BAYESIAN COMPUTATIONAL METHODS FOR SPARSE AUDIO AND MUSIC PROCESSING

S.J. Godsill[1]*, A.T. Cemgil*[1]*, C. Févotte*[2] *and P.J. Wolfe*[3]

[1]University of Cambridge
Cambridge, CB2 1PZ, UK
sjg@eng.cam.ac.uk
www-sigproc.eng.cam.ac.uk/sjg

[2]GET/Télécom Paris
75014 Paris France
fevotte@tsi.enst.fr
www.tsi.enst.fr/ fevotte/

[3]Harvard University
Cambridge, MA 02138-2901 USA
patrick@eecs.harvard.edu
www.eecs.harvard.edu/ patrick/

ABSTRACT

In this paper we provide an overview of some recently developed Bayesian models and algorithms for estimation of sparse signals. The models encapsulate the sparseness inherent in audio and musical signals through structured sparsity priors on coefficients in the model. Markov chain Monte Carlo (MCMC) and variational methods are described for inference about the parameters and coefficients of these models, and brief simulation examples are given.

1. INTRODUCTION

In applications such as coding, noise reduction, missing data, source separation and music transcription, models can be represented as sums of large numbers of 'atoms' drawn from some large and possibly over-complete dictionary. Within this framework we may write:

$$x(t) = \sum_{k \in \mathcal{K}} c_k g_k(t) \qquad (1)$$

where c_k are some unknown coefficients and $g_k(t)$ are elements of some dictionary $\mathcal{K}$ of 'atoms', or elementary components from which the signal is composed. We will consider the dictionary of basis functions to be fixed and known in this paper, although it is possible to consider learning these also from the data. The dictionary should contain all of the important component types present in an audio signal, including, at least, oscillating functions. To motivate the discussion, we consider the Gabor representation, where the atoms are windowed oscillating exponential functions of many different frequencies, aimed at modelling the non-stationary oscillations at different times and frequencies of an audio signal,

$$x(t) = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} c_{m,n}\, g_{m,n}(t), \quad g_{m,n}(t) = g\left(t - \frac{n}{N}L\right) e^{2\pi j \frac{m}{M} t},$$

where $g(t)$ is some appropriate window function (e.g. Gaussian, Hamming, Hanning, etc.) that is shifted to different locations in time and frequency by time-shift and multiplication with a complex exponential function. See [1, 2] for examples of its use in audio reconstruction. Many other dictionary types are possible. For example, see [3] for related models using a modified DCT (MDCT) basis, which is orthogonal, and [4, 5] for an approach which combines several MDCT bases with different time-frequency resolutions together in one single model. Other possible dictionaries could include wavelets, or other custom-made atoms suitable for modelling of audio.

Observed data are modelled as some function of $x(t)$, usually including a random noise disturbance. In the simple case of additive noise we have

$$y_t = \sum_{k \in \mathcal{K}} c_k\, g_k(t) + d_t,\ t \in \{0, ..., L-1\} \qquad (2)$$

Typically, the noise term d_t is modelled as independent and white Gaussian noise, $d_t \sim N(0, \sigma_d^2)$, although the general methods described here are adaptable to more realistic non-white and non-Gaussian settings. Another setting where such models can be usefully employed is in source separation. In this cases a number of sources, each modelled as $x(t)$ above, are mixed together, and the task is to reconstruct the sources from the observed noisy mixture, see [3] for details.

Generally the task will be to extract estimates of the underlying sourse $x(t)$, or equivalently, estimates of the coefficients c_k for each source. It is desired to favour *sparse* solutions, i.e. solutions for which many of the estimated coefficients c_k are zero or close to zero, but also that the coefficients should have physical interpretability (in our case, persistence over time and/or frequency). We here regard this sparse signal extraction task as a Bayesian inverse problem [1]. This requires a direct time series modelling of the data, rather than modelling in some 'transform domain', for example MDCT or discrete Fourier coeffients, although in cases where the atoms are an orthogonal basis (for example the MDCT or orthogonal wavelet bases), the two approaches will often coincide.

In the remainder of this paper we discuss choices of models and prior distributions that favour sparse, meaningful solutions, and discuss general computational methods for dealing with the complex models that arise. Owing to the space limitations of this format we have not been able to include a full list of relevant references from the literature beyond our own work, and other background material: for these see the bibliographies of the cited applications and methods papers [1, 3].

