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

VARIATIONAL BAYESIAN MIXTURE OF INDEPENDENT COMPONENT ANALYSERS FOR FINDING SELF-SIMILAR AREAS IN IMAGES

R.A. Choudrey and S.J. Roberts

University of Oxford
Robotics Research
Oxford, U.K.

ABSTRACT

Mixture modelling techniques such as Mixtures of Principal component and Factor analysers are very powerful in finding and representing Gaussian clusters in data. Meaningful representations may be lost, however, if these clusters are non-Gaussian. In this paper we propose extending the Gaussian-based analysers mixture model to an Independent Component Analysers mixture model. We employ recent developments in variational Bayesian inference and structure determination to construct a novel approach for modelling non-Gaussian clusters. We automaticaly determine the local dimensionality of each cluster and use variational Bayesian inference to calculate the most likely number of clusters in the data space. We demonstrate our framework by finding areas in images which are 'self-similar' under the independence assumption of ICA.

1. INTRODUCTION

The goal of pattern analysis and recognition is to extract information from some data. In order for this information to be useful, the distribution of data must be represented in some meaningful way. In many cases, insight may be gained by dividing the data into areas and analysing each of these clusters under some informative framework, for example using some understanding of the assumed data generating process. One such method is to model the data as being produced by a mixture of data generators (also called analysers or coordinate frames), where each component generator is responsible for generating a particular cluster. The problems to overcome in this mixture modelling are to decide how many generators are needed, where to place them, and how to adjust them to best represent the data. Each area is described by its own, local coordinate frame constructed under some fundamental assumption e.g. independent directions. Under such a framework, the area can be considered to be 'self-similar'.

Mixtures of Gaussians (MoG) are widely used throughout the fields of machine learning and statistics for data modelling, where each generator is a Gaussian density. Despite their popularity, however, MoGs suffer from two serious drawbacks. The first is that, as the dimensionality S of the data space increases, the size of each covariance matrix, S^2, becomes prohibitively large. This problem has been solved by Tipping and Bishop [1] who replaced each Gaussian with a probabilistic Principal Component Analyser (PCA). This allowed the dimensionality of each covariance to be effectively reduced whilst maintaining the richness of the model class. This mixture was modified into a Mixture of Factor Analysers (FA) [2] where variational Bayesian inference was used to infer the optimum number of analysers.

The second problem with MoGs is that each component is a Gaussian, a strong assumption which is often violated in many natural clustering and segmentation problems. It is this second problem which we address in this paper. A solution is reached by extending the mixtures of probabilistic PCA/FA model to a Mixture of Independent Component Analysers (ICA) model. We improve on previous work [3] [4] by incorporating a very flexible ICA model that can generate arbitrary densities using MoGs, and by bringing the formalism into the Bayesian arena. We use Bayesian inference to infer the optimum number of ICAs needed, and automatically determine their ideal dimensionalities.

1.1. Mixtures of ICA

Independent Component Analysis (ICA) assumes the observed sensor data vector $\boldsymbol{x}$, of dimension S, is generated by linearly transforming an unobserved source vector $\boldsymbol{s}$, of dimension L, which has independent components i.e. $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s}$ where $\boldsymbol{A}$ is the SxL *bases* or *mixing*[1] matrix. For the independence assumption to be meaningful, the source density - and, therefore, the sensor density - must be non-Gaussian. If the sensor density consists of various self-similar, non-Gaussian manifolds (i.e. clusters), then the data distribution can be segmented along the boundaries of these areas. An intuitive model for such data is a Mixture of Independent Component Analysers.

ICA mixture models were first formulated by Lee *et al.* in [3]. This model used the extended Infomax algorithm [5] to switch the source model between sub-Gaussian and super-Gaussian regimes. The model was learnt through a combination of maximum likelihood and gradient ascent. Although well demonstrated, the source model could only switch between Laplacian and bimodal densities and thus lacked flexibility. This was relaxed in [4] by utilising generalised exponential sources which can model a wide variety of kurtoses by the adjustment of a parameter. Although

[1] Please note that the term *mixing* refers to the linear mixing in the ICA model and *mixture* refers to stochastic generating method in a mixture model.

more flexible, the densities could only be unimodal and the learning scheme was also maximum likelihood.

