The role of sparsity and the behavior for mixture data in homeopathic independent component analysis (ICA)

W. Windig*, B.M Wise
Eigenvector Research, Inc.
196 Hyacinth Road,
Manson, WA 98831
USA

M.R. Keenan
8346 Roney Rd.
Wolcott, NY 14590
USA

*Corresponding author
6 Olympia Drive,
Rochester, NY 14615
Phone: +1 585 663 1139
Email: windig@eigenvector.com

A B S T R A C T

In a recent paper, we proposed Homeopathic ICA, which is a simple method to expand the use of independent component analysis (ICA). Adding sparsity enabled the resolution of data sets into the chemical constituents that could not be resolved with conventional ICA. The data sets used were dominated by the pure components to be resolved, however, one of the successfully resolved data sets contained mixtures in addition to the pure components.

This paper will show in more detail the role of sparsity in ICA and study tolerance for mixtures in addition to pure components. A diagnostic value is introduced to determine the optimal sparsity for a proper resolution of mixtures. As shown in the previous paper sparsity may mean noise level intensities. The relation between Homeopathic ICA and quartimax will be shown. Matlab code is included for efficiently performing Homeopathic ICA. In addition to simulated data we will discuss new results obtained on energy dispersive X-ray spectroscopy (EDS) of a CuNi diffusion couple and a braze interface. In addition, the results from magnetic resonance imaging (MRI) of a binary mixture will be shown.

1. Introduction

As we have shown in a previous paper it is possible to expand the use of ICA by adding sparsity to the data [#1039]. This can be done simply by adding zeros to the data set or by applying the Haar-wavelet transformation. The latter has the advantage that the size of the data matrix does not increase, which significantly reduces the calculation time. The term homeopathic was used for this version of ICA since we seemingly 'dilute' the data set with zeros, while improving the performance of ICA.

The method was explained for simple simulated data, where only pure components were present. However, one of the actual (not simulated) data sets did have mixtures of the two pure components in it and could be resolved properly. This paper will study how ICA can resolve data where mixtures are present in addition to the pure components. Simulated data will be used to explore the behavior of ICA with mixture data present. A diagnostic measure will be introduced to determine the optimal ICA solution. In addition to examples with simulated data we will discuss new results obtained on energy dispersive X-ray spectroscopy (EDS) of a CuNi diffusion couple and a braze interface. New results obtained on magnetic resonance imaging (MRI) of a binary mixture system will also be shown.

2. Theory

In the previous paper [#1039] we discussed how the fastICA algorithm (#1007) uses excess kurtosis (by default) to achieve non-gaussianity of the components (#1064). For an array x with n elements (for example, the scores of a principal component scores model of a data set) the excess kurtosis is calculated as follows, see Equation (1)

$$k_{excess} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^4}{\frac{1}{n}(\sum_{i=1}^{n}(x_i-\bar{x})^2)^2} - 3 \qquad (1)$$

In this paper, we will also use the raw kurtosis, defined as follows.

$$k_{raw} = k_{excess} + 3 \quad (2)$$

In addition, we will also discuss the use of the skewness.

$$s = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{\frac{1}{n}(\sum_{i=1}^{n}(x_i - \bar{x})^2)^{3/2}} \quad (3)$$

For an ideal uniform distribution, the value of the raw kurtosis is 1.80, for the excess kurtosis -1.2 and for the skewness the value is zero. An example of a (not ideal) uniform distribution is shown in Figure 1a. For an ideal normal distribution (mean of 10, standard deviation of 0.2) the raw kurtosis is 3.0. For an example see Figure 1b. The reason for subtracting the value 3 from the raw kurtosis is to have a value of zero for the excess kurtosis for a normal distribution. The skewness for this distribution is zero.

As discussed in the previous paper (#1039) adding zeros to a data set is the basis for Homeopathic ICA. As an example, we use the data on which the normal distribution of 10000 data points in Figure 1b is the basis and add 50000 zeros (5 blocks, where a block is defined as the size of the original data set). The distribution of the resulting sparse data is shown in Figure 1c.

The narrow peak results in a high value of the raw kurtosis (4.2) and the excess kurtosis (1.2). When the distribution is asymmetrical with a long left tail skewness has a negative value. Similarly, a long tail to the right produces a positive value for skewness, 1.79. When we have a sparse data set dominated by zeros, as discussed in the previous paper, it is like a right skewed distribution with a positive skewness.

The choice of the absolute value of the excess kurtosis for fastICA [#1007] is clear from this image. The goal of ICA is to make the distributions non-Gaussian. The value of the absolute excess kurtosis is zero for a Gaussian distribution and becomes positive when deviations of Gaussianity occur in the directions of wide (uniform) distributions and narrow (sparse) distributions.

A dataset with only pure components does not have independent contribution profiles. When adding zeros, the contribution profiles becoming more and more independent. A linear combination of these sparse independent contribution profiles will be less sparse. By optimizing for sparsity, using the absolute excess kurtosis, these narrow, sparse independent contribution profiles are found.

As was discussed before (1), sparsity could be introduced to a data set by actually adding zeros to the data set or by applying the Haar transformation. However, the effect of adding zeros on the variance-covariance matrix, which is the basis in the fastICA calculations, can also be calculated without actually adding zeros. This approach is obviously faster than actually adding zeros and also faster than performing the Haar transformation on the data set. A program that performs homeopathic ICA in this manner is available through the Matlab file exchange (*). The help files of this and additional programs are shown in Appendix A.

As shown in the previous paper and ‘infinite’ number of zeros must to be added to a data with mutually exclusive contributions of pure components to resolve the pure components. There is, however, for this

also a faster approach possible. FastICA works with centered data with, by definition, the mean as the origin. When adding an infinite number of zeros, the mean converges to zero. Resulting in the following equation

$$\lim_{n\to\infty} k_{excess} = \frac{\sum_{i=1}^{n} {x_i}^4}{\frac{1}{n}(\sum_{i=1}^{n} {x_i}^2)^2} - 3 \quad (4a)$$

Where n is the total number of elements.

which equals

$$\lim_{n\to\infty} k_{excess} = \frac{n\sum_{i=1}^{n} {x_i}^4}{(\sum_{i=1}^{n} {x_i}^2)^2} - 3 \quad (4b)$$

In the fastICA algorithm the internal results are length scaled. As a result, the denominator has a constant value of one. When n is very large the constant -3 is negligible, so the equation becomes.

$$\lim_{n\to\infty} k_{excess} = n\sum_{i=1}^{n} {x_i}^4 \quad (4c)$$

The quartimax criterion is as follows.

$$q = \sum_{j=1}^{n}\sum_{i=1}^{nspec} b_{ij}^4 \quad (5)$$

Where q is the quartimax value and n the number of orthogonally rotated vectors with *nspec* spectra, where the original vectors are, for example, scores (contributions) resulting from singular value decomposition (#3000). The rotated vectors in matrix **B** have q maximized. This rationalizes that the excess kurtosis for a high (infinite) number of zeros (Equation 4c) is proportional to the quartimax criterion in Equation 5.

3. Materials and methods

3.1 Simulated data set for mixture study

For the mixture study the following data sets were constructed. The data matrix **S** with the contributions for the data has the following composition.

$$\mathbf{S} = \begin{bmatrix} 1 & 1 & 1 & \cdots\cdots & 1 & 0 & 0 & 0 & \cdots\cdots & 0 \\ 0 & 0 & 0 & \cdots\cdots & 0 & 1 & 1 & 1 & \cdots\cdots & 1 \end{bmatrix} \quad (6)$$

The size of **S** is ncomp × nspec (number of pure components by number of spectra), which is in our case 2×2000. There are 1000 ones and 1000 zeros on the first row, and the reverse on the second row. The rows of **S** are the contributions of two spectra in matrix **A**, size nvar × ncomp**,** where nvar is

the number of variables, see Equation 7. The orientation of the data set is according to the conventions used in ICA, where the independent components are in rows (#1007).

$$\mathbf{X} = \mathbf{AS} \tag{7}$$

The first column of **A** contains a Gaussian shaped spectrum with mean at data point 25 and a standard deviation of 10 and the second column contains a spectrum with a mean at data point 75 and a standard deviation of 10. The total number of data points for each Gaussian is 100. The contributions in **S** and the spectral data set **A**, form the bases of dataset **X**, size nvar ×nspec. The term contributions is used instead of concentrations, since there is generally an unknown multiplicative factor between the actual concentrations and the values resulting from mixture analysis type of data analysis.

The maximum values of the matrices **A** and **S**, and thus of **X** are one. Uniformly distributed noise with a range from -.05 to 0.05, which is 10% of the maximum intensity in the data, was added to **X**. The matrix **A** is, in ICA terminology, called the mixing matrix. In our case **A** contains the spectra associated with the independent contributions in **S**.

