k}; \frac{\alpha_i}{2}, \frac{2}{\alpha_i\,\lambda_i^2}\right) dv_{i,k} \tag{7}$$

where $\mathcal{G}(x; \gamma, \beta)$ is the Gamma distribution, defined in Appendix A. Thus, using a t distribution is equivalent to introducing the auxiliary random variable $v_{i,k}$ and assuming that $\tilde{s}_{i,k}$ is conditionally Gaussian upon $v_{i,k}$ with

$$p(\tilde{s}_{i,k}|v_{i,k}) = \mathcal{N}(\tilde{s}_{i,k}; 0, v_{i,k}^{-1}) \tag{8}$$

$$p(v_{i,k}|\alpha_i, \lambda_i) = \mathcal{G}\left(v_{i,k}; \frac{\alpha_i}{2}, \frac{2}{\alpha_i\,\lambda_i^2}\right) \tag{9}$$

This characterisation is important for the variational approximation methods described later. In the following, we denote $\mathbf{v}_k \equiv [v_{1,k}, \dots, v_{n,k}]^T$ and $\mathbf{v} \equiv [\mathbf{v}_0, \dots, \mathbf{v}_{N-1}]$. We further let $V_k \equiv \mathrm{diag}\,(\mathbf{v}_k)$. The hyperparameters are denoted as $\boldsymbol{\alpha} = [\alpha_1, \dots, \alpha_n]$ and $\boldsymbol{\lambda} = [\lambda_1, \dots, \lambda_n]$.

2.2.4 Mutual independence

We assume that the sequences of source coefficients are mutually independent, such that $p(\tilde{\mathbf{s}}) = \prod_{i=1}^n p(\tilde{s}_i)$. As pointed out in [5], the assumption of mutual independence of the *coefficients* in the transform domain can be considered more realistic than the mutual independence of the sources in time domain, which is the standard assumption of ICA methods.

2.2.5 Noise properties

We assume that $\tilde{\mathbf{n}}$ is a i.i.d Gaussian noise with diagonal precision (inverse covariance) matrix

$$\begin{aligned} R &= \mathrm{diag}\,(r_1, \dots, r_j, \dots r_m) \\ r_j &\sim \mathcal{G}\,(r_j; a_{R,j}, b_{R,j}) \end{aligned}$$

We point out that when an orthonormal basis is used (*i.e*, $\mathbf{\Phi}^{-1} = \mathbf{\Phi}^T$), $\mathbf{n}$ is also a Gaussian i.i.d noise with precision R.

3. VARIATIONAL EM

Given the model, we can formulate source separation as the following Bayesian posterior inference problem:

$$\begin{aligned} \mathcal{P} &= \frac{1}{Z_x} p(\mathbf{x}|\tilde{\mathbf{s}}, \mathbf{A}) p(\tilde{\mathbf{s}}|\mathbf{v}) p(\mathbf{A}) p(\mathbf{v}|\boldsymbol{\lambda}) p(R) && (10) \\ &\equiv \frac{1}{Z_x} \phi(\tilde{\mathbf{s}}, R, \mathbf{A}, \mathbf{v}|\boldsymbol{\lambda}) && (11) \end{aligned}$$

where $Z_x = p(\mathbf{x})$ is the normalisation constant known as the *evidence*. Once this posterior is calculated, the desired quantities (e.g. sources) can be estimated from this posterior by marginalization:

$$p(\tilde{\mathbf{s}}|\mathbf{x}) = \int dR d\mathbf{A} d\mathbf{v}\; \mathcal{P}(\tilde{\mathbf{s}}, R, \mathbf{A}, \mathbf{v}|\boldsymbol{\lambda}) \tag{12}$$

Unfortunately, for the above model, the exact posterior is intractable. Therefore, it is necessary to resort to numerical approximation techniques such as MCMC [10].