2. SPARSE MODELS AND PRIOR DISTRIBUTIONS

The above models in their basic form do not assume any sparseness of the coefficients $\{c_k\}$. However, the dictionaries of atoms may be highly over-complete, that is there may be many alternative representations of a given signal using different values of $\{c_k\}$. Moreover, the direct inclusion of noise terms d_t means that many possible estimated signals may be consistent with the noisy data. It would seem natural to push estimates of the coefficients towards sparse sets, having as few non-zero elements as possible, while still being consistent with the data. A common approach to this idea is to pose the problem as a constrained optimisation, according to criteria such as least squares fitting, and L_1 minimisation of the coefficients. A disadvantage with such approaches is that they may not lead to physically meaningful or interpretable results, with the knock-on effect that undesired artefacts may be perceived in any reconstruction of the estimated waveforms. Anyone who has worked in the audio processing area will know just how susceptible audio signals are to the introduction of audible artefacts - owing to the ear and brain's very sensitive and nonlinear perception mechanisms.

Here we describe alternative approaches to sparsity that are based on Bayesian prior probability models. The principal unknowns in the above systems (ignoring for now the statistics of the noise sources σ_d^2 and any mixing coefficients for source separation), are the coefficients $\{c_k\}$. One might be tempted to consider a likelihood-based solution to the problem. In the above models the likelihood functions are routinely obtained from the independent

Gaussian noise assumption. Take for example the additive noise model (2). Then,

$$p(y_t|\{c_k\}) = p(y_t|x(t)) = \frac{1}{\sqrt{2\pi\sigma_d^2}} \exp\left(-\frac{1}{2\sigma_d^2}(y_t - x(t))^2\right)$$

and

$$p(y_{0:L-1}|\{c_k\}) = \prod_{t=0}^{L-1} p(y_t|\{c_k\})$$

where the notation $y_{0:L-1}$ is shorthand for $[y_0, y_1, ..., y_{L-1}]^T$.

However, for overcomplete dictionaries, the maximum likelihood solution is ill-defined, since many parameter combinations can have the same likelihood. Again, it is quite typical then to formulate a regularised likelihood problem, using for example the L_1 norm of the coefficients as a regulariser. Such an approach can equivalently be regarded as a Bayesian solution to the problem, as discussed below, where we consider more general Bayesian approaches to regularisation.

Suppose now that we know some prior probability distribution for the unknown coefficients, say $p(\{c_k\})$, then inference can be performed as a Bayesian inverse problem. The posterior distribution is defined as

$$p(\{c_k\}|y_{0:L-1}) = \frac{1}{Z_y} p(\{c_k\}) p(y_{0:L-1}|\{c_k\})$$

where

$$Z_y = \int_{\{c_k\}} p(y_{0:L-1}|\{c_k\}) d\{c_k\}$$

is a normalising constant independent of $\{c_k\}$.

Given a suitable form for the prior distribution the coefficients, and hence the signal, can then be inferred by computational methods. Consider now, however, the prior distribution, which is of primary importance in construction of an appropriate estimation scheme. The prior distribution in a Bayesian model expresses prior belief about the coefficients, in the form of a probability distribution. We would wish that the prior represented physical meaningful and interpretable knowledge about audio signals, which should include information about how coefficients are related to one another (dependence structures) and also how they are individually distributed (marginal distributions). The simplest choices of prior will, however, assume prior independence of all coefficients, and identical distribution for each coefficient, i.e.

$$p(\{c_k\}) = \prod_{k \in \mathcal{K}} p(c_k)$$

Quite a lot can be achieved with an independent model of this type. Perhaps the simplest useful choice would be a zero-mean Gaussian, $p(c_k) = \mathcal{N}(0, \sigma_c^2)$, and with σ_c^2 estimated somehow or known in advance. We will however describe below models that incorporate non-Gaussianity and dependence between coefficients, in order to exploit better the known characteristics of audio and musical signals.

