4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003), April 2003, Nara, Japan

BAYESIAN ICA WITH HIDDEN MARKOV MODEL SOURCES

R.A. Choudrey and S.J. Roberts

University of Oxford
Robotics Research
Oxford, U.K.
email: riz, sjrob@robots.ox.ac.uk

ABSTRACT

This paper seeks to incorporate temporal information into the Independent Component Analysis process by marrying ICA models to Hidden Markov Models (HMMs). HMMs are models for picking up dynamic changes of state in the underlying data generation process, and are therefore useful in capturing high-order temporal information. In previous work, we introduced Bayesian ICA with mixture of Gaussian sources learnt using variational methods. In such a model, HMM methodology can be incorporated by stipulating a Markov prior over the mixture coefficients in the variational Bayesian ICA source models. This results in ICA with flexible, dynamic sources. The proposed method is a piecewise approach to detecting dynamic movement, focussing on abrupt changes in the source model. The proposed model will be shown to be more powerful than stationary ICA at blindly separating very noisy mixtures of images.

1. INTRODUCTION

Virtually all methods for ICA consider observed data to be independently and identically distributed and, as such, any correlations across time are deemed unimportant. Considering the wide spread and successful application of ICA to time series data, dynamics are conspicuous by their absence; considering the extra complexity involved in incorporating a dynamic element, this is understandable. Resolute blindness to temporal information may not be suitable in all cases, however. In some observations, there may be dynamic changes in the sources, affecting the source pdf statistics. In a sense, this occurs in the vectorised representation of images. Images have localised features, each feature represented by a different part of the overall image pdf. As one moves along the image vector, a hypothetical data generator would draw from one feature 'sub-pdf' for a while, then suddenly draw from another as one moved from, say, an expanse of sky to the roof of a house, the subsequently to its brick-work etc. ICA models are generally blind to such temporal information. If ICA was carried out on the whole, ordered data-set, one would expect a model with temporal sensitivity to out perform a static one, particularly in noisy situations when *all* information is useful. Such a model could perform filtering (e.g. noise reduction), forcasting as well as more accurate classification.

Dynamic ICA was first demonstrated by Pearlmutter and Parra's 'contextual' ICA [1] for square, noiseless mixing, where simple temporal information was incorporated into the source densities. This was improved in [2] in which Generalised Auto-Regressive (GAR) sources were used to condition source signals on previous values. A different methodology was proposed by Attias in [3] based on non-stationary, non-instantaneous mixing of stationary sources. This could, for example, be used for blind deconvolution. Learning non-stationary, instantaneous mixing using particle filters has been proposed by Murata *et al.* [4] and Everson and Roberts [5], the latter also making some in-roads into non-stationay sources. Penny and Roberts proposed a simple extension to HMMs, incorporating square, noiseless ICA observation generators with inverse-cosh densities in [6], allowing abrupt changes in the mixing to be captured. This was extended by Penny *et al.* [7] to include ICA GAR sources. Attias extended his IFA framework [8] to include Hidden Markov Model sources in[9]. The unobservable (hidden) process underlying the generation of the source signals moves from state to state in a 'Markov' process, whereby the occupancy of a state at time t depends only on which state the process was in at some previous time, $t - \tau$. Unlike GAR sources, HMMs capture higher-order temporal statistics.

The Bayesian formulation of models is often desireable when dealing with noisy data and/or small amounts of data as the prior distributions act as natural regularisers. Furthermore, Bayesian learning averages over all possible parameter values giving unbiased estimates, robust predictions and allows the comparison of assumptions underlying different models. Bayesian ICA has been formulated previously [10, 11, 12, 13], as have Bayesian HMMs [14, 15]. The model presented in this paper combines the variational Bayesian ICA model presented in [13] with the variational Bayesian HMM introduced in [14] to produce a Bayesian treatment of [9], a variational Bayesian ICA with HMM sources (vbICA-HMM).

2. THE MODEL

In common with ICA in the literature, we choose a generative model to work with. The observed variables, $\boldsymbol{x}$, of dimension M are modelled as a linear combination of statistically independent latent variables, $\boldsymbol{s}$, of dimension L, with added Gaussian noise

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s} + \boldsymbol{n} \tag{1}$$

where $\boldsymbol{A}$ is an $M \times L$ mixing matrix and $\boldsymbol{n}$ is M-dimensional additive noise. In signal processing nomenclature, M is the number of (observed) sensors and L is the number of latent (hidden) sources.

