ARTEFACT DETECTION IN ASTROPHYSICAL IMAGE DATA USING INDEPENDENT COMPONENT ANALYSIS

Maria Funaro, Erkki Oja, and Harri Valpola

Neural Networks Research Centre, Helsinki University of Technology
P.O.Box 5400, 02015 HUT, Finland
{*maria.funaro,erkki.oja,harri.valpola*}*@hut.fi*

ABSTRACT

This paper is the first reported application of ICA on astrophysical image data. When studying far-out galaxies from a series of consequent telescope images, there are several sources for artefacts that influence all the images such as camera noise, atmospheric fluctuations and disturbances, and stars in our own galaxy. For this problem, the linear ICA model holds very accurately, because the independence of such artefacts is guaranteed. Using image data on the M31 Galaxy, it is shown that several clear artefacts can be detected and recognized based on their temporal pixel luminosity profiles and independent component images. Once these are removed, it is possible to concentrate on the real physical events like gravitational lensing. ICA might provide a very useful preprocessing for the large amounts of available telescope image data.

1. INTRODUCTION

In modern astrophysics, one of the main research directions is understanding the dark matter in the universe. Possible candidates include compact objects such as small black holes, dwarf stars, or planets. When such an object passes near the line of sight of a star, the luminosity of the star will increase – an effect called gravitational lensing, predicted by the general theory of relativity.

In studying other galaxies than our own, individual stars cannot be resolved, but a whole group of unresolved stars is registered in a single pixel element of a telescope CCD camera. In a new technique called pixel lensing (see [1]), the pixel luminosity variations over time are monitored, and using these time series the lensing events can yet be detected even in the case of unresolved stars.

A problem in the analysis of the images and luminosity variations is the presence of artefacts. One of the possible artefacts are the resolved or individual stars between the far-out galaxy and the camera, which emerge sharply from the luminosity background. Other artefacts are cosmic rays and noise in the CCD camera. Separating these artefacts from possible physical events is one of the steps in the analysis of pixel lensing data.

The new idea proposed here is to use ICA for artefact detection and removal. This is motivated by the fact that for astrophysical data, the independence of the artefacts is often theoretically guaranteed, and also the linear mixing model holds exactly. This is an almost ideal application for ICA. The ICA technique has been quite successful in artefact removal for biomedical signals [7]. The difference in the basic set-up between the biomedical signals and the astrophysical data is that in the latter case, the signals are digital images. Up to now, there have been few applications of ICA on the global analysis of image data, with functional MRI imaging the most advanced one [4]. From a mathematical point of view, our problem has similarities with the fMRI analysis.

The contents of this paper are as follows: our basic ICA approach for the spatial - temporal image data is outlined in Section 2. Section 3 describes the test for pixel lensing data and gives the results. Some conclusions are drawn in Section 4.

2. ICA FOR TIME-VARYING IMAGES

In the astrophysical data, we have a number of digital images, recorded over consequent nights when the conditions are favourable, and carefully calibrated for geometrical and photometric alignments. Let N be the number of pixels in an image and T be the number of image samples. Let $\mathbf{X} = [x_{tn}]$ be the $(T \times N)$ data matrix whose rows are the individual images, stacked row by row into vectors, and whose columns are the single pixel luminosity time series, here simply called light curves. In this case, formally similar to functional neuroimaging, we have two possibilities for performing both ICA and the preceding PCA decorrelation / compression: spatial or temporal [6].

The *spatial* ICA model is

$$\mathbf{X} = \mathbf{AS} \tag{1}$$

This work was supported by the Centre of Excellence Program of the Academy of Finland, project New Information Processing Principles, 44886. The corresponding author is E. Oja

where $\mathbf{A}$ is an $(T \times M)$ mixing matrix and $\mathbf{S}$ is an $(M \times N)$ matrix whose rows are M independent source images, with $M \leq T$. The *temporal* ICA model is

$$\mathbf{X}^T = \mathbf{A}'\mathbf{S}^T$$

where $\mathbf{A}'$ is another mixing matrix[1] and the rows of $\mathbf{S}^T$ are the individual light curves.

