Non-negative factor analysis supporting the interpretation of elemental distribution images acquired by XRF

Matthias Alfeld[1], Mirwaes Wahabzada[2], Christian Bauckhage[2], Kristian Kersting[2,3], Gerd Wellenreuther[4], Gerald Falkenberg[1]

[1] Photon Science, DESY, Notkestraße 85, 22607 Hamburg, Germany

[2] Fraunhofer IAIS, Schloss Birlinghoven, 53757 Sankt Augustin, Germany

[3] Computer Science Department, TU Dortmund University, Joseph-von-Fraunhofer-Str. 23, 44227 Dortmund, Germany

[4] European XFEL GmbH, Albert-Einstein-Ring 19, 22761 Hamburg, Germany

E-mail: `matthias.alfeld@desy.de`

Abstract.
Stacks of elemental distribution images acquired by XRF can be difficult to interpret, if they contain high degrees of redundancy and components differing in their quantitative but not qualitative elemental composition. Factor analysis, mainly in the form of Principal Component Analysis (PCA), has been used to reduce the level of redundancy and highlight correlations. PCA, however, does not yield physically meaningful representations as they often contain negative values. This limitation can be overcome, by employing factor analysis that is restricted to non-negativity. In this paper we present the first application of the Python Matrix Factorization Module (pymf) on XRF data. This is done in a case study on the painting *Saul and David* from the studio of Rembrandt van Rijn. We show how the discrimination between two different Co containing compounds with minimum user intervention and *a priori* knowledge is supported by Non-Negative Matrix Factorization (NMF).

1. Introduction

XRF imaging is one of the most versatile techniques for the acquisition of elemental distribution images. It is used with different instruments for the investigation of a wide range of objects. These objects range from microscopic samples, which are investigated with a lateral resolution of less than a micrometer, to historical paintings, of which several square meters can be imaged with a resolution of a fraction of a millimeter.

If in these samples components are present that contain multiple elements detectable by XRF, a high level of redundancy can be observed, as the lateral distribution of one component would be represented in multiple elemental distribution images. Further, components differing in their quantitative, but not qualitative elemental composition can be difficult to discriminate. For many years, factor analysis has been used to support the interpretation of XRF data by reducing redundancies and the dimensionality of a data matrix, composed of raw spectral data or elemental distribution images.

The underlying problem is shown in Figure 1. Factor analysis aims at the decomposition of a matrix $\mathbf{V} \in \mathbb{R}^{d \times n}$ into two lower ranked matrices $\mathbf{W} \in \mathbb{R}^{d \times k}$ and $\mathbf{H} \in \mathbb{R}^{k \times n}$. Here, the
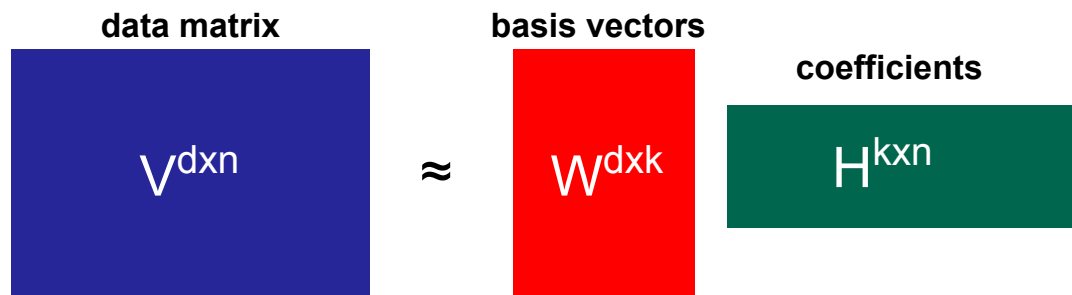
Figure 1. Factor Analysis: The data matrix $\mathbf{V}$ is described as the product of the basis vectors $\mathbf{W}$ and the coefficients $\mathbf{H}$.

matrix $\mathbf{V}$ represents an XRF data matrix, whose columns ($\mathbf{n}$) correspond to pixels represented by the recorded intensity of $\mathbf{d}$ sets of fluorescence lines of different elements. The basis vectors $\mathbf{W}$ describe the loading of the data set's variables on $\mathbf{k} < \mathbf{d}$ bases, while the coefficients in $\mathbf{H}$ represent the lateral distribution of the degree that bases describe the data matrix. They represent the coordinates of the input objects in a lower-dimensional space, spanned by the basis vectors in $\mathbf{W}$. A common approach for the factorization is Principal Component Analysis (PCA) that has been used for factor analysis in the interpretation of XRF data sets for a long time [1]. PCA is very effective in terms of data compression and noise reduction. Here, an approach is to compute the eigenvectors of the data covariance matrix, with largest eigenvalues as basis vectors $\mathbf{W}$. Here $\mathbf{W}$ is called the loading matrix and $\mathbf{H}$ the score matrix.

