Generalised Component Analysis

Neil D. Lawrence† Michael E. Tipping‡
neil@dcs.shef.ac.uk mtipping@microsoft.com

23rd May 2003

†Department of Computer Science, Regent Court, 211 Portobello Road, Sheffield, S1 4DP, U.K.
‡Microsoft Research, 7 J. J. Thomson Avenue, Cambridge, CB3 0FB, U.K.

Abstract

Principal component analysis is a well known approach for determining the principal sub-space of a data-set. Independent component analysis is a widely used technique for recovering the linearly embedded independent components of a data-set. In this paper we develop an algorithm that, for super-Gaussian sources, extracts the direction and number of independent components of a data-set and determines the principal sub-space of the remaining components. This is achieved through the use of a latent variable model. We refer to the approach as Generalised Component Analysis and demonstrate its ability to both extract independent and principal components, as well as to determine the number of independent components, on toy and real world data-sets.

1 Introduction

In independent component analysis (ICA) [5, 10] we assume that an observed data-point consists of a d dimensional vector, $\mathbf{x}_n = \left(x_n^{(1)}, \ldots, x_n^{(d)}\right)$, that is generated by linear mixing of q source signals, $\mathbf{s}_n = \left(s_n^{(1)}, \ldots, s_n^{(q)}\right)$,

$$\mathbf{x}_n = \mathbf{A}\mathbf{s}_n + \boldsymbol{\eta}_n,$$

where $\boldsymbol{\eta}_n$ represents additive noise, which we will assume to be Gaussian and isotropic, $\mathbf{A}$ is known as the mixing matrix and we have assumed that the set of all data-points, $\mathbf{X}$, is centred.

The most important characteristic of ICA is that the components of the random variables $s_n^{(j)}$ are assumed to have been generated independently:

$$p\left(\mathbf{s}_n\right) = \prod_{j=1}^{q} p\left(s_n^{(j)}\right)$$

where q is the dimensionality of the latent space. However, if these latent variables are Gaussian, then it can be shown that recovery of the mixing matrix is only possible to within an arbitrary rotation [16]. The space spanned by this rotation is known as the principal sub-space. If, on the other hand, we are interested in recovering the *independent components* it is necessary to select a representation for the latent distribution which is non-Gaussian. Common choices are Laplacian or hyperbolic secant distributions. In this paper we use the Student-t. The Student-t distribution is super-Gaussian, in other words it has heavier tails than a Gaussian. This makes it suitable for determining the independent components of source distributions which are also super-Gaussian, such as speech and EEG signals.

We describe the structure of our Student-t distributed latent variable model in Section 2. In Section 5 we show how the parameters of the model may be efficiently determined through variational inference, giving an overview of the implementation details in Section 6. Finally we demonstrate the algorithm's ability to automatically estimate the number and direction of the independent components, as well as the principal components, in Section 7.

2 The Probabilistic Model

A Student-t distribution is controlled by a scale parameter, σ, and its degrees of freedom, ν,

$$t(y|\nu,\sigma) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi\sigma^2}}\left(1+\frac{y^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}, \tag{1}$$

where $\Gamma(\cdot)$ is the Gamma function. For low degrees of freedom, the distribution is very heavy tailed and as such it is suitable for determining independent components which are super-Gaussian. In the limit as the $\nu \to \infty$ the Student-t tends to a Gaussian with standard deviation σ. Therefore, if we can optimise ν as part of the model fitting process, we may use the Student-t for recovering source distributions which are both Gaussian and super-Gaussian; in other words a model based around this distribution can be considered to perform 'Generalised' Component Analysis (GCA).

Our GCA model is based on a source distribution that is a Student-t, $p(s_n^j|\nu_j,\sigma_j) = t(s_n^j|\nu_j,\sigma_j)$, and related to our data by a Gaussian noise model,

$$p(\mathbf{x}_n|\mathbf{A},\beta,\mathbf{s}_n) = \prod_{i=1}^{d}\sqrt{\frac{\beta}{2\pi}}\exp\left(-\frac{\beta}{2}(\mathbf{x}_n-\mathbf{A}\mathbf{s}_n)^{\mathrm{T}}(\mathbf{x}_n-\mathbf{A}\mathbf{s}_n)\right),$$

where d is the data dimensionality and $x_n^{(i)}$ represents the ith component of data point n. We have also introduced the inverse noise variance, or precision, β. We will look to estimate the parameters of our model through optimisation of an objective function, an obvious choice for which is the log-likelihood which may be obtained through marginalising the latent variables

$$\ln p(\mathbf{X}|\mathbf{A},\beta) = \ln\prod_{n=1}^{N}\int p(\mathbf{x}_n|\mathbf{A},\beta,\mathbf{s}_n)\,p(\mathbf{s}_n)\,d\mathbf{s}_n.$$

Unfortunately the required marginalisation is in general non-analytic, in the next section we therefore consider the 'noiseless' case for which a likelihood function may be developed then in Section 5 we develop a variational approximation which allows us to make progress.

3 The Likelihood Model

For simplicity we first consider the 'noiseless' case. For the Student-t distribution, this model was first proposed by [13]. Taking the limit of $p(\mathbf{x}_n|\mathbf{A},\beta,\mathbf{s}_n)$ as β tends to zero gives

$$p(\mathbf{x}_n|\mathbf{A},\beta,\mathbf{s}_n) = \prod_{i=1}^{d}\delta\left(\mathbf{a}_i^{\mathrm{T}}\mathbf{s}_n - x_n^{(i)}\right),$$

where $\mathbf{a}_i$ is a column vector containing the ith row of the mixing matrix, $x_n^{(i)}$ is the element from the ith column and nth row of the design matrix and $\delta(\cdot)$ is the Dirac-delta function. As is shown in [13], if $q = d$ the likelihood function may now be written

$$\ln p(\mathbf{X}|\mathbf{A}) = -N\ln|\mathbf{A}| + \ln\prod_{n=1}^{N}p\left(\mathbf{A}^{-1}\mathbf{x}_n\right).$$