For astronomical data, the temporal model is not feasible because the spatial dimension is very much larger than the temporal dimension and reliable estimation of matrix $\mathbf{A}'$ would be difficult in this case [3]. Moreover, the spatial model is quite natural because, for instance, the interference caused by fixed individual stars between the desired far-out objects and the camera are superimposed on all the images, and they are even in theory totally independent of the extragalactic events.

Let us write the spatial ICA model of eq. (1) in the more conventional vector form as

$$\mathbf{x}_n = \mathbf{A}\mathbf{s}_n. \tag{2}$$

Now $\mathbf{x}_n$ is the T dimensional vector representing the n-th pixel light curve through all the T images, and $\mathbf{s}_n$ is the corresponding source vector with independent components. Written as

$$\mathbf{x}_n = \sum_{m=1}^{M} \mathbf{a}_m s_{mn} \tag{3}$$

we see that the M columns of $\mathbf{A}$ or mixing vectors $\mathbf{a}_m$ can also be interpreted as "virtual light curves", whose linear combinations give the observed light curves $\mathbf{x}_n$. The mixing vector characterizes the temporal behaviour of the m-th source, while the source image $(s_{mn}), n = 1, ..., N$ characterizes the spatial behaviour over the pixel field. Both of these can be used to interpret the physical meaning of a given term in the sum.

For ICA analysis, we have chosen to use the FastICA algorithm [2, 3] because of its appealing convergence properties. Preliminary sphering of the data is recommendable to simplify the algorithm and to reduce noise. This means transforming the vectors $\mathbf{x}_n$ into $\mathbf{z}_n = \mathbf{V}\mathbf{x}_n$ such that the new vectors $\mathbf{z}_n$ have uncorrelated and unit variance elements. One of the methods to accomplish this is classical PCA. Computing the eigenvalues of the data covariance matrix gives indications about the number of sources to be used in the model. After whitening, the mixing model becomes

$$\mathbf{z}_n = \mathbf{V}\mathbf{x}_n = \mathbf{V}\mathbf{A}\mathbf{s}_n = \mathbf{W}\mathbf{s}_n$$

where the matrix $\mathbf{W}$ is orthogonal.

To compute matrix $\mathbf{W}$ by the FastICA algorithm, its individual columns $\mathbf{w}_i$ are updated by the iteration

$$\mathbf{w}_i := E\{\mathbf{z}g(\mathbf{w}_i^T\mathbf{z})\} - E\{g'(\mathbf{w}_i^T\mathbf{z})\}\mathbf{w}_i$$

[1] $\mathbf{A}'$ should not be confused with the transpose of $\mathbf{A}$

followed by orthonormalization of the matrix $\mathbf{W}$ after each updating step. Function $g(.)$ in the update rule is an odd nonlinear function and $g'(.)$ is its derivative. The choice of a suitable function is discussed in detail in [3]. In our case, the function was $g(u) = \tanh u$.

When $\mathbf{W}$ has been estimated, the original mixing matrix is approximated from $\mathbf{A} = \mathbf{V}^T\mathbf{W}$. The independent components are obtained from $\mathbf{s}_n = \mathbf{W}^T\mathbf{z}_n,\ n = 1, ..., N$.

3. EXPERIMENTAL RESULTS

The digital images used in our study have been recorded over 35 unevenly sampled nights at the MDM observatory [5] with the telescope McGraw-Hill pointed toward the M31 Galaxy, equipped with a CCD camera of 2048 $\times$ 2048 pixels. The exposure time was 6 minutes for each of the frames.