Several versions of the data set are created in the form **Xn**, where n represents the number of mixtures added. The mixtures in a row are random uniformly distributed values between 0 and 1 and are added to the matrix **S**. The mixtures are ordered according to composition for plotting purposes. For a value of x in element j in the first row the value for the corresponding element in the second row is (1-x). The noise described above is also added to these mixtures. The four datasets created are **X0** (which is identical to **X**), **X600** and **X1600** and **X5000**.

3.2 Actual data

3.2.1 Cu-Ni diffusion couple

Cu-Ni is a simple binary system with complete solid solubility across the entire compositional range. This sample is a diffusion couple that contains pure copper and pure nickel, as well as, a concentration gradient through the diffusion zone. The sample was imaged with Energy Dispersive X-ray Spectrometry (EDS) using a 7 kV accelerating voltage. Under these conditions, only the highly overlapping L emission lines of Cu and Ni are excited, and the data set analyzed here consisted of 150-channel spectra covering the range of 0.15 to 1.64 keV acquired at each of 15360 pixels of the 128 × 120 pixel image. Complete details describing how the sample was fabricated and imaged can be found in [#979] together with an analysis of an expanded version this data set using multivariate curve resolution (MCR-ALS).

3.2.2 Braze

The braze data set is an EDS spectral image comprising 16384 spectra (pixels). Each spectrum was acquired with 0.01 keV resolution, and spanned the range from 0.2 to 10.24 keV (1005 channels total). The sample itself is braze joint between a copper-silver braze material and an iron-nickel-cobalt alloy, with titanium present at the interface to promote adhesion. The copper and silver in the braze are largely phase-separated leading to a sample consisting of four distinct chemical phases. Additional information about data acquisition conditions is contained in [#1013].

3.2.3 N*uclear Magnetic Resonance (NMR) of a binary mixture*

For details about the sample and its analysis see (#805). A simple phantom was constructed from PVC, see Figure 2, with two compartments separated by a thin plastic sheet, forming a wedge. The first compartment contained 0.5 mM $MnCl_2$ and the other compartment 8 mM $NiCl_2$. The phantom was scanned by NMR to obtain a series of 15 T_2 relaxation images of 256× 256 pixels. The images were acquired of a 5mm thick plane in the center plane, parallel to the bottom, of the phantom, resulting in mixture data in the area of the plastic sheet separating the pure components. The series of 15 T_2 relaxation images exhibit an exponential decay, which is different for the two components, and a mixture of the two decays for the mixtures. The T_2 value is the inverse of the exponential decay constant. The theoretical T_2 values for this sample are 29.6 ms for $MnCl_2$ and 150 ms for $NiCl_2$. The experimental values obtained in the paper are 28.2 ms and 137.7 ms, respectively.

3.3. Data analysis

All calculations were performed using Matlab Version 9.1 (Mathworks, 3 Apple Hill Drive, Natick, MA 01760-2098) on a MacBook Pro with the OSX El Captain. The ICA functions used in this paper were written by the authors. The help files of the functions are show the Appendix and are the functions available on the Matlab side with user contributions (https://www.mathworks.com/matlabcentral/fileexchange/). For the kurtosis option the results of our programs are virtually identical to the results of the FastICA package (Helsinki University of Technology, Laboratory of Computer and Information Science, P.O. Box 5400, FI-02015 HUT, FINLAND), available through the internet [#2000].

4. Results and discussion

4.1 Simulated data sets

It is important to realize that what we want to achieve is a solution that represents the components we created the mixtures from. We will call this solution *the correct solution.* If a solution provided by ICA is not the correct solution it does not mean that there is a problem with the ICA technique. It means that the components of the correct solution are not the most independent linear combination of the data in **X** (Equation 7).

For a proper pre-diagnostic value, we need to realize that fastICA is based on the same space as principal components (PCA), using whitened (autoscaled) variables (mean is zero and standard deviation is one) transformed by an orthonormal transformation (rotation) (#1007). Therefore, the two independent components (contributions) we need for the correctly resolved simulated data set will have a correlation coefficient of zero.

This means that the contributions of the pure component contributions used in the simulated data must also have a correlation coefficient of zero (with added zeros!). Since we work with simulated data we can easily check this.

One must realize that uncorrelated contributions are not necessarily independent. For example, a series of points randomly distributed on a circle has a correlation coefficient of zero, even though there is a well-defined relation between the data points. Therefore, the diagnostic of the correlation coefficient a

value of zero for the contributions of the correct solution is necessary to find a proper ICA solution (resulting in the correct solution), but it is not necessarily the most independent solution.

To study to effect of adding mixtures to a data set dominated by pure components the simulated data set described above was used. First, we will study the effect of adding zeros to a data set not containing mixtures, **X0**. For this we use the following diagnostics.

First, we use the correlation coefficient between the two (noiseless) pure component contributions, with the added zeros, which we know because this is a simulated data set.

Second, we will show a measure of the difference between the ICA scores (after zeros are deleted) and the two (noiseless) pure component contributions. The difference between the original data and the ICA results is based on the relative root square of the differences between two data sets. For matrices **M** and **N** the fit is calculated as follows:

$$rrssq = \frac{\sum_{i=1}^{nrows}\sum_{j=1}^{ncols}(m_{ij}-n_{ij})^2}{\sum_{i=1}^{nrows}\sum_{j=1}^{ncols}m_{ij}^2} \quad (8)$$

Matrix **M** is the reference matrix, in this case the contributions of the original data, and matrix **N** represents the ICA scores. The value falls between zero and one, where the value of zero is a perfect match.

The reason we use the *rrssq* instead of the more commonly used correlation coefficient to compare the contribution profiles is that linear combinations of the independent contributions in Equation 6 have the same shape as the original contributions (except for a 0.5/0.5 combination). Since the correlation coefficient only shows differences in shape (because the mean is subtracted in the calculation) it is not suited as a diagnostic tool for these contributions.

The third and fourth diagnostics are the mutual information (MI) and the sum of the absolute values of the excess kurtosis calculated for the two (noiseless) pure component contributions. As discussed in the previous paper the MI (#1039) has a value of 0 when variables are statistically independent, otherwise it is positive (#1007).

The mutual dependence (#1022) of the two variables *X* and *Y* is calculated as follows:

$$MI = \sum_{x\in X}\sum_{y\in Y} p_{x,y}(x,y)\log_2\frac{p_{x,y}(x,y)}{p_x(x)\,p_y(y)} \quad (9)$$

Where $p(x,y)$ is the true probability of *X* and *Y* and $p_1(x)p_2(y)$ the probabilities if *X* and *Y* are independent.

We will first show the diagnostic values to see under which conditions the correct solution arises, which occurs when the correlation coefficient of the input contributions is zero. Consequently, the rrssq

value should be low.

In this simulated data set and the actual mixtures that follow below the zeros are added in blocks, where each block has the size of the original data. For the data without mixtures, **X0**, the effect of adding zeros is shown in Figure 3a. Please note that different scales are used for the rrssq and the correlation coefficient. The lowest value for the rrssq is the last point on the graph. The highest value of the correlation coefficient is close to zero, which we expect for the correct solution. The lowest value for the *rrssq* and the highest, closest to zero, correlation coefficient coincide, which is exactly what one would expect. One may conclude that adding more zeros will result in a correlation coefficient closer to zero. This is indeed the case, as will be shown below. The equation for the correlation coefficient *r* for two vectors *x* and *y* with each n elements is.

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{(n\sum x^2 - (\sum x)^2)(n\sum y^2 - (\sum y)^2)}} \tag{10}$$

When adding zeros to x and y the products $\sqrt{xy}$, $\sqrt{x}$, $\sqrt{x^2}$, etc. stay the same. However, n increases when zeros are added. This means that the terms with n become larger and larger and the terms without n can be ignored, resulting in the following limit

$$\lim_{n\to\infty} r = \frac{n\sum xy}{\sqrt{(n\sum x^2)\,(n\sum y^2)}} \tag{11}$$

Which equals.

$$\lim_{n\to\infty} r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} \tag{12}$$

This is the inner product of the length scaled vectors x and y. The added zeros do not influence the value of Equation 12. For the simulated dataset **X0** the inner product of the data in Equation 6 is zero. The values of the inner products are indicated as the higher horizontal lines in Figure 3a,b,c. The lower line indicates the value of zero as reference for the correlation curve. For **X0** the zero-correlation reference line and inner product line coincide.

In Figure 3d we also see that the MI value is low when the correlation coefficient is close to zero. In other words, the correct solution, after adding 20 blocks of zeros, is close to independent, as the low MI value shows. The kurtosis value is inversely correlated with the MI, as expected.

In Figure 3b we see the results after adding 600 mixtures, data matrix **X600**, which is 23% of the number of total number of cases. Interestingly, the correlation coefficient crosses the value of zero. Again, the expected behavior of a low value for rrssq coincides with the correlation coefficient of (closest to) zero. This is indicated with a vertical line. The correlation in this case will converge to the inner product of the length scaled original data, the value of which is indicated in the plot.