But although very effective, it has been criticized for the lack of interpretability of the resulting basis vectors. XRF data consists only of non-negative values and it is desirable that the resulting representation retains this property, which is not the case for PCA. To overcome this limitation other approaches constrain the resulting factorization to non-negativity. Multivariate statistical analysis with non-negative constrains is well known for energy dispersive X-ray spectroscopy with electron beams [2, 3]. In this paper we employ Non-Negative Matrix Factorization (NMF) [4]. It is worth noting, if $\mathbf{k} = \mathbf{d}$, PCA yields an exact representation of the data matrix $\mathbf{V}$ (we only rotate the data along principal components), while NMF only provides an approximation. However, while a good representation of the data set is desirable, the real criterion for the quality of the decomposition of the data set is how well the different components are represented and separated.

NMF is a widely used factor analysis method with non-negative constrains that requires also as input a matrix with non-negative values. It defines simple update rules for the matrices $\mathbf{W}$ and $\mathbf{H}$ with non-negative constraints, based on the squared Frobenius norm or Kullback-Leibler Divergence as objective functions (see [4] for more details). NMF results in a sparse and part based representation, which is more interpretable than a PCA decomposition. A comparison of basis vectors are shown in Figure 2, where we applied these methods on the CBCL face database[1].

Each image represents a basis vector (column in matrix $\mathbf{W}$). Using NMF on this test data set aims at the decomposition of the faces into their different basic components (e.g. nose, ear, eyes). This follows from the constraints that allow only additive combinations and thus results in part based representations, as shown in Figure 2 (right). Therefore, they allow an easier interpretation of bases. In contrast, basis images found by PCA (left) can also contain negative entries (red parts) that are not easy to understand.

Other variations of NMF increase the interpretability using convexity constraints [5, 6].

[1] http://cbcl.mit.edu/software-datasets/FaceData.html

Figure 2. Resulting basis vectors (columns in **W**) for PCA (left) and NMF (right) from the CBCL face database. The data set is represented by a matrix, whose columns correspond to one particular face image. PCA basis vectors (eigenfaces) contain negative values (red parts) whereas those found by NMF do not.

Alternatively, one can compute a factorization by representing the data as combination of few actual data points (see [7] and references therein). This further increase the interpretability of resulting basis vectors (columns in matrix **W**), since they represent real data points from matrix **V**, and compared to NMF they are not part based.

In this paper we apply NMF to a XRF data set acquired on the painting *Saul and David*, which is shown in Figure 3. The painting is part of the collection of the Mauritshuis Museum in The Hague (NL) and is attributed to the studio of Rembrandt van Rijn (1606-1669) with a disputed attribution to the master himself. The investigation of the central part of the painting aimed on the distinction between original materials and materials used in later conservation treatments. The investigation of the painting has been described in detail in [8].

2. The painting

At the moment of the investigation (early spring 2011), the painting was covered by a thick layer of darkened varnish and possible overpaint and was being studied prior to eventual conservation treatment. The planning for this was not straightforward, as the extent of original paint still present in the painting was unknown, due to its complex history. Moreover, the authenticity of the curtain depicted in the area between the figures was questioned.

At an unknown moment in the past the painting was cut into two separate pieces. The reason for this is unknown. Based on dimensions listed in the auction records the two figures were rejoined between 1830 and 1869. The join between the two pieces of canvas is clearly discernible in the X-ray radiograph (XRR , see Figure 3). Examination with the stereomicroscope revealed, that much of the curtain visible in the background of the painting was added after the parts were rejoined to visually blend them together. Further, it was known from investigations using a light microscope and SEM-EDX studies of paint cross sections that the original paint is still present under the surface of the curtain. However, its exact extent was unknown as these samples only provided local information. The original paint was found to be rich in smalt, a blue pigment used in the 17th century, consisting of ground Co-rich blue glass.

