Algorithms for Exploiting Negative Correlation

Simon Durrant[†] Keith Kendrick[‡] Jianfeng Feng[†]
[†]Department of Informatics, Sussex University
Brighton BN1 9QH, UK
[‡]The Laboratory of Cognitive and Behavioural Neuroscience
The Babraham Institue, Cambridge CB2 4AT, UK

Abstract

Negative correlation has clear statistical benefits for noise reduction and data representation. This paper describes two new algorithms, negatively-correlated component analysis (NCCA) and negatively-correlated basis analysis (NCBA), which are designed to exploit the benefits of negative correlation. They build on the existing ICA approach, which can be seen as a special of these two algorithms. Examples of both algorithms are given, demonstrating their usefulness and superior performance to existing ICA algorithms.

1 Overview

1.1 Introduction

Since the development of algorithms for Independent Component Analysis (ICA) little more than a decade ago [1], it has seen a number of applications in neuroscience, most notably in cases of blind source separation, and in models which relate natural image statistics to the properties of the early visual system [2, 3].

In ICA, components are by definition assumed to be statistically independent, that is:-

$$\mathrm{E}\{g(x)h(y)\} = \mathrm{E}\{g(x)\}\mathrm{E}\{h(y)\} \quad (1)$$

This also means that components must be uncorrelated, since this is a weaker

condition subsumed by independence, where h(x) and g(x) are simply identity functions:-

$$\mathrm{E}\{xy\} \quad = \quad \mathrm{E}\{x\}\mathrm{E}\{y\} \tag{2}$$

Whilst this can be a useful working assumption, independence is often statistically not the optimal condition for a set of variables. In particular, negatively-correlated variables have some properties which make them preferable to positively-correlated and independent variables, especially in cases of noisy systems where negative correlation can help reduce the noise [4] and increase the storage capacity (space filling).

1.2 Benefits of Negative Correlation

Two specific benefits of negative correlation are demonstrated here. The first, shown in figure 1, is that negatively-correlated gaussian noise will tend to reduce to zero more rapidly than independent and positively-correlated gaussian noise as the number of instances of this noise increase. The utility of this result is shown in figure 2, where gaussian noise is added to a number of replications of the same image. Where the noise is positively-correlated the image is almost completely obscured. Independent noise also results in an image in which the detail has been completely lost to the noise. Only when the noise (of the same strength as the previous two cases) is negatively-correlated, does the image emerge, as the noise effectively cancels itself out.

[FIGURE 1 ABOUT HERE]

[FIGURE 2 ABOUT HERE]

The second benefit of negatively-correlated variables is their ability to fill a space better than positively-correlated or independent variables, because of their tendency to push each other away. Figure 3 shows a collection of data, some positively-correlated elements, which will be called basis functions, some independent basis functions, and some negatively-correlated basis functions. It is immediately obvious from the figure that the negatively-correlated basis functions are more evenly distributed throughout the space than the independent basis functions, which in turn are more evenly distributed than the positively-correlated basis functions. The specific benefit of this space-filling is that when the data need to be expressed in terms of the basis functions (as is the case in a typical linear model, including the ICA model), then the

residual error is minimised when the basis functions are negatively correlated, and maximised when they are positively correlated, provided that the coefficients are restricted to non-negative values. The need for this restriction arises because where basis functions are negatively-correlated with each other, by definition the most negatively-correlated function for basis function **a**, will be simply -**a**. It is clear from this that if coefficients to basis function **a** are allowed to take negative values, then there is no meaningful distinction between **a** and -**a** in the model, which effectively means that whenever **a** is present, -**a** is also implied as present. This means that a set of independent components will best cover the space under these circumstances, with their implied negatively-correlated components also being present; actual positively- or negatively-correlated basis functions under these circumstances will be to some extent redundant and suboptimal. However, where only non-negative coefficients are allowed, basis function **a** no longer implies -**a** as well, meaning that there is now a benefit to having real negatively-correlated basis functions, as the implies ones are no longer present. This non-negative coefficient restriction is increasingly popular in more recent work [5, 6] for other, principled, reasons, such as the fact that natural quantities cannot be negative, images cannot consist of negative amounts of constituent objects, neural firing rates cannot be negative etc., and so the non-negative constraint should not be regarded as a weakness of the existing approach.

[FIGURE 3 ABOUT HERE]

Given these benefits to negative correlation, it may be expected that some evolved systems would exploit this fact. Recently, it has been shown that negative correlation between neural firing of different neurons is leads to a decrease in the noise of the signal (not surprising in view of the demonstrations given earlier in this section), which therefore offers enhanced performance in a stochastic system [7]. The authors also found evidence confirming the existence of this negative correlation in in vivo experiments on the rat's olfactory bulb.

1.3 Two different approaches for negative correlation and ICA

We saw earlier that the basic ICA model, $\mathbf{X} = \mathbf{AS}$, therefore has two quite different sets of variables to estimate: the components themselves, which form the matrix $\mathbf{S}$, and the basis functions, which together form the mixing matrix $\mathbf{A}$. Both of these are candidates for negative correlation, and as such two complementary algorithms, negatively-correlated component anal-

ysis (NCCA) and negatively-correlated basis analysis (NCBA), have been developed and are presented in this paper. Both of them are generalisations of ICA, which can be seen as special case of either algorithm.

The next section will outline the theoretical framework for NCCA and NCBA, show the basic steps in the implementation of it, and highlight the benefits of the specific approach taken here. Following that is a section containing some simple examples demonstrating NCCA's ability to recover negatively-correlated components, and then a section showing examples of how NCBA takes full advantage of the benefits of negative correlation.

2 Algorithm

2.1 General Form

Both NCCA and NCBA use the same fundamental approach, which is to have an ICA core to find a set of components or basis functions, along with a lagrangian penalty term to encourage those components or basis functions to be negatively correlated.
We therefore start with the basic ICA model:-

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{3}$$
$$\mathbf{y} = \mathbf{W}\mathbf{x} \tag{4}$$

($\mathbf{x}$ are the mixed components, $\mathbf{A}$ is the mixing matrix, $\mathbf{S}$ are the original source components, $\mathbf{y}$ are the recovered source components, $\mathbf{W}$ is the demixing matrix.)

As stated above, we can find negatively-correlated components or basis functions by maximising independence, as under ICA, with an additional constraint to minimise the correlations (maximise the negative of the correlations) of the recovered components or basis functions using the technique of Lagrange multipliers.

The difference between the marginal distributions $f_{y_i}(y_i, \mathbf{W})$ and the joint distribution $f_{\mathbf{y}}(\mathbf{y}, \mathbf{W})$ of the independent components $\mathbf{y}$, can be expressed as the difference between the marginal differential entropies $\sum_{i}^{m} H(y_i)$ and the joint differential entropy $H(\mathbf{y})$ of these components. This in turn can be given by the Kullback-Leibler (K-L) divergence:-

$$
\begin{aligned}
D_{f\|\tilde{f}}(\mathbf{W}) &= \sum_{i}^{m} H(y_i) - H(\mathbf{y}) && (5)\\
H(\mathbf{y}) &= H(\mathbf{W}\mathbf{x}) = H(\mathbf{x}) + \log \mid \det(\mathbf{W}) \mid && (6)\\
\Rightarrow D_{f\|\tilde{f}}(\mathbf{W}) &= \sum_{i}^{m} H(y_i) - H(\mathbf{x}) - \log \mid \det(\mathbf{W}) \mid && (7)
\end{aligned}
$$

The correlation penalty term, including the different versions for the two different algorithms, will be outlined in the next section. For now, it will be represented by the lagrangian placeholder function $F(\mathbf{W})$, and the standard lagrangian coefficient, λ. This gives us the correlation penalty term:-

$$\lambda(F(\mathbf{W})) \tag{8}$$

This can be added to the K-L divergence to give a complete objective function to be minimised:-

$$D_{f\|\tilde{f}}(\mathbf{W}) = \sum_{i}^{m} H(y_i) - H(\mathbf{x}) - \log \mid \det(\mathbf{W}) \mid + \lambda(F(\mathbf{W})) \tag{9}$$

It is important to note that in an algorithmic implementation, this function is equivalent to the following two-step procedure:

$$
\begin{aligned}
\bar{D}_{f\|\tilde{f}}(\mathbf{W}) &= \sum_{i}^{m} H(y_i) - H(\mathbf{x}) - \log \mid \det(\mathbf{W}) \mid && (10)\\
D_{f\|\tilde{f}}(\mathbf{W}) &= \bar{D}_{f\|\tilde{f}}(\mathbf{W}) + \lambda(F(\mathbf{W})) && (11)
\end{aligned}
$$

This means that the standard K-L divergence function can be calculated in the first step, and the negative correlation penalty term can be applied in the second step, without the optimisation technique employed for both steps having to be the same. The result of this is that existing algorithms for the ICA core can be imported without any significant modification for the first step, and a simple gradient approach used to reduce the correlation between

the derived components or basis functions for the second step.