It is convenient to define $\mathbf{W} = \mathbf{A}^{-1}$, and then gradients with repsect to $\mathbf{W}$ may be determined as

$$\frac{\partial \ln p\left(\mathbf{X}|\mathbf{A}\right)}{\partial \mathbf{W}} = N\mathbf{A} + \sum_{n=1}^{N} \mathbf{x}_n \mathbf{z}_n^{\mathrm{T}} \tag{2}$$

where the elements of $\mathbf{z}_n$ are

$$z_n^{(j)} = \frac{\partial p\left(s_n^{(j)}\right)}{\partial s_n^{(j)}}$$

and $s_n^{(j)} = \mathbf{w}_j^{\mathrm{T}} \mathbf{x}_n$. The focus of [13] is a covariant algorithm (see also [1]) but also mentioned is the possibility of optimising the latent distributions through gradient based methods. Whilst [13] suggests optimisation of both the degrees of freedom parameter and the scale parameter, for reasons discussed in Section 5.1, we prefer a constrained optimisation, which restricts the variance of each latent distribution to unity. The variance of the Student-t distribution is given by $\frac{\nu_j \sigma_j^2}{\nu_j - 2}$ for $\nu_j > 2$. Implementing this constraint through Lagrange multipliers leads us to set $\sigma_j^2 = \frac{\nu_j - 2}{\nu_j}$ (see Appendix), substituting for σ_j^2 in the log likelihood we obtain

$$\begin{aligned} \ln p\left(\mathbf{X}|\mathbf{A}\right) &= -N \ln |\mathbf{A}| + N \sum_{j=1}^{q} \ln \Gamma\left(\frac{\nu_j + 1}{2}\right) - N \sum_{j=1}^{q} \ln \Gamma\left(\frac{\nu_j}{2}\right) \\ &\quad - \frac{N}{2} \sum_{j=1}^{q} \ln\left(\nu_j - 2\right) \pi - \sum_{j=1}^{q} \sum_{n=1}^{N} \frac{\nu_j + 1}{2} \ln\left(1 + \frac{s_n^{(j)\,2}}{\nu_j - 2}\right), \end{aligned}$$

Implementation of the constraint that $\nu_j > 2$ could be achieved through the introduction of slack variables, but we choose simply to take $\nu_j = \nu_{\min} + \gamma_j^2$ where $\nu_{min} > 2$. Gradients with respect to γ_j may then be determined as follows

$$\begin{aligned} \frac{\partial \ln p\left(\mathbf{X}|\mathbf{A}\right)}{\partial \gamma_j} &= \gamma_j N \psi\left(\frac{\nu_{\min} + \gamma_j^2 + 1}{2}\right) - \gamma_j N \psi\left(\frac{\nu_{\min} + \gamma_j^2}{2}\right) \\ &\quad - \frac{\gamma_j N}{\left(\nu_{\min} + \gamma_j^2 - 2\right)} - \gamma_j \sum_{n=1}^{N} \ln\left(1 + \frac{s_n^{(j)\,2}}{\nu_{\min} + \gamma_j^2 - 2}\right) \\ &\quad + \gamma_j \sum_{n=1}^{N} \frac{\nu_{\min} + \gamma_j^2 + 1}{\left(\frac{\nu_{\min} + \gamma^2 - 2}{s_n^{(j)}}\right)^2 + \nu_{\min} + \gamma_j^2 - 2}, \end{aligned} \tag{3}$$

where $\psi\left(\cdot\right) = \frac{\partial \ln \Gamma(\cdot)}{\partial \cdot}$ is the digamma function.

4 Results from the Noiseless Algorithm

A popular application of ICA is in the analysis of ECG data, to illustrate the application of the noiseless algorithm in this domain we considered a data-set (Figure 1) consisting of the cutaneous potential recordings of a pregnant woman [11]. The data consists of 2500 points (5 seconds of data at 500 Hz sampling rate) from five abdominal electrodes and three thoracic electrodes. We initialised the mixing matrix $\mathbf{A}$ through principal component analysis, each ν_j was initialised as 5 and $\nu_{\min}$ was set to 2.5. We then optimised the log likelihood through using the gradients given in eqns (3) and (2) in a scaled conjugate gradient algorithm. The resulting components, ordered by their 'Gaussianity', as determined by degrees of freedom parameters, are shown in Figure 2. The first six components appear to be non-Gaussian, the last two have very high degrees of freedom and were therefore assumed to be Gaussian. These two components span a sub-space of data known as the principal sub-space. They are not uniquely determined, there is a continuum of vectors which represent the same sub-space. In principal component analysis, to obtain a unique solution,

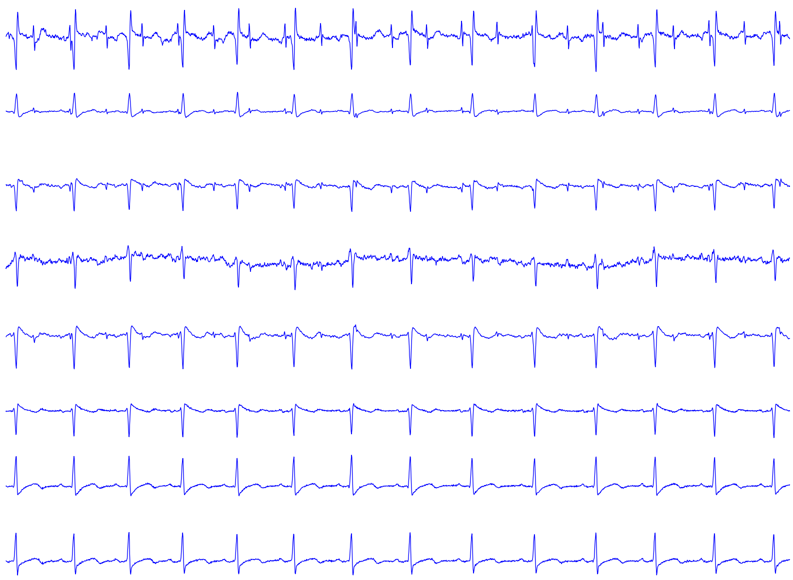

Figure 1: The five abdominal (top five plots) and three thoracic (bottom three plots) recordings.

these vectors are constrained to be orthogonal to one another. We therefore explored the principal sub-space spanned by the mixing matrix columns associated with the Gaussian components, $\mathbf{A}_{\mathrm{G}}$. We projected the sub-space spanned by this matrix onto the eigenvectors of $\mathbf{A}_{\mathrm{G}}^{\mathrm{T}}\mathbf{A}_{\mathrm{G}}$. The resulting matrix $\bar{\mathbf{A}}_{\mathrm{G}}$ was then substituted back into $\mathbf{A}$ to give $\bar{\mathbf{A}}$. The components associated with this mixing matrix are those shown in Figure 2.