The noise is assumed to be Gaussian, with zero mean and diagonal precision matrix $\boldsymbol{\Lambda}$.The probability of observing data vector $\boldsymbol{x}^t$ is then given by

$$p(\boldsymbol{x}^t|\boldsymbol{s}^t, \boldsymbol{A}, \boldsymbol{\Lambda}) = \left|\det(\frac{1}{2\pi}\boldsymbol{\Lambda})\right|^{\frac{1}{2}} \exp[-E_D^t] \tag{2}$$

where

$$E_D^t = \frac{1}{2}(\boldsymbol{x}^t - \boldsymbol{A}\boldsymbol{s}^t)^T \boldsymbol{\Lambda}(\boldsymbol{x}^t - \boldsymbol{A}\boldsymbol{s}^t) \tag{3}$$

Since the sources are mutually independent, the distribution over $\boldsymbol{s}$ for data point t can be written as

$$p(\boldsymbol{s}^t) = \prod_{i=1}^{L} p(s_i^t) \tag{4}$$

Using such a generative approach, one must stipulate the source density.

2.1. Source Model

The choice of a flexible and mathematically tractable source model is crucial if a wide variety of source distributions are to be modelled; in particular, the source model should be capable of encompassing distributions with a wide variety of kurtoses and complex, multi-modal distributions.

One such distribution is a factorised mixture of univariate Gaussians (MoG) with L factors (i.e. sources) and m_i components per source:

$$\begin{aligned} p(\boldsymbol{s}^t|\boldsymbol{\theta}) &= \prod_{i=1}^{L}\sum_{q_i=1}^{m_i} p(q_i^t = q_i|\boldsymbol{\pi}_i) p(s_i^t|q_i, \mu_{i,q_i}, \beta_{i,q_i}) \\ &= \prod_{i=1}^{L}\sum_{q_i=1}^{m_i} \pi_{i,q_i}\mathcal{N}(s_i^t; \mu_{i,q_i}, \beta_{i,q_i}) \end{aligned} \tag{5}$$

where the mixing proportions $\pi_{i,q_i} = p(q_i^t = q_i|\boldsymbol{\pi}_i)$, the prior probability of choosing component q_i of the i^{th} source. q_i^t is a variable indicating which component of the i^{th} source is chosen for generating s_i^t and takes on values of $\{q_i = 1, \ldots, q_i = m_i\}$. Such a source model is stationary in that there is no temporal coupling across source states. If the source signals contain dynamics, then this temporal information can be exploited by conditioning the probabilistic occupancy of source state at time t, $\boldsymbol{q}_i^t$, on a source state at some previous time, say $t-1$. Such a coupling gives a Hidden Markov model with Gaussian generators.

The Hidden Markov Model

The HMM presented here is for a 1-dimensional signal, $\boldsymbol{s}$. The equivalent model for L statistically independent signals is simply a product of L 1-dimensional HMMS.

In essence, an HMM is a finite-state machine that switches between different probability density functions (pdfs) which represent the observation generators. The state change is a Markov process, whereby the occupancy of state d at time t depends probabilistically *only* on the state at some previous time, $t - \tau$. The model presented below is a first-order Markov process, where $\tau = 1$. Each observation is generated by first moving from state c at time $t-1$ to state d at time t according to transition probability r_{cd}, then stochastically drawing (or picking) from the d^{th} pdf.

The model is defined as follows. Let the model have m states represented by $q = \{1, 2, \ldots, m\}$. Let variable q^t be an m-dimensional vector that indicates which state is chosen at time t. If the HMM is in state d at time t, then q^t has a 1 in the d^{th} entry and zeros everywhere else. Let $\boldsymbol{R}$ represent the matrix of state transition probabilities where the probability of moving from state c to d is given by r_{cd}

$$p(q^t = d|q^{t-1} = c, \boldsymbol{R}) = r_{cd} \tag{6}$$

The probability of the HMM starting in a given state is given by the vector $\boldsymbol{\pi} = \{\pi_1, \pi_2, ..., \pi_m\}$ such that

$$p(q^1 = d|\boldsymbol{\pi}) = \pi_d \tag{7}$$

Let the observation at time t be denoted s^t. The probability of generating s^t given state $q^t = d$ is given by

$$p(s^t|q^t = d, \theta_d) = p_d(s^t|\theta_d) \tag{8}$$

where θ_d represents the parameters of the d^{th} observation pdf.The observation densities can be anything, for example Gaussians or ICA models. The parameter set for the m pdfs is $\{\theta_1, \ldots, \theta_m\}$, and for the whole HMM the total parameter set is $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{R}, [\theta_1, \ldots, \theta_m]\}$.