2.1 Heavy-tailed prior distributions

While the above model is highly limited, and does not encourage sparsity in any particularly useful way, we will see that very useful results can be achieved through introduction of an unknown scale parameter σ_{c_k} for each normal component, i.e.

$$p(c_k|\sigma_{c_k}) = \mathcal{N}(0, \sigma_{c_k}^2)$$

and then assigning some prior distribution $p(\sigma_{c_k})$ in a *hierarchical* Bayesian scheme. It is now well known that coefficient distributions in speech and audio are non-Gaussian and heavy-tailed in many standard dictionaries, see e.g. [6, 3]. Supposing that scale parameters are mutually independent *a priori*, then the joint prior distribution becomes

$$p(\{c_k, \sigma_{c_k}\}) = \prod_{k \in \mathcal{K}} p(c_k|\sigma_{c_k}) p(\sigma_{c_k})$$

We are now in a much richer class of heavy-tailed prior distributions for coefficients, the *scale mixture of normals* (SMiN) [7] class, also commonly referred to as the *scale mixture of Gaussians* (SMoG) class. The implied prior distribution for each c_k is obtained by marginalising the scale parameter as follows:

$$p(c_k) = \int_0^\infty p(c_k|\sigma_{c_k}) p(\sigma_{c_k}) d\sigma_{c_k}$$

and now we can see that this is a (continuous) mixture of normal distributions, each with different scale parameter. The SMiN representation includes a wide range of important heavy-tailed distributions, such as the Student-t, the symmetric α-stable, the double exponential, the generalised Gaussian, and the generalised hyperbolic distribution, each obtained through a different choice of the mixing distribution $p(\sigma_{c_k})$. The reason we specify models in terms of scale mixtures rather then just in terms of the implied prior distribution itself, is mostly computational: there are very efficient versions of iterative algorithms such as EM, Markov chain Monte Carlo (MCMC) and variational approaches when a conditionally Gaussian structure such as this is maintained. Possibly the simplest member of the SMiN class to deal with (at least computationally) is the Student-t distribution, which is obtained when $\lambda_k = 1/\sigma_{c_k}^2$ has a gamma distribution, with parameters $\alpha > 0$ and $\beta > 0$,

$$p(\lambda_k) = \mathcal{G}(\lambda_k|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda_k^{\alpha-1} \exp(-\beta\lambda_k), \qquad (0 < \lambda_k < \infty) \tag{3}$$

(or equivalently, $p(\sigma_{c_k})$ has the square-root inverse gamma distribution). The resulting marginal prior distribution is then the zero-mean Student-t, with precision (inverse-scale) parameter α/β and degrees of freedom 2α,

$$p(c_k) = \mathcal{S}(c_k|0, \alpha/\beta, 2\alpha) \propto (1 + c_k^2/\beta)^{-(\alpha+1/2)}$$

By suitable choice of α and β it is possible to model more or less heavy-tailed distributions, which favour highly sparse or less sparse solutions. A quite extreme special case of this class takes the limit as both α and β tend to zero. In this case very sparse solutions are favoured and the mixing distribution is the improper Jeffreys' prior [8] $p(\lambda_k) \propto 1/\lambda_k$. This prior was used to good effect by [**?**] in a general setting for inducement of sparsity and by [9] in an audio enhancement setting.

2.2 Dependence structures for the coefficients

The use of independent scale mixture priors across the coefficients can achieve a certain amount and can certainly be used to estimate sparse signals consistent with the data. However, the independence assumption is a quite a severe limitation from several perspectives. First, when using an independent heavy-tailed prior, there is no strong penalisation of individual large components from the dictionary becoming isolated in the reconstructed coefficient map. This can lead to artefacts or unnatural sounding reconstructions of an audio signal. Second, it does not tally with the known structures of audio signals: we expect a certain degree of continuity of activity across time and/or frequency; to put it crudely, tonal components of a speech or music signal are expected to continue at the same or similar frequency for a number of time frames in the data, while transients are expected to be active across a range of frequency components but localised in time, see Fig. 1 for a typical musical example. Some kind of prior modelling of this effect will lead to more interpretable structures in the reconstructed coefficient maps, and also to fewer perceived artefacts in the reconstructed signals. Such dependence structures may be modelled at various levels of the Bayesian hierarchical models: directly at the level of the coefficients c_k, at the level of the scale parameters σ_{c_k}, or at a higher level in the hierarchy. Here we describe the latter approach, in which an 'activity' variable $\gamma_k \in \{0,1\}$ is associated with each coefficient c_k, in addition to its scale parameter σ_{c_k}. This activity variable is able to switch the coefficient to precisely zero with non-zero probability, which is not a feature of other models considered thus far in the paper. Hence we have a direct control over the sparsity of the estimated model.