Finally to show the nature of the distributions learned by the model we show the latent distribution associated with the first independent component ($\nu_1 = 2.5$) in Figure 3 and one of the latent distributions associated with a Gaussian direction ($\nu_8 = 832$).

5 Variational Inference

A constraint of the model described above is that we may not seek latent representations of the data with dimensionality $q < d$. However, by reintroducing the noise, β into our likelihood this limitation can be overcome. We are not the first to propose such a modification to the 'noiseless' algorithm. Bell and Sejnowski's original ICA algorithm was treated with versions which used a noise model firstly by [3] who proposed a Bayesian version of the algorithm and later by [9] who focused on its application in overcomplete representations.

Let us decompose the Student-t distribution into its hierarchical form which consists of a Gaussian whose precision (or inverse noise), τ, is sampled from a gamma distribution, thereby introducing an additional latent variable that represents the source distribution's precision, τ. We may write the hierarchical form of the Student-t as

$$p\left(s_n^{(j)}|\nu_j,\sigma_j\right) = \int p\left(s_n^{(j)}|\tau_n^{(j)}\right) p\left(\tau_n^{(j)}|\nu_j,\sigma_j\right) d\tau_n^{(j)} \tag{4}$$

where

$$\begin{aligned}
p\left(s_n^{(j)}|\tau_n^{(j)}\right) &= \sqrt{\frac{\tau_n^{(j)}}{2\pi}} \exp\left(-\frac{\tau_n^{(j)}}{2} s_n^{(j)\,2}\right) \\
p\left(\tau_n^{(j)}|\nu_j,\sigma_j\right) &= \frac{\left(\frac{\nu}{2}\sigma_j^2\right)^{\frac{\nu_j}{2}}}{\Gamma\left(\frac{\nu_j}{2}\right)} \tau_n^{(j)\,\frac{\nu_j}{2}-1} \exp\left(-\frac{\nu_j}{2}\sigma_j^2\tau_n^{(j)}\right).
\end{aligned}$$

Figure 2: The components uncovered by the GCA algorithm ordered according to the 'Gaussianity' of the source. The degrees of freedom parameters were 2.5, 2.5, 2.5, 3.57, 3.79, 6.06, 339 and 832. Components 1, 2, 4 and 5 appear to be associated with the mothers heartbeat, 3 and 6 arise from the fetus. The data appears to only contain six independent components, the other directions being Gaussian.

Here we have parameterised the gamma prior over the source precision such that when $\tau_n^{(j)}$ is marginalised in eqn (4) we recover eqn (1). We denote this form of parameterisation of the gamma distribution t–gam$(y|\nu, \sigma)$. Decomposing the Student-t still leaves us with an intractable problem; the advantage of this step, though, is that it renders our model amenable to the approximating 'variational inference' framework.

Decomposition of the Student-t leaves us with a model that is composed of distributions which are all members of the conjugate exponential family. This allows us to implement the machinery of variational inference without modification [8]. As a result, we are able to obtain a lower bound on the log-likelihood which may be maximised to estimate the mixing matrix.

Consider the following formulation of the log-likelihood

$$\ln p(\mathbf{X}|\theta) = \ln \frac{p(\mathbf{X}, \mathbf{S}, \mathbf{T}|\theta)}{q(\mathbf{S}, \mathbf{T})} - \ln \frac{p(\mathbf{S}, \mathbf{T}|\mathbf{X}, \theta)}{q(\mathbf{S}, \mathbf{T})},$$

where we have introduced the arbitrary distribution $q(\mathbf{S}, \mathbf{T})$ and have represented the set of all the models' parameters by θ. Taking expectations under this distribution leads to

$$\ln p(\mathbf{X}|\theta) = \int q(\mathbf{S}, \mathbf{T}) \ln \frac{p(\mathbf{X}, \mathbf{S}, \mathbf{T}|\theta)}{q(\mathbf{S}, \mathbf{T})} d\mathbf{S} d\mathbf{T} - \mathrm{KL}(q||p). \tag{5}$$

where $\mathrm{KL}(q||p)$ represents the Kullback Leibler divergence between $q(\mathbf{S}, \mathbf{T})$ and $p(\mathbf{S}, \mathbf{T}|\mathbf{X}, \theta)$. The KL divergence is known to be zero if the two distributions are identical and positive otherwise, thus the first term of eqn (5) is a lower bound on the likelihood,

$$\ln p(\mathbf{X}|\theta) \geq \int q(\mathbf{S}, \mathbf{T}) \ln \frac{p(\mathbf{X}, \mathbf{S}, \mathbf{T}|\theta)}{q(\mathbf{S}, \mathbf{T})} d\mathbf{S} \equiv \mathcal{L}(\theta), \tag{6}$$

the quality of which may be improved by minimising the KL divergence. If we consider all possible forms for the approximating distribution, we find the bound becomes equality when