We tested the performance of ICA for artefact removal in some windows of 100 $\times$ 100 or 101 $\times$ 101 pixels, randomly taken in each of the four fields in which the CCD camera is divided. In this section we report the results obtained on two of them. To have an idea of how the CCD camera images look like after preprocessing, see Fig. 1a-b.

Note that the temporally averaged intensity has been subtracted from each pixel since only the changes of intensity are of interest for lensing events, not the intensity per se. Dark areas in Fig. 1a-b correspond to pixels with an unusually low intensity for those particular pixels although the actual intensity might be higher than in some other pixels which are shown white in the figure.

In our ICA model (2), $\mathbf{x}_n$ is now a 35-dimensional vector representing the n-th pixel light curve. The sample size N is $100 \times 100 = 10000$ or $101 \times 101 = 10201$, according to the area we are examining. In the preliminary whitening, we have also reduced the dimension of the whitened vectors $\mathbf{z}_n$ to 10. This choice is justified by the fact that for all the tests we have done, the first ten eigenvectors always contribute to more than 90% of the energy content. Thus, $\mathbf{A}$ is a 35 $\times$ 10 matrix, and we have 10 independent components or source images. The mixing vectors are 35-dimensional and they can be plotted as temporal light curves over the 35 observation times.

The first results concern the 100×100 image window, whose coordinates are those of Fig. 1a. Looking at the ten mixing vectors and the corresponding source images, we find four mixing vectors whose temporal profiles, shown in Fig. 2, have a conspicuous spike. These are characteristic of cosmic rays. When a cosmic ray hits a pixel at time t, the luminosity value of this pixel has a sudden increase and also some other neighboring pixels are influenced and have a similar behaviour. Note that in each of these four mixing vectors, the luminous peak on the temporal axis occurs at a different time. The occurrence of cosmic rays is further evidenced by the corresponding independent component im-

ages, shown in Fig. 3. The cosmic rays have occurred at different times and at different locations of the image field and are thus fully independent events. Note also that from the original camera images, an example of which is Fig. 1a, these cosmic rays are not easily detected.

One might question whether these mixing vectors and source images could be in fact artificial results of the statistical estimation process and not any real physical events at all. Now that the locations of the cosmic rays have been detected, it is easy to check the original light curves at these pixel locations. The light curves have the same spiky behaviour as the mixing vectors, proving that the artefacts are real.

Looking at the other six mixing vectors for this image window, we have found that two of them show systematic increase and decrease in time which could be related to possible physical events like gravitational lensing. The other mixing vectors, instead, represent noisy fluctuations and their independent component images do not show evidence of any particular structure. This kind of random fluctuations over the 35 separate observations are due to atmospheric disturbances and camera noise.

The second results concern the 101×101 image window, whose coordinates are those of Fig. 1b. Now we have three mixing vectors which obviously characterize resolved stars, because they show large fluctuations around a constant value. In particular, one of them has larger values compared to the other two. We show this mixing vector in Fig. 4a. Its independent component image, shown in Fig. 4b, shows a particular circular structure, which is characteristic of resolved stars. This star can be distinguished also in the original image (Fig. 1b), but in the independent component image it has been accentuated and separated from the rest of the star field.

Of the remaining seven independent components, one shows a bright line on the CCD camera (Fig. 5b); we do not know exactly what it is but it is probably another artefact. Because the mixing vector is just noise except for one of the observation times, where there is a strong luminosity spike, this might be a lighted object passing throuh the sky.

Again, we have three other mixing vectors which show increase and decrease in time, related to possible interesting astrophysical events. Concerning the remaining three mixing vectors, they show noisy fluctuations and their independent component images do not indicate any particular structure.

4. CONCLUSION

We claim here that standard linear ICA can be used to detect artefacts present in astrophysical data. To the best of our knowledge, this is the first application of ICA on this specific problem. Especially, when studying far-out galaxies, the possible artefacts caused by camera noise, atmospheric fluctuations or disturbances, or stars in our own galaxy, are all theoretically independent of the real physical pixel lensing effects that the investigators are interested in.