One possible approximation method as an alternative to MCMC, that leads usually to a faster optimisation procedure is the *structured mean field* approach, also known as *variational Bayes* [11, 12]. Variational Bayes boils down to approximating the integrand $\mathcal{P}$ defined in (10) with a simpler distribution $\mathcal{Q}$ such that the integral (12) in becomes tractable. An intuitive interpretation of this technique is minimising the KL divergence [15] with respect to (the parameters of) $\mathcal{Q}$ where

$$KL(\mathcal{Q}||\mathcal{P}) = \langle \log \mathcal{Q} \rangle_{\mathcal{Q}} - \langle \log \mathcal{P} \rangle_{\mathcal{Q}} \tag{13}$$

Here, $\langle f(x) \rangle_{p(x)} \equiv \int dx p(x) f(x)$ denotes the expectation of f w.r.t. p. Using non-negativity of KL [15] we obtain a lower bound on the evidence

$$\log Z_x \geq \langle \log \phi(\tilde{\mathbf{s}}, R, \mathbf{A}, \mathbf{v}|\boldsymbol{\lambda}) \rangle_{\mathcal{Q}} - \langle \log \mathcal{Q} \rangle_{\mathcal{Q}} \tag{14}$$

It is clear that maximising this lower bound is equivalent to finding the "nearest" $\mathcal{Q}$ to $\mathcal{P}$ in terms of KL. For our model, we choose the approximating distribution $\mathcal{Q}$ of form

$$\mathcal{Q} \equiv q(\mathbf{A}) q(R) \prod_{k=0}^{N-1} q(\tilde{\mathbf{s}}_k) \prod_{i=1}^{n} q(v_{i,k})$$

Although a closed form solution for $\mathcal{Q}$ still can not be found, it can be easily shown, e.g. see [16], that each factor potential $\mathcal{Q}_\alpha$ of the optimal approximating distribution should satisfy the following fixed point equation

$$\mathcal{Q}_\alpha \propto \exp\left(\langle \log \phi(\tilde{\mathbf{s}}, R, \mathbf{A}, \mathbf{v}|\boldsymbol{\lambda}) \rangle_{\mathcal{Q}/\mathcal{Q}_\alpha}\right) \tag{15}$$

Here, $\mathcal{Q}/\mathcal{Q}_\alpha$ denotes the product of all factors excluding $\mathcal{Q}_\alpha$ and α is a set valued index running over clusters, i.e. $\alpha \in \{\{\mathbf{A}\}, \{v_{i,k}\}, \{r_j\}, \{\tilde{\mathbf{s}}_k\}\}$ with $i = 1 \dots n$, $j = 1 \dots m$ and $k = 0 \dots N-1$. Hence, the structured mean field approach leads to a set of fixed point equations that need to be iterated leading to a variational EM algorithm. The actual form of the $\mathcal{Q}_\alpha$ and update rules for the model are given at the appendix.

3.1 Maximisation over the hyperparameter $\boldsymbol{\lambda}$

The hyper-parameters $\boldsymbol{\lambda}$ defines the scale of the latent source coefficients and hence it is important to set these correctly for the separation performance. It is possible to integrate out $\boldsymbol{\lambda}$ by extending the approximating distribution $\mathcal{Q}$. In this study, however, we estimate $\boldsymbol{\lambda}$ simply by maximum likelihood II, i.e., by maximising directly the lower bound defined in (14) w.r.t. $\boldsymbol{\lambda}$.

Original matrix
$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 \\ -1 & 0.2679 & 3.7321 \end{bmatrix}$
MCMC
$\hat{\mathbf{A}} = \begin{bmatrix} 1 & 1 & 1 \\ -0.9849 & 0.2787 & 3.7213 \\ (\pm 0.0067) & (\pm 0.0025) & (\pm 0.0061) \end{bmatrix}$
Variational
$\hat{\mathbf{A}} = \begin{bmatrix} 1 & 1 & 1 \\ -0.9765 & 0.2768 & 3.7213 \\ (\pm 0.0011) & (\pm 0.007) & (\pm 0.0027) \end{bmatrix}$

Table 1: Estimates of $\mathbf{A}$

	$\hat{s}_1$			
	SDR	SIR	SAR	SNR
MCMC	6.3	15.0	7.3	20.4
Variational	6.4	15.4	7.3	21.7
	$\hat{s}_2$			
	SDR	SIR	SAR	SNR
MCMC	5.1	14.3	5.8	27.8
Variational	5.2	15.0	5.8	24.8
	$\hat{s}_3$			
	SDR	SIR	SAR	SNR
MCMC	16.6	23.7	17.8	29.8
Variational	16.6	25.3	17.5	29.7

Table 2: Performance criteria of estimated sources with both methods.

