Non-negative assisted principal component analysis: A novel method of data analysis for raman spectroscopy

Astrid L. Blee | John C. C. Day | Peter E. J. Flewitt | Alejandro Jeketo | David Megson-Smith

1Interface Analysis Centre, H.H. Wills Physics Laboratory, Bristol, UK
2Rolls-Royce plc, Derby, UK

Correspondence
Astrid Blee, Interface Analysis Centre, H.H. Wills Physics Laboratory, Bristol, BS8 1TL, UK,
Email: astrid.blee@bristol.ac.uk

Funding information
Engineering and Physical Sciences Research Council, Grant/Award Number: 17000030

Abstract
A novel method for the analysis of multivariate Raman spectroscopy data is presented. The method combines non-negative matrix factorisation and principal component analysis, integrating the advantages and combating the disadvantages of both techniques. It involves the derivation of physically realistic spectra and the analysis of chemical and spatial trends across a sample surface. Proof of concept is demonstrated through two investigations. The first is a set of Raman spectra taken from a powder sample containing potassium sulphate, calcium carbonate and sodium sulphate. A second uses Raman data taken from an artificially corroded sample of superalloy material commonly used in gas turbine engines. This successful proof of concept for samples with unknown surface content sets the way for future development of the technique.

KEYWORDS
data analysis, non-negative matrix factorisation, principal component analysis

1 | INTRODUCTION

Raman spectroscopy is a powerful optical technique used in the identification and analysis of chemical compounds.\(^1\) It is non-destructive and acquires spectral data relatively quickly, whilst requiring no sample preparation. A laser is shone on the sample surface and a spectrum of the scattered light is obtained. Most of the light is scattered elastically and so has the same wavelength as the incident laser. A small fraction of the incident photons may interact with vibrational modes of the sample and either impart energy to raise a molecular bond to a higher vibrational energy state or acquire energy from an excited vibrational state.\(^2\) These photons either lose or acquire energy from the sample and are consequently shifted to longer or, less commonly, shorter wavelengths. Because chemical bonds have characteristic vibrational modes with specific energies, the Raman spectrum acquired is characteristic of those bonds and provides a chemical fingerprint from which the composition of the sample may be determined. By comparing detected Raman spectra with literature spectra, the current chemical makeup of a material can be determined. This identification is relatively simple—provided appropriate reference spectra exist (to which measured data can be compared), and the sample composition is simple. If this is not the case, identification can be difficult.

It is often desirable, but much more difficult, to determine the prior chemistry that caused a particular chemical makeup—that is, what compounds or reactions came before the present surface chemistry? In many situations, nonhomogeneous reaction environments and uncontrolled stoichiometry can result in chemical reactions or processes terminating prematurely or incompletely. The
products and reagents of such reactions offer clues to the chemical history of a sample.

The coincidence or relative exclusivity of substances could offer insights into the reactions, physical trends and chemical relationships, which have occurred previously, thus allowing greater understanding of analysed samples and experimental conditions.

1.1 Raman spectroscopy: An industrial tool

Raman spectroscopy has been used as an industrial tool for many applications, particularly in pharmaceutical,[3,4] medical,[5] engineering[6,7] and forensic[8,9] areas of research. The transition of Raman spectroscopy from a purely academic tool to one applied in industrial settings demands a high level of expertise to interpret gathered data at present. It is therefore necessary to produce meaningful results understandable to all involved in the sample evaluation process, not just by ‘expert users’. Although it is relatively easy for computer software to automatically identify simple compounds of high purity with only one main spectral component, it becomes increasingly challenging to identify mixed samples—as are found in industrial settings.

One example of this would be the application of Raman for the detection of corrosion in jet engines. Under normal working conditions, engine components may undergo degradation as a consequence of the high operating temperatures, often in excess of 800°C.[10] One form of this is so-called ‘hot corrosion’.[11–13] Corrosion detection is extremely important and is currently managed through conservative estimates of component lifespans.[14] Early detection would improve lifespan evaluation and help both reduce cost and decrease waste by avoiding the premature scrapping of components.