Because ICA is an unsupervised technique, qualitative and domain-dependent expertise is necessary for interpretation of the found independent components. To corroborate our claim, we showed that the mixing vectors which characterize cosmic rays and resolved stars have a well-defined profile and their corresponding independent component images indicate particular structures (Figs. 2, 3 and 4). Some other mixing vectors characterize the artefacts due to noisy fluctuations in the camera and the atmosphere, and their independent component images do not show any typical structure.

Using ICA, the large amounts of telescope image data could be preprocessed automatically to pinpoint the exact locations and occurrence times of possible artefacts and to accentuate the physical pixel lensing events. This affirmation is also confirmed by the further tests we are performing on other areas of the CCD camera.

5. REFERENCES

[1] A. Gould, The theory of pixel lensing, *Astrophys. J.* 470, 1996, pp. 201 - 210.

[2] A. Hyvärinen and E. Oja, A fast fixed-point algorithm for independent component analysis, *Neural Comp.* 9, 1997, pp. 1483 - 1492.

[3] A. Hyvärinen, J. Karhunen, and E. Oja: *Independent Component Analysis*. Wiley Interscience, New York, 2001.

[4] T. P. Jung, S. Makeig, T. W. Lee, M. J. McKeown, G. Brown, A. T. Bell, and T. J. Sejnowski, Independent component analysis of biomedical signals, *Proc. 2nd Int. Workshop on ICA and BSS*, June 19 - 22, 2000, Helsinki, Finland, pp. 633 - 644.

[5] MDM Observatory: `http://www.astro.lsa.umich.edu/obs/mdm/`

[6] K. S. Petersen, L. K. Hansen, and T. Kolenda, On the independent components of functional neuroimages, *Proc. 2nd Int. Workshop on ICA and BSS*, June 19 - 22, 2000, Helsinki, Finland, pp. 615 - 620.

[7] R. Vigario, J. Särelä, V. Jousmäki, M. Hämäläinen, and E. Oja, Independent component approach to the analysis of EEG and MEG recordings, *IEEE Tr. Biomed. Eng.* 47, 2000, pp. 589 - 593.

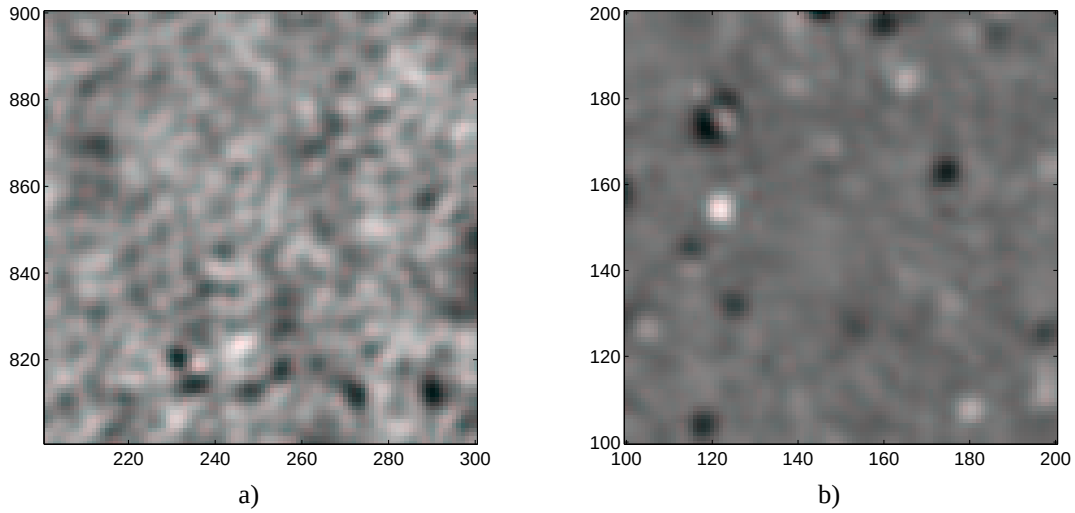

Fig. 1. a) Image of the pixel luminosity values for one of the 35 observation times in the fourth camera field, with coordinates $201 \leq x \leq 300$ and $801 \leq y \leq 900$. b) Another image of the pixel luminosity values for one of the 35 observations times, with the pixel coordinates $100 \leq x \leq 200$ and $100 \leq y \leq 200$.