4. AUDIO RESULTS

We study a mixture of $n = 3$ audio sources (speech, piano, guitar) with $m = 2$ observations. The mixing matrix is given in Table 1. The second row of $\mathbf{A}$ corresponds to $[\tan\psi_1 \tan\psi_2 \tan\psi_3]$ with $\psi_1 = -45$ deg, $\psi_2 = 15$ deg and $\psi_3 = 75$ deg. We set $R = (0.03)^2 \mathbf{I}_m$, which corresponds to 20dB and 26dB noise on each observation. The signals are sampled at 8kHz with length $N = 65356$ ($\approx 8s$). We used a MDCT orthonormal basis [17] to decompose the observations, with a sine window of length 64ms (512 samples).

We applied both the variational method described above and the Gibbs sampler described in [9, 10] to the MDCT coefficients of the observations. Both simulations were run with the same prior on $\mathbf{A}$ ($\mathbf{vec}\,\Omega_A = \mathbf{0}$ and $\Sigma_A = 10\ \mathbf{I}_{nm}$) and the same initialisations ($\forall i, \alpha_i = 1, \lambda_i = 0.1$, $R = (0.1)^2 \mathbf{I}_m$ and $\mathbf{A}$ drawn from the prior).

While the variational approach is deterministic, given the initial $\mathcal{Q}$ distribution and an update order of factors, the Gibbs sampler is stochastic and convergence relies on the particular seed of the random number generator. Over several runs of the sampler, convergence to the stationary distribution $p(\mathbf{A}, \tilde{\mathbf{s}}, R, \mathbf{v}, \boldsymbol{\alpha}, \boldsymbol{\lambda}|\tilde{\mathbf{x}})$ was not observed before 4000 iterations. Adding 1000 iterations used to compute MMSE estimates of the parameters of interest, the total 5000 iterations take 15 hours a Mac G4 cadenced at 1.25 GHz (with a MATLAB implementation).

In contrast, the variational method converges after 200 iterations, which requires 1.5 hours on the same computer. Means of the marginals of $\mathcal{Q}$ were used as estimates of the various parameters.

Estimates of the mixing matrix are reported in Table 1. Sources are reconstructed by inverse MDCT of the estimated source coefficients $\tilde{\mathbf{s}}$. The reconstructions are compared to the original sources using the source separation evaluation criteria described in [9]; basically, the SDR (Source to Distortion Ratio) provides an overall separation performance criterion, the SIR (Source to Interferences Ratio) measures the level of interferences from the other sources in each source estimate, SNR (Source to Noise Ratio) measures the error due to the additive noise on the sensors and the SAR (Source to Artifacts Ratio) measures the level of artifacts in the source estimates. The performance criteria are reported in Table 2. We point out that the performance criteria are invariant to a change of basis, so that figures can be computed either on the time sequences ($\hat{\mathbf{s}}$ compared to $\mathbf{s}$) or the MDCT coefficients ($\hat{\tilde{\mathbf{s}}}$ compared to $\tilde{\mathbf{s}}$). The estimated sources can be listened to at `http://www-sigproc.eng.cam.ac.uk/~cf269/eusipco05/sound_files.html`, which is perhaps the best way to assess the audio quality of the results.

5. CONCLUSIONS

Tables 1 and 2 show that the separation quality is identical with both methods, for a computation time roughly 10 times shorter with the variational approach.

However, in some simulations carried out, especially with mixing matrices where two columns are almost linearly dependent, the variational method consistently underestimated the number of sources. When two columns of $\mathbf{A}$ point at close directions (*e.g*, ψ_1 close to ψ_2), the variational approach tended to merge the corresponding sources into a single one, setting a column of $\mathbf{A}$ to zero, whilst MCMC was able to consistently locate the three components. The reason for this might be that the actual posterior landscape is multimodal, with each mode corresponding to a "possible explanation of data". Our over-smooth variational approximation is missing an important local maxima and thus favours a solution that corresponds to a simpler explanation. This point is however to be investigated by further simulation studies. Interestingly, when we clamp the matrix close to its true value, the sources are estimated quickly and reliably by variational EM.

These first conclusions of our work suggest that a hybrid method incorporating both MCMC and variational steps may be a viable approach for fast and robust source separation. In the future, we plan to design such an optimisation schema, that, while keeping the robustness of the MCMC approach in estimating the mixing matrix, makes use of variational steps for the estimation of latent sources for fast convergence.