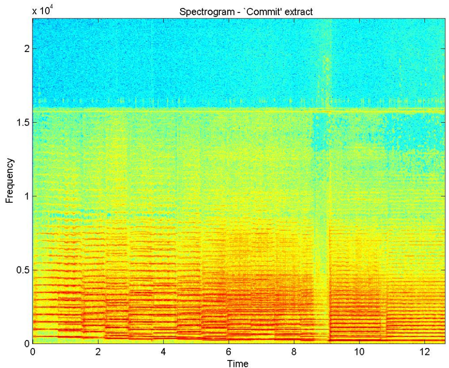


Figure 1: Spectrogram of musical extract. Note both horizontal and vertical structure in coefficient map

The basic structure can be summarised as follows,

$$p(c_k \,|\, \sigma_{c_k}, \gamma_k) = (1-\gamma_k)\delta_0(c_k) + \gamma_k \mathcal{N}(c_k \,|\, 0, \sigma_{c_k}^2), \;\; \gamma_k \in \{0,1\}$$

where $\delta_0(c_k)$ is the Dirac δ-function centred at zero.

Of course, as before, the activity variables could be considered as independent of one another. However, advantage can be gained from consideration of the dependence over time and frequency that is likely to be present in real audio signals. A general framework for formulating this is the *Markov random field* [10]. In such a model, *local* dependencies are specified through a prior conditional distribution of activity variables, conditioned on a local neighbourhood of surrounding activity variables:

$$p(\gamma_k \,|\, \gamma_{-k}) = p(\gamma_k \,|\, \gamma_{\mathcal{N}(k)})$$

where γ_{-k} refers to all γ_j variables except for γ_k itself, and $\mathcal{N}(k)$ is a local neighbourhood of coefficient indices which are typically chosen to be close to coefficient k in both time and frequency. Some standard special cases that fit into this framework are the Markov chain in time (favours continuity of 'tonal' coefficients across time but not frequency), and the Markov chain in frequency (favours continuity of 'transient' components across frequency, but not in time) - see [1, 2] for various examples of these.

In practice one does not know in advance whether to expect transient behaviour (persistence across frequency) or tonal behaviour (persistence across time), so a generic model should be able to adapt to the signals as they are found. An example of this is found in [1] where a Markov random field encouraging both transients and tonal components is shown to be able to estimate contiguous 'patches' of activity in speech signals. As a possible extension of these approaches which might consider a direct classification of activity into 'transient', 'tonal', or both, with different model behaviour in each we might consider extending these models to indicate different types of activity. The indicator variable might then take one of three possible values:

$$\begin{cases} \gamma_k = 0, & \text{inactive} \\ \gamma_k = 1, & \text{active - `tonal'} \\ \gamma_k = 2, & \text{active - `transient'} \end{cases}$$

with a Markov random field structure designed to model the different behaviours of each type of activity. Such an approach has not yet been investigated to our knowledge, but could potentially give a useful performance advantage, with in addition a classification of the time-frequency surface into different types of activity. As an alternative, dictionaries with different characteristics can be combined together - see for example [4] in which two or more orthogonal MDCT bases are combined together within a Bayesian scheme - one aimed at capturing 'tonal' components and another of capturing 'transient' components.

3. INFERENCE METHODS FOR SPARSE MODELS

The models as posed in the previous section have a large number of unknowns: $\{c_k\}$, $\{\sigma_{c_k}\}$ and $\{\gamma_k\}$. In addition there may be other parameters or hyperparameters which are unknown, for example α and β in the gamma distribution if we are using the Student-t version of the prior, and the noise variance σ_d^2. If source separation is being carried out the mixing matrix A is also likely to be an unknown parameter. We group these additional parameters into a vector θ. From the likelihood and prior modelling considerations reviewed above, Bayes' Theorem may once again be used to write a posterior distribution for all unknowns, conditioned on the data:

$$\mathcal{P} = p(\{c_k, \gamma_k\}, \theta | y_{0:L-1}) \propto p(y_{0:L-1} | \{c_k, \gamma_k\}, \theta) p(\{c_k, \gamma_k\} | \theta) p(\theta) \quad (4)$$