In this paper, we present a Mixtures of ICA model (MoICA) trained using Bayesian methods [6] [7]. In line with [8] and [9], we choose a fully-adaptable factorial Mixture of Gaussians as the source model for our ICA components. Essentially, each ICA component will model each cluster as a mixture of Gaussian sub-features. To overcome the heavy computational load associated with Bayesian learning, we use the variational framework explored in [10] to make assumptions about the posterior, giving tractability to the Bayesian model. The variational Bayesian method is carried through to the ICA mixture model, allowing model comparison, incorporation of prior knowledge, control of model complexity thus avoiding over-fitting. By monitoring the *variational free energy* of our models, we can compare model assumptions [10] allowing us, in particular, to infer the optimum number of ICA components - and therefore clusters - in the data distribution. We also employ Automatic Relevance Determination (ARD) [11] to suppress unsupported manifold dimensions and thus effectively infer the number of latent dimensions of each cluster as part of the learning process. This leads to the variational Bayesian Mixture of ICAs model (vbMoICA).

2. THE MODEL

A set of C clusters in the data distribution is considered generated by a mixture model with C components where each component is identified with each cluster. The probability of generating a data vector $\boldsymbol{x}^n$ from a C-component mixture model given assumptions $\mathcal{M}$ is

$$p(\boldsymbol{x}^n|\mathcal{M}) = \sum_{c=1}^{C} p(c|\mathcal{M}_0)p(\boldsymbol{x}^n|\mathcal{M}_c, c) \tag{1}$$

A data vector is generated by choosing one of the C components stochastically under $p(c|\mathcal{M}_0)$ and then drawing from $p(\boldsymbol{x}^n|\mathcal{M}_c, c)$. $\mathcal{M} = \{\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_C\}$ is the vector of component model assumptions, $\mathcal{M}_c$, and assumptions about the mixture process, $\mathcal{M}_0$. $p(\boldsymbol{x}^n|\mathcal{M})$ is known as the evidence for model $\mathcal{M}$ and quantifies the likelihood of the observed data under model $\mathcal{M}$. Monitoring the evidence allows us to identify the most appropriate assumptions, for example the ideal number of components needed to best 'explain' the data distrbution. The variable c indicates which component of the mixture model is chosen to generate a given data vector $\boldsymbol{x}$. If $p(c|\mathcal{M}_0)$ is a vector of probabilities and each component $p(\boldsymbol{x}^n|\mathcal{M}_c, c)$ is a Factor analyser, then (1) describes a Mixture of Factor analysers [2]. In our mixture model, however, each component has a non-Gaussian density derived from the ICA model presented in the next section.

THE ICA MODEL

The observed variables, $\boldsymbol{x}$, of dimension S are modelled as a linear combination of statistically independent latent variables, $\boldsymbol{s}_c$, of dimension L_c with added Gaussian noise

$$\boldsymbol{x} = \boldsymbol{A}_c\boldsymbol{s}_c + \boldsymbol{y}_c + \boldsymbol{e}_c \tag{2}$$

where $\boldsymbol{y}_c$ is an S-dimensional *bias* vector, $\boldsymbol{e}_c$ is S-dimensional additive noise and c represents the c^{th} ICA model. The columns of the mixing matrix, $\boldsymbol{A}_c$, are the independent directions used to represent the c^{th} manifold.

Equation (2) acts as a complete description for cluster c in the data density. The bias vector, $\boldsymbol{y}_c$, defines the position of the cluster in the S-dimensional data space, $\boldsymbol{A}_c$ describes its orientation and $\boldsymbol{s}_c$ describes the underlying manifold. The noise, $\boldsymbol{e}_c$, is assumed to be Gaussian and isotropic with precision $\lambda_c I$. The probability of observing data vector $\boldsymbol{x}^n$ under component c is then given by

$$p(\boldsymbol{x}^n|\theta_c, c) = \left(\frac{\lambda_c}{2\pi}\right)^{\frac{S}{2}} \exp\left[-E_c^t\right] \tag{3}$$

where $\theta_c = \{\boldsymbol{A}_c, \boldsymbol{s}_c^n, \boldsymbol{\lambda}_c, \boldsymbol{y}_c\}$ and where

$$E_c^t = \frac{\lambda_c}{2}(\boldsymbol{x}^n - \boldsymbol{A}_c\boldsymbol{s}_c^n - \boldsymbol{y}_c)^T(\boldsymbol{x}^n - \boldsymbol{A}_c\boldsymbol{s}_c^n - \boldsymbol{y}_c) \tag{4}$$