In this case, pure components with mixtures, the starting correlation of this normalized data set is -1.

The inner product of the length scaled data of the correct contributions is positive. The correct solution lies in between.

So in this case adding more and more zeros deteriorates the correct solution. Also, it is interesting that the MI value in Figure 3e of the correct solution is relatively high compared to Figure 3d. In other words, although ICA delivers the correct solution at the cursor, it is not really independent. This is an example where the correlation of the independent components is zero, while the MI is not zero.

Adding more mixtures, data matrix **X1400**, see Figure 3c, shows that the crossing of the zero point of the correlation coefficient is after adding ever fewer zeros. Apparently, adding more than 41% mixtures is the limit for homeopathic ICA of this type of data. For the correct solution, we now see an even higher value for MI in Figure 3f. Although we show only the MI values here for the known independent components, the ICA results show the same MI tendency.

Adding more mixtures leads to a deterioration of the solution since the correlation of (approximately) zero falls in the area where the rrssq is high.

Before discussing the effect of adding mixtures to in more detail we will first show results at different points in the range of solutions. The original **X0** data without zeros added is perfectly negatively correlated when no noise is added, so the rank is one. In our case, there is noise which will result in a second independent component (IC). This is meaningless, so the first solution will be ignored. The second solution, after doubling the size of the data with zeros, is shown in Figure 4a, e. It is clear from the rrssq plot in Figure 3a that this is a 'bad' solution from our point of view: the solution has no similarities with our reference data. This is clear from the spectra and contributions in Figure 4a, e. The results are labeled by reference to the ideal value of their correlation coefficient of zero in the titles of the top plots. This input for this solution has a negative correlation, leading to this 'bad' solution, so we refer to this as the corr<<0 solution to indicate the clear negative correlation. The results are very similar what one gets in PCA. One of the components is similar to the sum of the two components and the other one is similar to the difference of the two plots.

As a second example we take the 6^{th} point of the **X0** solution (Figure 3a) (5 blocks of zeros added), the results of which are shown in Figure 4b,f, labeled corr<0. The contributions do not have a part of zero values (within the range of the noise), see Figure 4b, indicating that they are positive linear combinations of the actual pure components. Since these contributions are basically mixtures, we call them under-resolved (#698). The associated spectra in Figure 4b show negative parts, in other words, they are linear combinations with negative contributions in them. For this, the term over-resolved is used. In this type of solutions, the over/under-resolution is always accompanied by the reverse under/over resolution of the other domain. Since this aim is to reproduce the original data as good as possible in a least square sense, a deviation in one domain of the solution must be compensated by the inverse deviation of the other domain.

The next example, the corr≈0 solution, is taken from the 6^{th} point (5 blocks of zeros added) of the **X600** mixture solution, see Figure 3b. From the rrssq and the correlation in Figure 3b at the sixth point we expect to get close to the expected ideal solution of the spectra and contributions, as is confirmed by Figure 4c,g.

As the last example, corr>0, the **X1400** data, we used the solution where 20 blocks of zeros added were

added. Here we see the opposite of the corr<0 solution. Now we have under-resolved spectra and over-resolved contribution, see Figure 4d,h.

According to the results presented above the homeopathic ICA approach cannot be used for simulated data sets with more than 1400 mixtures. However, one can wonder if the sum of the absolute values excess kurtosis is the proper criterion of our sparsity based approach. As mentioned above (Figure 1), the raw kurtosis has a negative value for a uniform distribution, a zero value for normal distribution and a positive value for sparse distributions. By taking the absolute values of the kurtosis there will be a maximum for uniform distributions and a maximum for narrow (sparse) distributions. The problems with the solutions where only a few zeros are added distributions with a relatively high amount of uniformity occur. Figure 5 shows the histograms of the contributions in Figure 4 with the added zeros. It is clear that the ICA solution in Figure 5a is based on approximately uniform distributions. The raw kurtosis values are below 3, indicating that the excess kurtosis is negative (see Equations 1,2), resulting in positive values for the absolute excess kurtosis. This is also clear from the plot of the absolute kurtosis values in Figure 3d,e,f. The absolute kurtosis values go through a minimum around 1 block of zeros, indicating the presence of negative values of the raw kurtosis before that minimum. The other histograms in Figure 5 show that the histograms clearly indicate sparsity.

In our case the objective function is narrow (sparse) solutions, solutions based on uniform to normal distributions are incorrect. A simple solution is to use the raw kurtosis (Equation 2) instead of the absolute values of the excess kurtosis. Another possibility is to use the skewness of a solution. The absolute values of the skewness will maximize for sparse solutions, which form basically highly skewed distributions. As an example, we will use the data set **X5000**, see Figure 6.

In Figure 6a plot of the rrssq values, as a function of the number of blocks of zeros, is shown for the sum of the absolute excess kurtosis values. The position of 1.12 blocks of zeros, indicated with a vertical line, is where the correlation of the known contributions, with added zeros, is zero. As we have seen above, this is where the correct solution occurs. It is clear that the correct solution cannot be obtained for **X5000** using the sum of the absolute excess kurtosis values, as was already concluded from Figure 3. However, when using the sum of the raw kurtosis the correct solution can be obtained with the fastICA code described in the Appendix, see Figure 6b. The raw kurtosis option is not available in the original fastICA code (#2000). Similarly, the sum of the skewness values reaches the correct solution, as shown in Figure 6c.

A more detailed picture of the behavior of the kurtosis and skewness is shown in Figure 6d,e,f. This figure shows polar plots in the space of the first two principal components (PC 1 and PC 2) of the centered data with 1.12 blocks of zeros added. FastICA determines the rotation where the objective function, sum of absolute kurtosis, sum of raw kurtosis, and the sum of the (absolute) skewness values, has a maximum. The polar plots show the sum of the sum of the three objective functions as a function of the angle of rotation of the two PC's.

Figure 6d shows the polar plot of the sum of the absolute kurtosis values. This gives the incorrect solution at 25 degrees, as the rrssq value in Figure 6a shows. Figure 6e shows the polar plot of the sum of the raw kurtosis values, which results in the correct solution at 71 degrees. Figure 6f shows the polar plot of the sum of the skewness values, which again shows the correct results at 71 degrees. The correct solution shows again the by now familiar Gaussian profiles and associated contribution profiles and is therefore not shown.

Comparing Figure 6d with 6e,f shows that the maximum in Figure 6e is a local minimum in Figure 6d. This is the effect of using the raw kurtosis instead of the absolute kurtosis.

Although the maxima in the solutions in Figure 6e,f are identical, it is clear that the sum of the absolute skewness value has the preference. The maxima in Figure 6f are better defined and this also causes the maximum to be found faster with iterative solutions, as was confirmed with tests. A comparison of the running time of the three objective functions for the braze data set, with the sum of absolute kurtosis defined as 100, we get 93 for raw kurtosis and 89 for skewness. These values are based on the average of 100 runs.

The conclusion is that when the objective is to find a sparse solution the skewness is the best objective function for fastICA.

Summarizing the behavior of homeopathic ICA for data with only pure components and data with mixtures in addition to the pure components the following.

a) Data with mutually exclusive components cannot be properly resolved with fastICA. After adding more and more zeros the contributions converge towards independence (and, as a consequence, with zero correlation). Using skewness as a measure of sparsity fastICA is able to resolve the data properly after adding a high number ('infinite') number of zeros.
b) Data with mutually exclusive components and mixtures of the components cannot be solved with fastICA. After adding zeros the contributions of the pure components reach a point where their correlation is zero without independence (adding more zeros the correlation deviates again from zero). At this point a rotation of PCA, which is achieved by fastICA, will resolve the data properly when sparsity is achieved through using skewness as the objective function.

4.3.Diagnostics for correct solution

With the examples of the simulated solutions we knew the solution. However, in practical cases the correct solution is not known. Therefore, a diagnostic is needed to determine the proper number of blocks to be added. Figure 7a,b,c shows a plot of the length scaled rows of matrix **S** (Equation 7) of the data set **X5000** for different numbers of blocks added to the data set, where Figure 7b show the best solution in this series. The results vary under-resolved to over-resolved IC's, contributions, as a function of the number of blocks of zeros added. The skewness was used as the objective function.

As a first step for the calculation of diagnostic values the data in Figure 7a,b,c the absolute values are calculated of matrix **S** and the columns are normalized to a sum of 1.

$$z_{i,j} = \frac{|s_{i,j}|}{\sum_{i=1}^{ncomp}|s_{i,j}|} \tag{13}$$

The results are shown Figure 7d,e,f. The absolute values are needed for the calculation for the diagnostics and to eliminate the sign ambiguity of multivariate methods.

