INDEPENDENT DATA DECOMPOSITION

Stephen Roberts and Rizwan Choudrey

Pattern Analysis & Machine Learning Research Group, University of Oxford, UK.
{sjrob,riz}@robots.ox.ac.uk

ABSTRACT

In many data analysis problems it is useful to consider the data as generated from a set of unknown (latent) generators or sources. The observations we make of a system are then taken to be related to these sources through some unknown function. Furthermore, the (unknown) number of underlying latent sources may be less than the number of observations. Recent developments in Independent Component Analysis (ICA) have shown that, in the case where the unknown function linking sources to observations is linear, such data decomposition may be achieved in a mathematically elegant manner. In this paper we extend the general ICA paradigm to include a very flexible source model, prior constraints and conditioning on sets of intermediate variables so that ICA forms one part of a hierarchical system. We show that such an approach allows for efficient discovery of hidden representation in data and for unsupervised data partitioning.

Keywords: Independent component analysis, latent variable models, variational Bayes, unsupervised partitioning, data representation.

1. INTRODUCTION

Independent Component Analysis (ICA) has been widely used in data analysis and decomposition in recent years (see, for example, [1] for an overview). In this paper we consider the case in which constraints exist in our beliefs regarding the mixing process and the sources. In particular we investigate the issue of enforcing *positivity* onto the ICA model. This is coupled with the building of hierarchical models and used to perform unsupervised data partitioning and decomposition. As the models are evaluated in a fully Bayesian manner model selection may be applied to infer the complexity of data representation in the solution space.

2. THEORY

In general, we have a set of observations vectors, $\mathbf{x} \in \mathbb{R}^N$ which are believed to be caused by a set of underlying *latent* sources, $\mathbf{s} \in \mathbb{R}^M$, which we cannot directly access. What we observe is some function, however, of these sources i.e. $\mathbf{x} = f(\mathbf{s})$ where $f : \mathbb{R}^M \mapsto \mathbb{R}^N$. The problem of inferring the sources, and f as well, from observations alone is typically an impossible task if we have no other knowledge of the mapping. If the mapping is linear via a (non-square) mixing matrix, $\mathbf{A}$, i.e. $\mathbf{x} = \mathbf{A}\mathbf{s}$ then it is still ill-posed but constraints allow for its solution. ICA makes the solutions to this Equation well-posed by forcing independence between the components, $\mathbf{s}$, of the basis. From the perspective of probabilistic modeling this independence means that the joint probability density over the basis sources factorizes.

2.1. Hierarchical models

We consider the ICA source estimates as some inverse process, f^{-1}, acting on the observations, i.e. $\mathbf{s} = f^{-1}(\mathbf{x})$. In a standard ICA model, this inverse function is a generalized linear function (a function of the pseudo-inverse of the mixing matrix and the inferred noise model). In a hierarchical approach we consider, however, the unknown sources to be some generalised linear function of a set of intermediate variables, which we denote $\boldsymbol{\phi}$, which are themselves functions of the observations. The mapping may hence be written as:

$$\mathbf{s} = f^{-1}(\boldsymbol{\phi}(\mathbf{x})). \tag{1}$$

Making the ICA assumption that f may be modeled as a linear process, we may see the intermediate variables as a ‘dummy’ observations set under the equation

$$\boldsymbol{\phi}(\mathbf{x}) = \mathbf{A}\mathbf{s} + \mathbf{n}. \tag{2}$$

where $\mathbf{n}$ is an unknown noise process. In principle we may treat this equation as that of a fully generative model, i.e. in which our objective in model learning is to maximize $p(\mathbf{x}|\mathbf{s}, \mathbf{A}, \mathbf{n})$ and hence the resultant sources are conditioned on the observations, i.e. $p(\mathbf{s}|\mathbf{x}, \mathbf{A}, \mathbf{n})$. This necessarily involves the inversion of $\boldsymbol{\phi}(\mathbf{x})$. This is is relatively straightforward if the intermediate variables are linear in the observations, $\mathbf{x}$. This, however, gives little in the way of modeling power over standard ICA. We choose, therefore, for the intermediate variables to be *non-linear* in $\mathbf{x}$. Inversion of this functionality then involves more complex procedures. In the examples we present in this paper we are concerned primarily in the inference of the latent sources and their number and we take the pragmatic step to condition on the intermediate variables rather than the observations themselves, i.e. $p(\mathbf{s}|\boldsymbol{\phi}, \mathbf{A}, \mathbf{n})$. The practical effect of this decision is that we may form the mapping from the observations to the intermediate variables, $\mathbf{x} \to \boldsymbol{\phi}$ in an independent step which is not adapted as part of the hierarchical model. The ICA model then is inferred using Equation 2. In this form the model becomes similar to that of *Radial Basis Function* (RBF) systems (see, e.g., [2]) in which $\mathbf{x} \to \boldsymbol{\phi} \to \mathbf{s}$.