The activation functions forming the update steps in an iterative algorithm for the two equations above can be formed by taking the gradient of the objective function with respect to the demixing matrix $\mathbf{W}$. For the various terms in the equations, this gradient is computed as follows:-

$\sum_{i}^{m} H(y_i)$ The marginal distributions are the most problematic, as the formation of the gradient requires a parametric estimation of the distributions. This can be achieved with reasonable accuracy using the Gram-Charlier expansion. However, ICA algorithms typically take advantage of a computationally much simpler approximation, where the objective function is simply given by an appropriate nonquadratic function. The most popular specific choice is $\log(\cosh(\mathbf{Wx}))$, which yields $\mathbf{x}\tanh(\mathbf{Wx})$ as the derivative term; more generally, the derivative is $\mathbf{x}\varphi(\mathbf{Wx})$.

$H(\mathbf{x})$ The first of the two terms which together make up the joint distribution is a function only of the mixture variables $\mathbf{x}$, which means that this term is a constant, not dependent on $\mathbf{W}$. It thus drops out of the gradient altogether.

$\log \mid \det(\mathbf{W}) \mid$ The second of the joint distribution terms clearly is dependent on $\mathbf{W}$. Some fixed-point ICA algorithms also drop this term, by pre-whitening the data (thus assuming zero correlation), which results in this term also being a constant. However, this is clearly not appropriate for negatively-correlated component analysis, and so the gradient of this term must be included. This is given by $\mathbf{W}^{-T}$ (the inverse transpose of the demixing matrix).

$\lambda(F(\mathbf{W})$ The abstract form of the correlation penalty term has a similarly abstract gradient: $\lambda\frac{dF(\mathbf{W})}{d\mathbf{W}}$. The detailed form of these functions is outlined in the next section.

Putting these gradient terms together, we have the complete gradient activation functions to be used in the iterative algorithm:-

$$\frac{d\bar{D}_{f\|\tilde{f}}(\mathbf{W})}{d\mathbf{W}} = \mathbf{x}\varphi(\mathbf{W}\mathbf{x}) - \mathbf{W}^{-T} \quad (12)$$

$$\frac{dD_{f\|\tilde{f}}(\mathbf{W})}{d\mathbf{W}} = \frac{d\bar{D}_{f\|\tilde{f}}(\mathbf{W})}{d\mathbf{W}} + \lambda\frac{dF(\mathbf{W})}{d\mathbf{W}} \quad (13)$$

This finally gives us iterative update steps for estimating the demixing matrix $\mathbf{W}$ based on maximising the negative gradient:-

$$\Delta\bar{\mathbf{W}} = \eta[\mathbf{W}^{-T} - \mathbf{x}\varphi(\mathbf{W}\mathbf{x}) \quad (14)$$

$$\Delta\mathbf{W} = \Delta\bar{\mathbf{W}} - \lambda\frac{dF(\mathbf{W})}{d\mathbf{W}} \quad (15)$$

These provide the central weight update steps in the most general form. Specific implementation involved the use of a chosen existing ICA technique for the first update step; several have been tested for use with the algorithms presented here, including a simple generic gradient method developed for testing these algorithms, the Bell-Sejnowski algorithm [8], Amari's natural gradient version of the Bell-Sejnowski algorithm [9], and Hyvärinen's FastICA algorithm [10], with the important caveat that the orthogonalisation step in a whitened domain must be removed (in order to allow components to be correlated at all), when the implementation and testing of the algorithms is described in more detail.

Specific implementation of the second update step involves a choice of negative correlation penalty function $F(\mathbf{W})$, and a method for optimising this function with respect to $\mathbf{W}$. This is the subject of the next section.

2.2 Negative Correlation Penalty Function

In order to encourage the derived components or basis functions to be negatively correlated, it is clearly necessary to minimise a function which measures the correlation between the components. As correlation is represented by the correlation matrix, minimisation of this matrix is the obvious choice. However, the function is not quite as straightforward as this, because the correlation matrix is a matrix of variables, and the gradient of this is a third-order (three-dimensional) matrix, whereas what is actually needed for the

update steps is a vector-values function of variables (which is in practice a matrix where the variables are represented by a vector of samples; this gives us a matrix to update either A or S in the ICA model, depending on which algorithm we are using). The reason for this discrepancy is that the correlation matrix really represents a matrix of separate correlation measures, rather than the single correlation measure that we need to minimise. The solution for this is to actually sum the elements of the correlation matrix. Hence for variables $\mathbf{a}$, (in practice represented by matrix of samples $\mathbf{A}$), the elements of the correlation matrix $\mathbf{a}\mathbf{a}^T$ are summed together. This provides the function which can be minimised with respect to each of the variables in $\mathbf{a}$, giving a vector-valued function which minimised the overall correlation of $\mathbf{a}$. A simple two variable example is given as follows:-

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \tag{16}$$

$$\mathbf{a}\mathbf{a}^T = \begin{bmatrix} a_1a_1 & a_1a_2 \\ a_2a_1 & a_2a_2 \end{bmatrix} \tag{17}$$

$$F(\mathbf{a}) = \sum \mathbf{a}\mathbf{a}^T = a_1a_1 + a_1a_2 + a_2a_1 + a_2a_2 \tag{18}$$

$$\frac{dF(\mathbf{a})}{\mathbf{a}} = \begin{bmatrix} a_1 + 2a_2 \\ a_2 + 2a_1 \end{bmatrix} \tag{19}$$

This gives us a vector (in practice a sample matrix) which can be used as the lagrangian penalty term, and results in the derived variables being more negatively correlated than would otherwise be the case. This gradient approach is completely stable given an appropriate learning rate, in common with other simple gradient algorithms. In the general case, the result is as follows:-

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_i \end{bmatrix} \tag{20}$$

$$\mathbf{a}\mathbf{a}^T = \begin{bmatrix} a_1a_1 & a_1a_2 & \dots & a_1a_i \\ a_2a_1 & a_2a_2 & \dots & a_2a_i \\ \vdots & \vdots & \ddots & \vdots \\ a_ia_1 & a_ia_2 & \dots & a_ia_i \end{bmatrix} \tag{21}$$

$$F(\mathbf{a}) = \sum \mathbf{a}\mathbf{a}^T = a_1a_1 + a_1a_2 + a_2a_1 + \ldots + a_ia_i \tag{22}$$

$$\frac{dF(\mathbf{a})}{\mathbf{a}} = \begin{bmatrix} 2\sum a_i \\ 2\sum a_i \\ \vdots \\ 2\sum a_i \end{bmatrix} - \mathbf{a} \tag{23}$$

This general gradient algorithm for reducing the correlation between a set of variables is used in both NCBA and NCCA. For NCBA, the variables whose correlation is to be minimised are the set of basis functions, which are the columns of the mixing matrix $\mathbf{A}$, which therefore means $\mathbf{A}^T$ gives the sampled variables in rows to be used in the above formulae. For NCCA, the components which are the rows of $\mathbf{S}$ are the variables for use.