Later added paint layers can be identified by the presence of anachronistic compounds, which were not used in the 17th century. This could not be done by conventional techniques such as XRR and Infrared reflectography as the lead white $((PbCO_3)_2Pb(OH)_2)$ rich ground that also contains black pigments and the dark, bone black and earth pigment (Fe and Mn oxides) rich pictorial layers yield only low contrast images with these methods [8, 9].
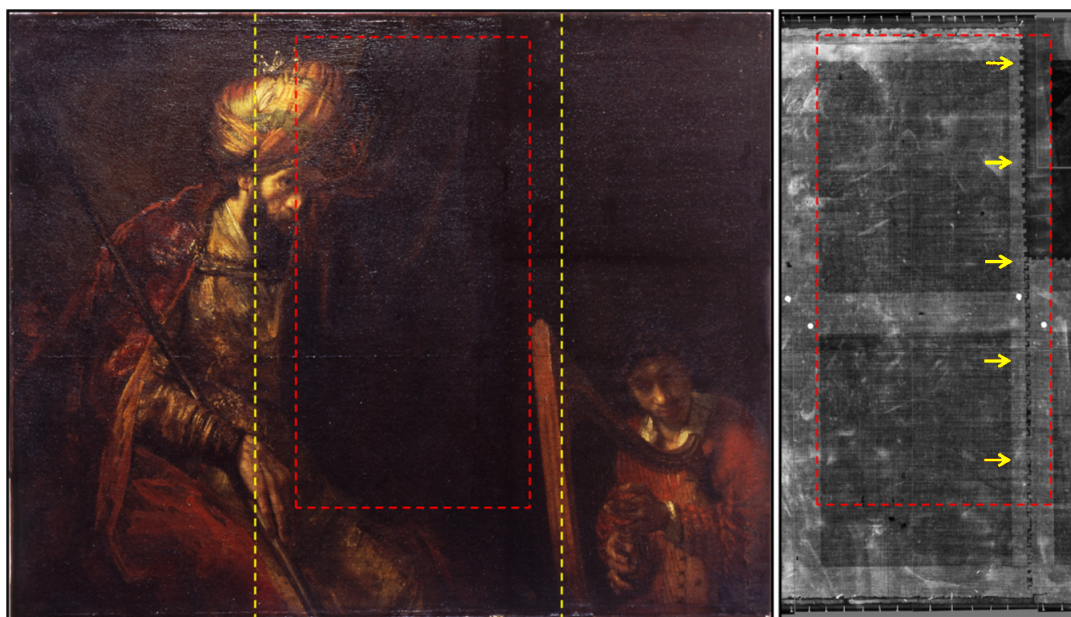
Figure 3. Rembrandt and/or Studio, Saul and David, c.1655, Mauritshuis inv. no. 621 / Bredius no. 526, oil on linen, 130.5 x 164.5 cm. In the XRR (right) the joins between the different pieces of canvas are clearly visible. The red dashed lines indicate the area investigated by MA-XRF, the yellow dashed ones the area of the painting shown in the XRR. The join is indicated in the XRR by yellow arrows.

3. Experimental

The painting was investigated in the conservation studio of the Mauritshuis Museum by means of a mobile XRF scanner. The scanner consists of a Mo-anode microfocus X-ray tube with a fixed polycapillary lens mounted with four energy dispersive SD-detectors on a motorized stage. This scanner allows for the investigation of an area of 60x25 cm^2 in a single scan. It is described in detail in [10] as "Instrument C". The painting was scanned with a step size of 1 mm and an average dwell time of 1.2 s per pixel, so that the central part of the painting (approx. 60x120 cm^2), which is discussed here, was investigated in roughly two weeks. The spectra of the four detectors were summed for each pixel prior to processing with the AXIL software package [11] and in-house written software.