If there are T data points, then $\boldsymbol{s} = \{s^1, ..., s^T\}$ and $\boldsymbol{q} = \{q^1, ..., q^T\}$. The joint probability of state path $\boldsymbol{q}$ generating observation sequence $\boldsymbol{s}$ given parameter set $\boldsymbol{\theta}$ is

$$\begin{aligned} p(\boldsymbol{s}, \boldsymbol{q}|\boldsymbol{\theta}) &= p(q^1|\boldsymbol{\pi}) \prod_{t=2}^{T} p(q^t|q^{t-1}, \boldsymbol{R}) \prod_{t=1}^{T} p(s^t|q^t, \theta_{q^t}) \\ &= p(\boldsymbol{q}|\boldsymbol{\theta}) p(\boldsymbol{s}|\boldsymbol{q}, \boldsymbol{\theta}) \end{aligned} \tag{9}$$

where

$$p(\boldsymbol{q}|\boldsymbol{\theta}) = p(q^1|\boldsymbol{\pi}) \prod_{t=2}^{T} p(q^t|q^{t-1}, \boldsymbol{R}) \tag{10}$$

$$p(\boldsymbol{s}|\boldsymbol{q}, \boldsymbol{\theta}) = \prod_{t=1}^{T} p(s^t|q^t = d, \theta_d) \tag{11}$$

The likelihood that the signals, $\boldsymbol{s}$, were generated by this model using parameters $\boldsymbol{\theta}$ irrespective of state path is

$$p(\boldsymbol{s}|\boldsymbol{\theta}) = \sum_{\{\boldsymbol{q}\}} p(\boldsymbol{q}|\boldsymbol{\theta}) p(\boldsymbol{s}|\boldsymbol{q}, \boldsymbol{\theta}) \tag{12}$$

where $\{\boldsymbol{q}\} = \{\boldsymbol{q}_1, ..., \boldsymbol{q}_{m^T}\}$, the space of all state paths. Note the mixture-model form of (12); HMMs are simply mixture models with Markov time dependencies across hidden states $\boldsymbol{q}$.

An HMM is learnt using the Forward-Backward algorithm [16], an efficient Expectation-Maximisation method. The most likely time-path of the hidden states can be inferred using the Viterbi algorithm [16]. Under a Bayesian formultion, an HMM can also be learnt using the variational Bayes approximation, also known as ensemble learning. This was derived by MacKay in [14] and was shown to be similar to the maximum-likelihood Forward-Backward algorithm. The Viterbi path is also similar.

3. BAYESIAN FORMALISM

A Bayesian scheme can be used to introduce prior knowledge, control complexity, perform model selection, and integrate out the dependency of a model on its parameters. As ever, a strict Bayesian treatment is computationally expensive and is often intractable. Therefore, a variational Bayesian approximation [17, 18] can be used to allow tractability and increase computational efficiency. The objective function to be maximised in this methodology is the negative free energy (NFE) $F(\boldsymbol{s}|\mathcal{M})$

$$\begin{aligned} F(\boldsymbol{s}|\mathcal{M}) &= \langle \log p(\boldsymbol{s},\boldsymbol{q},\boldsymbol{\theta}|\mathcal{M})\rangle_{p'(\boldsymbol{q},\boldsymbol{\theta})} \\ &+ \mathcal{H}[p'(\boldsymbol{q},\boldsymbol{\theta})] \end{aligned} \tag{13}$$