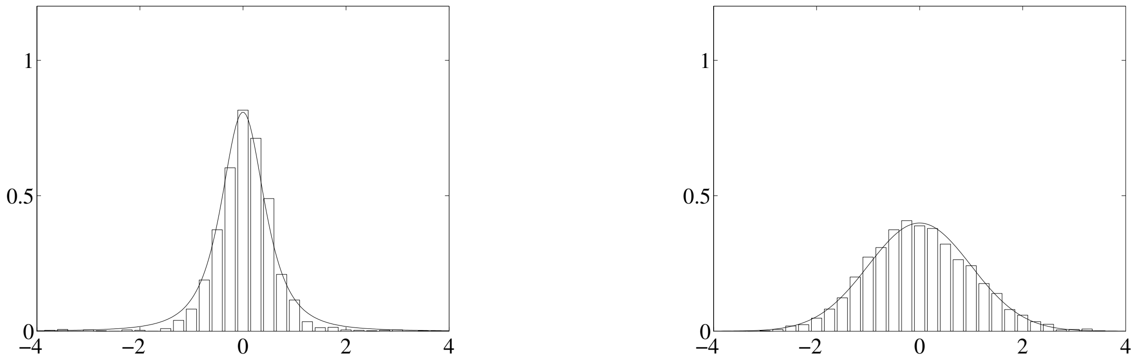

Figure 3: The Student-t distribution associated with the first independent component (left) and that associated with one of the Gaussian directions (right). Also included are histograms of the actual recovered sources for comparision.

$q(\mathbf{S},\mathbf{T}) = p(\mathbf{S},\mathbf{T}|\mathbf{X})$. However, in our model, this form of the q-distribution leads to no reduction in the computational complexity of the inference process. Instead we first of all constrain the class of our approximating distribution by assuming it separates, $q(\mathbf{S},\mathbf{T}) = q(\mathbf{S})q(\mathbf{T})$, [19]. The functional form of the approximating distributions, $q(\mathbf{S})$ and $q(\mathbf{T})$, which most tighten bound (6) can then be found [19] by inspection,

$$\begin{aligned} q(\mathbf{S}) &\propto \exp\left\langle \ln p(\mathbf{X},\mathbf{S},\mathbf{T}|\theta)\right\rangle_{q(\mathbf{T})}, \\ q(\mathbf{T}) &\propto \exp\left\langle \ln p(\mathbf{X},\mathbf{S},\mathbf{T}|\theta)\right\rangle_{q(\mathbf{S})}, \end{aligned}$$

where $\langle\cdot\rangle_{p(x)}$represents an expectation under the distribution $p(x)$. Our approximation to the posterior therefore consists of the following two distributions

$$\begin{aligned} q(\mathbf{S}) &= \prod_{n=1}^{N} \mathcal{N}\left(\mathbf{s}_n|\bar{\mathbf{s}}_n, \Sigma_s^n\right), \\ q(\mathbf{T}) &= \prod_{j=1}^{q}\prod_{n=1}^{N} t\!-\!\mathrm{gam}\left(\tau_n^{(j)}|\hat{\nu}, \hat{\sigma}_n^{(j)}\right) \end{aligned}$$

whose parameterisation may be found to be

$$\Sigma_s^n = \left[\mathrm{diag}\left\langle\tau_n\right\rangle + \beta\mathbf{A}^{\mathrm{T}}\mathbf{A}\right]^{-1}, \tag{7}$$

$$\bar{\mathbf{s}}_n = \Sigma_s^n\beta\mathbf{A}^{\mathrm{T}}\mathbf{x}_n, \tag{8}$$

$$\hat{\sigma}_n^{(j)\,2} = \frac{\nu_j\sigma_j^2 + \left\langle s_n^{(j)\,2}\right\rangle}{\nu_j+1}, \tag{9}$$

$$\hat{\nu}_j = \nu_j + 1. \tag{10}$$

Here we have dropped the subscript for the expectations where the appropriate distribution is thought to be obvious. The expectations of interest are related to the parameters in the following manner

$$\left\langle\mathbf{s}_n\right\rangle = \bar{\mathbf{s}}_n, \tag{11}$$

$$\left\langle\mathbf{s}_n\mathbf{s}_n^{\mathrm{T}}\right\rangle = \bar{\mathbf{s}}_n\bar{\mathbf{s}}_n^{\mathrm{T}} + \Sigma_s^n. \tag{12}$$

$$\left\langle\tau_n^{(j)}\right\rangle = \hat{\sigma}_n^{(j)\,-2} \tag{13}$$

$$\left\langle\ln\tau_n^{(j)}\right\rangle = \psi\left(\hat{\nu}_j\right) - \log\left(\frac{\hat{\nu}_j}{2}\right) - \log\hat{\sigma}_n^{(j)\,2}. \tag{14}$$

We may also obtain an expression for our variational bound on the log likelihood

$$\mathcal{L}(\theta) = \langle \ln p(\mathbf{X}|\mathbf{S},\mathbf{A},\beta)p(\mathbf{S}|\mathbf{T})p(\mathbf{T})\rangle + H\left(q(\mathbf{S})\right) + H\left(q(\mathbf{T})\right), \tag{15}$$

where the first term may be computed using the following results

$$\begin{aligned}
\langle \ln p(\mathbf{X}|\mathbf{S},\mathbf{A},\beta)\rangle &= \frac{Nd}{2}\ln\frac{\beta}{2\pi} - \sum_{n=1}^{N}\frac{\beta}{2}\left\langle (\mathbf{x}_n - \mathbf{A}\mathbf{s}_n)^{\mathrm{T}}(\mathbf{x}_n - \mathbf{A}\mathbf{s}_n)\right\rangle \\
\langle \ln p(\mathbf{S}|\mathbf{T})\rangle &= -\frac{Nq}{2}\ln 2\pi + \frac{1}{2}\sum_{n=1}^{N}\sum_{j=1}^{q}\left\langle \ln \tau_n^{(j)}\right\rangle - \sum_{n=1}^{N}\sum_{j=1}^{q}\frac{\left\langle \tau_n^{(j)}\right\rangle}{2}\left\langle s_n^{(j)\,2}\right\rangle \\
\langle \ln p(\mathbf{T})\rangle &= \sum_{j=1}^{q}\frac{N\nu_j}{2}\ln\frac{\nu_j\sigma_j^2}{2} - \sum_{j=1}^{q}N\ln\Gamma(\nu_j/2) \\
&\quad + \sum_{n=1}^{N}\sum_{j=1}^{q}\left(\frac{\nu_j}{2}-1\right)\left\langle \log\tau_n^{(j)}\right\rangle - \sum_{n=1}^{N}\sum_{j=1}^{q}\frac{\nu_j}{2}\sigma_j^2\left\langle \tau_n^{(j)}\right\rangle,
\end{aligned}$$