There are various inference tasks that might interest us. In signal reconstruction problems (noise reduction, interpolation of missing values, source separation) the signal $x(t)$ itself will be of interest. Since this can be obtained directly by a linear transformation of the coefficients, see (1), we will consider here just the task of estimation of the coefficients $\{c_k\}$. Suppose for example that a posterior mean estimate of the coefficients is required (which would yield a minimum mean-squared error performance under the assumed model and prior). This can be formulated as a high-dimensional integral:

$$E[\{c_k\}|y_{0:L-1}] = \int \{c_k\} p(\{c_k, \gamma_k\}, \theta | y_{0:L-1}) d\theta d\{\gamma_k, c_k\}$$

which cannot be calculated in closed form for the models considered here (except for the simple fixed Gaussian case, which is not of interest for sparse modelling). Alternatively, in applications such as source coding, we may require an estimate of the activity variables, in order to determine which coefficients are significant for coding purposes. This can be estimated in a similar fashion (here taking a *Maximum a posteriori approach* rather than a posterior mean for the discrete activity variables:

$$\{\hat{\gamma_k}\} = \text{argmax} \int p(\{c_k, \gamma_k\}, \theta | y_{0:L-1}) d\theta d\{c_k\} \quad (5)$$

where the integral is now over all variables except for the activity map $\{\gamma_k\}$. In either case above, numerical approximation techniques must be employed. We consider here Monte Carlo and deterministic (variational) approaches, which provide different trade-offs in performance and speed.

3.1 Monte Carlo Approaches to Computation

In the Monte Carlo approach a large number of (possibly dependent) samples is drawn from the joint distribution $p(\{c_k, \gamma_k\}, \theta | y_{0:L-1})$. Then, under suitable conditions, the posterior mean for each coefficient is estimated simply as the arithmetic mean of the samples:

$$\hat{c}_k = \frac{1}{N} \sum_{n=1}^{N} c_k^{(n)}$$

where $\{c_k^{(n)}\}_{n=1}^N$ are the N samples for variable c_k. This result is very simple and general, but the art lies in drawing of samples from the posterior distribution. This can be achieved by many means, for example rejection sampling or importance sampling. For this type of highly complex problem, however, Markov chain Monte Carlo (MCMC) methods are the most competitive approach, albeit at high computational expense. See [11, 12] for full details of Monte Carlo methodology, and MCMC methods in particular.

The two main components of a MCMC algorithm are the Gibbs' Sampler and the Metropolis-Hastings sampler. These two methods give recipes for construction of Markov chains which converge in the limit (as number of samples becomes large) to a desired 'target' distribution ($p(\{c_k, \gamma_k\}, \theta | y_{0:L-1})$ in our case).

The Gibbs' Sampler can be used when conditional distributions for unknowns, or preferably groups of unknowns, can readily be sampled from. In our models we might for example consider the following natural grouping, or partition, of the unknowns:

$$\{c_k\}, \{\sigma_{c_k}\}, \{\gamma_{c_k}\}, \theta$$

Then, the so-called *full conditional distribution* for each parameter - the distribution of that parameter conditioned on the data and on all remaining parameters - is calculated,

$$p(\{c_k\}|\{\sigma_{c_k}\}\{\gamma_{c_k}\}, \theta, y_{0:L-1}),\ p(\{\sigma_{c_k}\}|\{c_k\}, \{\gamma_{c_k}\}, \theta, y_{0:L-1})$$

$$p(\{\gamma_{c_k}\}|\{c_k\},\ \{\sigma_{c_k}\}, \theta, y_{0:L-1}), p(\theta|\{\sigma_{c_k}\}, \{c_k\}, \{\gamma_{c_k}\}, y_{0:L-1})$$

The algorithm is then initialised at some, essentially arbitrary, value $\{c_k\}^{(0)}$, $\{\sigma_{c_k}\}^{(0)}$, $\{\gamma_{c_k}\}^{(0)}$, $\theta^{(0)}$. Each iteration of the Gibbs' Sampler then simply cycles through the conditional distributions, sampling with replacement from each in turn. For example, moving from samples at iteration $n-1$ to iteration n, we have:

$$\begin{aligned}
\{c_k\}^{(n)} &\sim p(\{c_k\}|\{\sigma_{c_k}\}^{(n-1)}\{\gamma_{c_k}\}^{(n-1)}, \theta^{(n-1)}, y_{0:L-1}) \\
\{\sigma_{c_k}\}^{(n)} &\sim p(\{\sigma_{c_k}\}|\{c_k\}^{(n)}, \{\gamma_{c_k}\}^{(n-1)}, \theta^{(n-1)}, y_{0:L-1}) \\
\{\gamma_{c_k}\}^{(n)} &\sim p(\{\gamma_{c_k}\}|\{c_k\}^{(n)}, \{\sigma_{c_k}\}^{(n)}, \theta^{(n-1)}, y_{0:L-1}) \\
\theta^{(n)} &\sim p(\theta|\{\sigma_{c_k}\}^{(n)}, \{c_k\}^{(n)}, \{\gamma_{c_k}\}^{(n)}, y_{0:L-1})
\end{aligned}$$

where '$\sim$' denotes an independent random draw from the distribution to the right. The algorithm then iterates until convergence to the target posterior has been achieved (this is known as the 'burn-in' time), after which time samples can be used in Monte Carlo estimation.