One further trick is required in order to employ this gradient approach as the second step in our NCBA and NCCA algorithms. These algorithms, as seen in the earlier equations, require the update steps to be for the separating matrix $\mathbf{W}$ (although in practice, some ICA algorithms use the mixing matrix $\mathbf{A}$ in their update step). We therefore need to be able to give the negative correlation step update matrix, expressed above in terms of either $\mathbf{A}^T$ for NCBA or $\mathbf{S}$ for NCCA, in terms of $\mathbf{W}$. To do this we can employ another trick, noting the fact that the least-squares error inverse for a non-square matrix $\mathbf{A}^T$ is given by the Moore-Penrose pseudoinverse, $(\mathbf{A}^T)^+$. This therefore gives us the best esimate of $\mathbf{W}$ to be used directly in the update step, and has the added benefit of being simple and relatively efficient to calculate. Conversely, in order to first enter the $\mathbf{A}$ domain in order to calculate the gradient update step, the pseuodinverse can be used in the other direction, on the $\mathbf{W}$ separating matrix yielded by the first update step. We thus have a translation from $\mathbf{W}$ into $\mathbf{A}$ for the gradient update step, and then back again into $\mathbf{W}$ to yield the final updated $\mathbf{W}$ matrix for the current iterative pass. It should be noted that it is not possible to perform the gradient update directly in the $\mathbf{W}$ domain because minimising the correlation of $\mathbf{A}^T$ is equivalent to maximising the correlation of $\mathbf{W}$, which is unstable because the fixed point of perfect correlation does not invert to perfect negative correlation back in the $\mathbf{A}$ domain.

For the NCCA algorithm, the update translation operations are slightly different. Given the $\mathbf{W}$ matrix from the first update step, the components $\mathbf{S}$ can be easily calculated by noting that $\mathbf{S} = \mathbf{WX}$. Once the negative correlation update has been calculated in the $\mathbf{S}$ domain, the conversion back to the $\mathbf{W}$ domain is given by another simple calculation: $\mathbf{W} = \mathbf{SX}^+$. It should be noted that where S is constrained to be non-negative, which is not the inherently the case for NCCA as it is for NCBA but may be adopted for some purposes nonetheless, the calculation of S is more complex, and typi-

cally found using a constrained optimisation technique, which will generally be much slower than the methods given here.

2.3 Benefits of the Current Approach

By adopting a two-step update procedure, where the separating matrix is first calculated using an ICA update step, and then the resultant components or basis functions are made more negatively correlated using the new gradient step given in the previous section, there are a number of particular benefits:-

- The two update steps do not need to use the same, gradient-based optimisation procedure. This is especially important as the negative correlation gradient algorithm is not stable in the **W** domain in which ICA update steps typically operate.
- By having the ICA update step separately, existing ICA update steps can be used with almost no modification required. The existing algorithms do not even have to use a gradient optimisation approach to be usable; multiplicative or quadratic programming algorithms can also be used. The only constraint on ICA algorithm is that it must not contain an orthogonalisation step (so the FastICA algorithm is excluded or at least has to have that step removed, which can make it unstable for large problems). This is obviously necessary in order to allow components to be any other than uncorrelated.
- By translating into the **A** domain for NCBA and the **S** domain for NCCA where necessary, the ICA step can operate in either the **A** or the **W** domain, and still be compatible with these algorithms.
- Existing ICA algorithms do not need to estimate the components **S** in order to be used with these algorithms, although obviously ones that do are also compatible.
- By using a separate negative correlation update step, the effect of the negative correlation penalty term is both easy to asses, and easy to control through the strength of the parameter λ.
- The separate negative correlation update step ensures the stability of the algorithm, as the stability of existing ICA steps is not altered within the first update step, and the second update step also has guaranteed stability for a sufficiently low learning rate.

It can be seen that the current algorithms are an extension of, and in a real sense a generalisation of, ICA, combining the benefits of existing ICA algorithms with the the benefits of negative correlation. The following two sections give some brief demonstrations of how these combined benefits allow these two new algorithms to outperform ICA.

3 Examples of NCCA

The examples in this section show the NCCA algorithm in operation. As NCCA is designed to find negatively-correlated components, the demonstrations here focus on its ability to accurately recover source signals that are negatively-correlated. Its performance is contrasted with that of ICA on the same tasks.

3.1 Example 1: Basic performance

The first example (figure 4) clearly demonstrates the most important feature of NCCA - its superior ability to recover the original, negatively-correlated signals. While ICA has recovered signals that remain quite significantly mixed, and are not the original source signals, NCCA has successfully recovered the original, negatively-correlated source signals to a much greater extent.

[FIGURE 4 ABOUT HERE]

3.2 Example 2: ICA recovers original independent signals, NCCA recovers negatively-correlated source signals

The example here (figure 5) visibly demonstrates the difference between the independence goal of ICA and the negatively-correlated components goal of NCCA. The negatively-correlated source signals are recovered by NCCA, whilst ICA recovers independent signals. The signals recovered by ICA are actually closely related to those from NCCA, and can be explained in terms of the method used for generating the source signals. This was a standard technique of starting with independent source signals (such as a sine wave and a sawtooth function for the two-component example), and pre-mixing them with a negative correlation matrix to establish the original source signals for the algorithms to recover. After this, the pre-mixed source signals are mixed

together with the mixing matrix to produce the mixed data. Because both mixing and pre-mixing are linear operations, they can in fact be described by just a single mixing operation, as though the original independent signals were mixed together just once to produce the mixed data. Because of this, it is not surprising that ICA finds this combined mixing matrix and original independent source signals. It is important to note that this does not at all invalidate this test; on the contrary, it points to a specific weakness in this ICA algorithm when faced with correlated signals (which it is not designed for). It is desirable however, also to test the algorithms without this pre-mixing stage leading to this phenomenon. This is addressed in the next section.

[FIGURE 5 ABOUT HERE]

3.3 Two methods for generating negatively-correlated test signals

The most common method for generating negatively-correlated test signals is to first generate independent signals, and then to pre-mix them with a negative correlation matrix. An advantage of this method is that it allows easy and precise control of the correlation relationship between any number of components. However, it was seen in the previous example that under these circumstances, ICA will tend to recover the independent signals prior to pre-mixing, rather than negatively-correlated source signals. An alternative method for creating negatively correlated signals without pre-mixing by a correlation matrix is to use phase control. By adjusting the relative phase of two periodic signals, their correlation can be altered. Figure 6 shows two periodic signals along with a graph that shows how the correlation changes with phase shift. It is straightforward using this approach to set the correlation to a desired value, including a particular negative correlation, or alternatively simply to set the phase to the point of maximally negative correlation. The advantage of this approach is that the source signals remain in their original form, without being pre-mixed. Clean signals with a negative correlation provide a useful way of further testing the NCCA algorithm, and this is the method that is used in the remaining two examples.

[FIGURE 6 ABOUT HERE]

3.4 Example 3: ICA recovers independent "mixtures", NCCA recovers negatively-correlated clean signals

Using the technique of phase shift in generating the negatively-correlated source signals, this example (figure5) shows the superior performance of NCCA in recovering the original source signals. It is notable that ICA recovers signals that are statistically independent, but that do not take the precise shape of the original source signals. In finding independent rather than negatively-correlated signals, ICA is forced to find slight mixtures of the original signals, rather than the pure signals themselves.

[FIGURE 7 ABOUT HERE]

3.5 Example 4: Assessing the correlation penalty coefficient (λ)

The final example in this section looks at the role of the correlation penalty coefficient (λ). The value of the coefficient was systematically varied whilst the other experimental parameters (learning rate, epochs etc.) remained constant. It can be seen in figure 8 that the correlation of the derived components changes smoothly with the value of λ, which shows both the stability of the algorithm under changes to this value, and demonstrates that the negative correlation penalty step offers a way to systematically control the correlation of the components found by NCCA (including even making them positively correlated if so desired).