A. STANDARD DISTRIBUTIONS

Multivariate Gaussian and Gamma distributions are defined as

$$\begin{aligned} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= |2\pi\,\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\,\boldsymbol{\Sigma}^{-1}\,(\mathbf{x}-\boldsymbol{\mu}) \\ \mathcal{G}(x; a, b) &\equiv \frac{1}{\Gamma(a)} b^{-a} x^{a-1} \exp(-\frac{x}{b}) \mathbb{I}_{[0,+\infty)}(x) \end{aligned}$$

The sufficient statistics have the form

$$\begin{aligned} \langle x \rangle_{\mathcal{N}} &= \mu \qquad \left\langle x x^T \right\rangle_{\mathcal{N}} = \Sigma + \mu\mu^T \\ \langle x \rangle_{\mathcal{G}} &= ab \qquad \langle \log x \rangle_{\mathcal{G}} = \mathbf{digamma}(a) + \log(b) \end{aligned}$$

Here, the digamma function defined as $\mathbf{digamma}(x) \equiv d\log\Gamma(x)/dx$.

B. VARIATIONAL APPROXIMATION

The generative model is given as

$$\begin{aligned}
\tilde{\mathbf{s}}_k|V_k &\sim \mathcal{N}(\tilde{\mathbf{s}}_k; 0, V_k^{-1}) \\
x_k|\tilde{\mathbf{s}}_k, \mathbf{A}, R &\sim \mathcal{N}(x_k; \mathbf{A}\tilde{\mathbf{s}}_k, R^{-1}) \\
V_k &\sim \mathcal{G}(V_k; \mathbf{a}_Q, \mathbf{b}_Q) \equiv \prod_i \mathcal{G}(v_{k,i}; a_{Q,i}, b_{Q,i}) \\
R &\sim \mathcal{G}(R; \mathbf{a}_R, \mathbf{b}_R) \equiv \prod_j \mathcal{G}(r_j; a_{R,j}, b_{R,j}) \\
\mathbf{vec}\,\mathbf{A} &\sim \mathcal{N}(\mathbf{vec}\,\mathbf{A}; \mathbf{vec}\,\Omega_p, \Psi_p)
\end{aligned}$$

where we have defined the matrix $V_k \equiv \mathbf{diag}\,\mathbf{v}_k$. Note that we tie the parameters of $p(V_k)$ across k. The variational approximation has the form

$$\mathcal{Q} \equiv q(R)q(\mathbf{A})\prod_{k=1}^{K} q(\tilde{\mathbf{s}}_k)q(V_k)$$

The update rules for the variational factors are given as in the following:

- Hidden state

$$\begin{aligned}
q(\tilde{\mathbf{s}}_k) &= \mathcal{N}(\tilde{\mathbf{s}}_k; m_k, S_k) \\
S_k &= \left(\left\langle \mathbf{A}^T R\mathbf{A}\right\rangle + \langle V_k\rangle\right)^{-1} \\
m_k &= S_k \langle\mathbf{A}\rangle^T \langle R\rangle\, x_k
\end{aligned}$$

- Mixing matrix[1]

$$\begin{aligned}
q(\mathbf{A}) &= \mathcal{N}(\mathbf{vec}\,\mathbf{A}; \mathbf{vec}\,\Xi, \Phi) \\
\Phi &= \left(\left(\sum_k \left\langle \tilde{\mathbf{s}}_k\tilde{\mathbf{s}}_k^T\right\rangle \otimes \langle R\rangle\right) + \Psi_p^{-1}\right)^{-1} \\
\mathbf{vec}\,\Xi &= \Phi(\mathbf{vec}(\langle R\rangle \left\langle \sum_k x_k\tilde{\mathbf{s}}_k^T\right\rangle) + \Psi_p^{-1}\,\mathbf{vec}\,\Omega_p)
\end{aligned}$$

- Hidden state precision, $i = 1 \ldots n$

$$\begin{aligned}
q(V_k) &= \mathcal{G}(V_k; \bar{\mathbf{a}}_{V,k}, \bar{\mathbf{b}}_{V,k}) \equiv \prod_i \mathcal{G}(v_{k,i}; \bar{a}_{V,k,i}, \bar{b}_{V,k,i}) \\
\bar{a}_{V,k,i} &= a_{V,i} + \frac{1}{2} \\
\bar{b}_{V,k,i} &= \frac{b_{V,i}}{\left\langle \tilde{s}_{k,i}^2\right\rangle b_{V,i}/2 + 1}
\end{aligned}$$