2.1.1. Hierarchical models as data partitioning models

One of the aims of unsupervised data analysis lies in the partitioning (or clustering) of data into classes, or groups, which are self-similar in some sense. In our recent work on hierarchical partitioning models we argue for partitions themselves to be modeled as mixtures of underlying simpler functions (Gaussians, for example) [3]. The objective argued for in this approach was to minimize the entropy of the resultant partitioning. If we consider the canonical measure of independence in ICA models as the *mutual information* between the recovered source components then ICA models can be seen as minimizing this mutual information measure. We may rewrite the mutual information between the components of $\mathbf{s}$, in Equation 2 as (see, for example, [1], chapter 1):

$$MI[\mathbf{s}] = \mathcal{H}[\boldsymbol{\phi}] + \frac{1}{2}\log|\det \mathbf{A}^{\mathsf{T}}\mathbf{A}| + \sum_{m=1}^{M}\mathcal{H}[s_m] \quad (3)$$

in which $\mathcal{H}[.]$ is a Shannon entropy measure and $\mathbf{A}$ is the mixing matrix. As we are considering the case in which the intermediate variables are *fixed* with respect to the ICA algorithm, the term $\mathcal{H}[\boldsymbol{\phi}]$ is also fixed and minimization of the mutual information is tantamount to minimizing the entropies over the sources. These sources are treated as measures over the class partitioning. The effect of minimization of the mutual information between the sources (i.e. the objective of ICA) may hence be seen as similar to other objectives in unsupervised partitioning, in particular the minimization of partition entropy as derived and employed in [4, 3].

2.2. Prior constraints

In the formalism presented in the section above, we take the elements of $\mathbf{A}$ to be *non-negative.* We assume that the responses which form the intermediate variable set are combinations of the underlying sources via a mixing process which, following [3], we take to be strictly non-negative. This enables us, for example, to treat the resultant sources (partition measures), after normalization (i.e. such that they sum to unity) as estimates of the posterior partition probabilities.

2.3. Source Model

The choice of a flexible and mathematically attractive (tractable) source model is crucial if a wide variety of source distributions are to be modeled. As we believe each source is non-negative, we utilize a mixture of *rectified* Gaussians[1] for each source model [15]. The set of all parameters of this model source could be learnt through a maximum likelihood approach such as the Expectation-Maximization (EM) algorithm ([5], [6]) (see also [7] for a comprehensive derivation of the EM algorithm with regard to ICA/IFA). Maximum-likelihood approaches are well-documented (see [8], [9], [7] for an introduction), as are the pitfalls. We take, arguably, a more comprehensive approach in which a full Bayesian paradigm is employed. This is briefly discussed later in this paper.

[1] A rectified Gaussian is defined as $\mathcal{N}^r(y|\alpha^{-1}) = 2\mathcal{N}(y|0, \alpha^{-1})$ for $y \geq 0$ and zero for $y < 0$. The hyper-parameter α governs the precision (inverse variance) of the distribution.

2.4. Dimensionality inference

The prior over each element of the mixing matrix, $\mathbf{A}$ is a rectified Gaussian with precision α_i for each column. By using a rectified Gaussian we force non-negative solutions on the mixing process [12, 15]. By monitoring the evolution of the precision hyperparameters $\{\alpha_i\}$, the relevance of each source may be determined (this is referred to as *Automatic Relevance Determination* - ARD). If α_i is large, column i of $\mathbf{A}$ will be close to zero, indicating source i is irrelevant. The inferred values of $\{\alpha_i\}$ hence give an indication of the most-likely number of source structures within the data.