[FIGURE 8 ABOUT HERE]

4 Examples of NCBA

The NCCA algorithm has been shown to be effective in recovering components that are negatively-correlated. The NCBA algorithm has a complementary purpose, which is to utilise the noise-reduction and space-filling benefits of negative correlation. It was seen in earlier sections how negatively-correlated basis functions could offer a theoretical advantage over positively-correlated and independent basis functions in representing data with non-negative coefficients. This section contains two practical examples of this advantage in operation, inspired by the widespread use of ICA on natural image processing.

4.1 Example 1: A pre-whitened natural image

It can be seen that the original image has been preprocessed with a low-pass whitening filter. This image is actually one that has been used in examples of ICA, where such filtering is common to assist the ICA algorithm in useful basis functions. In order to give a fair trial to ICA, this pre-whitened image is used in the test here. Three different conditions were tested: positive correlation (where λ was given a negative value), independent (ICA) and negative correlation (NCBA, where λ was given a positive value). Figure 9 shows the basis functions found in the three conditions. It is immediately apparent that the positive correlation condition has obtained perfectly correlated basis functions, which is catastrophic for representing data points, as it is equivalent to only having one basis function. The independent and negatively-correlated conditions have found ten different basis functions. The correlation values are given for these, which show that the algorithm has indeed found positively-correlated, uncorrelated, and negatively-correlated basis functions respectively.

Figure 9 also shows the image reproduced by representing each 3x3 image patch as a non-negative linear combination of the basis functions for each of the three conditions, and placed in its appropriate position in the overall image. This technique allows an immediate evaluation of the performance of the algorithms. It is clear that the positively-correlated basis functions have allowed only a very poor representation of the image, not surprising in view of the perfect correlation between the basis functions. More significantly, however, the independent basis functions have also resulted in a rather noisy image reproduction, suggesting that they are suboptimal for this task. Only the negatively-correlated basis functions allow for a perfect reproduction. The reproduction error values are given for all three conditions, corroborating the visual evidence.

[FIGURE 9 ABOUT HERE]

4.2 Example 2: An unpreprocessed natural image

Whilst ICA algorithms prefer the data, in this case a natural image, to be preprocessed, in particular pre-whitened, it is worth investigating whether or not the NCBA algorithm performs any worse on an image which has not been preprocessed at all.

This example follows the same procedure as the previous one, with positively-correlated, independent, and negatively-correlated conditions. Figure 11

shows that once again, when positive correlation is encouraged, perfectly correlated basis functions are found, whereas the uncorrelated and negatively-correlated conditions find ten different basis functions.

The image reproductions in figure 11 also follow the pattern of the previous example, with the positively-correlated basis functions allowing the worst image reproduction, followed by the independent basis functions which still give a very noisy reproduction, and then the negatively-correlated basis functions which give a perfect, noise-free, reproduction. It can be seen from the error values as well that the lack of preprocessing of the image did not damage the performance of the algorithm at all, in contrast to that of the ICA algorithm, whose relative performance here was worse than in the previous example.

[FIGURE 10 ABOUT HERE]

5 Conclusions

Negative correlation has several benefits which can result in systems with lower noise, or more accurate representation of information with a limited set of resources. In particular, it has been shown that negatively-correlated noise is reduced in accordance with the central limit theorem much more effectively than independent or positively-correlated noise. It has also been shown that negatively-correlated basis functions allow a more accurate representation of a set of data with non-negative coefficients than the same number of independent or positively-correlated bases.

In this paper, we have outlined two algorithms to exploit these statistical benefits of negative correlation, both of which are developments of the relatively new ICA approach. NCCA finds components which are negatively correlated, whilst NCBA finds negatively-correlated basis functions. Both algorithms are based on an ICA core with a lagrangian penalty term encouraging negative correlation, but the algorithms make use of a number of special techniques in order to allow the penalty term to be applied separately, and in a different domain, to the main ICA update step. A number of advantages to this have been outlined, emphasising in particular the compatibility of these new algorithms with a wide variety of existing ICA approaches, as well as their relative efficiency and stability.

Several simple demonstration examples of NCCA and NCBA have been given here, each chosen to demonstrate a particular feature of the algorithms. These examples show that:-

- NCCA offers superior performance to ICA in recovering negatively correlated signals.
- ICA recovers uncorrelated versions of the signals, whilst NCCA recovers the actual negatively correlated signals.
- When clean, negatively correlated source signals are generated using a phase-shift technique, ICA tends to recover uncorrelated mixtures of these, whereas NCCA recovers the negatively correlated clean original signals.
- NCBA gives basis functions which allow more accurate representation of data (image data in the examples given here), allowing better recovery of that data, than ICA.
- NCBA appears to be less demanding in terms of required preprocessing of data for than ICA.
- For both algorithms, correlation of components/basis functions varies smoothly as a function of λ, the negative correlation penalty coefficient (shown as an NCCA example in this paper, but equally valid for NCBA also).

The examples presented in this paper are just very small demonstrations of what NCCA and NCBA can do. In particular, althought the NCBA examples were in this case given for image reproduction, it is important to note that there is nothing special about image data in this regard, and the result is equally applicable to any data whatsoever, including data in variables that are not themselves negatively-correlated. When non-negative coefficients are used, negative-correlated basis functions will always be on average at least as effective as independent basis functions, and usually more so, at representing any set of data whatsoever.

There are a number of possible further developments for the algorithms presented here. One possibility is to explore the effects of negatively-related higher-order moments, particularly in view of the higher-order, non-gaussian nature of ICA which is an important part of these algorithms. Whether or not the same benefits, perhaps to an ever greater extent, could exist for negative higher-order moments remains to be seen.

Another as-of-yet unexploited potential advantage of the NCBA algorithm also requires further development. This concerns the space-filling benefit of negative correlation. It can be shown that at present, the advantage conferred by the space-filling property of negatively correlated basis functions is

actually the result not of space-filling per se, but of the increased probability that the basis functions will surround the mean of the data, which therefore allows a more accurate non-negative coefficient representation. When all the basis functions lie in a similar direction from the data mean, as is more likely to happen with positively correlated and uncorrelated basis functions because they are more closely tied together, this will result in the suboptimal representation that is seen in the examples. What this means is that the actual space-filling itself, which results in negatively-correlated basis functions being on average closer to the data points they are representing and hence require on average lower coefficient values, is not yet being exploited by the algorithm. In fact, in systems where resources (which means coefficient values) are costly (including biological systems), this space-filling benefit is likely to be important. For example, in neural systems in may result in lower firing rates being needed because individual neurons may be more accurately attuned to individual stimuli. This intriguing idea requires further investigation.

Also related to biological systems, the notion of how negative correlation is actually implemented in such systems is another subject for research. For example, data from the olfactory bulb suggests that neural firing is negatively-correlated [7]. Whilst this result may be seen as supporting the above hypothesis that natural systems will exploit the benefits of negative correlation, it also raises the question as to what neural mechanisms can give rise to it. The issue is currently being investigated.

We hope that the NCCA and NCBA algorithma, as generalisations of ICA, allowing application to a wider group of problems and offering significant benefits in representing and reproducing data, offers a useful new statistical tool, as well as potentially offering insight into existing informating processing systems.

