Mixtures of Independent Component Analysers

Stephen J. Roberts[1] and William D. Penny[2]

[1]Robotics Research, Department of Engineering Science, University of Oxford, UK
[2]Wellcome Department of Cognitive Neurology, University College London, UK

1 Introduction

Gaussian Mixture Models (GMMs) are widely used throughout the fields of machine learning and statistics. Despite their popularity, however, GMMs suffer from two serious drawbacks. The first is that, as the dimensionality d of the problem space increases, the size of each covariance matrix, d^2, becomes prohibitively large. This can be dealt with by assuming isotropic Gaussians (ie. ignoring the covariance structure) but this greatly reduces the flexibility of the model class. This problem has been solved by Tipping and Bishop [12] who replaced each Gaussian with a probabilistic Principal Component Analysis (PCA) model. This allowed the dimensionality of each covariance to be effectively reduced whilst maintaining the richness of the model class. The second problem with GMMs is that each component is a Gaussian, an assumption which is often violated in many natural clustering problems [5]. It is this second problem which we address in this paper. A solution is reached by extending the mixtures of probabilistic PCA model to a mixtures of Independent Component Analysis (ICA) model.

2 Mixture Models

The probability of generating a data point $p(\mathbf{x}_n)$ from a K-component mixture model is:

$$p(\mathbf{x}_n) = \sum_{k=1}^{K} p(k)p(\mathbf{x}_n|k) \tag{1}$$

Furthermore, having observed datum $\mathbf{x}_n$ we can compute the probability that it was generated from component k as

$$p(k|\mathbf{x}_n) = \frac{p(k)p(\mathbf{x}_n|k)}{p(\mathbf{x}_n)} \tag{2}$$

Learning in mixture models takes place using an Expectation-Maximisation (EM) strategy where the objective is to maximise the joint log-likelihood of the observed and hidden variables, averaged over $p(k|\mathbf{x}_n) \stackrel{\text{def}}{=} \gamma_k(n)$. Using Jensen's inequality this maximisation can be shown to maximise the likelihood of the data (see eg. [2]). We therefore maximise

$$L = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_k(n) \log \left[p(k)p(\mathbf{x}_n|k)\right] \tag{3}$$

which splits up into two terms

$$L = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_k(n) \log p(k) + \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_k(n) \log p(\mathbf{x}_n|k) \tag{4}$$

Maximising the first term yields an update rule for the mixing coefficients $p(k) \stackrel{\text{def}}{=} \pi_k$, namely:

$$\pi_k = \gamma_k / N \tag{5}$$

where we define

$$\gamma_k \stackrel{\text{def}}{=} \sum_{n=1}^{N} \gamma_k(n) \tag{6}$$

Maximising the second term gives update formulae for the parameters of the component densities. In this paper, the component densities derive from ICA models which are described in the next section.

3 ICA

In an ICA mixture model each component has a density derived from an ICA model. For the kth component the data are generated according to

$$\mathbf{x} = \mathbf{A}_k \mathbf{s}_k + \boldsymbol{\mu}_k + \mathbf{e}_k \tag{7}$$

where $\mathbf{e}_k$ is zero-mean Gaussian observation noise having an isotropic covariance matrix with variance σ_k^2, $\mathbf{A}_k$ is the mixing matrix and $\boldsymbol{\mu}_k$ is the mean of component k. Essentially, $\boldsymbol{\mu}_k$ determines the location of the cluster and $\mathbf{A}_k \mathbf{s}_k$ its shape. Given particular source values the *conditional* likelihood of a data point is

$$p(\mathbf{x}|k, \mathbf{s}_k) = \mathcal{N}(\mathbf{x}; \mathbf{A}_k \mathbf{s}_k - \boldsymbol{\mu}_k, \sigma_k^2 I) \tag{8}$$

where $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. The likelihood of a data point is hence

$$p(\mathbf{x}|k) = \int p(\mathbf{x}|k, \mathbf{s}_k) p(\mathbf{s}_k) d\mathbf{s}_k \tag{9}$$

in which, as the underlying sources $\mathbf{s}_k$ are taken as statistically independent,

$$p(\mathbf{s}_k) = \prod_{i=1}^{M} p(s_{k,i}) \tag{10}$$

where M is the hypothesised number of sources.