3. BAYESIAN INFERENCE AND VARIATIONAL LEARNING

We choose to take a fully Bayesian approach and infer the posterior distributions over unknowns in the model (including the sources and all prior hyperparameters). 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 [10] for an excellent tutorial). In particular, we take the *variational Bayes* approach detailed in [11].

3.1. Variational Bayesian Learning

In the variational Bayes framework, the objective function to be maximized is the *negative free energy, F*

$$F = \langle \log p(\mathcal{D}, \mathbf{U}) \rangle_{p'(\mathbf{U})} + \mathcal{H}\left[p'(\mathbf{U})\right] \qquad (4)$$

where $\mathbf{U}$ consists of the set of all unknowns in the model, $\mathcal{M}$, and $\mathcal{D}$ is the visible data. The first term in (4) is the expectation of the joint density of hidden and observed variables with respect to an approximating posterior $p'(\mathbf{U})$. The second term is the entropy of $p'(\mathbf{U})$. The negative free energy forms a strict lower bound on the evidence, $p(\mathcal{D}|\mathcal{M})$, of the model, with the difference being the Kullback-Leibler (KL) divergence between the true and approximating posteriors. Maximising this function is equivalent to minimising the KL divergence between the true and approximate posteriors. A wide variety of models and assumptions can be compared and contrasted by calculating the free energy of each model. The higher the (negative) free energy, the higher the likelihood of the data under that model, and, therefore, the better that model is at 'explaining' the data.

By choosing $p'(\mathbf{U})$ such that it factorizes over the set of unknown model parameters, terms in each hidden variable can be maximized individually. More details of this approach may be found in [12, 13, 14, 15]. Once we have obtained the negative free energy, F, of our model one may proceed by specifying functional forms for each of the approximating parameter posteriors and use these in Equation (4) as shown by [16]. As shown in [17], however, there is no need to specify functional forms for (all) the approximating posteriors as they 'fall-out' of the maximization process, helped by the factorized form of $p'(\mathbf{U})$. The optimal form for each posterior is simply given by

$$p'(U_k) \propto p(U_k) \exp\left[\langle \log p(\mathcal{D}, \mathbf{U}) \rangle_{\prod_{l \neq k} p'(U_l)}\right] \qquad (5)$$

where the index k refers to the k-th parameter in $\mathbf{U}$.

This can be fully applied allowing free-form optimization giving the ensemble learning algorithms presented in [18] which is the approach taken in the experiments presented in this paper. Alternatively, a full mixture posterior can be used but this no longer allows full free-form learning. This approach is detailed in [13]. Full details of the algorithms and parameter update equations are given in [18, 13, 14]. All the derived posteriors require solving a set of coupled hyper-parameter update equations which are cycled 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 the source esimates under their respective posteriors over the whole data-set, and given the (now fixed) model parameter posteriors.

4. RESULTS

4.1. Initialization

We set vague priors for all hyper-parameters (which govern the prior distributions over parameters) [15]. Three component mixtures of rectified Gaussians were used for all source models. Model learning was terminated when the relative change in free energy dropped below 10^{-5}. This took between 10 and 50 iterations of the model for the examples shown in this paper. Fifty iterations took just under 1 minute of CPU time running under Matlab on a 1.4GHz processor. It is worth noting that, as a full Bayesian learning paradigm is taken, we avoid the need for user-specified parameters in the model.

4.2. Toy data

Figure 1(a) shows a simple data set consisting of two data stuctures. Whilst the inner region is Gaussian distributed the ring surrounding it is highly non-Gaussian and simple Gaussian mixtures, for example, fail to detect this structure. Also shown in the plot are the locations of a set of 20 Gaussian kernels fitted using the EM algorithm. Plot (b) shows the first five α^{-1} from the ICA model. We note that only two significant 'sources' have been inferred. Back projecting onto the data allows for the partitioning shown in plot (c).