The above conditional sampling steps may still not be feasible in practice - for our model this depends on the choice of dictionary (orthogonal/non-orthogonal) and the prior on the activity variables. Instead the parameters must be partitioned further into smaller sub-partitions. Even so, it may not be feasible to perform all of the conditional draws required. At this point the conditional draws may be replaced by random draws according to a Metropolis-Hastings rule, adopting a so-called Metropolis-within-Gibbs, or 'hybrid', sampling scheme, see [11] for full details. In our models above most steps can be achieved by Gibbs sampling, but a few of the hyperparameters, such as α in the gamma distribution may require a Metropolis-Hastings update. Many additional complexities and alternative schemes have of course been overlooked in this brief overview.

3.2 Variational (Deterministic) Methods

Monte Carlo methods described in the previous section rely on generating samples from the intractable target posterior density, typically via simulation of a Markov chain with the desired target posterior density. An alternative class of methods, based on deterministic approximations, has been applied extensively, notably in machine learning and statistical physics for inference in large models. While lacking theoretical guarantees of MCMC methods (which are nevertheless only valid when substantial computational resources are available), with only a fixed amount of CPU power, variational methods have proven to be viable alternatives in many practical situations.

There are various deterministic methods that can be viewed as variational methods [13]. One particularly simple variational method, that is algorithmically very similar to the Gibbs sampler, is the *structured mean field*, also known as *variational Bayes*, see [14, 15, 13] and references therein.

The main idea in variational Bayes is to approximate the exact posterior distribution $\mathcal{P} = \frac{\Psi_y}{Z_y}$ defined in (4) with a simple distribution $\mathcal{Q}$. The variational distribution $\mathcal{Q}$ is chosen such that its expectations can be obtained in easily, preferably in closed form. One such distribution is a factorised one $\mathcal{Q} = \prod_\alpha \mathcal{Q}_\alpha$ where α is an index that runs over disjoint clusters of variables, much as the parameters are partitioned into blocks for the Gibbs' sampler described above. For example, for the variable selection problem in (4), neglecting the parameter `theta` for now, we can take $\mathcal{Q} = \prod_k \mathcal{Q}(c_k)\mathcal{Q}(\gamma_k)$. In this case, the set of clusters would be $\mathcal{C} = \{\mathcal{C}_\alpha\} = \{\{c_1\} \dots \{c_K\}, \{\gamma_1\} \dots \{\gamma_K\}\}$. Alternatively, we could choose a clustering $\{\{c_{1:K}\}, \{\gamma_1\} \dots \{\gamma_K\}\}$ where $\mathcal{Q} = \mathcal{Q}(c_{1:K}) \prod_k \mathcal{Q}(\gamma_k)$ or $\{\{\gamma_1, c_1\} \dots \{\gamma_K, c_K\}\}$ where $\mathcal{Q} = \prod_k \mathcal{Q}(\gamma_k, c_k)$, and where K is the total number of atoms in the dictionary $\mathcal{K}$.

An intuitive interpretation of the mean field method is minimising the Kullback-Leibler (KL) divergence with respect to (the parameters of) $\mathcal{Q}$ where

$$KL(\mathcal{Q}||\mathcal{P}) \quad = \quad \langle \log \mathcal{Q} \rangle_{\mathcal{Q}} - \left\langle \log \frac{1}{Z_y} \psi_y \right\rangle_{\mathcal{Q}} \qquad (6)$$

Using non-negativity of KL, [16], one can obtain a lower bound on the normalisation constant Z_y

$$\log Z_y \geq \langle \log \psi_y \rangle_{\mathcal{Q}} - \langle \log \mathcal{Q} \rangle_{\mathcal{Q}} \quad \equiv \quad F[\mathcal{P}, \mathcal{Q}] + H[\mathcal{Q}]$$