4. Results

4.1. Elemental distribution images

In Figure 4 the elemental distribution images obtained from the central part of the painting are shown. Ca is present as a minor component in the black pigment bone black and as major constituent in chalk, which was used as a paint extender, and to fill the gaps in the join, probably in the form of chalk or gypsum. The distribution of Fe visualizes the earth pigments, which in general have a brownish/reddish color and were used in the original as well as in the later added paint. Since the curtain was mainly executed in these pigments, its folds are clearly visible in the Fe image. Mn is also present in earth pigments of darker tone. Cu is present in a green or blue pigment, which is present in the turban and to a minor amount in the ground of the painting. Pb is present in lead white. Hg is present in the red pigment vermilion (HgS) and was used in Saul's turban and to cover the join. Ti, Cr, Zn and Ba are mainly present in anachronistic materials, used to cover the joins. Cd is present in low abundance and it is unclear whether it
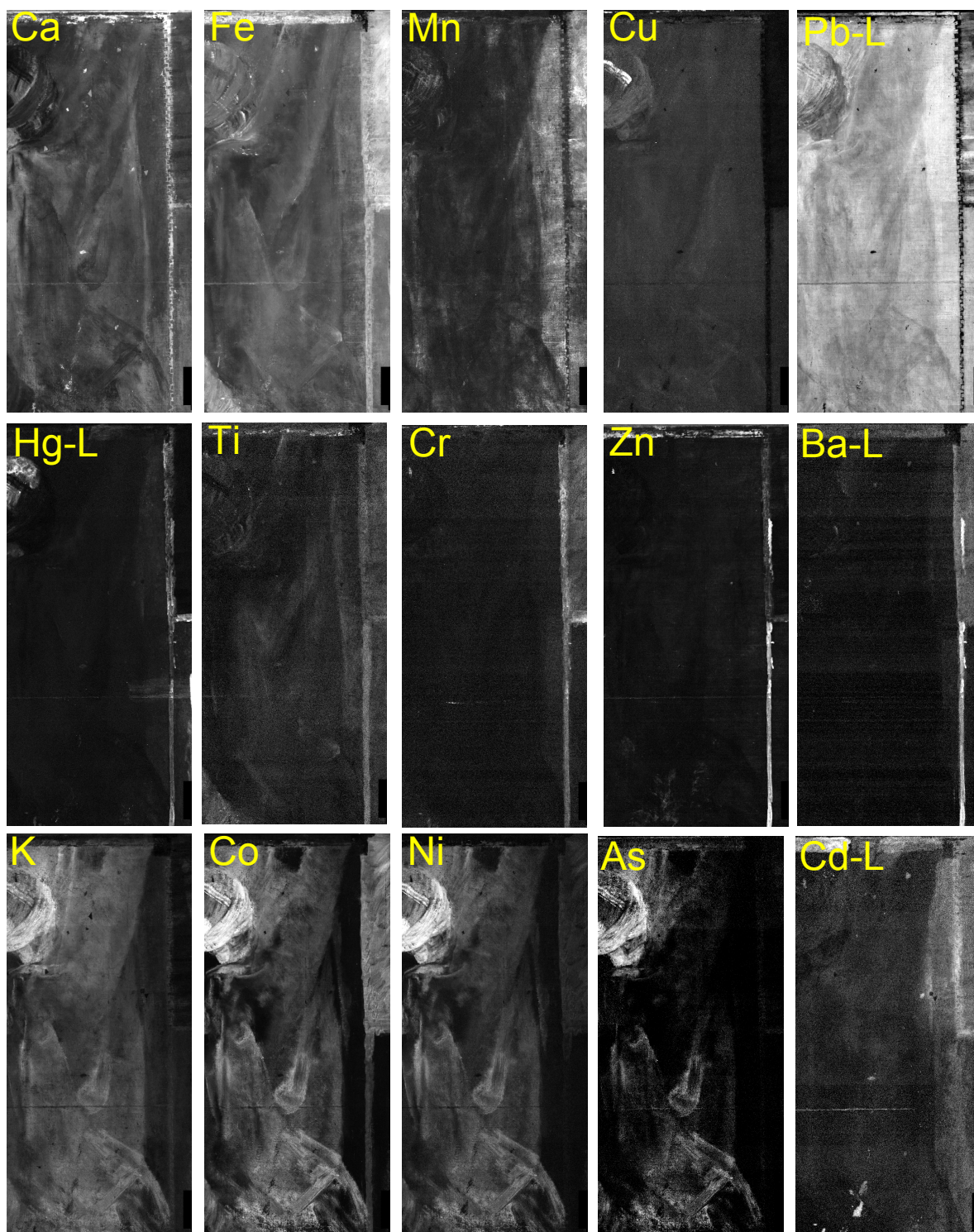
Figure 4. Elemental distribution images of the central part of the painting (approx. 60x120 cm^2). The elements were detected by their K or L level fluorescence radiation, as indicated in the labels.
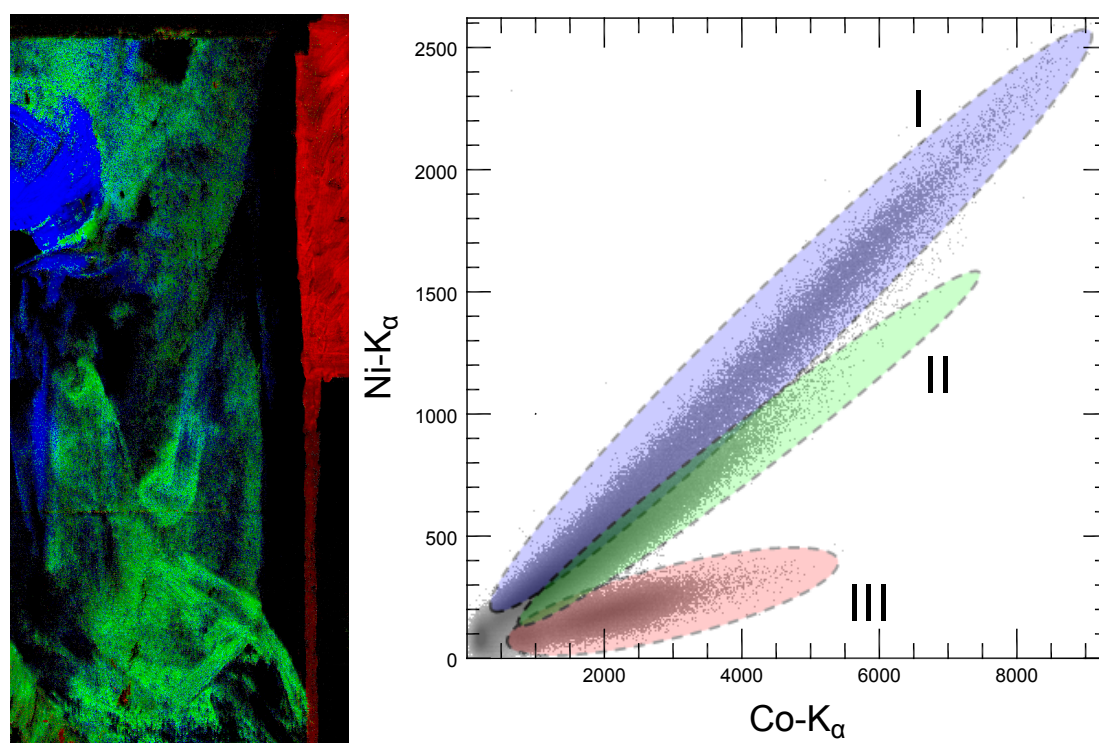
Figure 5. Co map colorized based on the Co/Ni ratio to visualize the extent of smalt (groups I and II) and drier (group III).