This can be shown to be a strict lower bound on the marginal data likelihood which is dependent only on model assumptions $\mathcal{M}$. By evaluating F for competing models, the model best 'explaining' the data can be inferred. The term $p'(.)$ is a factorised approximation to the true posterior and is chosen to confer tractability. If $\boldsymbol{W} = \{\boldsymbol{q},\boldsymbol{\theta}\}$, and the observation densities are Gaussian, $p_d(s^t|\theta_d) = \mathcal{N}(s^t;\mu_d,\beta_d)$, then the following factorisation is chosen

$$p'(\boldsymbol{W}) = p'(\boldsymbol{q})p'(\boldsymbol{R})p'(\boldsymbol{\pi})p'(\boldsymbol{\mu})p'(\boldsymbol{\beta}) \tag{14}$$

where $\boldsymbol{\mu},\boldsymbol{\beta}$ are vectors of m parameters. The prior distributions over the HMM parameters are

$$p(\boldsymbol{\pi}) = \frac{\Gamma(m\lambda_0)}{\Gamma(\lambda_0)^m}\prod_{c=1}^{m}\pi_c^{\lambda_0-1} \tag{15}$$

$$p(\boldsymbol{R}) = \prod_{c=1}^{m}\Gamma(\sum_{d'}\iota_{cd'})\prod_{d=1}^{m}\frac{r_{cd}^{\iota_{cd}-1}}{\Gamma(\iota_{cd})} \tag{16}$$

$$p(\boldsymbol{\mu}) = \prod_{c}^{m}\mathcal{N}(\mu_c;m_0,\tau_0) \tag{17}$$

$$p(\boldsymbol{\beta}) = \prod_{c}^{m}\mathcal{G}(\beta_c;b_0,c_0) \tag{18}$$

where $\Gamma(.)$ is the gamma function and $\mathcal{G}(.)$ is a Gamma distribution. Substituting the priors and (35) into the negative free energy (34) gives the following posteriors for the HMM parameters

$$p'(\boldsymbol{\pi}) = \Gamma(\sum_{c'}\hat{\lambda}_{c'})\prod_{c=1}^{m}\frac{\pi_c^{\hat{\lambda}_c-1}}{\Gamma(\hat{\lambda}_0)} \tag{19}$$

$$p'(\boldsymbol{R}) = \prod_{c=1}^{m}\Gamma(\sum_{d'}\hat{\iota}_{cd'})\prod_{d=1}^{m}\frac{r_{cd}^{\hat{\iota}_{cd}-1}}{\Gamma(\hat{\iota}_{cd})} \tag{20}$$

$$p'(\boldsymbol{q}) = \frac{1}{Z_{\boldsymbol{q}}}\left[\tilde{\pi}_{q^1}\prod_{t=2}^{T}\tilde{r}_{q^{t-1}q^t}\prod_{t=1}^{T}\tilde{p}_{q^t}(s^t|\theta_{q^t})\right] \tag{21}$$

$$p'(\boldsymbol{\mu}) = \prod_{c}^{m}\mathcal{N}(\mu_c;\hat{m}_c,\hat{\tau}_c) \tag{22}$$

$$p'(\boldsymbol{\beta}) = \prod_{c}^{m}\mathcal{G}(\beta_c;\hat{b}_c,\hat{c}_c) \tag{23}$$

Where the posterior parameters are given by

$$\hat{\lambda}_c = \lambda_0 + \gamma_c^1 \tag{24}$$

$$\hat{\iota}_{cd} = \iota_{c0} + \sum_{t=1}^{T-1}\xi_{cd}^t \tag{25}$$

$$\tilde{\pi}_c = \exp\left[\Psi(\hat{\lambda}_c) - \Psi(\sum_{c'}\hat{\lambda}_{c'})\right] \tag{26}$$

$$\tilde{r}_{cd} = \exp\left[\Psi(\hat{\iota}_{cd}) - \Psi(\sum_{d'}\hat{\iota}_{cd'})\right] \tag{27}$$

$$\tilde{p}_c(s^t|\theta_c) = \tilde{\beta}_c^{\frac{1}{2}}\exp\left[-\frac{\langle\beta_c\rangle}{2}\langle\left(s^t-\mu_c\right)^2\rangle\right] \tag{28}$$

$$\tilde{\beta}_c = \hat{b}_c\exp\left[\Psi(\hat{c}_c)\right] \tag{29}$$

$$\hat{m}_c = \frac{1}{\hat{\tau}_c}\left(\tau_0 m_0 + \langle\beta_c\rangle\sum_{t=1}^{T}\gamma_c^t s^t\right) \tag{30}$$