Here, F is interpreted as a negative *energy* term and $H[\mathcal{Q}]$ is the entropy of the approximating distribution. The maximisation of this lower bound is equivalent to finding the "nearest" $\mathcal{Q}$ to $\mathcal{P}$ in terms of KL divergence and this solution is obtained by a joint maximisation of the entropy H and F (minimisation of the energy) [17]. The solution is in general not available in closed form but is obtained as a result of a deterministic fixed point iteration. It can be easily shown, e.g. see [13, 18], that each factor $\mathcal{Q}_\alpha$ of the optimal approximating distribution should satisfy the following fixed point equation

$$\mathcal{Q}_\alpha \quad \propto \quad \exp\left(\langle \log \psi_y \rangle_{\mathcal{Q}_{\neg \boldsymbol{\alpha}}} \right) \qquad (7)$$

where $\mathcal{Q}_{\neg \boldsymbol{\alpha}} \equiv \mathcal{Q}/\mathcal{Q}_\alpha$, that is the joint distribution of all factors excluding $\mathcal{Q}_\alpha$. Hence, the mean field approach leads to a set of (deterministic) fixed point equations that need to be iterated until convergence.

The above fixed point equation can be viewed as a generalisation of the Expectation-Maximisation (EM) algorithm and Iterative Conditional Modes (ICM) [19]. Now suppose we choose some factors α in the variational distribution $\mathcal{Q}$ as degenerate point mass distributions with location parameter $\mathcal{C}_\alpha^*$ and we denote this set of indices as M. We have

$$\mathcal{Q} \quad = \quad \left(\prod_{\alpha \notin M} \mathcal{Q}_\alpha \right) \left(\prod_{\alpha \in M} \delta(\mathcal{C}_\alpha - \mathcal{C}_\alpha^*) \right)$$

Without loss of generality assume that we have two clusters $\mathcal{Q} = \mathcal{Q}_1(\mathcal{C}_1)\mathcal{Q}_2(\mathcal{C}_2) = \mathcal{Q}_1(\mathcal{C}_1)\delta(\mathcal{C}_2 - \mathcal{C}_2^*)$ where $M = \{2\}$. The fixed point equation implies the (E step)

$$\begin{aligned}
\mathcal{Q}_1(\mathcal{C}_1) \quad &\propto \quad \exp\left(\langle \log \psi_y(\mathcal{C}_1, \mathcal{C}_2) \rangle_{\mathcal{Q}_2(\mathcal{C}_2)} \right) \\
&= \quad \psi_y(\mathcal{C}_1, \mathcal{C}_2^*)
\end{aligned}$$

Note that this quantity is proportiOnal to the full conditional $p(\mathcal{C}_1|\mathcal{C}_2^*, y)$. Similarly, the M step is equivalent to finding the location parameter $\mathcal{C}_2^*$ as

$$\begin{aligned} \mathcal{Q}_2(\mathcal{C}_2) &\propto \exp\left(\langle \log \psi_y(\mathcal{C}_1, \mathcal{C}_2)\rangle_{\mathcal{Q}_1(\mathcal{C}_1)}\right) \\ \mathcal{C}_2^* &= \arg\max_{\mathcal{C}_2} \mathcal{Q}_2(\mathcal{C}_2) \end{aligned}$$

The latter step can be also seen as the minimiser of $KL(\delta(\mathcal{C}_2 - \mathcal{C}_2^*)||\mathcal{Q}_2(\mathcal{C}_2))$. Finally when we choose all variational distributions to be degenerate, we obtain $\mathcal{Q} = \mathcal{Q}_1(\mathcal{C}_1)\mathcal{Q}_2(\mathcal{C}_2) = \delta(\mathcal{C}_1 - \mathcal{C}_1^*)\delta(\mathcal{C}_2 - \mathcal{C}_2^*)$ a simple coordinate ascent algorithm that is known as ICM.

$$\begin{aligned} \mathcal{C}_1^* &= \arg\max_{\mathcal{C}_1} \psi_y(\mathcal{C}_1, \mathcal{C}_2^*) \\ \mathcal{C}_2^* &= \arg\max_{\mathcal{C}_2} \psi_y(\mathcal{C}_1^*, \mathcal{C}_2) \end{aligned}$$