Since the sources $\boldsymbol{s}_c = \{s_{c,1}, .., s_{c,i}, .., s_{c,L_c}\}$ are - by definition - mutually independent, the distribution over $\boldsymbol{s}_c$ for data point n can be written as

$$p(\boldsymbol{s}_c^n|\mathcal{M}_{\boldsymbol{s}_c}, c) = \prod_{i=1}^{L_c} p(s_{c,i}^n|\mathcal{M}_{s_{c,i}}, c) \tag{5}$$

where the product runs over the L_c sources of component c and $\mathcal{M}_{\boldsymbol{s}_c}$ is the vector of source model assumptions.

$p(\boldsymbol{s}_c^n|\mathcal{M}_{\boldsymbol{s}_c})$ is the source model for ICA component c. Traditionally an inverse-cosh [12] or unimodal density [13] has been used. These, however, have limited flexibility either in kurtosis-representation or capturing multiple modes. More recently, Attias introduced a Mixture of Gaussians (MoG) source model into his ICA formalism [8], allowing complex and potentially multi-modal distributions to be modelled. The source model for each component in our ICA mixture model is a factorised mixture of 1-dimensional Gaussians with L_c factors (i.e. sources) and m_i components per source

$$p(\boldsymbol{s}_c^n|\boldsymbol{\varphi}_c) = \prod_{i=1}^{L_c}\sum_{q_i=1}^{m_i} \pi_{i,q_i}\mathcal{N}(s_{c,i}^n; \mu_{i,q_i}, \beta_{i,q_i}) \tag{6}$$

where, for brevity, the ICA component subscript c has been dropped from parameters which can be seen to belong to ICA c from context. From now on, all subscripted parameters should be assumed to belong to the c^{th} ICA model, unless otherwise stated. Equation (6) essentially describes the local features of cluster c - μ_{i,q_i} is the position of feature q_i w.r.t. the cluster centre, β_{i,q_i} is its size, and π_{i,q_i} its 'prominance' w.r.t. other features.

The mixture proportions $\pi_{i,q_i} = p(q_i^n = q_i|\boldsymbol{\pi}_i)$ are the prior probabilities of choosing component q_i of the i^{th} source (of the c^{th} ICA model etc.). q_i^n is a variable indicating which component of the i^{th} source is chosen for generating $s_{c,i}^n$ and takes on values of $\{q_i = 1, \ldots, q_i = m_i\}$ (where m_i is, of course, depends on ICA model c). The mean and precision of Gaussian q_i in source i are μ_{i,q_i} and β_{i,q_i} respectively. The parameters of source i are $\varphi_{c,i} = \{\boldsymbol{\pi}_{c,i}, \boldsymbol{\mu}_{c,i}, \boldsymbol{\beta}_{c,i}\}$ where bold face indicates the vector of m_i parameters. The complete parameter set of

the source model is $\boldsymbol{\varphi}_c = \{\varphi_{c,1}, \varphi_{c,2}, \ldots, \varphi_{c,L_c}\}$. The complete collection of possible source states is denoted $\boldsymbol{q}_c = \{\boldsymbol{q}_{c,1}, \boldsymbol{q}_{c,2}, \ldots, \boldsymbol{q}_c, \boldsymbol{m}\}$ and runs over all $\boldsymbol{m} = \prod_i m_i$ possible combinations of source states.