Raman spectroscopy could be used as an analysis tool in this environment. It is non-destructive and has the potential to be deployed on-wing, which would have the added advantage of avoiding the cost and delay of removing the engine from the airframe. At present, the downside of this technique is the comprehensive data analysis required to produce meaningful results from collected data.

There are a number of different techniques, which can be applied to analyse Raman spectra. Common techniques include principal component analysis (PCA),[15] classic least squares (CLS),[16,17] partial least squares regression (PLSR),[18] multivariate curve resolution (MVC R)[19] and non-negative matrix factorisation (NMF).[20] The result of the analysis, regardless of technique, carried out on a data set will then be a set of component ‘spectra’. For Raman spectroscopy, results can be compared with literature spectra, either within a data set or series or by comparison with a relevant database.

For many applications, these databases already exist, either online (e.g., the RRUFF™ Project[21]) or as ‘addons’ for analysis software (e.g., the spectrum search function in WiRE 5.0 from Renishaw plc., UK). For non-standard samples though, these databases are minimal to non-existent.

There is also no known analytical method for inferring the chemical evolution of a sample from a simple data set taken at one point in time, post reaction. The majority of current data analysis techniques can show ‘heat maps’, revealing where derived components are found and where the signal is strongest. Ratios of relative intensity are possible, but only where at least some of the surface components of a sample are already known. A comprehensive method for showing chemical relationships is as yet lacking.

In this paper, we propose a novel data analysis technique that combines PCA and NMF for the purpose of producing realistic Raman spectra, whilst simultaneously suggesting analytical trends to allow further interpretation.

2 DATA ANALYSIS TECHNIQUES

The purpose of data analysis is to understand the useful information found in a larger data set. This is often done by condensing the information into a smaller, more easily interpretable, version.[15,22] This then allows easier interpretation of these data, to discover useful information and inform conclusions. As mentioned previously, there are many different techniques available for analysing multivariate data; a select few are discussed here.

2.1 Singular value decomposition

Many techniques rely on a form of singular value decomposition (SVD), which was developed in the late 1800s by a number of mathematicians.[21] They postulated that a matrix containing the initial data can be linearly transformed to a reduced ‘frequency space’ with two or three smaller matrices, each with its own eigenvalues and eigenvectors.[22] The mathematical approaches vary, and the difference between techniques is then dependent on the criteria applied to the resulting matrices.

The techniques (PCA and NMF) are both variants of SVD. PCA is a decomposition of the input data, with one matrix calculated to contain the maximum amount of
variance in the original data set, whilst NMF has a user-defined number of components, all of which are constrained to be positive.

### 2.2 Principal component analysis

PCA is a technique often used to analyse multivariate data, in many different fields, from medicine and pharmaceutical to the physical sciences.\(^{[24]}\) Initially proposed in 1901 by Pearson, who described the technique as ‘finding lines and planes of closest fit to systems of points in space’. Hotelling\(^{[25]}\) formalised PCA into its current format, coining the term ‘principal component’ (PC). The goal of PCA is to extract the most information possible from a data set, whilst compressing the size of that set to contain only the most important information. This allows the description of the data set to be simplified and means that the structure of the observations and variables can be analysed.

The finer details of PCA can be found elsewhere,\(^{[26-29]}\) but the following gives a brief overview of the technique. Initially, the data are stored as a matrix $Y$, which contains $i$ observations of $j$ variables and thus has dimensions $i \times j$. The aim of PCA is to express the maximum amount of the variance in $Y$ as a linear summation of $k$ components, where $k \leq j$ (i.e., the maximum number of components is the number of variables measured in a data set). This combination is given by

$$
\sum_{k=1}^{j} p_k x_k = Y p
$$

where $p$ is a new matrix with dimensions $(j \times k)$. A new term is defined, $\text{var}(Yp)$, which is the variance explained by the combination of $Y$ and $p$. It is given as

$$
\text{var}(Yp) = tp^T
$$

where $t$ is the weightings matrix required to reproduce the original data set (with dimensions $i \times k$) and $p^T$ is the transpose of the matrix $p$. The next step is to express as much variance as possible from $Y$ in $p$. This is because $p$ is a smaller matrix than $Y$ and is much easier to process, thus making it a more useful representation of the original data.