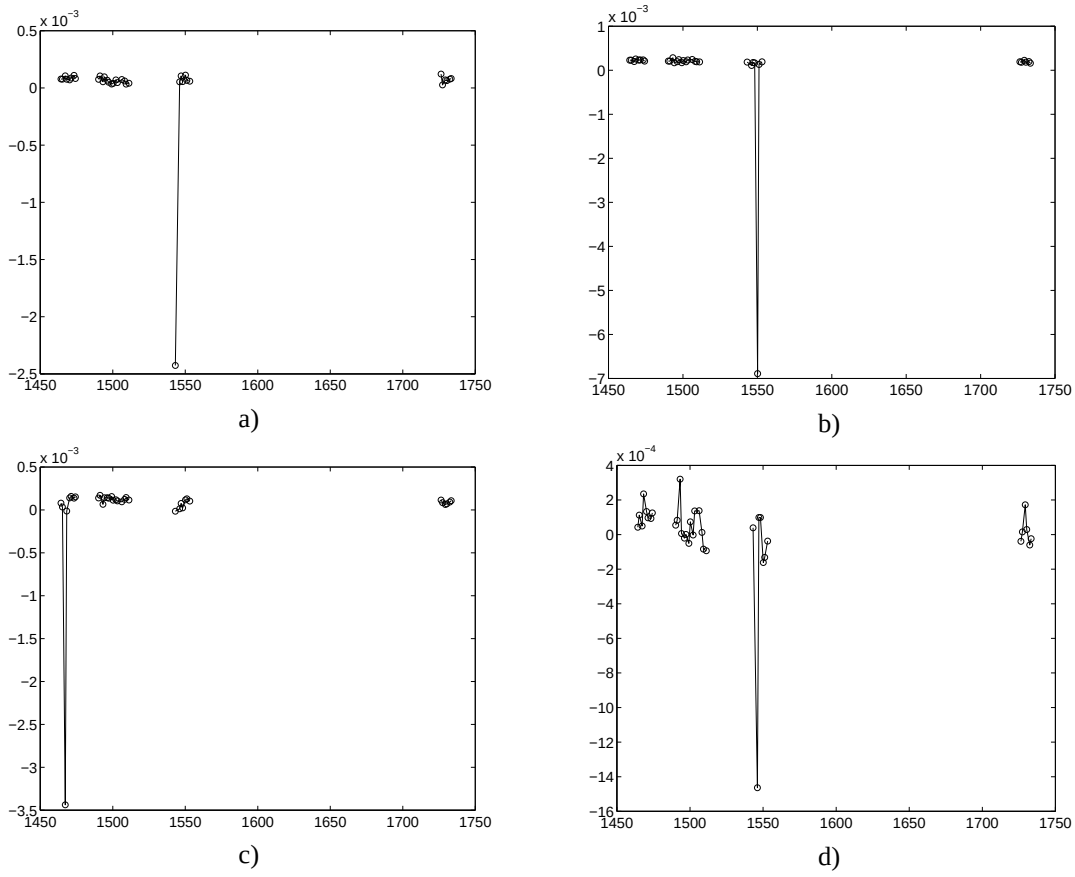

Fig. 2. The 35 elements of four mixing vectors. On the horizontal axis are the observation times. These profiles are typical for cosmic rays.