By integrating and summing over the hidden variables, $\{\boldsymbol{s}_c, \boldsymbol{q}_c\}$, the likelihood of the IID data $\boldsymbol{X} = \{\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^N\}$ given the model parameters $\Theta_c = \{\boldsymbol{A}_c, \boldsymbol{y}_c, \lambda_c, \boldsymbol{\varphi}_c\}$ can now be written as

$$p(\boldsymbol{X}|\Theta_c, c) = \prod_{n=1}^{N} \sum_{\boldsymbol{q}=1}^{\boldsymbol{m}} \int p(\boldsymbol{x}^n, \boldsymbol{s}_c^n, \boldsymbol{q}_c^n|\Theta_c, c) d\boldsymbol{s}_c \qquad (7)$$

where $d\boldsymbol{s}_c = \prod_i ds_{c,i}$. If we stipulate a form for $p(c|\mathcal{M}_0)$ in (1) where $p(c|\boldsymbol{\kappa}) = \{p(c=1) = \kappa_1, p(c=2) = \kappa_2, \ldots, p(c = C) = \kappa_C\}$, then (3) and (6) can be substituted into (1) to yield a maximum likelihood model. This can then be learnt through an iteratrive process such as the EM-algorithm [4] or gradient descent [3]. However, we take the Bayesian route by also integrating out $\{\kappa, \Theta_c\}$.

3. BAYESIAN INFERENCE AND VARIATIONAL LEARNING

The maximum likelihood approach to learning the parameters of a model is well documented, as are the pitfalls (e.g. over-fitting, no quantification of uncertainty in the model or model comparison). We choose to take the Bayesian approach and integrate out the parameters $\{\kappa, \Theta_c\}$ and hidden variables $\{\boldsymbol{s}_c, \boldsymbol{q}_c\}$. First, we will state the prior distributions over the hidden variables and model parameters. The priors are chosen to be appropriately conjugate to allow tractibility. All ICA component parameter priors factorise across the C components.

The prior over the ICA mixture indicator variables, $\boldsymbol{c} = \{c^1, c^2, \ldots, c^N\}$, simply factorises over the N data vectors $p(\boldsymbol{c}|\boldsymbol{\kappa}) = \prod_{n=1}^{N} \kappa_{c^n}$. The prior over the ICA mixture coefficients $\boldsymbol{\kappa}$ is a C component Dirichlet. It follows from (6) that the distribution over the MoG component indicator variables $\boldsymbol{q_c}$ and sources $\boldsymbol{s_c}$ for each ICA component is a product over all N data vectors with the summation in 6 removed. The prior over the source model (MoG) parameters, $\boldsymbol{\varphi}_c$, is a product of priors over $\boldsymbol{\pi}_c, \boldsymbol{\mu}_c, \boldsymbol{\beta}_c$. The priors over each of these factorises across the L_c sources. The prior for each $\boldsymbol{\pi}_{c,i}$ is a m_i component Dirichlet The prior over the MoG component means, $\boldsymbol{\mu}_{c,i}$, for source i is a product of m_i Gaussians. The prior over the associated precisions, $\boldsymbol{\beta}_{c,i}$, is a product of m_i Gamma distributions The prior over the bias vector, $\boldsymbol{y}_c = \{y_1, y_2, \ldots, y_S\}$, is a product over S zero-mean Gaussians for each ICA component. The prior over the sensor noise precision, λ_c, is a Gamma distribution for each component. The prior over each element of the mixing matrix, $\boldsymbol{A}_{ji}$ is a zero-mean Gaussian with precision α_i for each column. By monitoring the precisions $\boldsymbol{\alpha}_c = \{\alpha_1, \ldots, \alpha_{L_c}\}$, the relevance of each source may be automatically determined (ARD). If α_i is large, column i of $\boldsymbol{A}_c$ will be close to zero, indicating source i is irrelevant. Finally, the prior over each $\alpha_{c,i}$ is a Gamma.

Bayesian inference in such a model is computationally intensive and often intractable. An important and efficient tool in approximating posterior distributions is the variational method (see [14] for an excellent tutorial). In particular, we take the *variational Bayes* approach detailed in [10].