To do this, the scores matrix $t$ is chosen to be optimum (by which is meant maximising the variance contained in $p$). To avoid this variance being arbitrarily large, $t$ must also be normalised—nominally achieved by setting its Euclidean norm to one.

This same problem can then be rephrased as a calculation to find the most representative version of $t$ as a replacement for $Y$. By using standard linear regression equations,

$$
Y = tp^T + E
$$

where $p$ is now a normalised vector containing the regression coefficients required to reproduce $Y$, $p^T$ is the transpose of that matrix and $E$ is a matrix of residuals. Therefore, the optimisation problem can be written as

$$
\arg\min \|Y - tp^T\|^2
$$

$t$ is referred to as the scores matrix and $p$ as the components—or loadings—matrix. This problem is then solved computationally.

### 2.3 Advantages and disadvantages of PCA

PCA has many advantages.\(^{[27,30]}\) It is computationally light and produces components that are orthogonal and independent of each other. It is an unsupervised technique and thus requires no user input to produce results. It should be noted that PCA can create $j$ total components, where $j$ is the total number of measured variables, and therefore deciding on a $k$ value does require some level of user input. However, with the additional metric of how much variance is explained by each component, it is easy to quantitatively define limits on how many components should be analysed in a final version (e.g., sum of variance explained by components should be greater than 99% or each component should explain more than 0.5% variance).

The main disadvantage of PCA is the format of the output components.\(^{[31]}\) By transforming the data to a new set of components, its readability is reduced. The linear combination required to produce each PC contains both positive and negative weighting, which can then partially cancel each other out.\(^{[32,33]}\) This results in physically unrealistic results—for example, when analysing spectral data (such as is obtained from Raman spectroscopy), the produced PCs often contain negative values, inverse peaks and other features that make comparison with literature spectra very difficult. In general, identifying PCs as realistic spectra is achievable by an expert user, but this can be impossible if the sample composition in question is unknown and there are no databases with which to compare derived components.

It is possible to vary the algorithm used for PCA, to reduce the effect of these issues. Both non-negative PCA, sparse PCA and a combination of both criteria have been
used with relative success. Non-negative PCA was demonstrated by Han in 2010 when applied to spectra taken from both cancerous and healthy cells, where the data representation was forced to contain only positive entries. The method involves introducing a penalty parameter $\alpha$ into the optimisation of the PCA components, thus forcing the output components towards positivity. Zass and Shashua also tried this approach and added in a second penalty parameter, $\beta$, but this was found to sometimes result in a non-convergent result due to the play-off between the two parameters.

Support vector machine (SVM) learning was used by Han to identify and classify spectra, thus proving the realistic appearance of the output spectra. However, although this technique successfully identified ‘global features’, it struggled to identify local features (‘global’ features being curves and features that bear strong resemblance to spectral parts found throughout the data set, whereas ‘local features’ are those that only appear at specific points and only as a set group/on their own).

The constraining of the PCA algorithm always results in a trade-off between the ‘statistical fidelity’ (or actual variance) contained in the output and the interpretability of the output. This can result in computationally intensive work, results that do not fully represent local features or statistically unreliable results. These constraints on PCA are therefore not necessarily applicable to all data sets on all occasions and do not represent a cast-iron solution to the issues described above.

2.4 | Non-negative matrix factorisation

NMF was initially proposed by Lee and Seung as a method of deconstructing images into their component parts, as an alternative to both PCA and Vector Quantization. It has since become a highly useful tool for analysing multivariate data. In particular, it can be applied to the analysis of spectral data—such as Raman—where the constraint of non-negativity is physically realistic, rather than a purely analytical constraint.