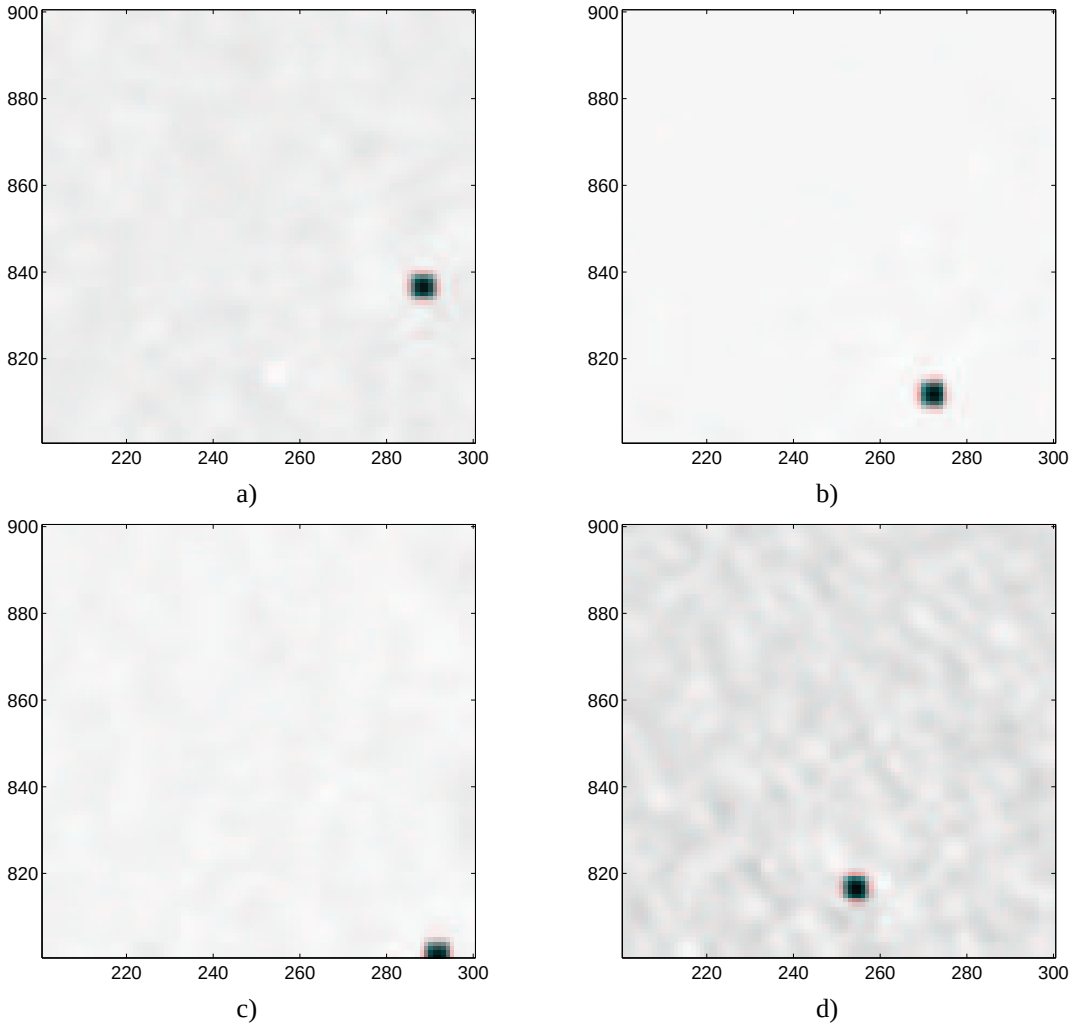

Fig. 3. Independent component images related to cosmic rays. Note that dark values here correspond to high intensity because the mixing vectors have their signs inverted. This can happen because the sign of independent components cannot be determined without prior knowledge about the physical setting.