is present as a drier or as a pigment, but as Cd salts were not used until the 1840s they are, as aforementioned elements, good tracers for later restoration treatments and indicate heavy reworking of the area along the join.

K, Co, Ni and As are mainly present in the blue pigment smalt. However, from paint cross sections investigated by SEM-EDX it is known that Co and Ni are also present in a second compound: Co salts added as drier to paint in the 19th century. These two compounds cannot be differentiated by the Co signal alone. However, since Co ores also contain Ni the compounds synthesized from these minerals feature a characteristic Co/Ni ratio.

In the Co/Ni scatter plot in Figure 5, three different groups can be distinguished: group I and II represent smalt, while group III represents the drier. The fact that the smalt signal is distributed over two groups may suggest that two different kinds of smalt were used in the creation of the painting. It is obvious that the statistics of the scatter plot are insufficient to fully resolve the groups I and II.

However, the Co associated with smalt and that used as a drier (group III) can be clearly distinguished; the distribution of these Co containing compounds can be visualized by coloring the Co distribution image in Figure 5 accordingly. The smalt distribution image reveals the folds of the original curtain and shows that under the surface a considerable amount of the original paint is still present, except in the area directly along the join, where much of the original paint was removed.

4.2. NMF
The study of scatter plots can allow for the distinction between compounds of similar elemental composition. However, this requires a high degree of user intervention. Especially, if not a single
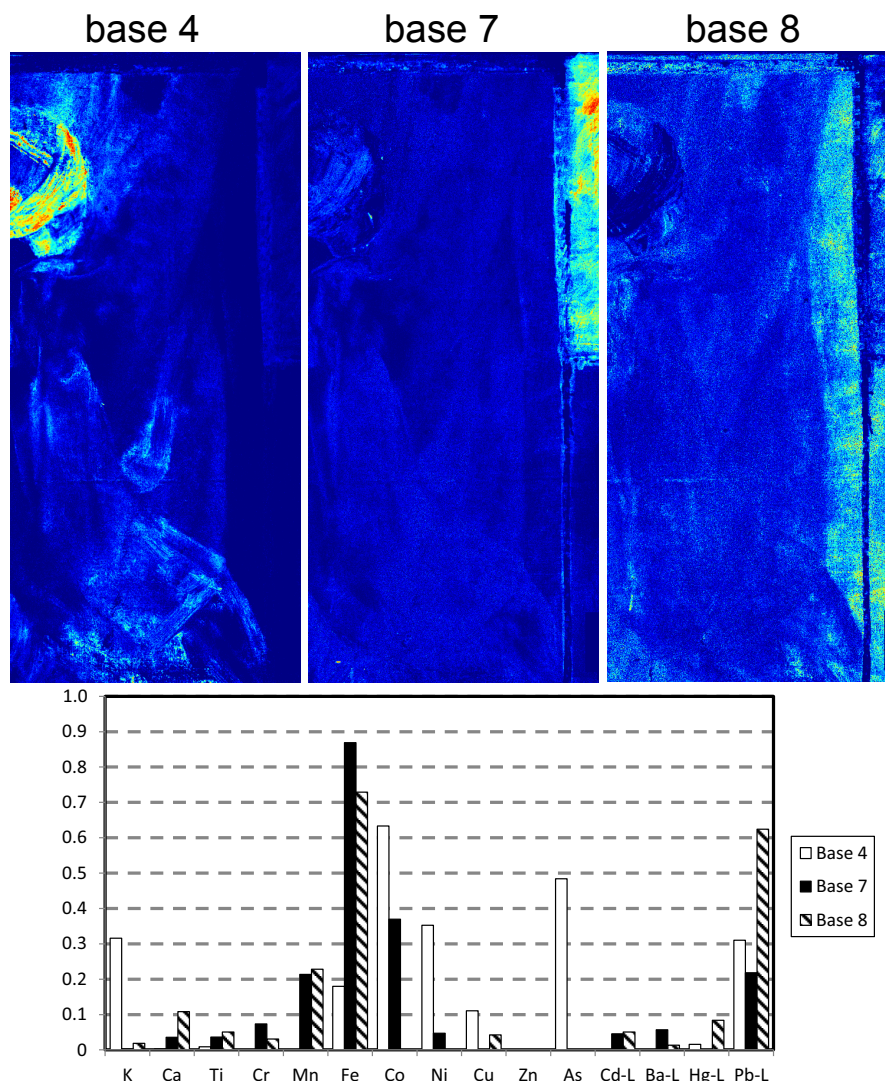
Figure 6. Three bases obtained by NMF. Base 4 represents smalt, base 7 the Co salt driers, while base 8 highlights the area in that no original paint is left. The scale of the bar plot is normalized intensity.

data set, such as in this case, but multiple from a series of measurements are studied or the correlation of several elements is of interest. In the case of *Saul and David* several regions of heavy reworking were known from the XRR, so that it was suspected from the beginning that the Co at the joins was not present in the pigment smalt but a different compound, which was confirmed by the SEM-EDX investigation of paint samples. In more complex or subtle cases this might have gone unnoticed.

The NMF[2] was applied on the stack of 15 elemental distribution images, each featuring 495x1074 pixels, shown in Figure 4. Using processed XRF data allowed to reduce the volume of data and so allowed for fast processing. Only signals containing information on the paint layers of the painting were considered; Ar signals, which originate from the air, were not included, as