$$\hat{\tau}_c = \tau_0 + \langle\beta_c\rangle\sum_{t=1}^{T}\gamma_c^t \tag{31}$$

$$\hat{b}_c = \left[\frac{1}{b_0} + \frac{1}{2}\sum_{t=1}^{T}\gamma_c^t\langle\left(s^t-\mu_c\right)^2\rangle\right]^{-1} \tag{32}$$

$$\hat{c}_c = c_0 + \frac{1}{2}\sum_{t=1}^{T}\gamma_c^t \tag{33}$$

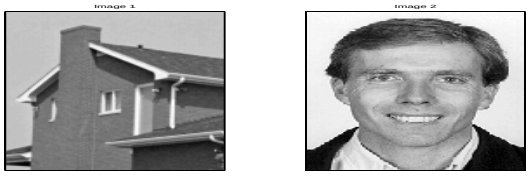
where $\langle.\rangle$ indicates expectations w.r.t. p'(.) and $\Psi(.)$ is the digamma function. The two auxillary variables ξ_{cd}^t and γ_c^t are calculated from the standard forward-backward algorithm in [16], with the 'tilded' estimates above, (26)-(28), substituting the maximum-likelihood estimates. This also ensures (21) is normalised. The update equations above are coupled so must be solved iteratively. Initial values for $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$ are set and a forward-backward pass is made. The calculated auxillary variables are then used to cyclically iterate through the updates above. The forward-backward pass is repeated and the process continues until convergence. The most likely hidden state-path can also be inferred using the Viterbi algorithm, again with the 'tilded' estimates replacing the maximum-likelihood ones.

It is a simple matter to incoporate L independent HMM sources into a variational Bayesian ICA model. If the ICA model parameters are $\boldsymbol{\Theta} = \{\boldsymbol{A},\boldsymbol{\Lambda},\boldsymbol{R},\boldsymbol{\pi},\boldsymbol{\theta}\}$, where $\boldsymbol{\theta}$ now contains the HMM parameters for all L sources, then these can be learnt using a variational approach, and the source signals can be similarly inferred. The priors over the ICA mixing matrix and noise precision are the same as those in [13], with a product of Gaussians over the $M \times L$ mixing matrix elements, and a Gamma distribution over the noise precision. The posterior over the mixing matrix, noise covarience and source signals is straight forward to derive using the following NFE

$$\begin{aligned} F(\boldsymbol{X}|\mathcal{M}) &= \langle \log p(\boldsymbol{X},\boldsymbol{S},\boldsymbol{q},\boldsymbol{\Theta}|\mathcal{M})\rangle_{p'(\boldsymbol{S},\boldsymbol{q},\boldsymbol{\Theta})} \\ &+ \mathcal{H}[p'(\boldsymbol{S},\boldsymbol{q},\boldsymbol{\Theta})] \end{aligned} \tag{34}$$

If $\boldsymbol{W} = \{\boldsymbol{S},\boldsymbol{q},\boldsymbol{\Theta}\}$, then the following factorisation is chosen

$$p'(\boldsymbol{W}) = p'(\boldsymbol{A})p'(\boldsymbol{\Lambda})p'(\boldsymbol{S}|\boldsymbol{q})p'(\boldsymbol{q})p'(\boldsymbol{\theta}) \tag{35}$$



(a) Source images

(b) Mixtures: 5% noise

(c) Mixtures: 50% noise

Figure 1: Original images and mixtures

where

$$p'(\boldsymbol{\theta}) = p'(\boldsymbol{R})p'(\boldsymbol{\pi})p'(\boldsymbol{\mu})p'(\boldsymbol{\beta}) \quad (36)$$

The posterior over sources and source states is also chosen to factorise over the L sources to reduce the computational load. The update equations for the ICA parameters are the the same as vbICA with MoG sources (vbICA-MoG) [13], with the source updates similar to the HMM updates given above.