and the second two terms, which represent the entropies of the variational distributions, can be found as

$$\begin{aligned}
H\left(q(\mathbf{S})\right) &= \sum_{j=1}^{q}\left[N\ln\left\{\Gamma\left(\frac{\hat{\nu}_j}{2}\right)\right\} - N\ln\left(\frac{\hat{\nu}_j}{2}\right)\right. \\
&\quad \left. -N\left(\frac{\hat{\nu}_j}{2}-1\right)\psi\left(\frac{\hat{\nu}_j}{2}\right) + \frac{N\hat{\nu}_j}{2} - \sum_{n=1}^{N}\ln\hat{\sigma}_n^{(j)\,2}\right] \\
H\left(q(\mathbf{T})\right) &= \sum_{n=1}^{N}\left[\frac{q}{2}(\ln 2\pi + 1) + \frac{1}{2}\ln|\Sigma_s^n|\right].
\end{aligned}$$

To estimate the parameters, $\beta, \mathbf{A}, \nu, \in \theta$, we look to optimise the lower bound on the likelihood in the hope that the true likelihood will also be raised. Through differentiation of (6) with respect to β and $\mathbf{A}$ and setting the result to zero we may obtain the following fixed point equations

$$\mathbf{A} = \left[\sum_{n=1}^{N}\left\langle \mathbf{s}_n\mathbf{s}_n^{\mathrm{T}}\right\rangle\right]^{-1}\sum_{n=1}^{N}\langle \mathbf{s}_n\rangle\mathbf{x}_n^{\mathrm{T}}, \tag{16}$$

$$\beta = Nd\left(\sum_{i=1}^{d}\sum_{n=1}^{N}\left\langle\left(x_n^{(i)} - \mathbf{a}_i^{\mathrm{T}}\mathbf{s}_n\right)^2\right\rangle\right)^{-1}. \tag{17}$$

Unfortunately, for optimisation with respect to the parameters of the Student-t latent distribution, no fixed point equations may be found. Furthermore, we must consider that there is redundancy in our representation: variations in scale may be accounted for either through scaling columns of the mixing matrix or through scaling a factor of the source distribution. To remove this redundancy we first considered constraining the scale parameters $\{\sigma_j\}$ to unity, the degrees of freedom parameters may then be found through gradient based optimisations, the gradient of the likelihood bound with respect to the degrees of freedom being

$$\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\nu_j} &= \frac{\partial}{\partial\nu_j}\langle \ln p(\mathbf{T})\rangle \\
&= \frac{N}{2}\ln\frac{\nu_j}{2} + 1 - \psi\left(\frac{\nu_j}{2}\right) + \ln\sigma_j^2 + \frac{1}{2}\sum_{n=1}^{N}\left(\left\langle \ln\tau_n^{(j)}\right\rangle - \sigma_j^2\left\langle \tau_n^{(j)}\right\rangle\right)
\end{aligned}$$

and used in a scaled conjugate gradient algorithm.

In implementation, however, we found that convergence of this algorithm, as monitored by the lower bound on the log likelihood, was slow unless particular orderings of the updates were undertaken. This was considered to be as a result of any change in the degrees of freedom parameter at each source requiring corresponding changes in a column of the mixing matrix so that the predictive variances associated with mixtures of that source was conserved. We therefore implemented a constraint on the source distribution whereby its variance in each direction was fixed to one.

5.1 Whitening the Source Distribution

The variance of the each component of the source distribution is given by $\frac{\nu_j}{\nu_j - 2}\sigma_j^2$, therefore the variance of the constrained Student-t distribution described above is given by $\frac{\nu_j}{\nu_j-2}$. This means that any change in the values of ν_j can lead to a large change in the predicted variance for each output. This change must be counteracted by corresponding changes in the mixing matrix $\mathbf{A}$ and as a result convergence can be very slow. Rather than constraining σ_j^2 we therefore constrained the variance associated with the source distribution to unity. As for the noiseless case above this constraint was implemented through a Lagrange multiplier and through setting $\nu_j = \nu_{\min} + \gamma_j^2$ where $\nu_{\min} > 2$. The likelihood bound as a function of $\boldsymbol{\gamma}$ may then be written

$$
\begin{aligned}
L(\boldsymbol{\gamma}) \quad = \quad & N\sum_{j=1}^{q}\left(\frac{\nu_{\min}+\gamma_j^2}{2}\right)\ln\left[\frac{\nu_{\min}+\gamma_j^2-2}{2}\right] - N\sum_{j=1}^{q}\ln\Gamma\left(\frac{\nu_{\min}+\gamma_j^2}{2}\right) \\
& + \sum_{j=1}^{q}\left(\frac{\nu_{\min}+\gamma_j^2}{2}-1\right)\sum_{n=1}^{N}\left\langle \log\tau_n^{(j)}\right\rangle \\
& - \sum_{j=1}^{q}\frac{\nu_{\min}+\gamma_j^2-2}{2}\sum_{n=1}^{N}\left\langle \tau_n^{(j)}\right\rangle . \qquad (18)
\end{aligned}
$$

Differentiating this function with respect to γ_k gives

$$
\begin{aligned}
\frac{\partial}{\partial\gamma_k}L(\boldsymbol{\gamma}) \quad = \quad & N\gamma_k\left[\ln\left(\frac{\nu_{\min}+\gamma_k^2-2}{2}\right) + \frac{\left(\nu_{\min}+\gamma_k^2\right)}{\left(\nu_{\min}+\gamma_j^2-2\right)} - \psi\left(\frac{\nu_{\min}+\gamma_k^2}{2}\right)\right] \\
& + \gamma_k\sum_{n=1}^{N}\left[\left\langle\log\tau_n^{(k)}\right\rangle - \left\langle\tau_n^{(k)}\right\rangle\right]. \qquad (19)
\end{aligned}
$$

This equation may be used in combination with eqn (18) in a line minimiser for determining the optimum value for each γ_k.