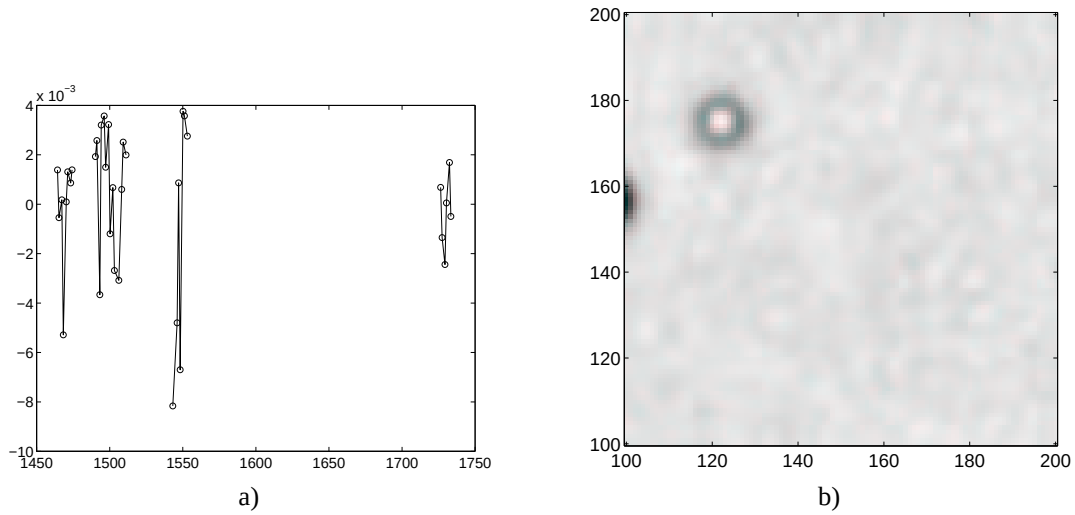

Fig. 4. One of the three mixing vectors (a) corresponding to a resolved star (b).

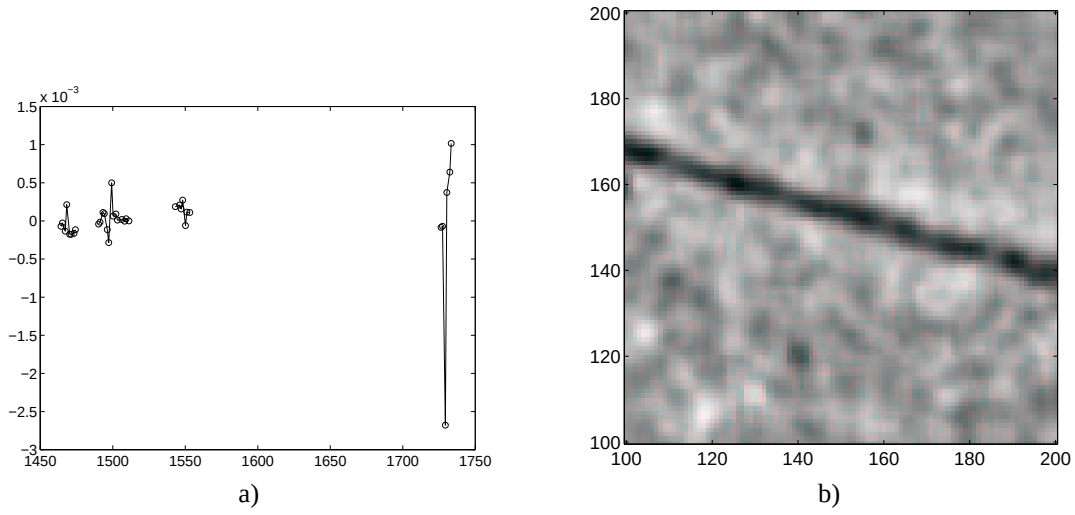

Fig. 5. The mixing vector (a) corresponding to a bright line (b). This is probably another artefact.