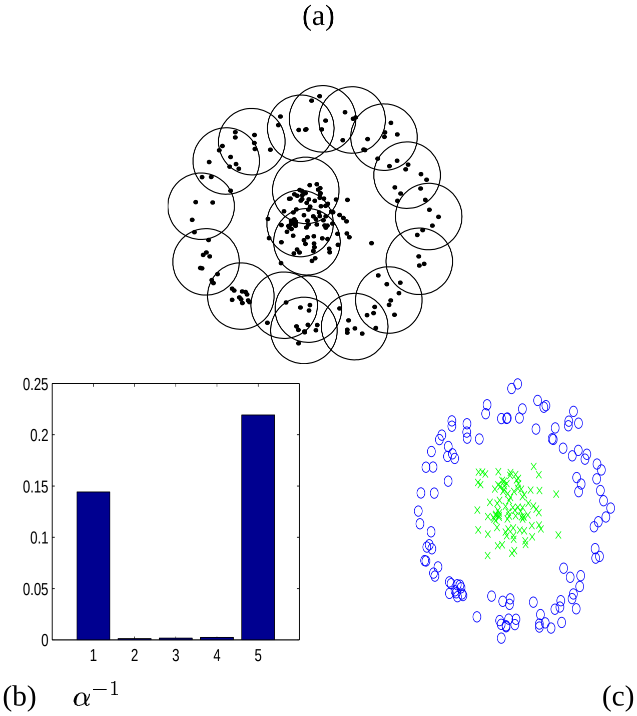

Figure 1: ***Ring data:*** *(a) data set and 20-component Gaussian kernel set, (b) ARD hyper-parameters showing support of three cluster model, (c) the resultant data partitioning.*

4.3. Wine data

As a second example we present results from a wine recognition problem. In our example the data set consists of 178 2-dimensional exemplars which are a set of chemical analyses of three types of wine. We fit an initial 20-kernel set using the Expectation-Maximization (EM) algorithm [5, 2]. These kernels, which form the basis for the ϕ intermediate variables, are shown along with the data set in Figure 2.

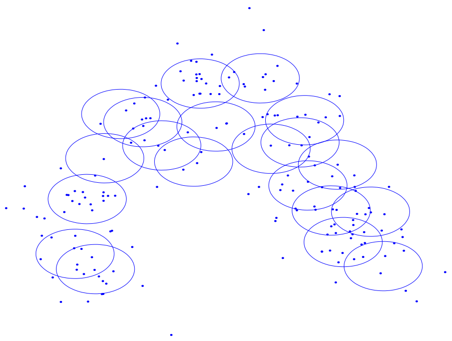

Figure 2: ***Wine data:*** *Gaussian kernel basis. The responses from this kernel set form the ϕ observation variables for the ICA model.*

Figure 3(a) shows the first five hyper-parameters, α_i, governing the scale of the mixing process. Three significant 'sources' have been determined. This is verified in plot (b) which shows the elements of the unmixing matrix, $\mathbf{W}$. Plot (c) shows the true class partitions (projected onto the first two dimensions of the data) and (d) the resultant partitioning. We note that the few differences between (c) and (d) are due to 'outlying' points from the class distributions and as such cannot be determined correctly by an unsupervised method. In this example there are 5 errors, corresponding to an equivalent classification performance of 97.2% and identical to the results achieved in [3]. We note that our approach is entirely unsupervised. Other analysis has been reported for this data set and our result is surprisingly good considering that supervised first-nearest neighbour classification achieves only 96.1%, and supervised multivariate linear-discriminant analysis 98.9% [19]. We note, furthermore, that standard ICA (without positivity priors) and PCA (which assumes Gaussian form for the sources) both fail to infer the correct structure in this data.

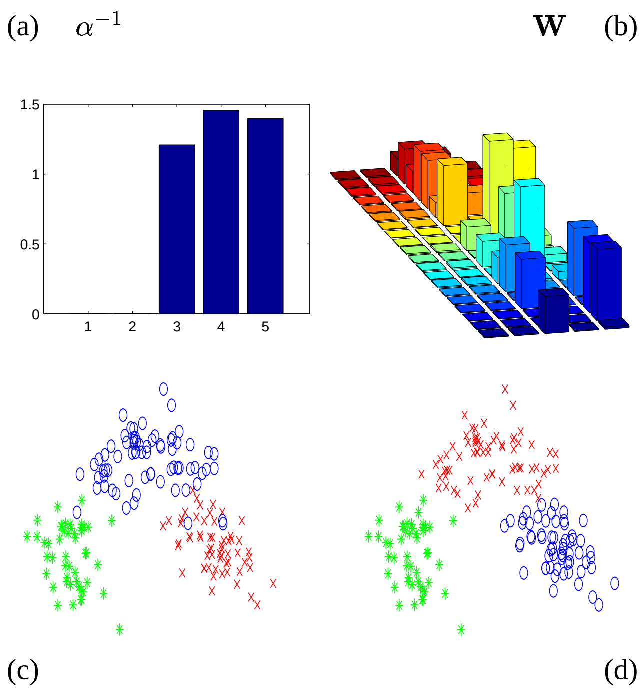

Figure 3: ***Wine data:*** *(a) ARD hyper-parameters showing support of three cluster model, (b) weights in the unmixing matrix, (c) the true data partitioning (from supervised labels) and (d) the resultant data partitioning.*