In Figure 7g,h,i the reciprocal values of the quarter-norm values of all the columns of S are shown where the norm of an array x is defines as

$$\|x\|_p = (\sum |x|^p)^{1/p} \tag{14}$$

For the quarter norm p =1/4. For normalized data the maximum is one.

Comparing the graphs in Figure 7g,h,i show that the reciprocal quarter norm of the columns shows significantly higher values for the columns in **S** where pure components are present. Even minor deviations, such as in the first part of Figure 7g, show significantly lower values.

A summary of these values can be calculated by calculating the inverse of the norm of the unfolded matrix of **Z** symbolized by z.

$$diagn = \frac{1}{\left(\frac{\sum_{k=1}^{ncomp \times nspec} z_k^{1/4}}{ncomp}\right)^4} \tag{15}$$

The scaling factor ncomp results in a maximum value of one. Therefore, the values of the diagnostic are between zero and one. For a noiseless dataset with only mutually exclusive pure components the diagnostic value will be one.

The presence of noise will make these values lower. The presence of mixtures will also lower the values. The maximum in the diagnostic value as a function of the number of blocks of zeros added to a data set will indicate the best solution of homeopathic ICA.

Equation 15 was tested on a limited number of data sets, as will be shown below. Some fine tuning may result from testing more data sets.

A Matlab program implementing this diagnostic is describe in Appendix A. Examples of this diagnostic measure are shown below.

4.4.1. Simulated data

Figure 8 shows the diagnostic curves resulting from Equation 15 for the simulated data sets. For fastICA the skewness was used, which is an option of the Homeopathic_ICA code described in the Appendix. It is not an option of the original fastICA code. As we have seen in Figure 3 there is no correct solution for **X0** in the range shown. For **X600** the correct solution (within the resolution of the plot) is after adding 5 blocks of zeros and for **X1400** the correct solution is after adding 3 blocks of zeros. Figure 6 shows that the correct solution got **X5000** is approximately after adding 1.2 blocks of zeros (plotted with a finer resolution the maximum is at 1.1, which we used in Figure 7b). In Figure 8 a horizontal line indicates the diagnostic value for adding in infinite number of zeros, calculated from the contributions resulting from the quartimax rotation (Equation 5). It is important to notice that the diagnostic values are significantly different from the quartimax values as an additional diagnostic feature.

As is clear from Figure 8 the diagnostics of Equation 15 indicate the correct solution with a maximum in the diagnostic values. It is also clear that the overall diagnostic values of mixtures decrease when more mixtures are present.

4.4.2 Actual data

4.4.2.1 Braze data and CuNi data

The next test of this diagnostic is the actual data sets described in the previous paper (#1039). The braze data set was given as an example of data with only pure components. Thus, we expect the diagnosis to keep increasing with adding zeros, as we have seen for the simulated data in Figure 8a. However, it appeared there was a maximum for the braze data set. The same is true for the CuNi (not shown). This is not so surprising. Even when there are only areas of pure components due to the observation area of each pixel two components will be registered on interface areas. It is not clear, though, if the interface mixtures are due to an instrument artifact, due to the size of the observation area, or if there are indeed some interactions.

Some minor differences can be observed between the solution at the maximum and the quartimax results. Furthermore, the maximum diagnostic value and the diagnostic value for the quartimax solution are very similar. Since we do not have additional information about these data sets it is not possible to judge the minor differences objectively.

The conclusion that data sets with only pure components may not be as pure as one would expect is an important one. This leads to the important realization that, since data sets with only pure components may be rare, one must be careful with expecting that adding more blocks of zeros will lead to increasingly better solutions. An algorithm that includes a diagnostic value as described here will likely help to determine the optimal fastICA solution and avoid under and over-resolved results.

4.4.2.2. NMR data of binary mixture

The results of this sample are originally 256×256. In order to increase the relative amount of mixtures the top 100 rows and the bottom 70 rows and the background part of the image were deleted, resulting in an 86×92 pixel image. There is a maximum at 1.9 blocks of zeros, see Figure 9a. The maximum of the diagnostics curve is clearly higher than the quartimax value, which was not the case for the braze and CuNi data mentioned above. For this data set additional information available. The theoretical T_2 values of the first and second component are 150 and 29.6, respectively. The results of expressing this data set in terms of exponential profiles (#805) resulted in T_2 values of 137.8 and 28.2, respectively. The T_2 values of the IC's extracted by homeopathic ICA as a function of the number of blocks added was calculated from the ratio of successive points in the exponential curves (#805) and are shown in Figure 9b,c. The T_2 values of the quartimax solution were also calculated and indicated in Figure 9b,c. The maximum of the diagnostic values indicates solutions close to or within the previously calculated T_2 values, while the T_2 quartimax values are not close to the T_2 reference values. This indicates that the diagnostic value is a promising measure to determine the proper number of blocks of zeros that need to be used in homeopathic ICA.

Result between the solution at 1.9 blocks of zeros and the quartimax solution are shown Figure 10 and

11.

Figures 10a,b,c are the first resolved image of $NiCl_2$.The side profile in Figure 10a shown the pixel values of all the rows of the image in Figure 10b and their mean. In Figure 10c the T_2 profile of NiCl2 is shown, together with a reference profile based on the known T_2 value. Figures 10d,e,f shows the resolved results of $MnCl_2$. Although the $MnCl_2$ deviates somewhat from the reference profile, the results clearly show a successful resolution of the data. The results could not be improved by analyzing the whole image.

In Figure 11 the results of the quartimax solution are shown. There are clear differences. The images of the solution clearly show negative intensities, typical for over resolved results, as was shown and discussed before in Figure 4h. The resolved T_2 profiles in Figure 11c,f clearly show signs of under-resolved results, shown in Figure 4d. Comparing Figure 11c with 11d shows that the T_2 curve of $NiCl_2$ in Figure 11c has a faster decay because of a small contribution of the faster decaying $MnCl_2$. Vice versa, the $MnCl_2$ in Figure 11f has a slightly slower decay than the decay in Figure 10f. This example clearly shows the behavior discussed for the simulated mixtures and the advantage of applying homeopathic ICA with an optimal number of blocks of zeros added.

A program to determine the optimal number of blocks to resolve the data, using the diagnostic value described above, is described in the Appendix under ICA_optimize.

The robustness of the diagnostic value was tested by analyzing different versions of the NMR data set. The first sample analyzed is the complete image of 225 rows. The mixture data are not exactly in the middle of the image, therefore 35 rows of the NiCl2 compartment were deleted to obtain the largest possible symmetrical image of 190 rows. After that a succession of symmetrical image data were analyzed which each one 10 fewer rows that the previous image up to the last image with 10 rows. The results are shown in Figure 12.

The graphs shown are the T_2 values of the exponential curves resulting from homeopathic ICA solution and also the T_2 values resulting from the quartimax solution.

The figure show that the solutions, in terms of the T_2 value calculated from the exponential decays resulting from homeopathic ICA, is stable up to about 90 rows. After this point the T_2 values start to deviate from the expected values. This is where there are no pure components are present anymore and where homeopathic ICA cannot find a proper solution anymore.

The curves resulting from the quartimax solution are close to a solution within the known T_2 values for the largest image, which is not so surprising because that is the image with relatively the fewest mixtures. When the images get smaller, and relatively more mixtures are present, quartimax starts to deviate more and more from the expected T_2 values

The results of the images in Figures 10 and 11 were done on the image of 86 rows, which is on the edge of the stable area. This 86 row image was chosen to show a clear difference between the homeopathic ICA solution and the quartimax solutions.

5. Tolerance for mixtures

Finally, the question is what the limits are for the number of observations of pure components in a mixture data set in order to resolve it properly. To study this a data set was created with 1000 mixture samples of two components. A varying number of observations of the same of two components in pure form were added. Noise was added as discussed before for the simulated data sets. The noise was different for all the data sets in the series. The optimal number of blocks of zeros to be added to get the best result according to the diagnostics in Equation 15 were calculated. The rrssq values (Equation 8) of the optimal solutions were determined to judge the quality of the solution. See Figure 15.

The results show that good results (rrssq ≤ 0.02) after adding about 120 observations of pure components.. The total number of spectra in that case was 1100, so the percentage of observations of pure components was 11%. Of course, these are simulated data sets, so the value of this number is limited. Nevertheless, this shows that the number of observations of pure components that need to be present for Homeopathic ICA is limited. This is confirmed by the results in Figure 11, where the profile if the resolved images show a relatively small area of pure components.