VARIATIONAL BAYESIAN LEARNING

Let the weights $\boldsymbol{W}$ be a vector of all hidden variables and unknown parameters. Variational Bayesian learning involves assuming some factored form for the posterior over weights, denoted $p'(\boldsymbol{W})$. The objective function to maximise in variational Bayesian learning [10] is then given by

$$F[\boldsymbol{W}] = \left\langle \log \frac{p(\boldsymbol{X}, \boldsymbol{W})}{p'(\boldsymbol{W})} \right\rangle_{p'(\boldsymbol{W})} \qquad (8)$$

The quantity F is called the *negative free-energy* (NFE) for model $\mathcal{M}$ and can be shown [10] to be a strict lower bound to the log evidence, with the difference being the Kullback-Leibler (KL) divergence between the true and approximating posterior. By maximising F, not only do we minimise the KL-divergence between the approximating and true posterior, we also implicitely integrate out the unknowns $\boldsymbol{W}$. By choosing an appropriate form for the approximation $p'(\boldsymbol{W})$, we perform tractable Bayesian learning. As F is a strict lower bound to the model (log) evidence, a wide variety of models and assumptions can be compared and contrasted by calculating F for each model. The higher F is, the higher the likelihood of the data under that model, and, therefore, the better that model is at 'explaining' the data.

In our model, $\boldsymbol{W} = \{\boldsymbol{c}, \boldsymbol{s}_c, \boldsymbol{q}_c, \boldsymbol{\kappa}, \Theta\}$ where the lack of subscript c indicates the collection of C weights. By choosing $p'(\boldsymbol{W})$ such that it factorises, terms in each hidden variable can be maximised individually. We choose the following factorisation

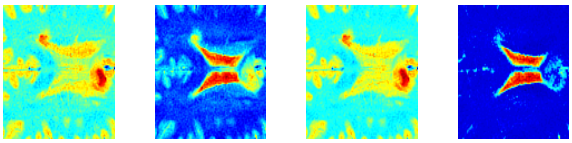
$$\begin{aligned} p'(\boldsymbol{W}) &= p'(\boldsymbol{c}) p'(\boldsymbol{s}_c|\boldsymbol{q}_c, c) p'(\boldsymbol{q}_c|c) p'(\boldsymbol{\kappa}) p'(\boldsymbol{y}) \\ &\times p'(\boldsymbol{\lambda}) p'(\boldsymbol{A}) p'(\boldsymbol{\alpha}) p'(\boldsymbol{\varphi}) \end{aligned} \qquad (9)$$

where $p'(\boldsymbol{\varphi}) = p'(\boldsymbol{\pi}) p'(\boldsymbol{\mu}) p'(\boldsymbol{\beta})$ and $p'(a|b)$ is the approximating density of $p(a|b, \boldsymbol{X})$. We will also stipulate that the posteriors over the sources factorise such that

$$p'(\boldsymbol{s}_c^n, \boldsymbol{q}_c^n|c) = \prod_{i=1}^{L_c} p'(q_i^n|c) p'(s_{c,i}^n|q_i^n, c) \qquad (10)$$

This additional factorisation allows efficient scaling of computation with the number of hidden sources.

By substituting $p(\boldsymbol{X}, \boldsymbol{W})$ and (9) into (8), we obtain expressions for the negative free energy, F, of our model. The bound F is maximised using the free-form approach of [15].All the derived posteriors require solving a set of coupled parameter update equations. In practice, this is best achieved by first initialising the posterior component responsibilities ($p'(\boldsymbol{c})$), use these to initialise each ICA component then commence learning on each ICA component. These components are then used to calculate the new posterior responsibilities and the learning 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 dataset and the (now fixed) model parameter distributions by calculating $\langle \boldsymbol{c} \rangle$, $\langle \boldsymbol{q}_c \rangle$ and $\langle \boldsymbol{s}_c \rangle$ under their respective posteriors.



(a) Original data (b) ICA features

We initialise the model by first running C-component k-means on the data. The class responsibilities generated are used to partition the data into C segments. These segments are used to initialise C ICA models at random. One must also choose the number of components in each source MoG. Although this can be determined by monitoring the contibution to F by each source ($F[\varphi_{c,i}]$) as a function of components, this is time consuming. We have found that 3 components are enough for most datasets, although 5 should be used for image data.