5. DISCUSSION & CONCLUSIONS

ICA models may be formed under the framework of generative Bayesian modeling. The model, hence, allows for priors over all parameters. This has the benefit that, not only is 'correct' Bayesian inference performed but that the priors allow components of the model which are not supported by the data (i.e. do not explain the data) to collapse thus giving a principled manner in which to infer the number of underlying latent data 'sources'. The use of priors also allows for constraints to the model space, and in this paper we explore the use of positivity priors on the mixing process from sources to observations. The non-negativity of the mixing process naturally fits with a hierarchical approach in which a set of intermediate (non-linear) function responses to the observed data is formed and the resultant latent structure within the data is formed as a linear decomposition in this intermediate space. This allows, for example, for unsupervised clustering to take place in a very flexible manner. As commented on earlier in this paper, ideally the functional mapping between the observations and the intermediate variables (which we may take to be non-linear, e.g. Gaussian kernel responses) would be inferred along with the ICA model. This is an ongoing area of research, and leads naturally to extensions of non-linear ICA.

6. REFERENCES

[1] S. Roberts and R. Everson. *Independent Component Analysis: principles and practice.* Cambridge University Press, 2001.

[2] C.M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, Oxford, 1995.

[3] S. Roberts, C. Holmes, and D. Denison. Minimum entropy data partitioning using Reversible Jump Markov Chain Monte Carlo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6), June 2001.

[4] S.J. Roberts, R. Everson, and I. Rezek. Maximum Certainty Data Partitioning. *Pattern Recognition*, 33(5):833–839, 2000.

[5] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. Roy. Stat. Soc.*, 39(1):1–38, 1977.

[6] R. M. Neal and G. E. Hinton. A view of the EM algorithm that justifies incremental, sparse and other variants. In Jordan [20], pages 355–368.

[7] H. Attias. Independent Factor Analysis. *Neural Computation*, 11:803–851, 1999.

[8] B. Pearlmutter and L. Parra. A Context-Sensitive Generalization of ICA. In *1996 International Conference on Neural Information Processing.*, 1996.

[9] J.-F. Cardoso. Blind signal separation: statistical principles. *IEEE Transactions on Signal Processing*, 9(10):2009–2025, 1998.

[10] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. In Jordan [20].

[11] T. S. Jaakkola and M. I. Jordan. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10:25–37, 2000.

[12] J. Miskin and D. MacKay. *Ensemble learning for blind source separation*, chapter 8 in *Independent Component Analysis: Principles and Practice*, S. Roberts & R. Everson (Eds.). Cambridge University Press, 2001.

[13] R. Choudrey and S. Roberts. Flexible Bayesian Independent Component Analysis for Blind Source Separation. In *Proceedings of ICA-2001*, San Diego, December 2001. (to appear).

[14] R. Choudrey and S. Roberts. Variational Mixture of Bayesian Independent Component Analysers. *Neural Computation*, 15(1), 2003. To appear.

[15] S. Roberts and R. Choudrey. Data Decomposition using Independent Component Analysis with Prior Constraints. *Pattern Recognition*, 2002. To appear.

[16] H. Lappalainen. Ensemble learning for Independent Component Analysis. In *Proceedings of ICA'99,Aussois, France.*, 1999.

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

[18] R. Choudrey, W. Penny, and S. Roberts. An ensemble learning approach to Independent Component Analysis. In *Proceedings of Neural Networks for Signal Processing*, Sydney, Australia, December 2000.

[19] S. Aeberhard, D. Coomans, and O. de Vel. Comparative-Analysis of Statistical Pattern-Recognition Methods in High-Dimensional Settings. *Pattern Recognition*, 27(8):1065–1077, 1994.

[20] M. I. Jordan, editor. *Learning in Graphical Models*. MIT Press, Cambridge, MA, 1999.