6 Algorithm Overview

The updates of parameters and expectations given above are used together to optimise the model, unfortunately it is not clear in what order these updates should be made. The algorithm can be viewed as an approximate expectation maximisation (EM) algorithm [6]. In standard EM algorithms it is common to update the expectations (the E-step) first, eqns (11), (12), (13) and (14), to tighten the lower bound on the likelihood before an update of the parameters (the M-step) through eqns (16) and (17). We are not, however, required to stick with this convention to obtain a convergent algorithm [15]. In our implementation, which is given in Algorithm 1, we chose to perform the updates in a random order.

The computational complexity of the algorithm is dominated by the matrix inverse in eqn (8) which must be performed for each data point. Each iteration has, therefore, $O\left(Nq^3\right)$ complexity. There is, however, scope to reduce this complexity by assuming that the variational posterior factorises over each latent dimension as is done by [14].

Algorithm 1 The GCA algorithm.

Require: A centred data-set. A minimum value for ν_j.

Initialise $\mathbf{A}$ with small random values, $\beta = \frac{d}{\sum_{i=1}^{d} \mathrm{var}\left(\mathbf{x}^{(i)}\right)}$, $\nu_j = 5$ and $\sigma_j^2 = \frac{\nu_j - 2}{2}$.

Update $\Sigma_s^n$and $\bar{\mathbf{s}}_n$ using (7) and (8).

Update values of $\hat{\nu}_j$and ${\hat{\sigma}_n^{(j)}}^2$ using (10) and (9).

repeat

  repeat

    Sample without replacement one of the following updates.

    Update $\Sigma_s^n$and $\bar{\mathbf{s}}_n$ using eqn (7) and eqn (8).

    Update values of $\hat{\nu}_j$and ${\hat{\sigma}_n^{(j)}}^2$ using eqn (10) and eqn (9).

    Update β using eqn (17).

    Update $\mathbf{A}$ according to eqn (16).

    Perform a line minimisation for each γ_j using eqn (18) and eqn (19).

  until All updates have been sampled.

until Change in bound on log-likelihood is less 1×10^{-5}.

GCA	1.09	1.03	1.01	0.913	0.927	0.124	0.0626	0.0586	0.0321	0.0354
PPCA	0.999	0.999	0.999	0.995	0.995	0.135	0.0674	0.0580	0.0349	0.0348

Table 1: Comparision of the norm of the eigenvectors' projections on the sub-space of the mixing matrix found by the GCA algorithm. The exact sub-space is not recovered, this is likely to be as a result of a limited number of data points (1000) in a ten dimensional space.

7 Examples

To demonstrate the efficacy of the variational algorithm we now turn to the study of some artificial and real world examples.

7.1 Toy Problems

Let us first consider the behaviour of the variational GCA algorithm when presented with data which is Gaussian in nature, under these conditions we expect the algorithm to estimate the principal sub-space in the manner of probabilistic PCA (PPCA), [16].

7.1.1 GCA on Gaussian data

A toy Gaussian data-set was obtained by sampling 1000 data points from a ten dimensional diagonal covariance Gaussian, the elements of which took the following values 10, 9, 8, 7, 6, 5, 4, 3, 2, 1. These samples were then rotated by an orthogonal matrix $\mathbf{R}$ and modelled with a GCA model containing five latent distributions. The model was initialised and optimised as described in Section 6 above. The norms of the eigenvalues' projections on to the sub-space of the discovered mixing matrix were computed. If this matrix spanned the sub-space the results would have been five ones followed by five zeros. The actual results are summarised in Table 1, for comparison a set of PPCA results are included with the GCA results.

7.1.2 GCA on non-Gaussian data

The second toy problem involved sampling 1000 points from six Student-t distributions with the following degrees of freedom: 3, 3, 3, 100, 100 and 100. The scale parameter was set to 1 for each distribution. These samples were then subject to a randomly generated affine transform and spherical Gaussian noise with a variance of 0.01 was added. The algorithm estimated the following degrees of freedom parameters: 3.31, 3.70, 3.91, 62.2, 80.2 and 141. The poor estimation of the

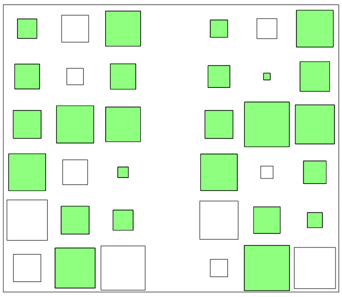

Figure 4: Hinton diagrams of the true independent components (left 3 columns) and those recovered by the algorithm (right 3 columns) for the non-Gaussian toy problem. A Hinton diagram represents the numbers in a matrix, in this case the mixing matrix, as a set of squares. The position of the squares corresponds to each number's position in the matrix and their size and colour correspond to the magnitude and sign of the number respectively .

last three degrees of freedom is typical of the algorithm, perhaps because in this regime the shape of the Student-t changes only marginally with large changes in the degrees of freedom parameters. The recovered independent components (the columns of the mixing matrix associated with the three non-Gaussian directions) are shown in Figure 4 along with the true values.

7.2 GCA on MEG data

One advantage of the probabilistic formulation of the model is that we may compute an undercomplete representation of our data. This may be useful in situations where the data is of higher dimensionality. We explored an MEG data-set [18] containing 122 signals, a subset of which are shown in Figure 5. We constrained ourselves to a twenty second portion of the data-set twenty-four seconds from the start of the sequence, during which the subject had been asked to blink. The resulting components, ordered by their 'Gaussianity' as determined by degrees of freedom parameters, are shown in Figure 6. We considered the first five components to be non-Gaussian. Their degrees of freedom parameters varied between 2.82 and 14.5. The following two components have degrees of freedom parameters of 110 and 951. As for the noiseless case we explored the principal sub-space spanned by the mixing matrix columns associated with the Gaussian components, $\mathbf{A}_{\mathrm{G}}$, projecting the sub-space onto the eigenvectors of $\mathbf{A}_{\mathrm{G}}^{\mathrm{T}}\mathbf{A}_{\mathrm{G}}$ the resulting matrix $\bar{\mathbf{A}}_{\mathrm{G}}$ again being substituted back into $\mathbf{A}$ to give $\bar{\mathbf{A}}$. The components associated with this mixing matrix are those shown in Figure 6.