References

[1] Hyvärinen, A., Karhunen, J. and Oja, E. (2001): *Independent Component Analysis* (Wiley)

[2] Olshausen, B.A. and Field, D.J. (1997): Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research* 37:3311-3325

[3] Hyvärinen, A. and Hoyer, P.O. (2000): Emergence of phase- and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation* 12:1705-1720 23

[4] Feng, J. and Tirozzi B. (2000): Stochastic resonance tuned by correlations in neuronal models. *Phys. Rev. E.* 61:4207-4211

[5] Lee, D.D. and Seung, H.S. (2001): Algorithms for non-negative matrix factorization *Advances in Neural Information Processing* 13

[6] Hoyer, P.O. (2002): Non-negative sparse coding feature subspaces. *Neural Networks for Signal Processing* 12: 557-565

[7] Nicol, A., Feng, J. and Kendrick, K. (in preparation): Negative Correlation Yields Computational Vigour in a Mammalian Sensory System

[8] Bell, A.J. and Sejnowski, T.J. (1995): An information maximization algorithm that performs blind separation. *Advances in Neural Information Processing Systems* 7:456-474 (MIT Press)

[9] Amari, S. (1999): Natural gradient learning for over- and under-complete bases in ICA. *Neural Computation* 11:1875-1883

[10] Hyvärinen, A. and Oja, E. (1997): A fast fixed-point algorithm for independent component analysis. *Neural Computation* 9,7:1483-1492

6 Figures

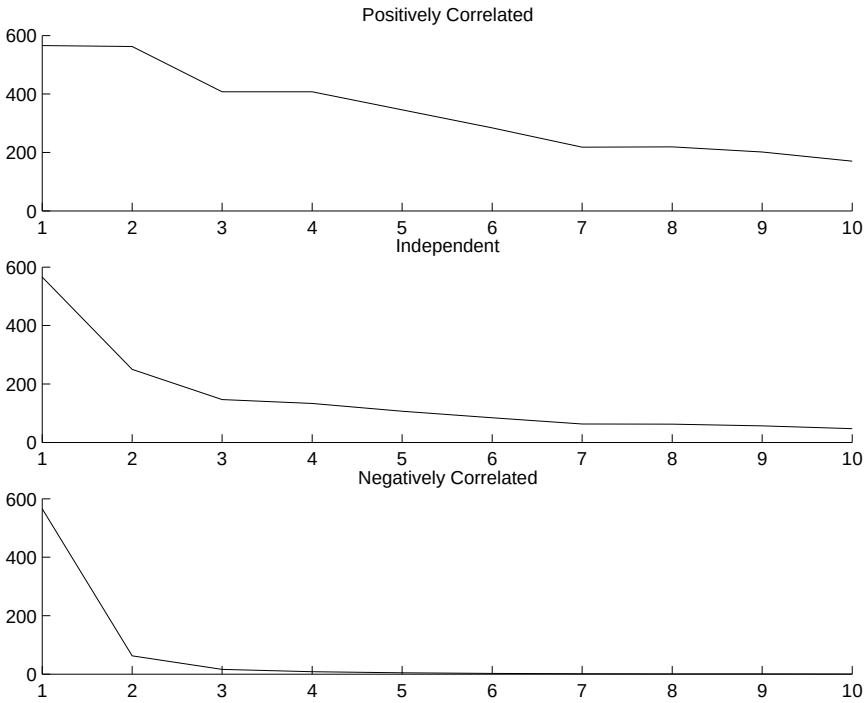

Figure 1: Central limit shrinkage of negatively-correlated noise. As the number of noise instances (samples) increases, negatively-correlated noise shrinks to zero quickly, whereas independent and positively-correlated noise require more instances for their values to decrease, with positively-correlated noise potentially having a non-zero asymptote. This shows how negative correlation can eliminate noise both more quickly, and more completely. The positively-correlated noise here has a correlation of 0.1 between each of the ten instances of noise, whilst the negatively-correlated noise has the opposite value of -0.1.

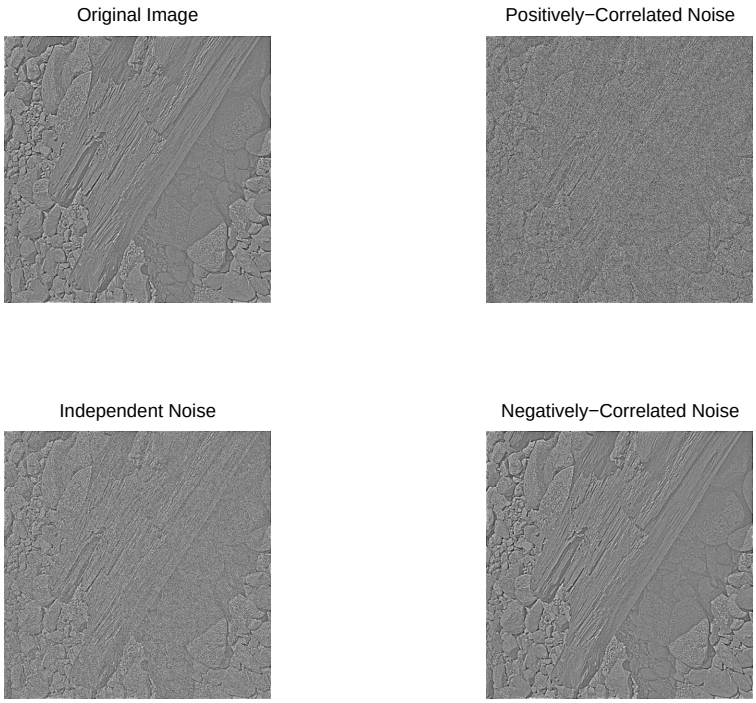

Figure 2: Benefit of negatively-correlated noise. With ten separate samples of noise added to the original image, the differing effects of the correlation of the noise can clearly be seen here. In particular, negatively-correlated noise largely disappears leaving original image clearly visible. Here, the positively-correlated noise again has a correlation of 0.1 between each of the ten instances of noise, whilst the negatively-correlated noise has the opposite value of -0.1.

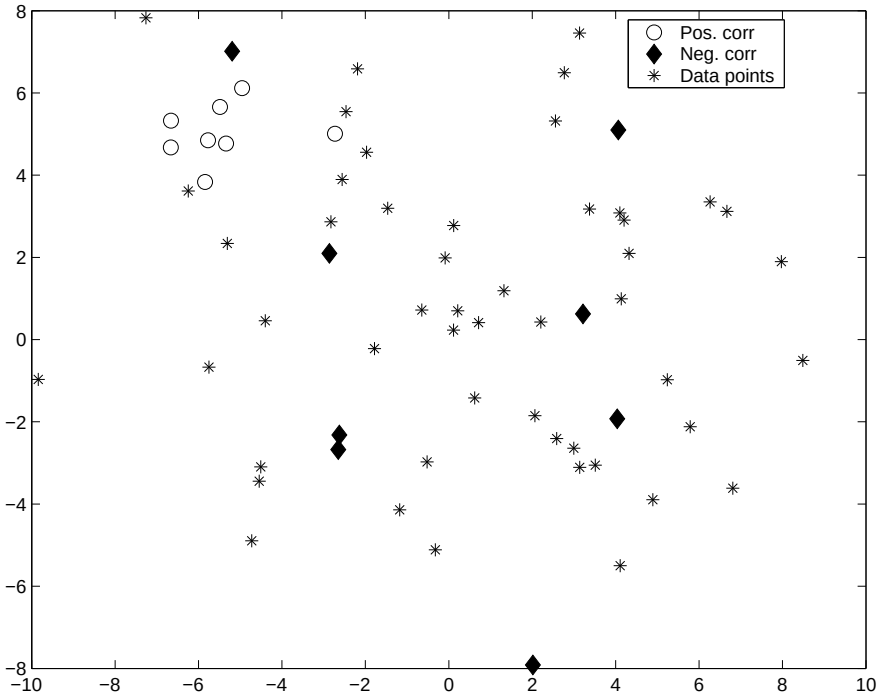

Figure 3: Efficacy of negatively-correlated basis functions. The negatively-correlated basis functions (black diamonds) are more widely distributed than the positively-correlated basis functions (white circles), and offer a more useful basis for representing the data points (crosses). The non-negative least-squares error for representing the data is 0 for the negatively-correlated bases, but 122.94 for the positively-correlated bases. In this example, the positively-correlated basis functions have a correlation with each other of 0.9 , whilst the negatively-correlated basis functions have the lowest possible correlation for eight basis functions, -0.14286.

