Independent component analysis: An approach to clustering.

Jamal B. Bugrien* and John T. Kent

University of Leeds

1 Introduction

Independent component analysis (ICA) (Hyvarinen *et al.*, 2001), and projection pursuit (PP) (Jones and Sibson, 1987), are closely related techniques, which try to look for "interesting" directions (projections) in the data. ICA assumes a model, $\mathbf{x} = \mathbf{A}\mathbf{s}$, where $\mathbf{x}$ is a vector of observed random variables, $\mathbf{A}$ is a $(d \times d)$ "mixing" matrix, and $\mathbf{s}$ is a vector of independent latent variables. The task then is to find $\mathbf{A}$ to recover $\mathbf{s}$. A key assumption is usually that the $\{s_j\}$ have different kurtosises $\{\kappa_j\}$, in order to separate the different independent components. In practice ICA usually measures "interestingness" of a linear combination $\mathbf{a}^T\mathbf{x}$ in terms of the size of its absolute kurtosis $|\kappa|$ or some related measures. Since for a Gaussian random variables the kurtosis is zero, this criterion measures to some extent, non-Gaussianity.

In this poster, we are interested in finding a clustering procedure that can be applied for exploratory analysis in large datasets. One specific use for ICA is to pick out clusters from multi-dimensional data via projection. The objective is to find one or more "interesting" directions. It turns out that in the clustering direction, kurtosis is usually negative. Hence it is useful in practice to use a modified version of ICA in which we minimise $\kappa(\mathbf{a}^T\mathbf{x})$ rather than maximise $|\kappa(\mathbf{a}^T\mathbf{x})|$. In this approach, the one-dimensional projection of the multi-dimensional data is used to provide the most interesting view for clustering from the full-dimensional data. However, this approach based on kurtosis is not robust to outliers, and hence a variety of other robust alternative approaches have been suggested (e.g. negentropy and general contrast functions). Then, by means of "density based" approaches to clusterings (e.g. kernel density estimation clustering (Silverman, 1986) or scale-space clustering (Lindeberg, 1994)) we can plot and explore the one-dimensional projected data, to obtain potential clusters.

2 Some properties of kurtosis

We look at some properties of kurtosis arise from the ICA model. First, let us assume some necessary notations: If $\mathbf{x}$ is a random variable with mean μ and variance σ^2, then define the fourth cumulant (non-standardised kurtosis) by $\tilde{\kappa}(x) = E[(x-\mu)^4] - 3\big(E[(x-\mu)^2]\big)^2$, and the standardised kurtosis by

$$\kappa(x) = \frac{\tilde{\kappa}(x)}{\big(E[(x-\mu)^2]\big)^2} = \frac{E[(x-\mu)^4]}{\big(E[(x-\mu)^2]\big)^2} - 3\,. \tag{2.1}$$

We can distinguish three cases:

$\kappa = 0$ for Gaussian random variables.

$\kappa > 0$ called super-Gaussian.

$\kappa < 0$ called sub-Gaussian.

In turn this leads us to define three versions of ICA: "super-ICA" (maximise κ), "standard-ICA" (maximise $|\kappa|$) and "sub-ICA" (minimise κ). Since clustering is typically associated with

"sub-Gaussianity" ($\kappa < 0$), our suggested method for cluster finding is to look for a linear combination $\mathbf{a}^T\mathbf{x}$ with minimal kurtosis.

To formalise this intuition, we first need to introduce some definitions and arguments: Recall the ICA model for a centred and whitened random variable $\mathbf{y}$ takes the form $\mathbf{y} = \mathbf{As}$, where $E(y_i) = 0$, var$(y_i) = 1$, cov$(y_i, y_j) = 0$, $E(s_i) = 0$, var$(s_i) = 1$ and the s_i's are independent, and define $\kappa_i = \kappa(s_i)$. It can be seen that $\mathbf{A}$ is orthogonal, $\mathbf{A}^T = \mathbf{A}^{-1} = \mathbf{B}$, say, so that $\mathbf{s} = \mathbf{By}$. The goal is then to find $\mathbf{B}$ from $\mathbf{y}$. To solve the problem, we assume the kurtosises κ_i are distinct, and without loss of generality they can be in descending order $\kappa_1 > \ldots > \kappa_d$. To begin with we find the first row of $\mathbf{B}$, $\mathbf{b}$ say, where $\mathbf{b}^T\mathbf{b} = 1$. For theoretical purposes, transform $\mathbf{b}$ to $\boldsymbol{\gamma} = \mathbf{A}^T\mathbf{b}$. Then using the ICA model $\mathbf{y} = \mathbf{As}$ and the additivity and scaling properties of kurtosis, the kurtosis $\kappa(\mathbf{b}^T\mathbf{y})$ takes the form