Once again, to show the nature of the distributions learned by the model we show the latent distribution associated with the first independent component ($\nu_1 = 2.82$) in Figure 7 and one of the latent distributions associated with a Gaussian direction ($\nu_8 = 951$).

8 Discussion

We have presented an algorithm that automatically estimates the number of independent components in a data-set and explains the remaining data through principal component analysis and a noise model. This 'automatic independence determination' (AID) is a characteristic shared with other models, [12, 14, 4], but most of them treat the mixing matrix in a Bayesian manner to achieve this goal. The method uses a latent variable which is Student-t distributed and exploits the distribution's ability to 'interpolate' between a heavy tailed distribution and a Gaussian distribution. Recall, as noted earlier, that this distribution is appropriate for extracting independent components that are super-Gaussian, but not sub-Gaussian.

The decomposition of the Student-t distribution we use can be viewed as an infinite mixture of zero mean Gaussians. One alternative representation [2, 14, 12, 7] for the source distribution is

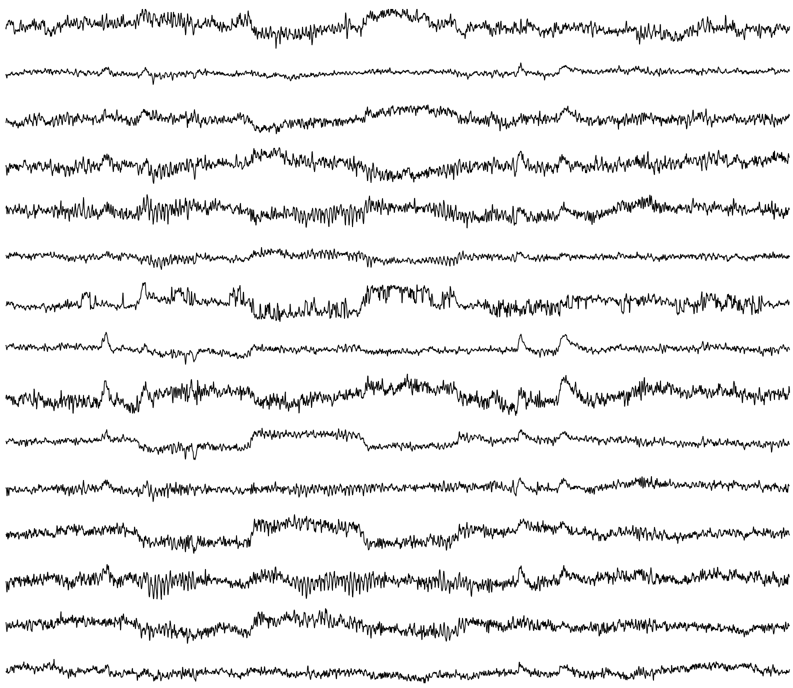

Figure 5: Fifeteen of the orginal 122 MEG signals.

that of a finite mixture of Gaussians. This type of source distribution may also be treated tractably using the variational inference techniques we used and the parameters of the source distributions can also be optimised by maximising the resulting lower bound on the log-likelihood. However, these source distributions do not have the attractive quality of the Student-t which can interpolate between the Gaussian distribution and a heavy tailed distribution. Finally, it is straightforwad to develop a Bayesian version of the algorithm, through the mechanism of variational inference, which not only extracts the number of independent components but also determines the dimensionality of the principal sub-space.

Software for running the experiments we've described is available at `http://www.dcs.shef.ac.uk/~neil/gca/`.

References

[1] S.-I. Amari, A. Cichocki, and H. H. Yang. A new learning algorithm for blind signal separation. In Touretzky et al. [17], pages 757–763.

[2] H. Attias. Independent factor analysis. *Neural Computation*, 11:803–851, 1998.

[3] H. Attias. Inferring parameters and structure of latent variable models by variational Bayes. In K. B. Laskey and H. Prade, editors, *Uncertainty in Artificial Intelligence*, volume 15, San Francisco, CA, 1999. Morgan Kauffman.

[4] R. A. Choudrey and S. J. Roberts. Flexible Bayesian independent component analysis for blind source separation. In *Proceedings of the 3rd International Conference on Independent Component Analysis and Blind Signal Separation, ICA 2001, San Diego, California*, 2001.

[5] P. Comon. Independent component analysis: a new concept? *Signal Processing*, 36:287–314, 1994.

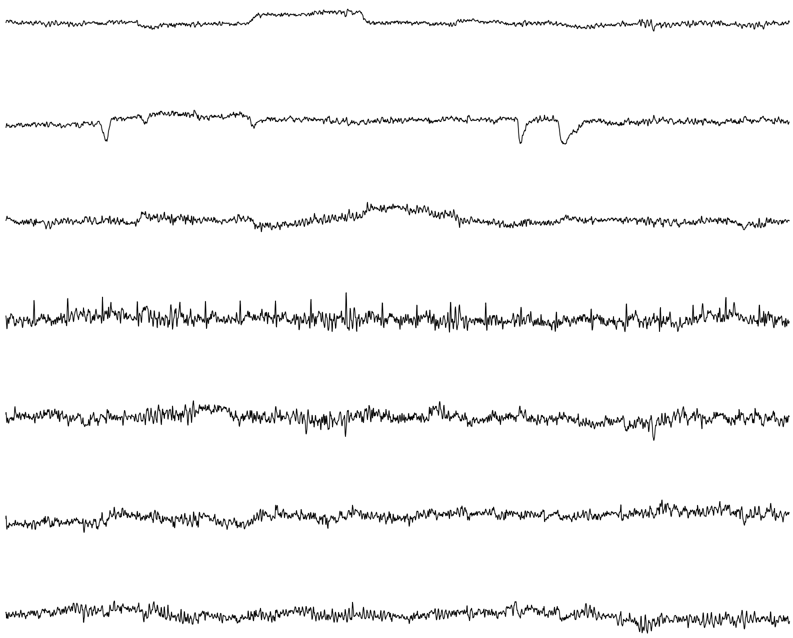

Figure 6: The 'raw' components uncovered by the GCA algorithm ordered according to the 'Gaussianity' of the source. The degrees of freedom parameters were 2.82, 3.12, 4.30, 8.95, 14.5, 110 and 951. As well as blinking artifacts a heart artifact is visible in the fourth component.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B*, 39(1):1–38, 1977.