All the derived posteriors require solving a set of coupled hyper-parameter update equations. In practice, this is best achieved by first cycling through $p'(\boldsymbol{S})$ and $p'(\boldsymbol{A})$ until convergence. These values are then passed to $p'(\boldsymbol{\theta})$, whose constituent updates are cycled with a forward-backward pass, until convergence. The hyper-parameters for $p'(\boldsymbol{\Lambda})$ are updated, then the whole process is repeated until convergence.

Once trained, the model can be used to reconstruct hidden source signals (to within a scaling and permutation) given a data set by calculating $\langle q_i \rangle$ and $\langle s_i \rangle$ under their respective posteriors over the whole data-set, and given the (now fixed) model parameter posteriors.

4. RESULTS

To assess the performance of vbICA-HMM, the model was used to separate the image mixtures in Figure 1. Both vbICA-HMM and vbICA-MoG models with 1-3 dimensionalities were trained on two sets of mixtures, one with 5% added Gaussian noise and the other with 50%, as shown in Figure 1. Each network was given 5 components per source and trained on the whole 16129-vector dataset. The ICA hyper-parameters for vbICA-HMM were set to values corresponding to vague priors (see, for example, [13]). The source HMM parameters were similarly vague, with m_0, b_0, c_0 initialised using K-means for all sources, and $\lambda_0 = 1$, $i_{c \neq d} = 5$ for all sources. Three different values for $i_{dd} = 10, 5000, 10^6$ were set to assess the effect of weak, medium and strong temporal priors.

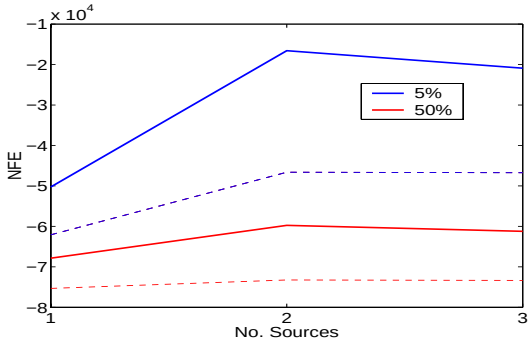
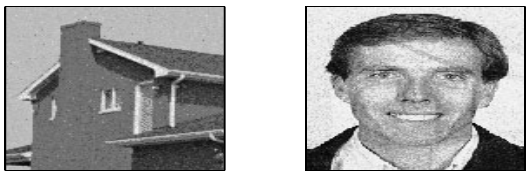


Figure 2: Plots of NFE. Dotted: vbICA-MoG Solid: vbICA-HMM

Figure 4 plots the NFE across model orders ($i_{dd} = 10$ for vbICA-HMM). For both vbICA-MoG and vbICA-HMM, a model with 2 sources is deemed most likely. Figure 3 shows the image reconstructions for vbICA-MoG and vbICA-HMM. The benefits reaped from exploiting temporal information is clear: the vbICA-HMM images are cleaner than those of vbICA-MoG. Although vbICA-MoG has been able to separate the images, the images are clearly noisier than those recovered by the dynamic source models. This is particularly evident at the higher noise level. Even though vbICA-HMM is more complex, this improved modelling is reflected in the higher NFE of vbICA-HMM over vbICA-MoG indicating the presence of source dynamics. In principle, the NFE can also be used to infer the most likely number of HMM states, although this can be cumbersome.

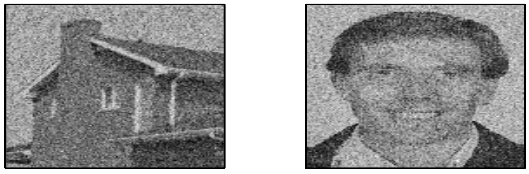
Shown in Figure 4 are the differences in mean square error and cross-talk between vbICA-MoG and vbICA-HMM reconstructions, for different strengths of prior over the diagonal of the transition matrix. The MSE of vbICA-HMM is nearly half that of vbICA-MoG in the low noise case, and nearly a third in the high-noise case. Similarly, the cross-talk is some 50% lower. Of the three strengths, the weaker prior has the highest NFE, although the medium model has the lowest MSE and the strongly constrained model has the lowest cross-talk. Note that the MSE of the strong prior is worse than that of the medium prior. This is a case of the prior being *too* strong. The overly-strong prior causes the HMM source models to stay in particular states longer than they should, giving rise to excessive streaking in the case of vectorised images. Care must be taken in setting priors for specific purposes, for example for finding localised features.