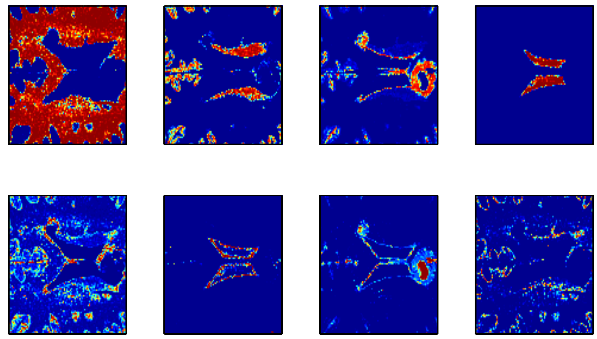
4. RESULTS

Areas in the data density described by a single coordinate frame are self-similar in that all points that fall within that area are generated by the same ICA process. This section will demonstrate that self-similarity in image density is interpretable as self-similar areas in the images. The vbMoICA algorithm was used to find self-similar areas in fMRI images and a scanned colour photograph of the Roman bath in Bath, U.K. For the former, vbMoICA was trained on two images taken at different spin sequences, while vbMoICA was trained on the vectorised RGB images of the latter.

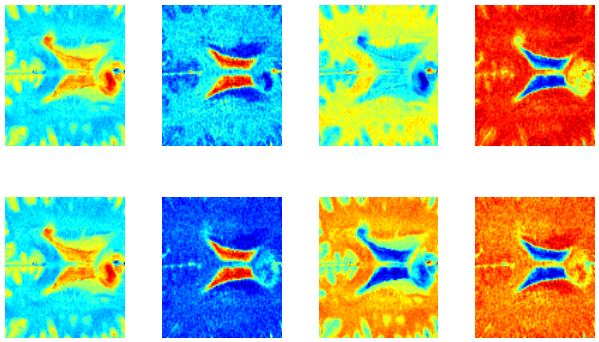
4.1. fMRI Images

We used the vbICA, variational Bayesian Mixture of Factor Analysers (vbMFA), gedecICA-MM and vbMoICA algorithms on fMRI images of a slice through a tumour patient's brain. Data was collected using both T2 and proton density (PD) spin sequences, which are used directly to form a two-dimensional feature space. Two 100x100 pixel images were vectorised, and 2000 samples were randomly drawn from the subsequent 2d data. Models with a range of latent dimensions were trained on the 2000 data vectors, and the most likely models were then used to unmix the complete 10000 points dataset into the corresponding independent features.

Figure 1 shows the two original (colour) images used along with the feature extracted by vbICA and the gedecICA mixture model. Due to the inherent limitations of ICA, no more than two features could be extracted from the fMRI images. The vbICA algorithm favoured a 2-source ICA model, shown in Figure 1(b). The left-hand feature is simply a copy of one of the original images. The right-hand image has separated out a local feature, which is, in fact, cerebro-spinal fluid. The vbMFA algorithm infers an 8 component model with $F = -14813.75$, giving the features in 1(c). The Bayesion Information Criterion of the gedecICA-MM chose a 4-component model with 2 sources



(c) vbMFA features

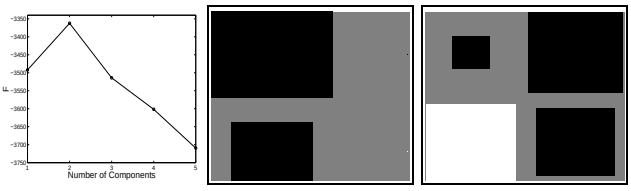


(d) gedecICA-MM features

Figure 1: fMRI images and features extracted

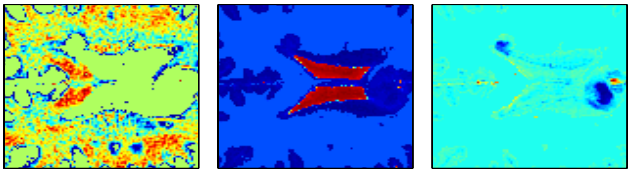
per ICA component, giving the eight overall features in Figure 1(d). Although the gedecICA-MM has managed to separate out the cerebro-spinal fluid, most features are simply scaled copies of each other and, therefore, the gedecICA-MM over-represents the fMRI images.