The observed data are stored in a matrix, $X$, which contains $m$ spectra, all with intensity measurements at $n$ wavenumbers. Two smaller matrices are then found, $A$ and $S$, such that

$$X = AS$$

where $A$ has dimensions $(m \times p)$ and $S$ $(p \times n)$. $p$ is defined as the number of derived spectra—also known as components—and is a user-defined term. Each row of $S$—the derived spectra matrix—represents one component spectrum with $n$ wavenumbers. $A$ (the weightings matrix) contains the mixing coefficients to produce each spectrum in $X$. To solve this then, the problem becomes one of optimisation—that is, we wish to minimise the distance, $d$, between $X$ and $AS$. Multiple formulations of the optimisation problem are possible, all of which are constrained by the non-negativity condition, $A, S \geq 0$.

2.5 | Advantages and disadvantages of NMF

The primary advantage of NMF is its ability to produce positive outputs, which for spectroscopy based applications are physically realistic. This makes identification of output spectra more intuitive. However, the number of output components has to be user defined. This is a significant disadvantage for Raman data sets of unknown components: If a data set actually contains five component spectra, and the algorithm is selected to extract three, there will be overfilled spectra, each with too many peaks (i.e., the data set has been under-fitted), whereas if 10 spectra are extracted, the majority will feature only certain peaks from each spectrum (i.e., the data set has been over-fitted). Determining this ‘magic number’ of components is a problem yet to be solved.

Efforts have been made to improve the usage of NMF for data analysis applications. For example, Geng et al. used NMF to analyse hyperspectral data, which were pre-treated by applying PCA to the data set. This was then used to identify hyperspectral signatures from a large mixed data set. The application of PCA prior to the NMF stage of analysis is very helpful, as the PCA essentially performs a dimension reduction on the data, making large unwieldy data sets much easier to handle and increases the signal-to-noise ratio (SNR). However, the method described is primarily aimed at reducing the dimensionality of the data set such that pre-ordained components can be more easily extracted from the original data by identification with literature spectra, and as such would be of less use for identifying wholly unknown spectra.

3 | NON-NEGATIVE ASSISTED PCA

A new data analysis technique is proposed, which combines the advantages of both NMF and PCA and combats the disadvantages. The results take the form of novel data plots, which allow for the interpretation of both chemical and physical trends within the original data set. The methodology is thus.
The initial data are prepared in the same way as required for NMF, as given in Section 2.4. An arbitrarily large number of components (derived spectra) are intentionally extracted—this number is defined as \( p \) as above, resulting in the two previously defined matrices, \( A \) and \( S \). Here, \( p \) is chosen to be large and therefore to deliberately over-fit the product \( AS \) to the original data, \( X \), that is, \( p \) is chosen to be arbitrarily large.

Next, PCA is performed upon the spectral weightings matrix, \( S \), to determine where the maximum variance lies. To understand the rationale behind this approach, consider the following. If the spectral composition of an analysed sample were randomly distributed, that is, with no pattern to the distribution of different chemical species across the surface, then the spectral weightings would also be randomly distributed—there would be no distinguishable trends in the distribution of spectral weightings. If this is not the case, as expected for the vast majority of samples, then there will be trends in the distribution found in \( S \).

Consider a sample containing two chemicals, A and B. Chemical A may be found, for example, only in the presence of Chemical B, or only when Chemical B is absent. These trends can be of key interest when determining the chemical history of a sample, by inferring chemical relationships between different derived spectra. For example, if a third chemical, Chemical C, is produced by the reaction between Chemicals A and B, this would be visible as a trend in the data set. Thus, these trends are found in the spectral weightings matrix. PCA identifies trends and variance within data sets and is therefore applied to the spectral weightings recovered from NMF.

If a derived spectrum contributes most to the variance within a particular component, it will have a large PCA score (either positive or negative). On the other hand, if it does not contribute to the component, it will have a small absolute PCA score. These results are visualised on a new type of data analysis plot—an illustrated example is shown in Figure 1. Each derived spectrum is plotted at the point on the ordinate corresponding to its (NMF-)PCA score. The horizontal axis here represents the spectral range to allow the spectra to be displayed in their entirety. Hence, the first point in each spectrum is plotted at the corresponding height on the vertical axis and the rest of the spectrum plotted to scale. All the spectra are scaled by a set factor prior to this step, so each NMF-PCA plot gives an intuitive representation of how the individual spectra relate to one another.