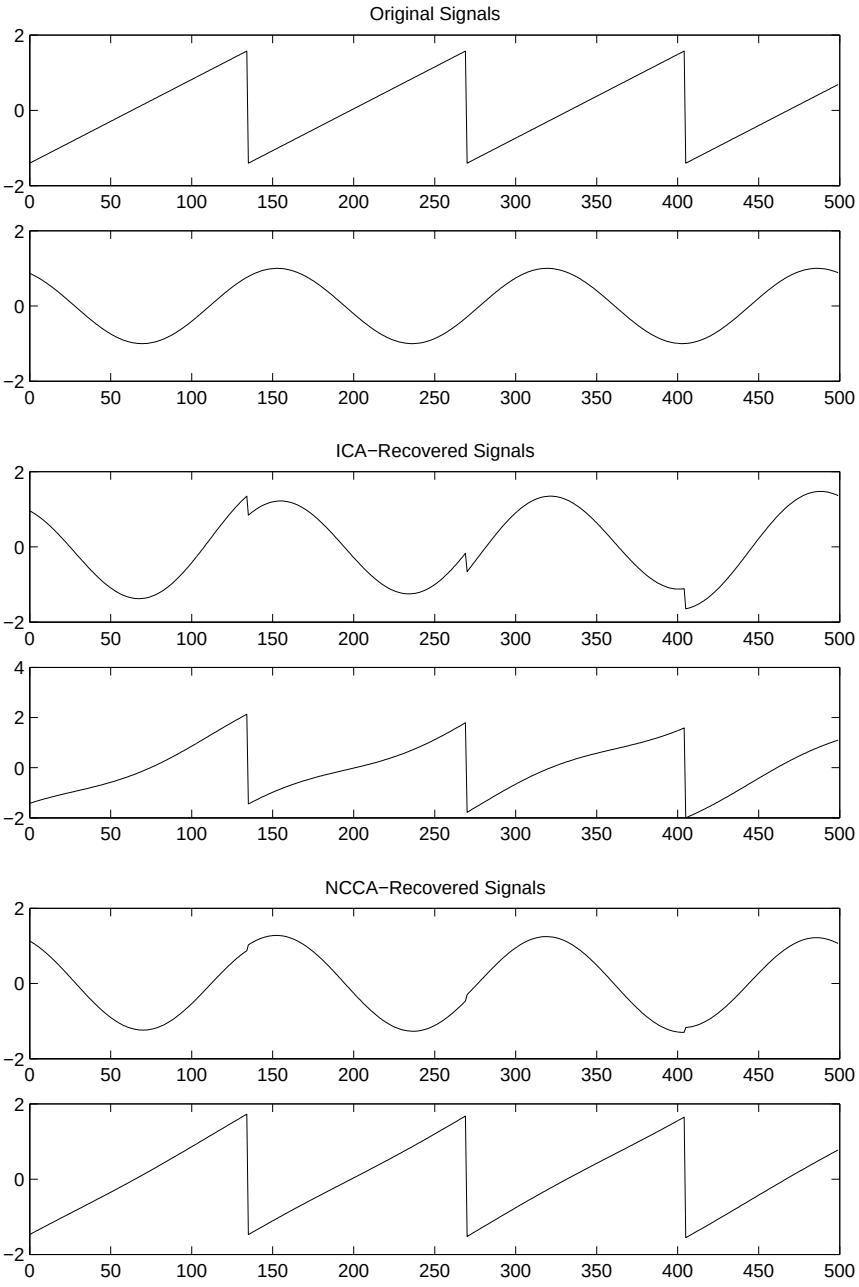

Figure 4: NCCA vs ICA for signal recovery. This shows the central benefit of the NCCA algorithm, which gives a better recovery of the original, negatively-correlated signals than ICA. The correlation of the ICA components was 0.0082009, whereas that of the NCCA components was -0.35562, much closer to the original signals correlation of -0.4.

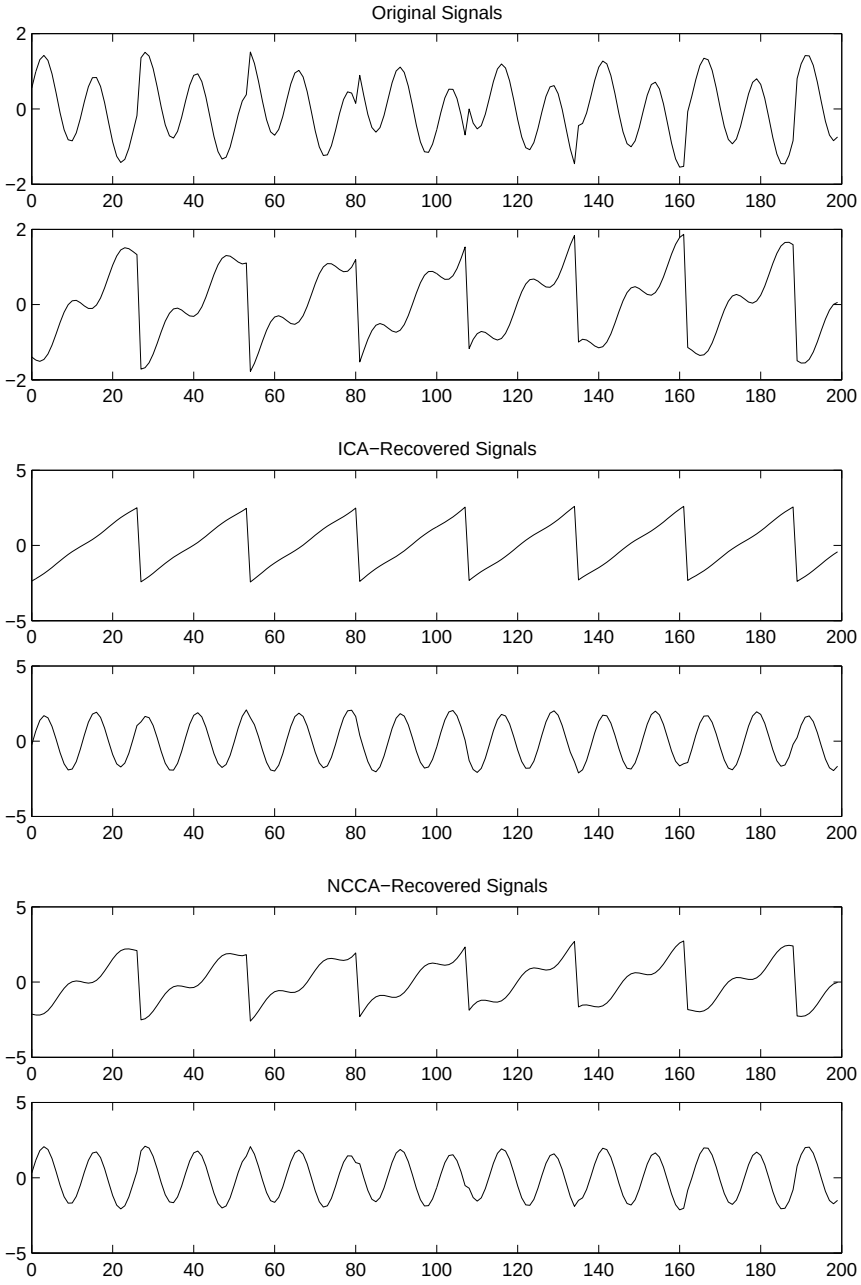

Figure 5: ICA recovering uncorrelated signals. Where the signals are pre-mixed to be negatively-correlated, before the main mixing stage to produce the mixed data, ICA recovers the first, uncorrelated versions of the signals, whereas NCCA recovers the desired, negatively-correlated signals. The correlation of the ICA components was 0.038128, whereas that of the NCCA components was -0.42265, again much closer to the original signals correlation of -0.4.