$$\kappa(\mathbf{b}^T\mathbf{y}) = \kappa(\mathbf{b}^T\mathbf{As}) = \kappa(\boldsymbol{\gamma}^T\mathbf{s}) = \kappa\left(\sum_{i=1}^{d}\gamma_i s_i\right) = \sum_{i=1}^{d}\gamma_i^4\kappa_i \tag{2.2}$$

Since $\mathbf{b}$ is standardised and $\mathbf{A}$ is orthogonal, $\boldsymbol{\gamma}$ is also standardised, $\boldsymbol{\gamma}^T\boldsymbol{\gamma} = 1$.

The following three lemmas can be proved to justify the three versions of ICA.

Lemma 1: If $\kappa_1 > 0$, and $\kappa_1 > \kappa_2 \geq \ldots \geq \kappa_d$, then $\max\big(\kappa(\boldsymbol{\gamma}^T\mathbf{s})\big) = \kappa_1$, and is attained for $\boldsymbol{\gamma} = \mathbf{e}_1 = [1, 0, \ldots, 0]^T$ (i.e. maximising over $\boldsymbol{\gamma}$ such that $\boldsymbol{\gamma}^T\boldsymbol{\gamma} = 1$).

Lemma 2: If the absolute kurtosises $|\kappa|$'s are distinct and satisfy $|\kappa_1| > |\kappa_2| \geq \ldots \geq |\kappa_d|$, then $\max\big(|\kappa(\boldsymbol{\gamma}^T\mathbf{s})|\big) = \kappa_1$, and is attained for $\boldsymbol{\gamma} = \mathbf{e}_1 = [1, 0, \ldots, 0]^T$.

Lemma 3: If $\kappa_1 \geq \ldots \geq \kappa_{d-1} > \kappa_d$ and $\kappa_d < 0$, then $\min\big(\kappa(\boldsymbol{\gamma}^T\mathbf{s})\big) = \kappa_d$, and is attained for $\boldsymbol{\gamma} = \mathbf{e}_d = [0, \ldots, 0, 1]^T$.

The signs of κ_1 and κ_d are critical to the results of Lemma 1 and Lemma 3:

1. From Lemma 1, if $\kappa_1 > 0$ then one-step super-ICA picks out IC$_1$ (first independent component, s_1).

2. From Lemma 3, if $\kappa_d < 0$ then one-step sub-ICA picks out IC$_1$ (last independent component, s_d).

For clustering we are interested in the latter case (2) where typically $\kappa_d < 0$ in the clustering direction. Then (one-step) sub-ICA will pick out the clustering direction.

3 Sub-ICA clustering in practice

Suppose that we initially have an $(n \times d)$ data matrix $\mathbf{X}$ and we are interested in finding the optimal direction to identify clusters from this given data matrix $\mathbf{X}$. Then in practice to find the optimal linear combination we proceed as follows:

1. Centre and whiten the given data matrix $\mathbf{X}$ to find the whitened data matrix $\mathbf{Y}$.

2. Using kurtosis in (2.1), define an objective function of the form

$$\Phi(\mathbf{b}) = \frac{\sum_{i=1}^{n}(\mathbf{Y}_i\mathbf{b})^4/n}{(\mathbf{b}^T\mathbf{b})^2} - 3 + (\mathbf{b}^T\mathbf{b} - 1)^2\,, \tag{3.3}$$

$$\text{Objective function} = \quad \text{Kurtosis} \quad + \quad \text{Penalty} \quad .$$

3. Minimise (3.3) numerically using a black box procedure in **R**. Repeat the procedure from several starting points in an attempt to find the global minimum.