We trained 1-5 component models using vbMoICA, with each ICA component having the maximum 2 allowed sources. The negative free-energy (i.e. F) plot in Figure 2(a) shows that a 2-component model is preferred, while the Hinton plots in Figures 2(b)-(c) infer 1-source and 2-source ICA components. Compared to vbMFA, vbMoICA had $F = -3618.8$, a substantial improvement. The 3 features extracted by the most likely model are presented in Figure 3 along with their learnt distributons. The single source of the first ICA component - shown in Figure 3(a) - is global 'background' brain-tissue detail. The second ICA component represents more local features where the central part of Figure 3(b) is, once again, the cerebro-spinal fluid. More interestingly, however, the second source has extracted 2 dark 'blobs'. These are the tumors, which neither vbICA, vbMFA or gedecICA-MM picked out. The features' respective MoGs can be interpreted as the distribution of colours in each feature. The tumors' distribution is heavily peaked around blue, with the left-hand tail capturing the yellow-green information. The other distributions similarly repre-

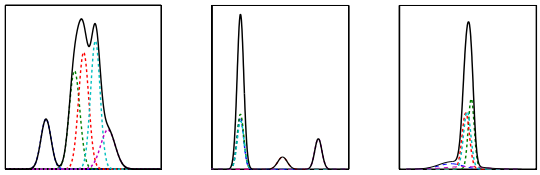


(a) vbMoICA negative free-energy (b) 1d cluster (c) 2d cluster

Figure 2: Inferred latent structure for fMRI images



(a) Background (b) C-S fluid (c) Tumours



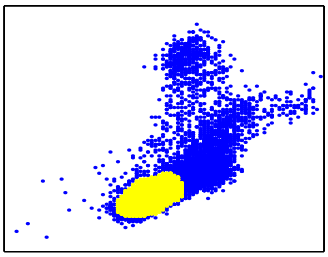
(d) ICA 1 (e) ICA 2

Figure 3: vbMoICA features extracted from fMRI images and respective ICA component source distributions

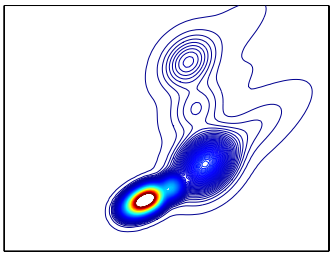
sent blue-red from left to right. The ability of vbMoICA to capture multi-modal feature distributions is pivotal in allowing the complex background distribution to be separated from the rest. This representation is thus much more interpretable and efficient than either simple ICA or gedecICA-MM. Figure 4(a) shows the data partition, while Figure 4(b) is a contour plot data likelihood under the model.

4.2. Roman Bath

The vbMoICA algorithm was used to find self-similar areas in a scanned photograph of a Roman bath taken on ISO200 film by an entry-level SLR camera. Figure 5 shows the original image and its RGB components in grey-scale. The 3 167×256 images formed a 3-dimensional data set, 42752 vectors long. The algorithms used in section 4.1, together with vbMoICA, were trained on 5000 vectors drawn at random. The learnt models were then used to unmix the whole data set. The vbMFA and gedecICA-MM algorithms found 23 and 13 areas respectively, none which were readily inter-



(a) vbMoICA partition



(b) vbMoICA data model

Figure 4: vbMoICA model for fMRI images

pretable. The vbMoICA algorithm inferred 3 self-similar areas in the data cluster, each of which was 2-dimensional.

Figure 5 shows the source reconstructions of the learnt ICA components. The three components can be broadly characterised as background buildings, main building and pool. Each is described by two sources, one which seems to represent chromanance information, and the other brightness. Although the three sets of sources are unlikely to be independent (particularly background and main building which share similar coloured stone), the mixture model nature of vbMoICA has allowed these areas to be separated, while the multi-modal nature of the sources have allowed the rich colour information to be coded. The segmentation is not perfect, but is interpretable. Thie separation maybe stronger if a Markov prior is stipulated over the mixture coefficients (see [16]).