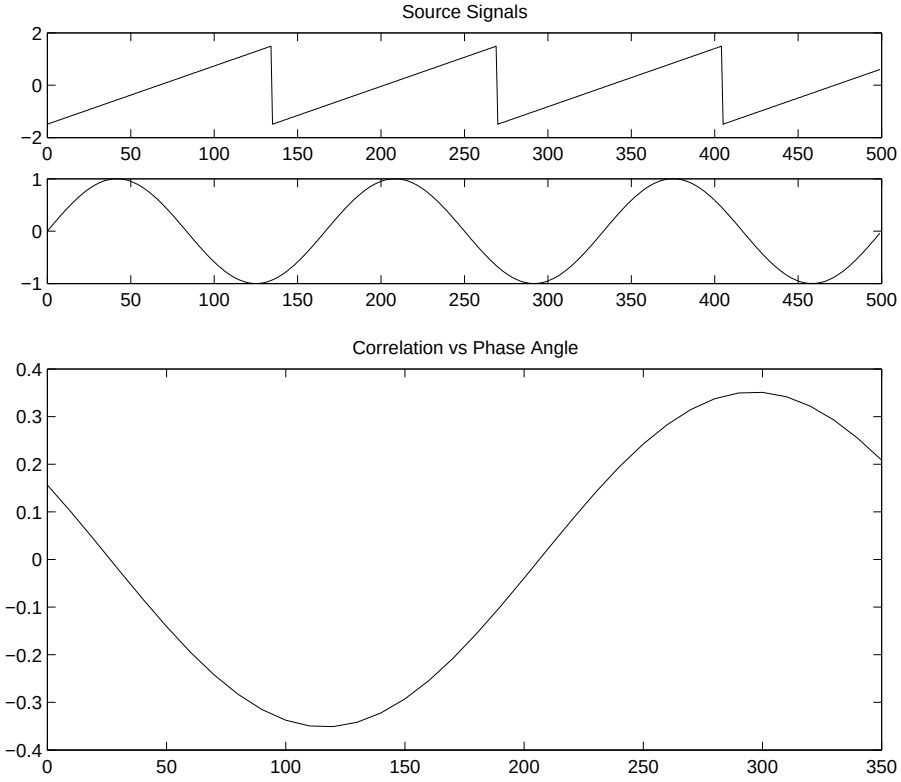

Figure 6: Correlation control using phase shift. Sinusoidal and saw-tooth source signals are shown, along with the correlation between these two signals as a function of the phase between them.

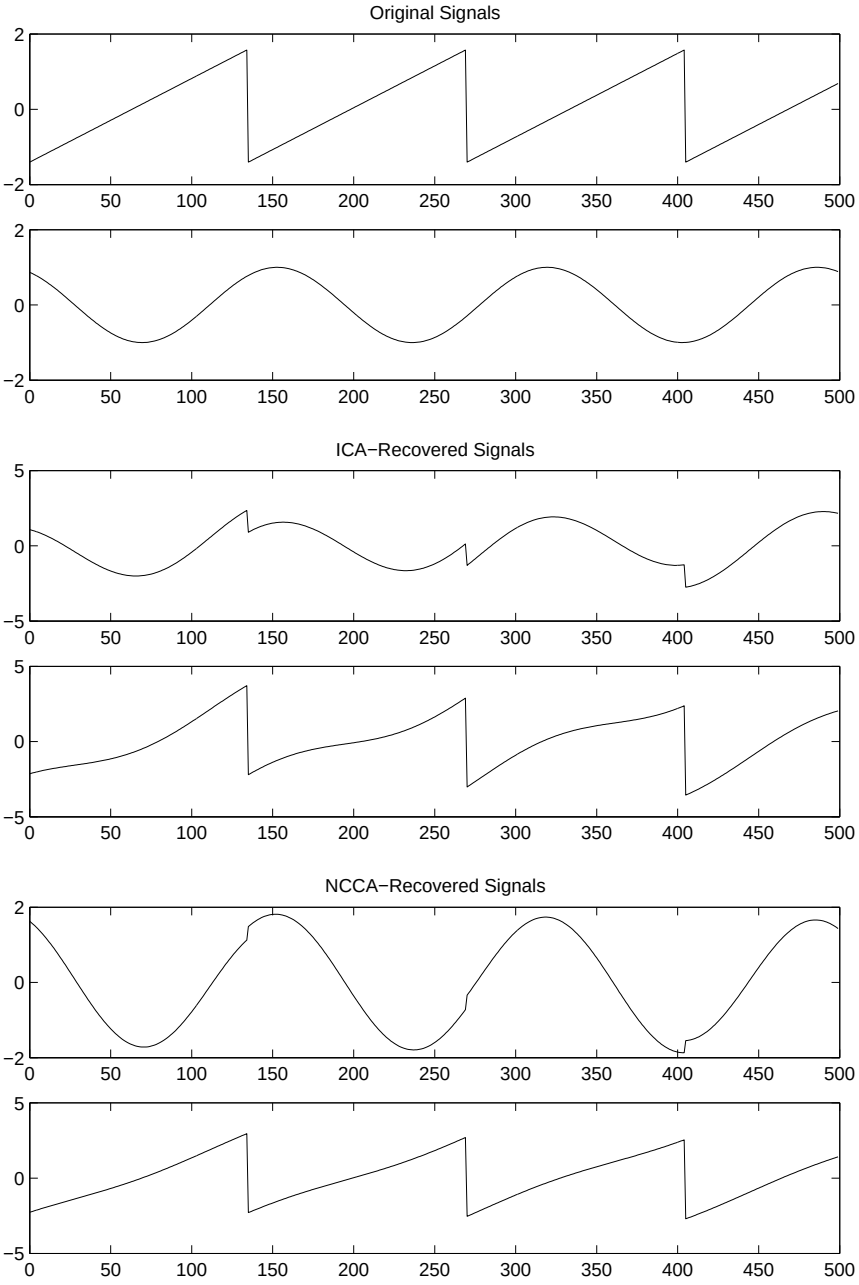

Figure 7: ICA recovering independent mixtures. This example shows that when clean negatively-correlated source signals are created using the phase-shift technique, ICA tends to recover independent versions of these signals, which are mixtures of the original source signals, whereas NCCA tends to recover the actual source signals. The correlation of the ICA components was 0.25919, somewhat positively correlated, whereas that of the NCCA components was -0.31384, once again much closer to the original signals correlation of -0.35089.

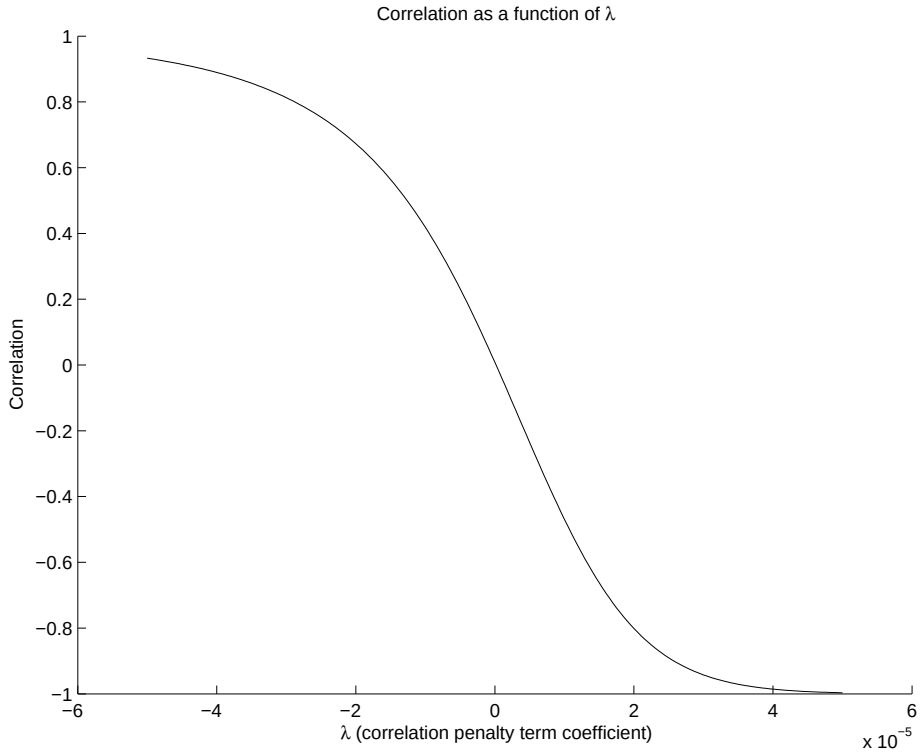

Figure 8: Correlation as a function of the penalty coefficient λ. The sigmoid curve in this graph highlights the robust and stable nature of the NCCA and NCBA algorithms, with correlation varying smoothly with the strength of the negative correlation penalty.