4. For simplicity we allow $\mathbf{b} \in \mathbb{R}^d$ to be unconstrained. However, due to the homogeneity of the first term and to the penalty in the second term, the optimal $\mathbf{b}$ will satisfy the constraint $\mathbf{b}^T\mathbf{b} = 1$.

4 An example of the sub-ICA algorithm

To illustrate this proposed sub-ICA clustering algorithm, we consider the Fisher's iris data. The results show that the algorithm works successfully to pick out the correct clustering direction.

The classical iris dataset first used by Fisher, (1936) consists of measurements made on 50 observations from each of three species (one quite different from the other two) of iris flowers. Four measurements were made on each flower, and there were 150 flowers in the entire dataset. We applied the sub-ICA clustering algorithm proposed above, to the entire four dimensional dataset. The result for the obtained linear combination at smoothing parameter $t = 0.1$ is shown in Figure 1. As can be seen; the density curve shows clear separation into two well defined clusters of 50 (one species) and 100 (two unresolved species) points.

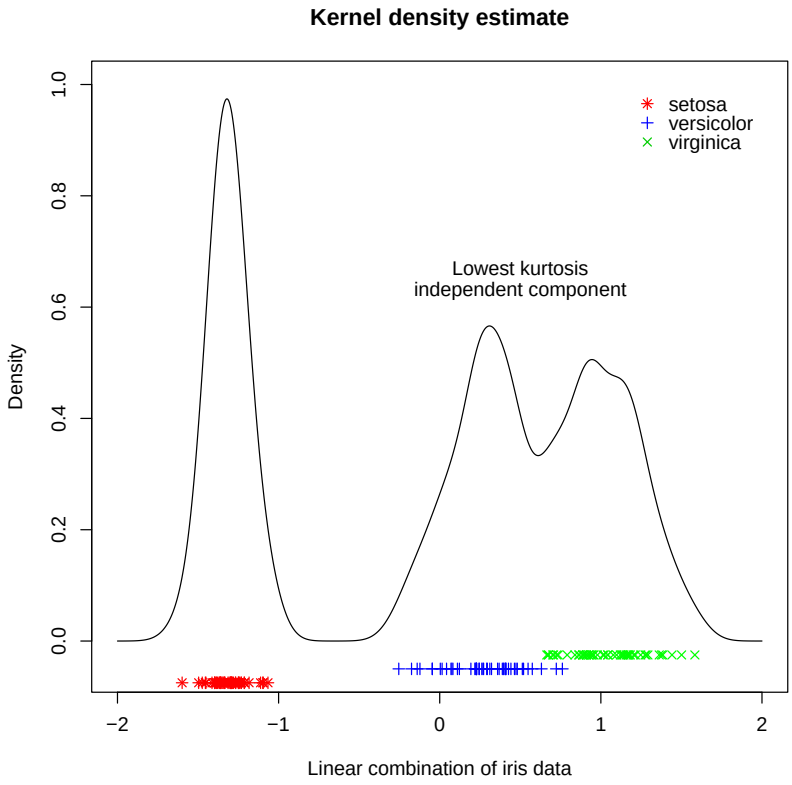

Figure 1: Plot of density estimates for the linear combination with minimal kurtosis at scale parameter $t = 0.1$, one separated cluster "setosa" and two overlapped clusters "versicolor and virginica".

There are also other more "robust" versions of the sub-ICA clustering algorithm can be based on more general contrast functions (see Hyvarinen *et al.*, 2001). These have also been applied to the four-dimensional iris data for comparison. Similar results to the linear combination displayed in Figure 1 are obtained.

5 Conclusions

In the bases of this and other examples we have shown that sub-ICA is a powerful method to pick out clusters. It is hoped that future work will benefit from and further enhance the thrust of this research.

References

Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, **7**, 179-188.

Hyvarinen, A. and Karhunen, J. and Oja, E. (2001). *Independent Component Analysis*. Wiley, New York.

Jones, M. C. and Sibson, R. (1987). What is projection pursuit? *Journal of Royal Statistical Society*, **A**, 1-36.

Lindeberg, T. (1994). *Scale-space Theory in Computer Vision*. Kluwer Academic Publishers, Dordrecht, Netherlands.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.