[7] R. M. Everson and S. J. Roberts. ICA: A flexible non-linearity and decorrelating manifold approach. *Neural Computation*, 11(8):1957–1983, 1999.

[8] Z. Ghahramani and M. J. Beal. Propagation algorithms for variational Bayesian learning. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13, Cambridge, MA, 2001. MIT Press.

[9] M. Girolami. A variational method for learning sparse and overcomplete representations. *Neural Computation*, 13(11):2517–2532, 2001.

[10] A. Hyvärinen and E. Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4–5):411–430, 2000.

[11] L. D. Lathauwer, B. D. Moor, and J. Vandewalle. Fetal electrocardiogram extraction by source subspace separation. In *Proc. IEEE-SP Workshop on Higher-order Statistics (HOS'95)*, pages 134–138, 1995.

[12] N. D. Lawrence and C. M. Bishop. Variational Bayesian independent component analysis. Technical report, 2000.

[13] D. J. C. MacKay. Maximum likelihood and covariant algorithms for independent component analysis. Unpublished manuscript, available from `http://wol.ra.phy.cam.ac.uk/mackay/homepage.html`, 1996.

[14] J. W. Miskin. *Ensemble Learning for Independent Component Analysis*. PhD thesis, Selwyn College, Cambridge, 2001.

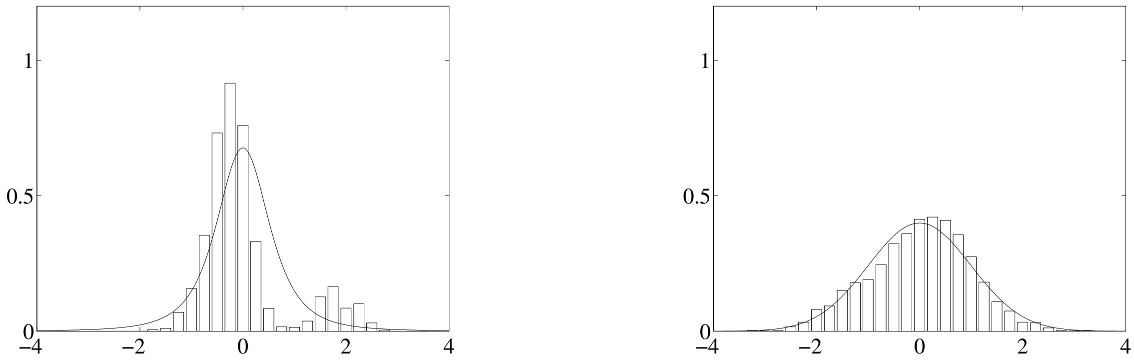

Figure 7: The Student-t distribution associated with the first independent component (left) and that associated with one of the Gaussian directions (right). Note that despite a strong mismatch between the actual source in the case of the first independent component, the Gaussian direction is still approximates the data.

[15] R. M. Neal and G. E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan, editor, *Learning in Graphical Models*, volume 89 of *Series D: Behavioural and Social Sciences*, pages 355–368, Dordrecht, The Netherlands, 1998. Kluwer.

[16] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, B*, 6(3):611–622, 1999.

[17] D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors. *Advances in Neural Information Processing Systems*, volume 8, Cambridge, MA, 1996. MIT Press.

[18] R. N. Vigário, V. Jousmäki, M. Hämäläinen, R. Hari, and E. Oja. Independent component analysis for identification of artifacts in magnetoencephalographic recordings. In M. I. Jordan, M. J. Kearns, and S. A. Solla, editors, *Advances in Neural Information Processing Systems*, volume 10, pages 229–235, Cambridge, MA, 1998. MIT Press.

[19] S. Waterhouse, D. J. C. MacKay, and T. Robinson. Bayesian methods for mixtures of experts. In Touretzky et al. [17], pages 351–357.

A Constraining the variance of the latent distribution

In probabilistic principal component analysis, the latent distribution is of fixed variance. Allowing the variance of the latent distribution to change leads to a redundancy in the models representation. It therefore seems appropriate to constrain the variance of the latent distribution in our model. We thus implement the constraint $\frac{\nu_j}{\nu_j-2}\sigma_j^2 = 1$. When maximising log likelihoods, or lower bounds on the log likelihoods, with respect to the parameters of the source distribution we develop a Lagrangian which incorporates this constraint:

$$L\left(\boldsymbol{\nu}, \boldsymbol{\sigma}^2, \boldsymbol{\lambda}\right) = F\left(\boldsymbol{\nu}, \boldsymbol{\sigma}^2, \boldsymbol{\lambda}\right) + \lambda_j \left(\frac{\nu_j}{\nu_j - 2}\sigma_j^2 - 1\right), \tag{20}$$

where $F(\cdot)$ is the original function to be optimised and we have introduced Lagrange multipliers λ_j. Differentiating with respect to λ_k we obtain

$$\frac{\partial}{\partial \lambda_k} L\left(\boldsymbol{\nu}, \boldsymbol{\sigma}^2, \boldsymbol{\lambda}\right) = \left(\frac{\nu_k}{\nu_k - 2}\right)\sigma_k^2 - 1.$$

implying $\sigma_j^2 = \frac{\nu_j-2}{\nu_j}$which we may substitute back into our function $F(\cdot)$ to impose the constraint.