The most likely hidden state path giving rise to the source signals can be inferred using the Viterbi algorithm with the tilded variables. This has the effect of segmenting the source distribution. In the image mixture example there were 5 states per source, so each image has been segmented



(a) vbICA-MoG: 5%

(b) vbICA-HMM: 5%



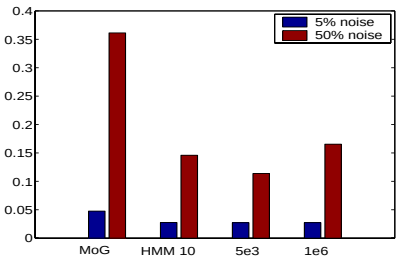
(c) vbICA-MoG: 50%

(d) vbICA-HMM: 50%

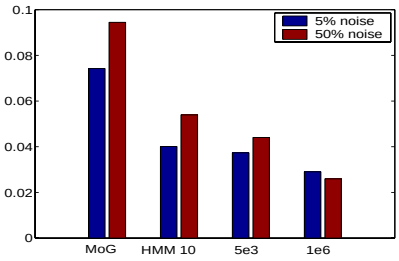
Figure 3: vbICA-MoG and vbICA-HMM reconstructions

into 5 classes. The segmentation carried out under the weak prior is shown in Figure 5(a). It is clear that each of the 5 HMM Gaussian generators corresponds to a brightness level, where - for example - the blue in 'House' representing the bright sky and white wood on the house, and the orange in 'Bloke' are the highlights on his skin. The colours do not correlate in the two images as the inferred state labels have an arbitrary ordering. Nontheless, the simultaneous (blind) image separation and segmentation carried out by vbICA-HMM is very impressive.

The learnt models can also be used to denoise images. Figure 5(b) shows an image of 'house' corrupted by 50% Gaussian noise. This was triplicated and fed to the vbICA-HMM model learnt on the relatively clean data in Figure 1(b), which inferred the most appropriate source signals giving rise to this data. One of these sources relates to 'bloke'



(a) MSE



(b) Cross-talk

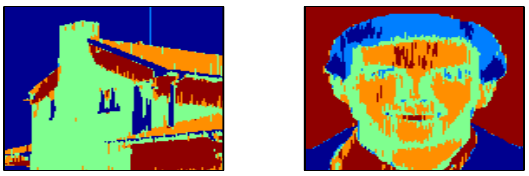
Figure 4: Performance of vbICA-HMM compared with vbICA-MOG

so can be discounted. The reconstruction under the other source is shown in Figure 5(b), showing a much cleaned-up image.

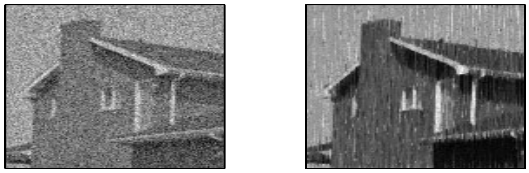
5. DISCUSSION

We have shown how the explicit use of temporal information in data can be used to improve the decompositional and representational power of stationary ICA. This temporal information is exploited by utilising Hidden Markov models to capture high-order dynamics in the statistics of the data. The vbICA-HMM model is more robust at finding independent components in noisy data as demonstrated by the image separation of very noisy image mixtures. An added benefit is the simultaneous segmentation of these images into areas of similar grey-scale. The learnt vbICA-HMM models can subsequently be used to denoise images modelled by its source HMMs.

The most obvious application of vbICA-HMM is in the BSS of time-series data. Music signals, for example, have temporal structure in a similar way to images, so BSS would be more robust using vbICA-HMM, exploiting temporal information to obtain cleaner reconstructions. This would be particularly applicable to speech recognition. HMMs are the current cutting edge method for recognising words and syllables. Although these work exceptionally well in lab conditions, the real world is noisy. When speaking into a (stereo) microphone, the word being spoken and the rubbish in the backround are resolutely independent. ICA with 2 sources, one HMM and the other MoG, could be used to



(a) Image segmentation



(b) Noise reduction

Figure 5: Image processing by vbICA-HMM

blindly siphon off background noise, giving a cleaner signal to the HMM thereby improving performance in real situations. If this HMM is trained ahead of time then installed in the ICA network, the improvemnet would be even more accute.