---

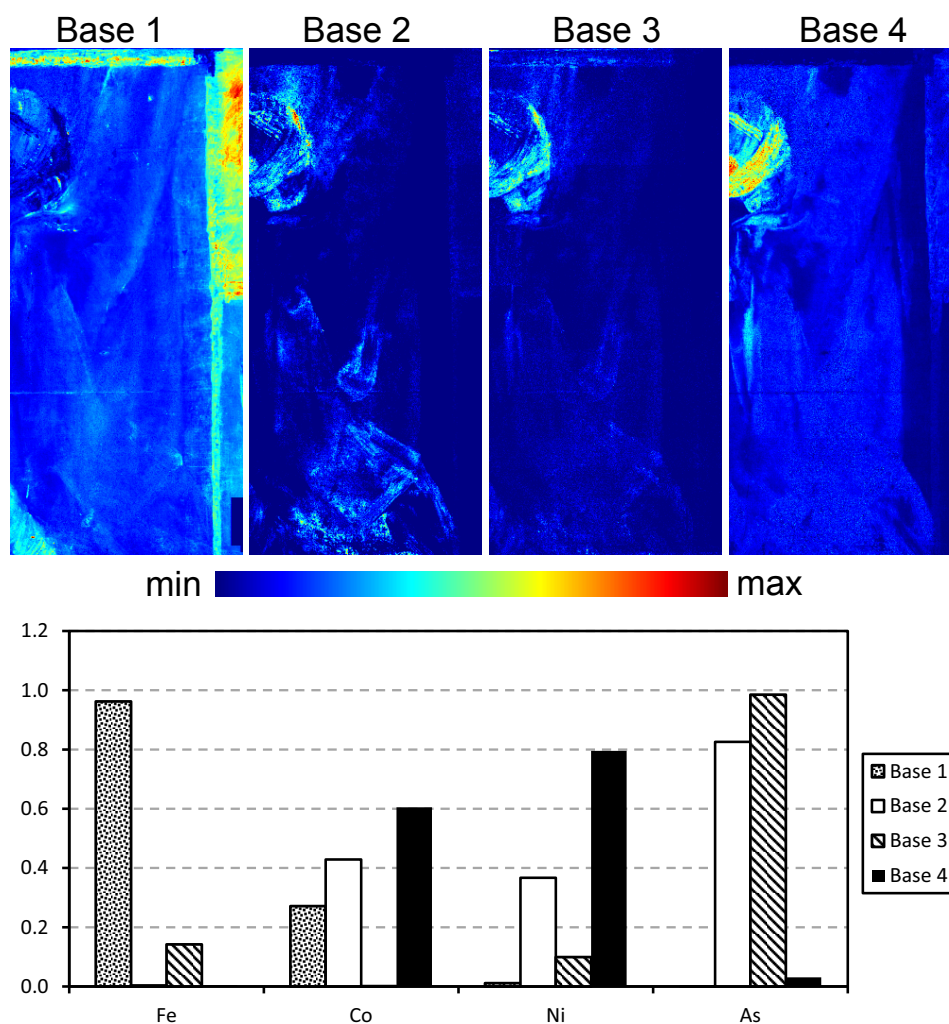[2] We used the NMF implementation of the Python Matrix Factorization Module (pymf [12]).

Figure 7. Four bases obtained by NMF on the elemental distribution images of Fe, Co, Ni and As. The bases 1, 2 and 4 are equivalent to the groups I, II and III in Figure 5. Base 3 illustrates the heterogeneous distribution of the As signal.

was the intensity of scattered primary radiation, which contains strong contributions from the wooden stretcher behind the painting. In case of multiple fluorescence lines of the same element only the most intense one was taken into account.

For the NMF investigation the data was scaled by dividing each distribution image by the square root of its mean value. This scaling step can be considered as a way of weighing the contribution of each image using the uncertainty of the average intensity as weight [1]. The results shown were obtained with 8 bases, in order to reduce the dimensionality of the data set by a factor of two. The calculation was aborted after 100 iterations, which was found sufficient for the decomposition of the data set. These 100 iterations took ca. 35 seconds on a standard Intel QuadCore CPU with 3.3 Ghz and 8 GB installed memory. As NMF starts with a random set of bases to iteratively improve the degree to that the variance of the data set is explained, the chosen seed for the random number generator has an considerable influence on the results. So, the processing in NMF was repeated 10 times with 100 iterations each and different seeds but otherwise identical parameters.

In Figure 6 three of the bases obtained are shown. Base 4 represents smalt and contains next to Co the elements K, Fe, Ni and As. These are typical minor components of smalt. Also Pb, which is virtually omnipresent in the painting, can be found. Base 7 represents the Co salt drier and contains next to Co and low amounts of Ni a considerable amount of Fe. Base 8 highlights the area of paint loss and contains the elements of the pigments used to fill the gap. Mainly Mn from dark earth pigments and Ca and Cd. The enhanced contribution of Pb to this base is assumed to be the result of a thinner covering layer, absorbing less of the Pb-L radiation emitted from the ground layer(s). The Co in the join is represented in another base (not shown). Given the heterogeneous distribution of material in the join area, no accurate representation of it was obtained. In all 10 calculations using different seeds the discrimination between smalt and drier was possible. The identification of the paint loss zone (base 8) was possible in 9 of 10 calculations. In the 100 iterations used no consistency between $\mathbf{W}$ and $\mathbf{H}$ was achieved for the different seeds, which was not the aim of the decomposition, but a correct decomposition of different components.