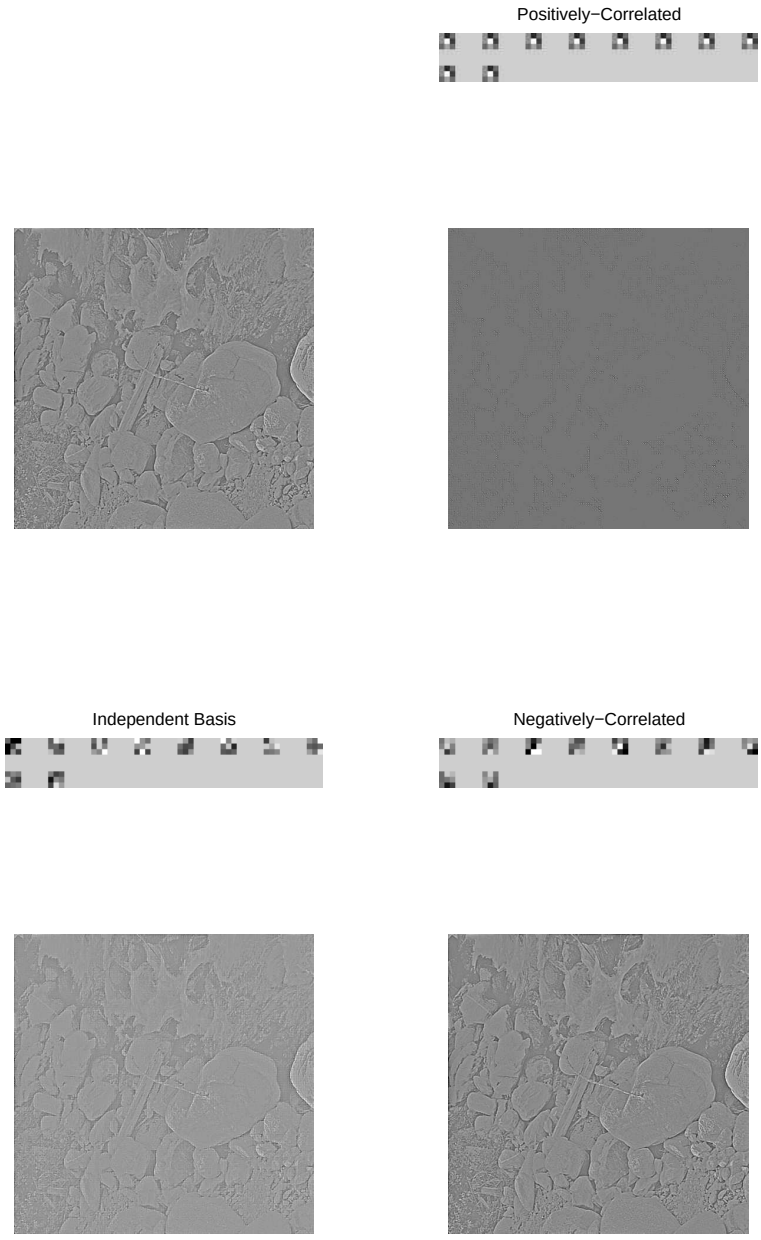

Figure 9: Basis functions and image representation. This figure shows the positively-correlated, independent, and negatively-correlated basis functions recovered by using the NCBA algorithm (with a negative penalty coefficient to obtain positively correlated basis functions and a zero coefficient to obtain the independent basis functions, which is therefore ICA in effect). Image data has also been represented using these basis functions, and the clear benefit of negative correlation is apparent. The correlation values for positively-correlated, independent and negatively-correlated basis functions respectively are 1 (the maximum), -0.03666 (close to uncorrelated), and -0.10717 (near to the lowest possible of -0.11111. The respective image representation LSE values are 1.6625, 0.36347 and 0.

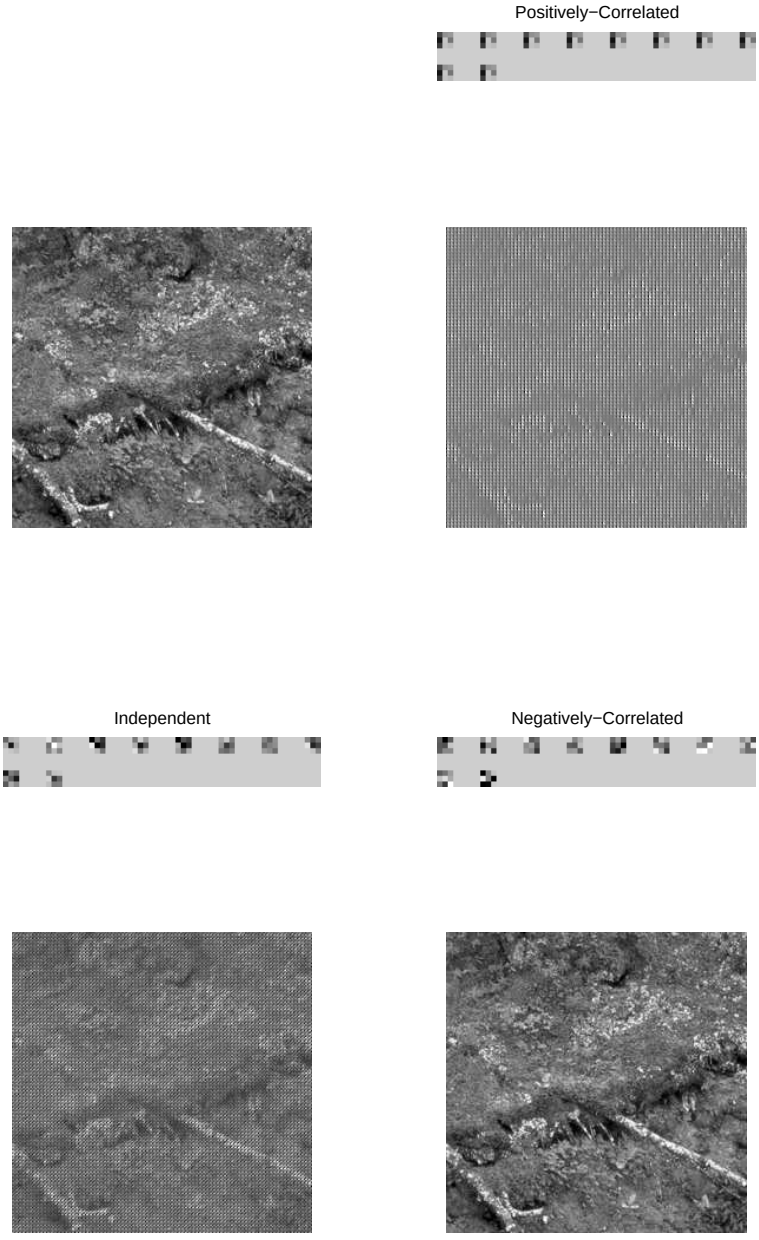

Figure 10: Basis functions and image representation for non-preprocessed image data. Similar to figure 9, except for the fact that in this case the image data has not been subject to any preprocessing. Again, negative correlation offers by far the best basis functions for representing the image data. The correlation values for positively-correlated, independent and negatively-correlated basis functions respectively are 1 (the maximum), 0.20188 (slightly positively correlated), and -0.098868 (near to the lowest possible of -0.11111. The respective image representation LSE values are 1.7583, 1.1481 and 0.