- Observation noise precision, $j = 1 \ldots m$

$$\begin{aligned}
q(R) &= \mathcal{G}(R; \bar{\mathbf{a}}_R, \bar{\mathbf{b}}_R) \equiv \prod_j \mathcal{G}(r_j; \bar{a}_{R,j}, \bar{b}_{R,j}) \\
\mathbf{z}_k &= \mathbf{diag}\left(\left\langle \mathbf{A}\tilde{\mathbf{s}}_k\tilde{\mathbf{s}}_k^T\mathbf{A}^T\right\rangle - 2x_k\langle\tilde{\mathbf{s}}_k\rangle^T\langle\mathbf{A}\rangle^T + x_k x_k^T\right) \\
\bar{a}_{R,j} &= a_{R,j} + \frac{K}{2} \\
\bar{b}_{R,j} &= \frac{b_{R,j}}{\left(\sum_k z_{j,k}\right) b_{R,j}/2 + 1}
\end{aligned}$$

The algorithm proceeds by updating the parameters of these factors iteratively. When a factor $\mathcal{Q}_\alpha$ is updated, the expectations depending upon parameters of $\mathcal{Q}_\alpha$ also change, hence factors of $\mathcal{Q}/\mathcal{Q}_\alpha$, that depend upon these expectations need to be updated e.t.c.

[1] The operator $\otimes$ denotes the Kronecker product.

REFERENCES

[1] B. A. Olshausen and K. J. Millman. Learning sparse codes with a mixture-of-Gaussians prior. In S. A. Solla and T. K. Leen, editors, *Advances in Neural Information Processing Systems*, pages 841–847. MIT press, 2000.

[2] M. S. Lewicki and T. J. Sejnowski. Learning overcomplete representations. *Neural Computations*, 12:337–365, 2000.

[3] M. Girolami. A variational method for learning sparse and overcomplete representations. *Neural Computation*, 13(11):2517–2532, 2001.

[4] T.-W. Lee, M. S. Lewicki, M. Girolami, and T. J. Sejnowski. Blind source separation of more sources than mixtures using overcomplete representations. *IEEE Signal Processing Letters*, 4(4), Apr. 1999.

[5] M. Zibulevsky, B. A. Pearlmutter, P. Bofill, and P. Kisilev. Blind source separation by sparse decomposition. In S. J. Roberts and R. M. Everson, editors, *Independent Component Analysis: Principles and Practice*. Cambridge University Press, 2001.

[6] M. Davies and N. Mitianoudis. A simple mixture model for sparse overcomplete ICA. *IEE Proceedings on Vision, Image and Signal Processing*, Feb. 2004.

[7] R. Gribonval. Sparse decomposition of stereo signals with matching pursuit and application to blind separation of more than two sources from a stereo mixture. In *Proc. ICASSP*, Orlando, Florida, May 2002.

[8] A. Jourjine, S. Rickard, and O. Yilmaz. Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures. In *Proc. ICASSP*, volume 5, pages 2985–2988, Istanbul, Turkey, Jun. 2000.

[9] C. Févotte, S. J. Godsill, and P. J. Wolfe. Bayesian approach for blind separation of underdetermined mixtures of sparse sources. In *Proc. 5th International Conference on Independent Component Analysis and Blind Source Separation (ICA 2004)*, pages 398–405, Granada, Spain, 2004.

[10] C. Févotte and S. J Godsill. A Bayesian approach to blind separtion of sparse sources. *Technical Report of Cambridge University Engineering Dept*, (CUED/F - INFENG/TR.511), Jan 2005.

[11] Z. Ghahramani and M. Beal. Propagation algorithms for variational Bayesian learning. In *Neural Information Processing Systems 13*, 2000.

[12] M. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. Technical Report 649, Department of Statistics, UC Berkeley, September 2003.

[13] P. J. Wolfe, S. J. Godsill, and W.-J. Ng. Bayesian variable selection and regularisation for time-frequency surface estimation. *J. R. Statist. Soc. Series B*, 2004.

[14] D. F. Andrews and C. L. Mallows. Scale mixtures of normal distributions. *J. R. Statist. Soc. Series B*, B(36):99–102, 1974.

[15] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., New York, 1991.

[16] W. Wiegerinck. Variational approximations between mean field theory and the junction tree algorithm. In *UAI-2000 (16-th conference)*, pages 626–633.

[17] S. Mallat. *A wavelet tour of signal processing*. Academic Press, 1998.