6. Concluding remarks

It was shown above that the presence of mixtures, next to pure components, can be resolved by homeopathic ICA after adding the proper number of zeros to the data set. A diagnostic value was shown to be a promising tool to indicate the proper number of zeros to be added. This can be done automatically using an optimization procedure in a program, which is described in the Appendix. Although for data of samples with only pure components present quartimax can be used to obtain the proper solution, even for such data mixtures may be present due to the observation area of the analytical technique. A procedure using the solution determined by the diagnostic value will give likely the correct and better results. For data sets with more mixtures, such as the MRI data, the optimal results are clearly better than the quartimax results. Therefore, such a technique may be a good general application for data sets where pure components are present next to mixtures. One may wonder what alternative techniques are available. One such technique is MCR with contrast (#994, #1938). An ICA based technique has the advantage that negative intensities are dealt with properly. For noisy data, the resolved data where intensities are (close to) zero the noise is likely to cause negative intensities, since the noise cannot be properly modelled with a limited number of factors. As was shown in our previous paper, the non-negativity constrained MCR technique may cause problems (#1039).

Conflict of interest

The authors declare no conflict of interest.

Figure captions

Figure 1. The raw, absolute kurtosis values and the skewness values of (a) a uniform distribution, (b) a normal distribution and (c) an example of a sparse distribution.

Figure 2. A schematic representation of the phantom sample used for the NMR two component sample.

Figure 3. Diagnostic values during process of adding zeros. The correlation of the known contributions (with added zeros) is shown in (a,b,c) for the data sets shown in title. The rrssq value (Equation 8) shows how well resolved contributions are reproduced by ICA. In (d,e,f) the MI values are shown of the known contributions and the sum of the absolute kurtosis values. The lower horizontal line in (a, b, c) is at zero correlation for reference. The higher horizontal line, representing the limit for the correlation values, is calculated according to Equation 12.

Figure 4. The resolved spectra are shown in (a,b,c) and the contributions (without added zeros) in (c,d,e). In in (a) and (e) the results correspond to the 2^{nd} data point (adding 1 block of zeros) in Figure 3a. In (b) end (f) the results are shown for the 6^{th} data point (adding 5 blocks of zeros) in Figure 3a. In (c) and (g) the results of the 6^{th} (adding 5 block of zeros) data point in Figure 3b is shown and in (d) and (h) the results of the 21^{st} data point in Figure 3c are shown.

Figure 5. The histograms of the resolved contributions in Figure (4) (with added zeros).

Figure 6. The rrssq values (Equation 8) as a function of number of blacks of zeros added for (a) the sum of the excess kurtosis values (Equation 1) as the objective function in ICA, (b) the sum of the raw kurtosis values (Equation 2) and (c) the sum of the skewness values (Equation 3). The vertical lines at 1.12 blocks of zeros added indicated where the correct solution can be found.
In (d), (e) and (f) a plot of values the same objective functions plotted as a function of rotation angle of the principal components for the data set with 1.12 blocks of zeros added.

Figure 7. The resolved results for **X5000** are shown in (a,b,c) for a number of blocks of zeros added as indicated in the title. In the corresponding (d,e,f) the normalized absolute values are shown. In (g,h,i) the inverse quarter norm values are shown with the diagnostic values (Equation 15) shown in the titles.

Figure 8. A plot of the diagnostic values as a function of the number of blocks of zeros added for the simulated data sets. The horizontal line indicates the value the curves converge to. These values are calculated on the resolved results obtained with the quartimax rotation (Equation 5).

Figure 9. The plots of the diagnostic values of the NMR data set. In (a) the diagnostics values are shown, with the dotted horizontal line indicating the diagnostic value of the quartimax results. The vertical line in the three plots indicates the number of blocks of zeros to be added, 1.9, according to the diagnostic value, to obtain a proper solution. In (b) the T_2 values of first the solutions are plotted together with the dashed lines indicating the theoretical value of 150 and the previously found solution of 137.8 (#805). The dotted horizontal line indicates the quartimax value of 125.7, which is the value the graph converges to. Similarly, (b) shows the results for the second component with the theoretical value for T_2 of 29.6, the previously calculated value of 28.2 and the quartimax value of 36.2.

Figure 10. A plot of the resolved results obtained from the NMR data set with 1.9 blocks of zeros added. In (a) and (d) on overlay plot of the columns of the resolved images in (b) and (e) are plotted, together with their mean. The resolved T2 values with references are shown in (c) and (d).

Figure 11. A plot of the resolved results obtained from the NMR data set with quartimax. This is equivalent to adding an infinite number of zeros. In (a) and (d) on overlay plot of the columns of the resolved images in (b) and (e) are plotted, together with their mean. The resolved T2 values with references are shown in (c) and (d).

Figure 12. The rrssq values of a data set with 1000 mixtures as a function of the number of pure components added.

Figure 13. The tolerance for mixtures. The x-axis shows how many pure components were added to a dataset of 1000 mixtures. The y-axis shows the rrssq values of the resolved contributions and the contributions used for the simulations.

References

[#1039]W. Windig, M.R. Keenan, Homeopathic ICA: A simple approach to expand the use of independent component analysis (ICA), Chemom. Intell. Lab. Syst. 142 (2015) 54-63

[#1007]AA. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, John Wiley & Sons, New York, NY, 2001.

[#2000] Helsinki University of Technology, Laboratory of Computer and Information Science, P.O. Box 5400 FI-02015 HUT FINLAND. Algorithm is described in (#978) http://research.ics.aalto.fi/ica/fastica/ (last accessed October 2016)

[#979]P. G. Kotula, M. R. Keenan, J. R. Michael, Automated Analysis of SEM X-Ray Spectral Images: A Powerful New Microanalysis Tool, Microsc. Microanal. 9 (2003), 1– 17.

[#698]W. Windig, The use of second-derivative spectra for pure-variable based self-modeling mixture analysis techniques,Chemom. Intell. Lab. Syst. 23 (1994), 71-86.

[#978]A. Hyvärinen, E. Oja, Independent Component Analysis: Algorithms and Applications, Neural Networks, 13 (2000) 411-430.

[#1013]C.S. Stork, M.R. Keenan, Advantages of Clustering in Phase Classification of Hyperspectral Materials images, Microscopy and Microanalysis, 16 (2010) 810-820.

[#805] W. Windig, J.P. Hornak, B. Antalek, Multivariate Image Analysis of Magnetic Resonance Images with the Direct Exponential Curve Resolution Algorithm (DECRA), Part 1: Algorithm and Model Study, JMR, 132 (1998) 298-306.

[#1010]Aapo Hyvärinen, Jarmo Hurri, Patrick O.Hoyer, A Probabilistic Approach to Early Computational Vision, Series: Computational Imaging and Vision, Vol. 39, Springer Verlag. 2009. (Free preprint version: http://www.naturalimagestatistics.net/nis_preprintFeb2009.pdf, last accessed October 2016)

[#1036] H. Parastar, M. Jalali-Heravi, Roma Tauler, Is independent component analysis appropriate for multivariate resolution in analytical chemistry? T.R.A.C., Vol. 31, 2012

[#3000] J.E. Jackson, A User's guide to principal components, John Wiley & Sons , 1991 New York, pp 161.

[#994] W.Windig, M.R. Keenan, Angle-Constrained Alternating Least Squares, Appl. Spectr. Volume 65, Number 3, 2011, 349-357

[#1038]W. Windig, J. M. Shaver, M. R. Keenan, B. M. Wise, Simplification of alternating least squares solutions with contrast enhancement, Chemom. Intell. Lab. Syst. 117 (2012) 159?168

#1064 Aapo Hyvärinen and Erkki Oja

Independent Component Analysis: Algorithms and Applications
Neural Networks, 13(4-5):411-430, 2000

#1022
Thomas M. Cover, Joy A. Thomas, Elements of Information Theory, John Wiley & Sons, New York, 1991

Appendix

1) The function simple_fastica performs homeopatic ICA and is the core of this package.

```
I/O:  [S,A,W] = simple_fastica(D, zmult, ICAcriterion)

       D is the ncomp x npixel data/scores matrix
       zmult is the number of zeros to add as a multiple of npixel
       ICAcriterion: { 1='abskurtosis', 2='rawkurtosis', 3='skew' }
       S contains the ncomp estimated independent components
       A is the mixing matrix
       W is the unmixing matrix
```

2) The function Homeopathic _ICA is a user friendly function that performs commonly used pre- and post processing related to the proper use of simple_fastica.

```
Homeopathic_ICA performs independent component analysis with novel options
[S,A]=Homeopathic_ICA (data, ncomp, z_mult, ICA_criterion, flag_poisson, offset);
The function calls simple_fastica. The function takes care of the data
reduction, changes sign of output to obtain positive results and has
Poisson scaling as an option.
INPUT: data: data matrix, independent components have to be in rows
      ncomp: number of independent components
      zmult (optional): number of zeros to add as multiple of # of columns,
          may be a fraction this is important for homeopathic ICA (1),
          default is 0.
      ICA_criterion (optional): 1 uses absolute value of kurtosis, often
          used as a default criterion, 2 uses raw kurtosis, which has
          advantages for sparse data and 3 uses skewness, default is 3
      flag_Poisson (optional): performs Poisson scaling, 0 no Poisson
      scaling, 1 Poisson scaling, default is 0.
      offset (optional): offset for Poisson, down weigths variables with
          low intentisy, typical values 1-3, default is 0.
OUTPUTS:S contains the length scaled resolved independent components (rows)
        A contains the resolved columns
        data is reproduced as follows: A*S
Author: Willem Windig, windig@eigenvector.com
Functions used: simple_fastica
References:
  W. Windig, M.R. Keenan,
  Homeopathic ICA: A simple approach to expand the use of independent
  component analysis (ICA), Chemom. Intell. Lab. Syst. 142 (2015) 54-63
  W. Windig, M. R. Keenan and B. M. Wise
  The effects of pre-processing of image data on self-modeling image
  analysis, J. Chemometrics 2008; 22: 500-509
```