4. APPLICATION ISSUES

The above modelling structures and computational methods are very general and may be used to construct algorithms for denoising, interpolation, source separation, multi-resolution modelling and coding, to name but a few. The choice of computational methodology between deterministic (variational or EM) and Monte Carlo (MCMC) is largely a matter of taste, and very often the code structure will be quite similar for the two approaches (the same partitioning of the parameter space may well be adopted for both a Gibbs' Sampler and a variational method, for example), and computational burden will be very similar *per iteration* of the algorithms. Various comparative simulations have been made, using both types of inference method. To summarise, variational methods are generally faster converging, but MCMC has been found to outperform it in testing scenarios, for example source separation models with near-degenerate mixing scenarios. One comparative simulation is shown in Figs. 2.a and 2.b . Here a set of damped sinusoidal functions forms the dictionary and we attempt to estimate the activity variables according to (5). Synthetic data are generated from the true prior model. We have generated 100 independent cases from the model. In figure 2.a, we compare the likelihood of the configuration found by both methods. In figure 2.b, we show the distribution of edit distance errors, where we count the number of mismatches between the true and estimated activity variable configuration. This is constructed as a testing problem and we see that the MCMC generally finds higher likelihood solutions with fewer errors in the activity variables.

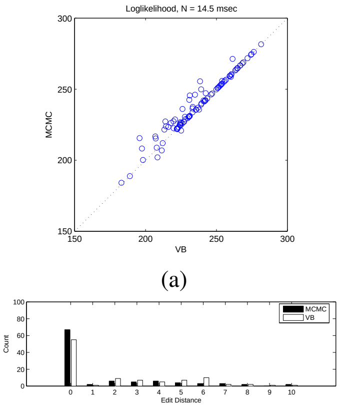
Other applications of the methods span source separation, denoising, interpolation of missing data and general inference for coefficient structure, see [1, 3, 2, 5] for details. Particularly impressive results have been obtained for the interpolation of missing data [2]. We anticipate that developments of these models and methods will have many more interesting applications across the musical audio area and the methods will also find use in other areas of science and engineering, such as for financial or biomedical data analysis.

REFERENCES

[1] P. J. Wolfe, S. J. Godsill, and W.J. Ng. Bayesian variable selection and regularisation for time-frequency surface estimation. *Journal of the Royal Statistical Society, Series B*, 66(3):575–589, 2004. Read paper (with discussion).

[2] P.J. Wolfe and S. J. Godsill. Interpolation of missing data values for audio signal restoration using a Gabor regression model. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2005.

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



(a)

(b)

Figure 2: Comparison of Gibbs sampler and Variational approximation on 100 independent cases in terms of edit distance with $t = 0 \ldots 319$ and $k = 1 \ldots 21$. (a) Likelihood comparison (b) Edit distance comparison

[4] C. Févotte and S. Godsill. Sparse linear regression in unions of bases via Bayesian variable selection. *IEEE Signal Processing Letters*, 13(7):441–444, Jul. 2006.

[5] C. Févotte, L. Daudet, S. J. Godsill, and B. Torrésani. Sparse regression with structured priors: Application to audio denoising. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Toulouse, France, May 2006.

[6] R. Martin. Speech enhancement based on minimum mean square error estimation and supergaussian priors. *IEEE Trans. Speech and Audio Processing*, 13(5), 2005.

[7] D. F. Andrews and C. L. Mallows. Scale mixtures of normal distributions. *J. Roy. Statist. Soc. Ser. B*, 36:99–102, 1974.

[8] H. Jeffreys. *Theory of Probability*. Oxford University Press, 1939.

[9] P. J. Wolfe and S. J. Godsill. Bayesian modelling of time-frequency coefficients for audio signal enhancement. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15, Cambridge, MA*. MIT Press, 2002.

[10] J. Besag. On the statistical analysis of dirty pictures. *J. Roy. Statist. Soc. Ser. B*, 48:259–302, 1986.

[11] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, New York, 2nd edition, 2004.

[12] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors. *Markov Chain Monte Carlo in Practice*. Interdisciplinary Statistics Series. Chapman & Hall, 1996.

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

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

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

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

[17] Radford M. Neal and Geoffrey E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. MIT Press, 1999.

[18] J. Winn and C. Bishop. Variational message passing. *Journal of Machine Learning Research*, 6:661–694, 2005.

[19] J.E. Besag. On the statistical analysis of dirty pictures (with discussion). *Jr. R. Stat. Soc. B*, 48:259–302, 1986.