Every PCA component is now referred to as an individual, orthogonal, trend, which explains some percentage of variance in the analysed data set. Within each trend, every derived spectrum has an associated PCA score, which represents how much it contributes to that component. These provide an insight into correlations between derived spectra and therefore the sample surface in question. For example, when two derived spectra are often found together, they will both have large PCA scores (either both positive or both negative). When actively not found together, they will have opposing sign PCA scores. If they are non-contributing or randomly contributing derived spectra, they will have near zero PCA scores.

Each PCA component is individually interpreted to explain the unique, physical and/or chemical trend that it represents. Some trends indicate that chemical species are associated with one another, whereas other trends indicate which species are found only when others are absent (i.e., they are mutually exclusive). Each trend—as the PCA components are now referred to—has an associated percentage variance that explains how prevalent that trend is and therefore how commonly it is found in the data set.

Another feature of these trends is that they are orthogonal—and therefore independent. This means that the information contained in each component does not infringe on information in others and will instead refer to different parts of the data set (or physical regions of the mapped data).

This approach has many advantages. In particular, the problem of determining the desired number of

---

**FIGURE 1** An illustration of the new type of data analysis plot produced by non-negative matrix factorisation-principal component analysis (NMF-PCA). Spectra that contribute most to the plot (‘present’ spectra) appear with the highest absolute scores (shown as [a] above in dark blue and red), whilst spectra which are not found there (‘anti-present spectra’) have large scores of the opposite sign ([c], green above). ‘Non-contributing’ spectra—and therefore randomly appearing compounds—have scores close to 0 (blue/purple or [b] above) [Colour figure can be viewed at wileyonlinelibrary.com]
derived spectra, $p$, to calculate is solved by deliberately over-fitting. In such a situation, derived spectra that are merely noise- or non-specific background spectra will have low PCA scores across all of the trends. If the data are over-fitted, there will be multiple derived spectra representing one fundamental spectrum. On these plots though, the copies will be highly correlated and therefore easy to identify. In addition, this technique only provides physically realistic derived spectra, as produced by NMF, thus meaning that they can be easily compared to literature spectra for reference. Together with the insight into trends in the data provided by the PCA scores, this facilitates both the identification of chemical species and the identification of relationships/correlations between those species.

4 | EXPERIMENTAL APPLICATIONS

The following data analysis was conducted using Python, in Spyder 4.1.3 from Anaconda. Both the NMF and PCA algorithms used were developed by scikit-learn. To demonstrate the value of the new approach, two specific investigations are presented.

4.1 | Investigation 1

A sample was prepared to demonstrate and validate the new analytical approach. Three powdered chemical compounds were arranged in a sample tray, such that there were two specific regions. A mix of potassium sulphate ($K_2SO_4$) and calcium carbonate ($CaCO_3$) was on one side, whilst sodium sulphate ($Na_2SO_4$) was alone on the other side, though some accidental mixing also occurred, producing regions of calcium carbonate on the other side with the sodium sulphate. All the powders had a purity of greater than 99%, and the layers were at least 5 mm thick so that there would be no accidental contamination of the signal from the tray below. In Figure 4, a traditional Raman map is presented, with peak intensity plotted for all three compounds across the mapped area, which shows the layout of the sample. It should be noted that this sort of map is only possible when the individual spectral fingerprints are known for all the chemicals involved—that is, it is suitable for a wholly artificial sample, but not for anything more complex.

Using a Renishaw inVia Raman spectrometer with a 532-nm green laser and a 20X objective, 1296 spectra were measured at 15-μm intervals on the sample. This formed a rectangular $72 \times 18$ array, which covered both regions of the sample area. Each spectrum contained measurements at 1011 wavenumbers. Figure 2 shows the fundamental spectra of the three compounds collected using the same experimental conditions.