3) The function diagnostics_ICA calculated diagnostics values for ICA results

```
INPUT: data to be evaluated, independent components in rows
OUTPUT: diagnostics according to:
    W. Windig, B.M. Wise, M.R. Keenan,
    The role of sparsity and the behavior for mixture data in homeopathic
    independent component analysis (ICA), in preparation
Author: Willem Windig. windig@eigenvector.com
```

4) The function Homeopathic_ICA_optimize automatically determines the optimal number of blocks of

zeros to be added.

```
Homepathic_ICA_optimize finds the optimal number of zeros to add
[S,A]=Homeopathic_ICA_optimize(data,ncomp,ICA_criterion,flag_poisson,offset,range);
INPUT: data: data matrix, independent components have to be in rows
      ncomp: number of independent components
      range (optional): range of the number of blocks used for
      optimization. May have to be adjusted, for example when the optimum
      optimum is around 1 block of zeros to avoid a local maximum.
      Default is [0 10];
      ICA_criterion (optional): 1 uses absolute value of kurtosis, ofte
          used as a default criterion, 2 uses raw kurtosis, which has
          advantages for sparse data and 3 uses skewness, default is 3
      flag_Poisson (optional): performs Poisson scaling, 0 no Poisson
      scaling, 1 Poisson scaling, default is 0.
      offset (optional): offset for Poisson, down weigths variables with
          low intentisy, typical values 1-3, default is 0.
      range (optional): range of the number of blocks used for
      optimization. May have to be adjusted, for example when the optimum
      optimum is around 1 block of zeros to avoid a local maximum.
      Default is [0 10];
OUTPUT:S contains the resolved lengthscaled independent components (rows)
      A contains the resolved columns
      data is reproduced as follows: A*S
      diagnostics contains the diagnostic values during optimization. The
      first column contains the blocks of zeros added, the second column
      contains the diagnostic values. An useful plot is generated as
      follows:
      [x,index]=sort(diagnostics(:,1));y=diagnostics(index,2);plot(x,y,'-*');
Author: Willem Windig,windig@eigenvector.com
Functions used: Homeopathic_ICA
```