To compute the likelihood of an observation we must be able to calculate the integral in equation 9. If we assume that the distribution over sources is dominated by a single peak, $\hat{\mathbf{s}}_k$, then the integral can be performed using Laplace's method. This gives [8, 10]

$$\log p(\mathbf{x}|k) = \frac{-(d-M)}{2}\log\left(2\pi\sigma_k^2\right) - \frac{1}{2\sigma_k^2}(\mathbf{z}_k - \mathbf{A}_k\hat{\mathbf{s}}_k)^\top(\mathbf{z}_k - \mathbf{A}_k\hat{\mathbf{s}}_k)$$
$$+ \log p(\hat{\mathbf{s}}_k) - \frac{1}{2}\log|\det(\mathbf{A}_k^\top\mathbf{A}_k)| \tag{11}$$

where $|\cdot|$ denotes the absolute value and $\mathbf{z}_k = (\mathbf{x} - \boldsymbol{\mu}_k)$. The likelihood for the kth ICA component in a mixture model is given by the second term in equation 4,

$$L(k) = \sum_{n=1}^{N}\gamma_k(n)\log p(\mathbf{x}_n|k) \tag{12}$$

which may be combined with equation 11 to give:

$$L(k) = \frac{-\gamma_k(d-M)}{2}\log\left(2\pi\sigma_k^2\right) - \frac{1}{2\sigma_k^2}E_k$$
$$+ \sum_{n=1}^{N}\gamma_k(n)\log p(\hat{\mathbf{s}}_k(n)) - \frac{\gamma_k}{2}\log|\det(\mathbf{A}_k^\top\mathbf{A}_k)| \tag{13}$$

in which we have summed over N data points and we define

$$E_k \stackrel{\text{def}}{=} \sum_{n=1}^{N}\gamma_k(n)(\mathbf{z}_k(n) - \mathbf{A}_k\hat{\mathbf{s}}_k(n))^\top(\mathbf{z}_k(n) - \mathbf{A}_k\hat{\mathbf{s}}_k(n)) \tag{14}$$

as the total reconstruction error. Maximising the above likelihood term results in an update equation for the mean of the k-th component as

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N}\gamma_k(n)\mathbf{x}_n}{\gamma_k} \tag{15}$$

3.1 The decorrelating manifold

In previous work we have constrained the unmixing matrix to be a decorrelating matrix [3]. The motivation for this is that, for sources to be statistically independent, they must be at least linearly decorrelating. This has the further advantage that the unmixing matrix lies on a lower-dimensional manifold which makes the approach considerably faster. The corresponding mixing matrix is defined as follows.

Consider $\mathbf{Z}_k$ as an $N \times T$ matrix of zero-mean data vectors, with each entry weighted by $\sqrt{\gamma_k(n)/\gamma_k}$. The Singular Value Decomposition (SVD) of $\mathbf{Z}_k$ is given by [3]

$$\mathbf{Z}_k = \mathbf{U}_k\boldsymbol{\Lambda}_k\mathbf{V}_k \tag{16}$$

hence $\mathbf{U}$ contains the principal components of the weighted observation covariance matrix and $\boldsymbol{\Lambda}_k = [\lambda_1(k), \lambda_2(k), \ldots\lambda_d(k)]$ contains the standard deviations of the corresponding principal components. The mixing matrix is then given by

$$\mathbf{A}_k = \mathbf{U}_k^M\boldsymbol{\Lambda}_k^M\mathbf{Q}_k^\top \tag{17}$$

where $\mathbf{U}_k^M$ and $\boldsymbol{\Lambda}_k^M$ are the first M columns of $\mathbf{U}_k$ and $\boldsymbol{\Lambda}_k$. The transform is also parameterised by an orthonormal matrix $\mathbf{Q}_k$ of dimension $M \times M$ which constitutes the ICA transform proper[1].