5. DISCUSSION

We have presented an algorithm for modelling complex non-Gaussian data distributions. The vbMOICA algorithm splits the distribution into self-similar areas, and uses ICA components to learn representations of these areas. The ICAs model these local manifolds by forming independent directions in the underlying distribution. ARD selects the appropriate dimensionality of each manifold, and variational Bayes allows the optimum number of ICA components to be inferred. We have demonstrated the algorithm by finding self-similar areas in images, separating them into interpretable features, which other mixture models failed to do.

(a) Image (b) Image-red

(c) Image-green (d) Image-blue

Figure 5: Original image and RGB components.

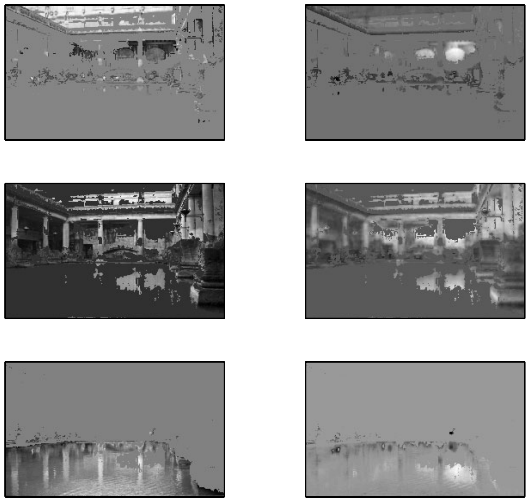


Figure 6: vbMoICA source reconstructions, top-bottom ICA models 1-3.

6. REFERENCES

[1] M.E. Tipping and C.M. Bishop, "Mixtures of probabilistic principal component analyzers," *Neural Computation*, vol. 11, no. 2, pp. 443–482, 1999.

[2] Z. Ghahramani and M. Beal, "Variational inference for Bayesian mixtures of factor analysers," in *Advances in Neural Information Processing Systems*, 2000, vol. 12, pp. 449–455.

[3] T.W. Lee, M.S. Lewicki, and T.J. Sejnowski, "ICA mixture models for unsupervised classification and automatic context switching," in *International Workshop on Independent Component Analysis*, 1999, pp. 209–214.

[4] W.D. Penny and S.J. Roberts, "Mixtures of Independent Component Analysers," in *Artificial Neural Networks - ICANN2001*. International Conference on Artificial Neural Networks, 2001, pp. 527–534.

[5] T.W. Lee, M. Girolami, and T.J. Sejnowski, "Independent Component Analysis using an extendend infomax algorithm for mixed sub-Gaussian and super-Gaussian sources," *Neural Computation*, vol. 11, no. 2, pp. 409–433, 1999.

[6] R. Choudrey and S.J. Roberts, "Variational mixture of Bayesian independent component analysers," Tech. Rep., University of Oxford, October 2001, Extended version to appear in *Neural Computation* 15(1).

[7] K. Chan, T-W. Lee, and T. Sejnowski, "Variational learning of clusters of undercomplete nonsymmetric independent components," in *Proceedings of ICA2001*, 2001.

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

[9] R.A. Choudrey, W.D. Penny, and S.J. Roberts, "An ensemble learning approach to Independent Component Analysis," in *Proceedings of Neural Networks for Signal Processing X, Sydney, December 2000*. IEEE Signal Processing Society, December 2000.

[10] 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).

[11] D.J.C. MacKay, "Probable networks and plausible predictions - a review of practical Bayesian methods for supervised neural networks," *Network: Computation in Neural Systems*, vol. 6, pp. 469–505, 1995.

[12] A.J. Bell and T.J. Sejnowski, "An information maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, pp. 1129–1159, 1995.

[13] J.W. Miskin and D.J.C. MacKay, "Application of ensemble learning to infra-red imaging," in *Proceedings of ICA2000*, 2000.

[14] M.I. Jordan, Z. Ghahramani, T.S. Jaakkola, and L.K. Saul, "An introduction to variational methods for graphical models," in *Learning in Graphical Models*, M.I. Jordan, Ed. The MIT Press, Cambridge, Massachusetts, 1999.

[15] 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.

[16] R.A. Choudrey and S.J. Roberts, "Learning hierarchical dynamics using Independent Component Analysis," Tech. Rep., University of Oxford, 2002, Submitted to ICA2003.