Figure 1

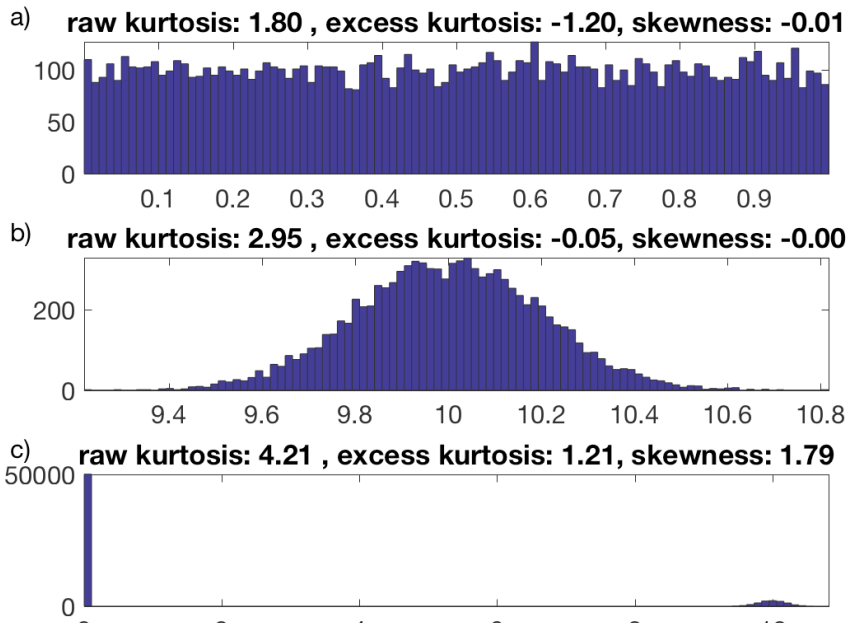


Figure 2

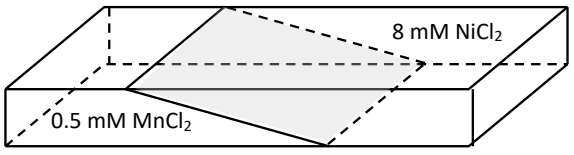


Figure 3

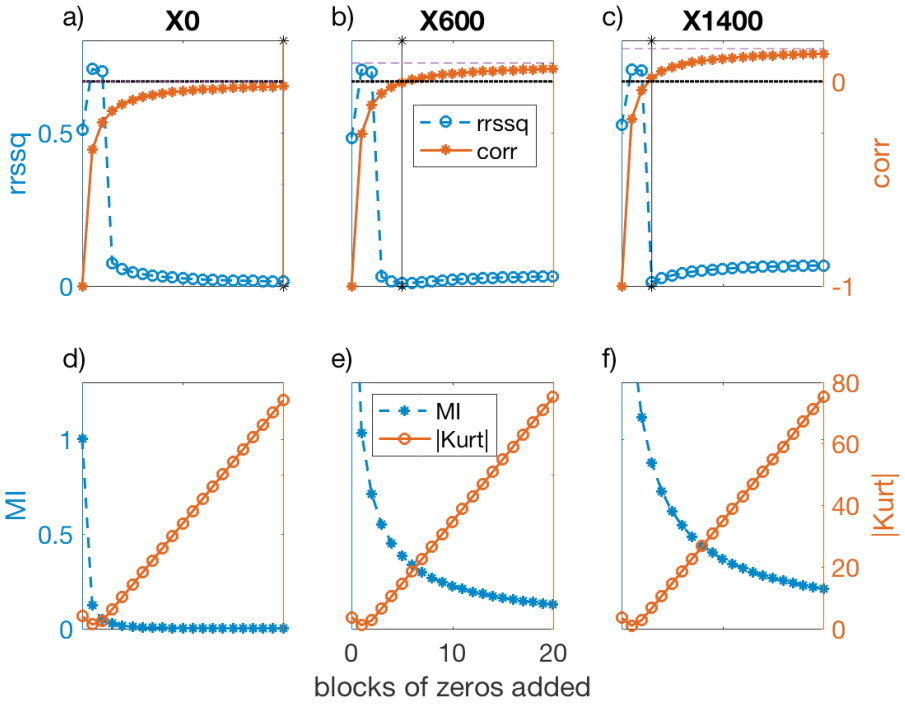


Figure 4

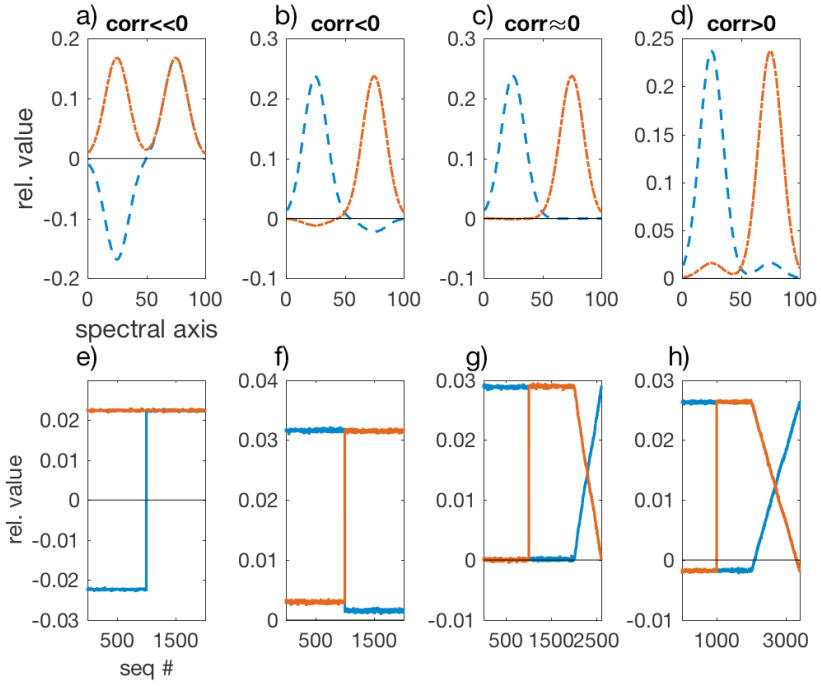


Figure 5

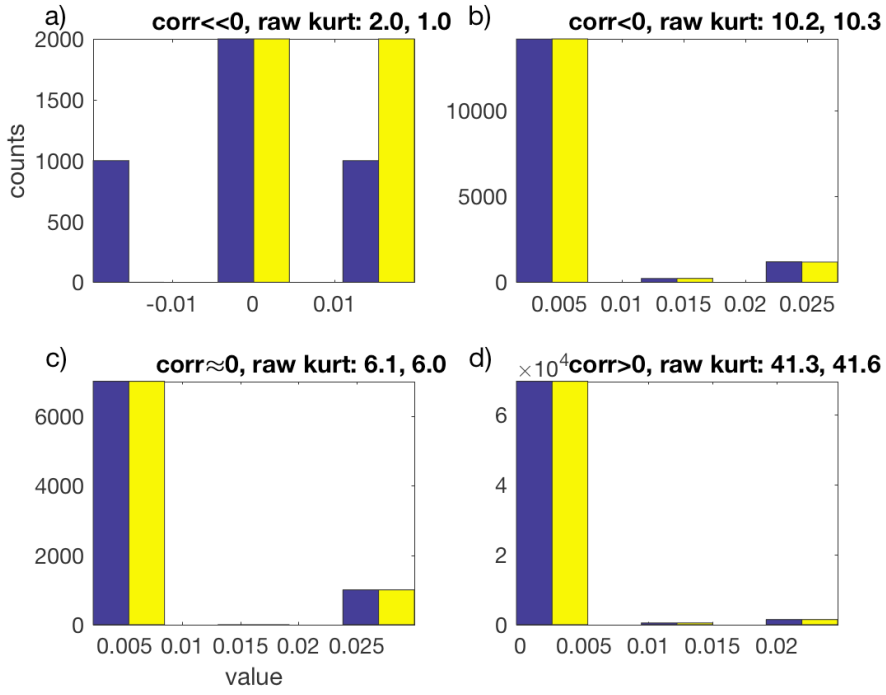


Figure 6

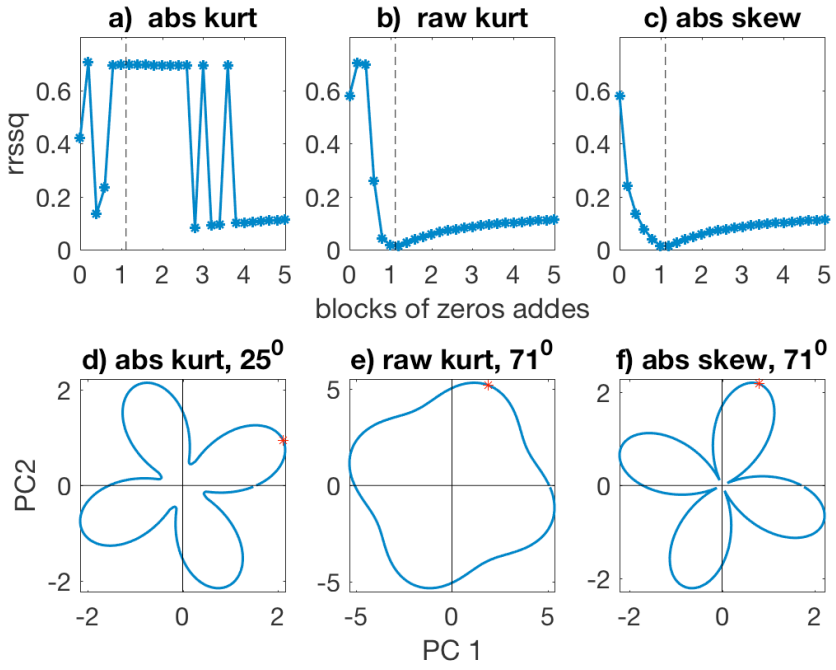


Figure 7

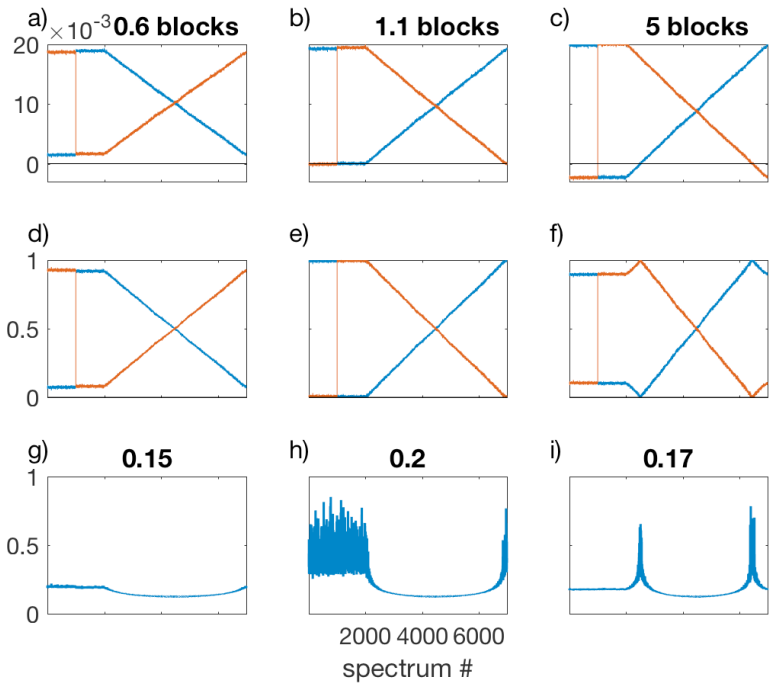
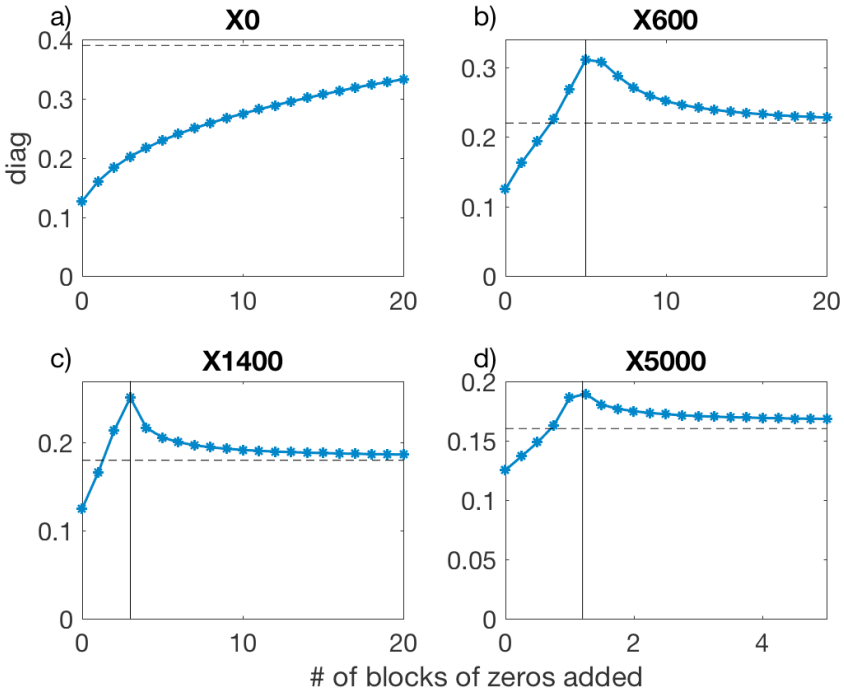