3.2 Source and noise estimation

We estimate the sources at their Maximum Likelihood (ML) values. These are recovered via an unmixing matrix given by the pseudo-inverse of the mixing matrix

$$\mathbf{W}_k = (\mathbf{A}_k^\top \mathbf{A}_k)^{-1} \mathbf{A}_k^\top = \mathbf{Q}_k (\boldsymbol{\Lambda}_k^M)^{-1} (\mathbf{U}_k^M)^\top \tag{18}$$

and the sources are then estimated as

$$\mathbf{s}_k = \mathbf{W}_k \mathbf{z}_k \tag{19}$$

By defining

$$\hat{\mathbf{z}}_k \stackrel{\text{def}}{=} (\boldsymbol{\Lambda}_k^M)^{-1} (\mathbf{U}_k^M)^\top \mathbf{z}_k \tag{20}$$

we can then write $\mathbf{s}_k = \mathbf{Q}_k \hat{\mathbf{z}}_k$. The reconstructed observations are hence given by:

$$\mathbf{A}_k \mathbf{s}_k = (\mathbf{U}_k^M \mathbf{U}_k^M)^\top \mathbf{z}_k \tag{21}$$

As the average reconstruction error (under an L_2 norm) is equal to the variance of the data not explained by the first M components [8]

$$E_k = \sum_{i=M+1}^{d} \lambda_i(k)^2 \tag{22}$$

so the ML estimate of the noise variance (found from the first derivative of equation 13) is [10]

$$\hat{\sigma}_k^2 = \frac{E_k}{d-M} \tag{23}$$

hence,

$$\hat{\sigma}_k^2 = \frac{1}{d-M} \sum_{i=M+1}^{d} \lambda_i(k)^2 \tag{24}$$

This is the same as the noise estimate in the probabilistic PCA model [11]. Therefore, if ICA is constrained to the decorrelating manifold and ML source estimates are used, the observation noise level is not dependent on $\mathbf{Q}_k$ ie. the ICA transform proper. It can therefore be calculated ahead of optimising $\mathbf{Q}_k$ and fixed to its calculated value.

[1] In this paper, because we have a flexible source model - see later - the rows of our mixing matrix can be constrained to unit length, so it is not necessary to also include a diagonal scaling matrix [3].

3.3 Mixing matrix estimation

The matrix $\mathbf{Q}_k$ is constrained to be orthonormal by writing it as

$$\mathbf{Q}_k = \exp(\mathbf{S}_k) \tag{25}$$

where $\mathbf{S}_k$ is a skew-symmetric matrix ($\mathbf{S}_k^\top = -\mathbf{S}_k$) whose upper off-diagonal entries are known as Cayley coordinates [6].

Substituting the maximum likelihood estimates for the noise variance from equation 23 and the unmixing matrix from equation 18 into equation 13 and noting that since $\mathbf{W}$ is the psuedo-inverse of $\mathbf{A}$ we have $\det(\mathbf{W}^\top\mathbf{W}) = 1/\det(\mathbf{A}^\top\mathbf{A})$, so the log-likelihood term becomes:

$$\begin{aligned} L(k) = &\frac{-\gamma_k(d-M)}{2}\log\left(2\pi\hat{\sigma}_k^2\right) - \frac{\gamma_k(d-M)}{2} \\ &+\sum_{n=1}^{N}\gamma_k(n)\log p(\mathbf{s}_k(n)) + \frac{\gamma_k}{2}\log|\det(\mathbf{W}_k^\top\mathbf{W}_k)| \end{aligned} \tag{26}$$

The second term has been simplified due to the introduction of the ML noise variance estimate. The last term has the effect of preventing the unmixing matrix from becoming co-linear (ie. it prevents the components from pointing in the same direction).

The only term in the likelihood that depends on $\mathbf{Q}_k$ is the log source density where the dependence is introduced via equations 18 and 19. In previous work [3] we show how to compute the derivative of this term and combine it with an expression for $d\mathbf{Q}_k/d\mathbf{S}_k$. In this paper we use the same expression save for weighting the data with $\gamma_k(n)$. This then gives the gradient of the likelihood with respect to the Cayley coordinates. Fitting the ICA model therefore corresponds to following this gradient using, for example, a quasi-Newton optimisation.

3.4 Generalised Exponential Sources

In Bell and Sejnowski's original ICA formulation [1] the source densities were (implicitly) assumed to be reciprocal-cosh densities. More recent work (see, for example, [3]) however, shows that the inverse-cosh form cannot adequately model sub-Gaussian densities[2]. A more general parametric form which can model super-Gaussian, Gaussian *and* sub-Gaussian forms is the 'Exponential Power Distribution' or 'Generalised Exponential' (GE) density

$$G(s;R,\beta) = \frac{R\beta^{1/R}}{2\Gamma(1/R)}\exp(-\beta|s|^R) \tag{27}$$

where $\Gamma(\cdot)$ is the gamma function [9]. This density has zero mean, a kurtosis determined by the parameter R and a variance which is then determined by $1/\beta$. This source model has been used in an ICA model developed in [3].