In the NMF of the whole data set no distinction between the different smalt groups (I and II in Figure 5) was obvious, albeit some bases indicated a separation between different groups of smalt. The *a priori* knowledge of two smalt groups could be reproduced, by limiting the NMF to the elemental distribution images of Fe, Co, Ni and As and increasing the number of iterations to 500 with four bases (see Figure 7). K, which is also a minor component of smalt, was omitted in this approach, as it is considerably lighter than the other elements and suffers stronger from absorption artifacts. The data was binned to 247x537 pixels to accelerate processing and normalized to its mean value. This results in all four elemental distributions having the same weight, independent of uncertainty and sensitivity of the instrument.

Base 1 represents earth pigments and the Co salts used as drier and are equivalent to group I in Figure 5. Base 2 is the equivalent to group II, while base 4 is equivalent to group III, which featured the highest Ni content relative to Co. Base 3 shows something not apparent in the scatter plot: the distribution of As and Co do not completely agree, but in some regions As is enhanced. If this difference is due to compositional variations or absorption effects has to be investigated further. It is a curious oddity that in base 3 no Co is present, albeit its distribution correlates with that of As in the elemental distribution images and in many other NMF treatments both elements were found in the same base.

While the elements in Figure 7 were chosen based on *a priori* knowledge to reproduce the manual results, it is quite reasonable that a researcher, experienced with the composition of smalt and XRF might select these elements. For an automatic separation of two smalt groups, improved NMF approaches such as those mentioned above might be helpful [5–7].

5. Conclusions

We have successfully shown in a case study how the NMF function of the pymf module supports the interpretation of a complex data set with a moderate degree of redundancy (Co, Ni, K and As being all present in smalt) and compounds with similar elemental composition (smalt and Co salts added as drier, both containing Co and Ni). A correct decomposition of components of the data set was possible with only minimal amounts of *a priori* knowledge.

Albeit the bases obtained suggested the presence of more than one smalt component, it could only be visualized by using *a priori* knowledge from scatter plots of the Co/Ni ratio. It is expected that NMF with more elaborate selection rules, considering e.g. sparseness or convexity constraints, would considerably improve the bases obtained.

While the manual interpretation of the scatter plot allowed for the discrimination between three Co containing compounds, it was rather time consuming. Further, it was also limited to the correlation of two elements and so the chances for overlooking the presence of Co in more than one component were present. An application of NMF at that moment would have

simplified the interpretation of the data. In the original study the distinction between original and later added paint made it possible for the conservator to make a more accurate description of the painting's true condition and an informed conservation treatment proposal taking this new information into account.

We expect that in the future NMF and other routines of the pymf module with an easy to use graphical user interface will play a role in the interpretation of XRF data, as might NMF routines from other software packages, since an application of NMF can considerably accelerate the interpretation of the data.

References
[1] Vekemans B, Janssens K, Vincze L, Aerts A, Adams F and Hertogen J 1997 *X-Ray Spectrom.* **26** 333–346

[2] Kotula P G, Keenan M R and Michael J R 2003 *Microsc. Microanal.* **9** 1–17

[3] Friel J J and Lyman C E 2006 *Microsc. Microanal.* **12** 2–25

[4] Lee D D and Seung H S 1999 *Nature* **401** 788–791

[5] Ding C, Li T and Jordan M 2010 *IEEE Trans. Pattern Anal. Mach. Intell.* **32** 45–55

[6] Cutler A and Breiman L 1994 *Technometrics* **36** 338–347

[7] Thurau C, Kersting K, Wahabzada M and Bauckhage C 2012 *Data Min. Knowl. Disc.* **24** 325–354

[8] Noble P, van Loon A, Alfeld M, Janssens K and Dik J 2012 *Technè* **35** 36–45

[9] Alfeld M 2013 *Development of Scanning macro-XRF for the Investigation of Historical Paintings* Ph.D. thesis University of Antwerp, Belgium, and University of Hamburg, Germany, p. 199–202

[10] Alfeld M, Janssens K, Dik J, De Nolf W and Van der Snickt G 2011 *J. Anal. At. Spectrom.* **26** 899–909

[11] Vekemans B, Janssens K, Vincze L, Adams F and Van Espen P 1994 *X-Ray Spectrom.* **23** 278–285

[12] Pymf: Python matrix factorization module, `https://code.google.com/p/pymf/`, last visited 02.12.13