Figure 8



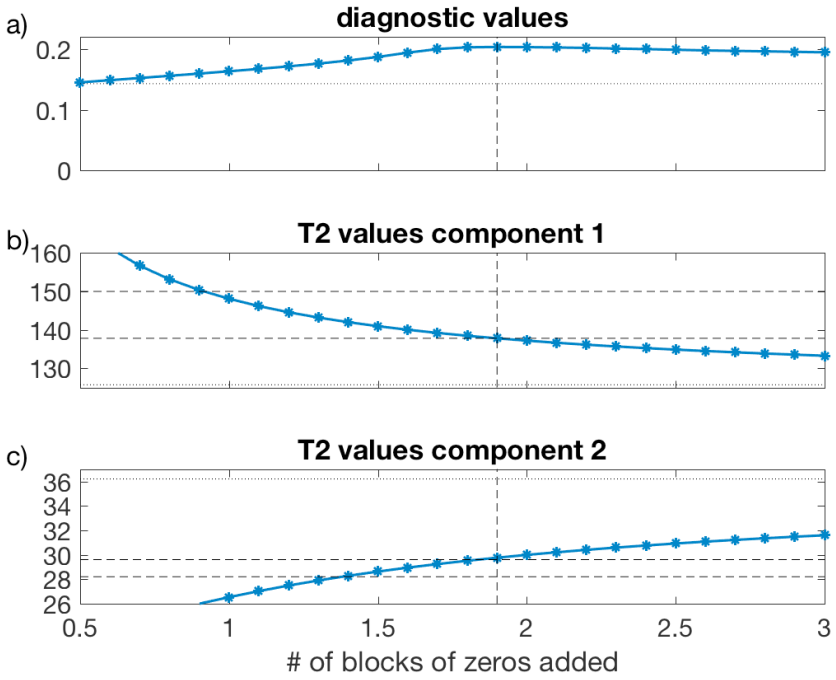


Figure 9

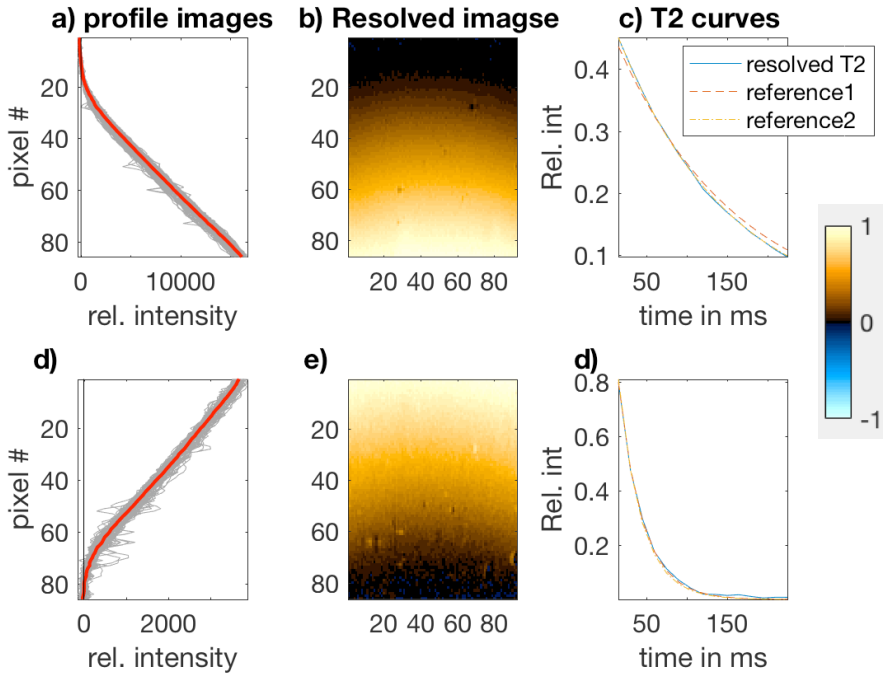


Figure 10

Figure 11

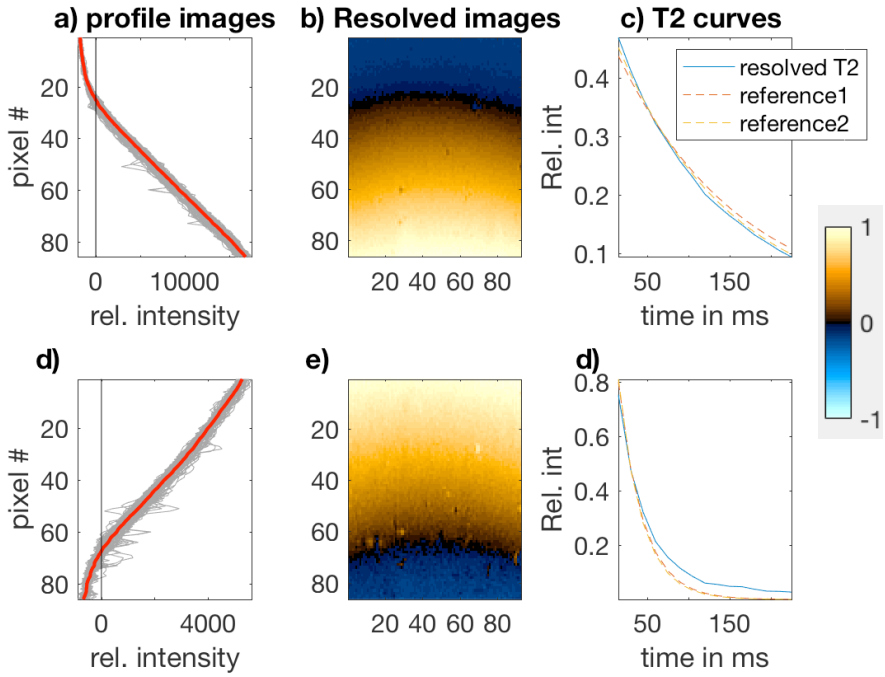


Figure 12

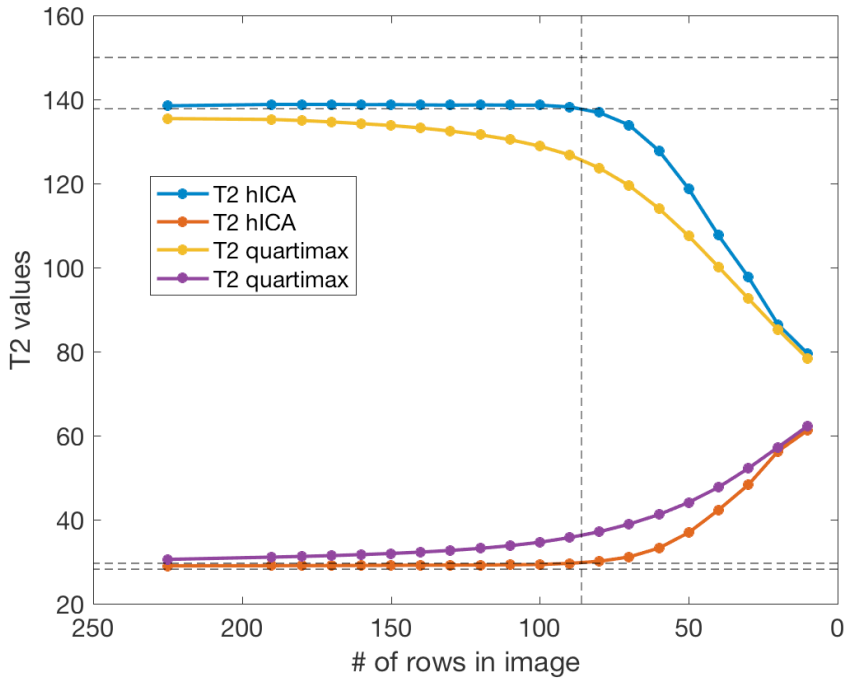
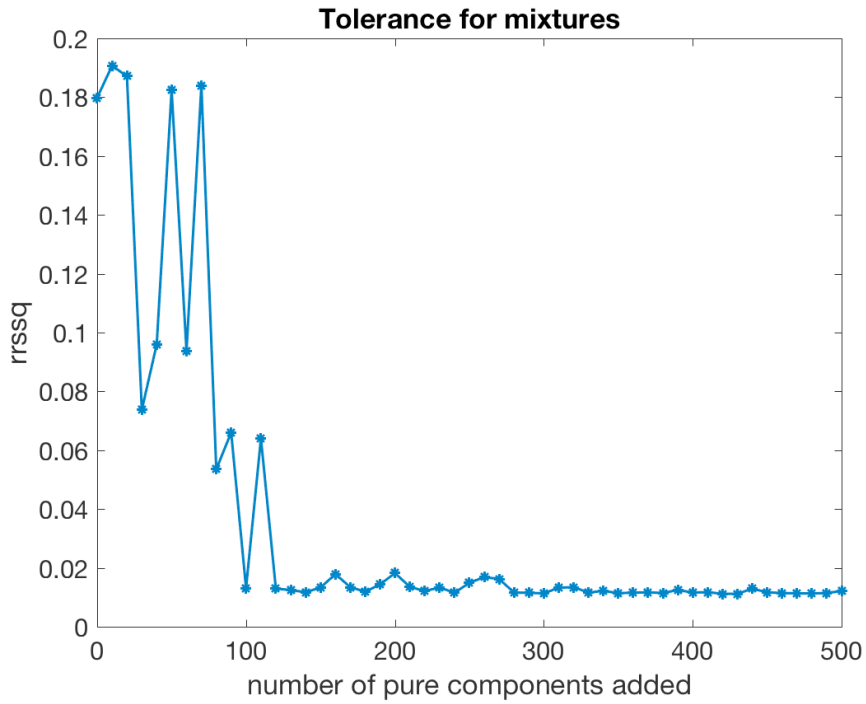


Figure 13


Tolerance for mixtures
rrssq
number of pure components added
0
0.02
0.04
0.06
0.08
0.1
0.12
0.14
0.16
0.18
0.2
0
100
200
300
400
500