[2] Sub/super densities have lighter/heavier tails than Gaussians.

The log-likelihood of the ith source model is given by the third term in equation 26. By dividing by γ_k we obtain the normalised log-likelihood

$$L_i(k) = \sum_{n=1}^{N} \frac{\gamma_k(n)}{\gamma_k} \log p(s_{k,i}(n)) \tag{28}$$

Substituting in the GE density gives

$$L_i(k) = \frac{1}{\gamma_k} \sum_{n=1}^{N} \gamma_k(n) \log \left[\frac{R_i \beta_i^{1/R_i}}{2\Gamma(1/R_i)} \right] - \frac{1}{\gamma_k} \sum_{n=1}^{N} \gamma_k(n) \beta_i |s_{k,i}(n)|^{R_i} \tag{29}$$

The source density parameters for each state can now be estimated component-wise by evaluating the derivatives $\frac{\partial L_i(k)}{\partial \beta_i}$ and $\frac{\partial L_i(k)}{\partial R_i}$ and setting them to zero. This gives

$$\frac{1}{\beta_i} = \frac{R_i}{\gamma_k} \sum_{n=1}^{N} \gamma_k(n) |s_{k,i}(n)|^{R_i} \tag{30}$$

and R_i can be found by solving the one-dimensional problem

$$\frac{\partial L_i(k)}{\partial R_i} = \frac{1}{R_i} + \frac{1}{R_i^2} \log \beta_i + \frac{1}{R_i^2} \varphi(1/R_i) - \frac{\beta_i}{\gamma_k} \sum_{n=1}^{N} \gamma_k(n) |s_{k,i}(n)|^{R_i} \log |s_{k,i}(n)| = 0 \tag{31}$$

where $\varphi(x) = \Gamma^{'}(x)/\Gamma(x)$ is the digamma function [9]. Equations 30 and 31 are generalised versions of the equations in Appendix A of [3] and are applied iteratively until a consistent solution is reached (this whole process is identical to the optimisation of GE sources in a Hidden Markov ICA context [7]).

4 Practical Issues

The ICA-mixture model is initialised using a K-means clustering routine [2]. If at the end of this process any cluster is assigned less than 2 points, the K-cluster model is deemed unviable.

Otherwise, the K-means process produces K data sets. A PCA model order selection criterion [6] is then applied to each to determine the optimal latent dimension, M_k, and an ICA model is then fitted to each group. The responsibilities are then calculated (E-step) and the full EM optimisation of the ICA-mixture model begins. This consists of repeated application of equations 5, 14, 30, 31 and setting the mixing matrix so as to maximise equation. 26.

It is also possible to perform model order selection on the number of clusters. This is easily achieved by substituting the ICA-mixture likelihood into a Bayesian Information Criterion (BIC) as described in [8].

5 Results

We applied our algorithm to two data sets[3]. The first consists of a set of spectra from quasar emissions. The data is 4-D and the algorithm selects a two-component ICA model (Fig. 1(a)) with each component having a latent dimension of 2. The resultant contours are shown in Fig. 1(b). The two components correspond to background and line emission spectra. The second dataset we

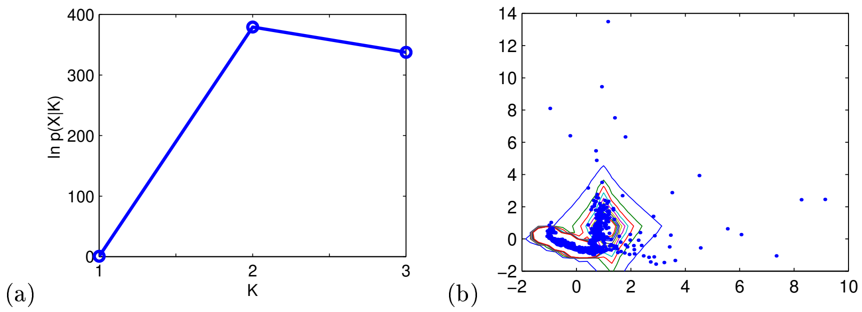

Fig. 1. Quasar data: *(a)* $\ln p(X|K)$ *against number of components,* K *and (b) resultant* $K = 2$ *density plot. Note the non-Gaussianity of the sources.*