The models could be extended in a number of ways. The dynamics of the HMM part could be made more sophisticated. The HMM presented in this paper is a 1st-order Markov process, whereby the statistics at time t depend on the hidden state at time $t-1$. This could be generalised to $t-\tau$. The independent components could be made dependent by coupling the hidden states using coupled HMMs [19]. This is an important method in fusing information extracted from disparate data, for example combining auditory speech and visual lip information to make speech recognition more reliable.

6. REFERENCES

[1] B. Pearlmutter and L. Parra, "A context-sensitive generalization of ICA," in *ICONIP '96*, 1996, pp. 151–157.

[2] B. Pearlmutter and L. Parra, "Maximum likelihood blind source separation: A context-sensitive generalization of ICA," in *Advances in Neural Information Processing Systems*, M.C Mozer, M.I Jordan, and T. Petsche, Eds. 1997, vol. 9, pp. 613–619, MIT press.

[3] H. Attias, "Blind source separation and deconvolution: the dynamic component analysis algorithm," *Neural Computation*, vol. 10, pp. 1373–1412, 1998.

[4] N. Murata, S. Ikeda, and A. Ziehe, "Adaptive online learning in changing enviroments," in *Advances in Neural Information Processing Systems*, M.C Mozer, M.I Jordan, and T. Petsche, Eds. 1997, vol. 9, pp. 599–605, MIT press.

[5] R Everson and S.J. Roberts, "Non-stationary Independent Component Analysis," in *Proceedings of the International Conference on Artificial Neural Networks (ICANN'99)*. IEE, 1999.

[6] W.D. Penny and S.J. Roberts, "Hidden Markov models with extended observation densities," Tech. Rep., Imperial College of Science Technology and Medicine, 1998.

[7] W. Penny, S. Roberts, and R. Everson, "Hidden Markov Independent Components Analysis," in *Advances in Independent Component Analysis*, Mark Girolami, Ed. Kluwer Academic Publishers, 2000.

[8] H. Attias, "Independent Factor Analysis," *Neural Computation*, vol. 11, pp. 803–851, 1999.

[9] H. Attias, "ICA and graphical models," in *Independent Component Analysis: Principles and Practice*, S Roberts and R Everson, Eds., chapter 3, pp. 95–112. Cambridge university press, 2001.

[10] K.H. Knuth, "Bayesian source seperation and localisation," in *Proceedings of SPIE'98*, A. Mohammad-Djafari, Ed., 1998, vol. 3459, pp. 147–158.

[11] H Lappalainen, "Ensemble learning for Independent Component Analysis," in *Proceedings of the First International Workshop on Independent Component Analyis*, 1999, pp. 7–12.

[12] J.W. Miskin and D.J.C. MacKay, "Ensemble learning for blind source seperation," in *Independent Component Analysis: Principles and Practice*, S.J. Roberts and R.M. Everson, Eds., chapter 7. Cambridge University Press, 2001.

[13] R.A. Choudrey and S.J. Roberts, "Flexible Bayesian Independent Component Analysis for blind source separation," in *Proceedings of ICA2001*, 2001.

[14] D.J.C. MacKay, "Ensemble learning for hidden Markov models," Tech. Rep., University of Cambridge, 1997.

[15] I. Rezek and S.J. Roberts, "Variational inference for hidden Markov models," *Electronic Letters*, 2001.

[16] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 257-286, 1989.

[17] D.J.C. MacKay, "Developments in probabilistic modelling with neural networks - ensemble learning," in *Proceedings of the third Annual Symposium on Neural Networks*, Nijmegen, The Netherlands, 1995, pp. 191–198, Springer.

[18] H. Attias, "Learning parameters and structure of latent variable models by variational Bayes," in *Electronic Proceedings of the Fifteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI-1999)*, http://www2.sis.pitt.edu/~dsl/UAI/uai.html, 1999, Association for Uncertainty in Artificial Intelligence (AUAI).

[19] I. Rezek, M. Gibbs, and S.J. Roberts, "Maximum a posteriori estimation of coupled hidden Markov models," *Journal of VLSI Signal Processing-Systems for Signal, Image, and Video Technology*, 2001, To appear.