consider is a 13-D set of wine composition data. Fig. 2(a) shows that a 3 component model is (correctly) favoured. Each component has a latent dimension of two. The resultant unsupervised partitioning is shown in Fig. 2(b) and is 97% accurate (in this case we know the true labels).

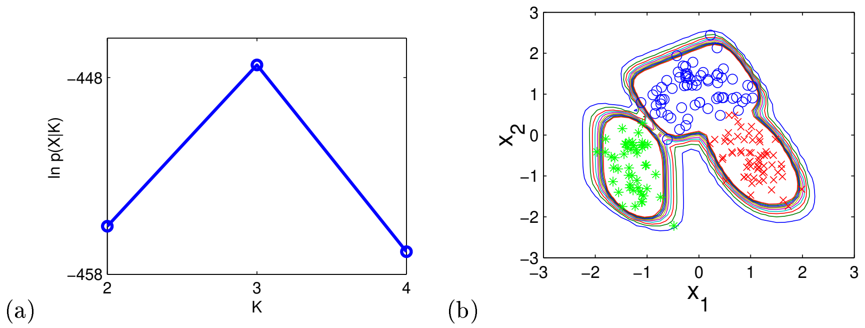

Fig. 2. Wine data: *(a)* $\ln p(X|K)$ *against number of components,* K *and (b) resultant* $K = 3$ *data partition and density.*

[3] Both from the UCI archives.

6 Discussion

We have proposed a clustering algorithm in which each cluster may be non-Gaussian. This has been achieved by extending the mixtures of probabilistic PCA model to a mixtures of ICA model. The model is a special case of Hidden Markov ICA [7], which itself may thence be viewed as MoICA for time-series data. An ICA mixture model has also been proposed in [5]. This paper extends their approach by allowing the dimensionality of the latent space to be smaller than the observation space and for the ICA components to be evaluated on the decorrelating manifold. Furthermore, the appropriate latent dimension is set using a model order criterion.

References

1. A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.
2. C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, 1995.
3. R. Everson and S.J. Roberts. Independent Component Analysis: A flexible non-linearity and decorrelating manifold approach. *Neural Computation*, 11(8), 1957–1984, 1999.
4. R.E. Kass and A.E. Raftery. Bayes factors and model uncertainty. Technical Report 254, University of Washington, 1993. http://www.stat.washington.edu/tech.reports/tr254.ps.
5. T. Lee and M.S. Lewicki. Image Processing methods using ICA mixture models. In S. Roberts and R. Everson, editors, *Independent Component Analysis: Principles and Practice*. Cambridge University Press, 2001.
6. T.P. Minka. Automatic choice of dimensionality for PCA. Technical Report 514, MIT Media Laboratory, Perceptual Computing Section, 2000.
7. W.D. Penny, R.M. Everson, and S.J. Roberts. Hidden Markov Independent Components Analysis. In M. Girolami, editor, *Advances in Independent Component Analysis*. Springer, 2000.
8. W.D. Penny, S.J. Roberts, and R. Everson. ICA: model order selection and dynamic source models. In S.J. Roberts and R. Everson, editors, *Independent Component Analysis: Principles and Practice*. Cambridge University Press, 2001.
9. W. H. Press, S.A. Teukolsky, W.T. Vetterling, and B.V.P. Flannery. *Numerical Recipes in C*. Cambridge, 1992.
10. S.J. Roberts. Independent Component Analysis: Source assessment and Separation, a Bayesian Approach. *IEE Proceedings on Vision, Image and Signal Processing*, 145(3):149–154, 1998.
11. M. E. Tipping and C. M. Bishop. Probabilistic Principal Component Analysis. Technical report, Neural Computing Research Group, Aston University, 1997.
12. M. E. Tipping and C. M. Bishop. Mixtures of Probabilistic Principal Component Analyzers. *Neural Computation*, 11(2):443–482, 1999.