Each measured spectrum covered the same spectral range divided into the same number of bins. The average spectrum (shown in Figure 3) contains at least 12 individual peaks, and any analysis of this data set must include the separation of peaks into their components. Without prior knowledge of the sample, it would be difficult to match these peaks to literature spectra; that is, it is
difficult to determine whether the peaks are a result of 12 compound ‘fingerprints’ with one peak each, or one compound with 12 peaks, or something in between. Individual spectra taken from the map data can be analysed, but this would be very time-consuming. However, the combination of NMF and PCA is able to analyse the spectra from the sample. In particular, we predict that the tool should be able to identify the following trends:

- Potassium sulphate and calcium carbonate are found together.
- These two chemicals are found where sodium sulphate is not found.
- In small regions, calcium carbonate is found with sodium sulphate.
- There are rare occasions where a combination of any of the three chemicals may be found.

The measured spectra were arranged in an $m \times n$ matrix $X$ as described above. NMF was performed with a $p$ value of 12, producing a $1296 \times 12$ derived spectra matrix (A) and a $12 \times 1011$ spectral weightings matrix (S). PCA was then performed on the spectral weightings matrix S.

As explained previously, trends (i.e., PCA components) were then extracted, which explained the greatest percentages of the variance in the data set. These trends are visualised graphically in the plots shown in Figure 5a–e. In these plots, sodium sulphate spectra are blue, potassium sulphate spectra are green and calcium carbonate spectra are red, as was shown previously in Figures 2 and 4. In addition, mixed spectra containing peaks from multiple compounds are given in orange and fluorescence spectra in grey.

In the results plots given, only the derived spectra that contributed most to each particular trend are plotted (i.e., spectra, which have an absolute NMF-PCA score of greater than ±0.5). This is to show the highest contributing spectra for each trend and remove extraneous data. Each of the plots in Figure 5 provides the percentage variance explained by that PCA component, which can be interpreted as a measure of the significance of a particular trend. Trends displaying less than 3% of the total variance have been excluded for brevity. This is also arguably an ‘arbitrary’ limit, the like of which was dismissed earlier as a flaw in the NMF algorithm. However, it is guided by a quantitative metric (the percentage variance explained by each trend) and is therefore a more reliable approach than the trial and error method previously used for defining a $p$ value for NMF. For reference, all 12 of the derived spectra are displayed in the supporting information (Figure S9).

### 4.1.1 Results

**Trend 0: Figure 5a**

The highest rated spectrum in this PCA component is calcium carbonate with a score of +1, with a mixed spectrum containing peaks from both calcium carbonate and potassium sulphate just below, with a score of +0.87. The lowest rated spectra in this component are both sodium sulphate, with scores of −0.95 and −0.63, respectively. This shows that where we find calcium carbonate and potassium sulphate, we actively do not find sodium sulphate. Trend 0 explains 48.8% of the variance in the data set and is therefore the predominant chemical trend.

**Trend 1: Figure 5b**

Sodium sulphate is the highest spectrum again, whilst the second rated spectrum (with a score of approximately 0.55) is calcium carbonate. We also have a spectrum for potassium sulphate with a score of approximately −0.7.
Therefore, on 20.8% of occasions, when sodium sulphate is present, it is accompanied by calcium carbonate, but potassium sulphate is not detected. This is as expected from the traditional Raman map, where we can see small areas of calcium carbonate mixed in with sodium sulphate (on the opposite side of the sample area to where it should have been). This is also consistent with the heat maps for the individual derived spectra, as shown in the supporting information (Figure S9).

There is also a fluorescence spectrum, shown by Spectrum 2 in the figure with a score of around −0.8. This spectrum is shown in the supporting information as Figure S9c: NMF Derived Spectrum 2 and corresponds to a point of high intensity, most likely caused by...
contamination in the sample. It is of a much lower intensity than the other derived spectra—note the much lower signal-to-noise ratio of the spectrum. All spectra were normalised, thus making this weak signal appear of a similar magnitude to the actual Raman spectra.

**Trend 2: Figure 5c**

This PCA component is similar to Trend 1. The highest rated spectrum is sodium sulphate, and next, with a score of 0.7, is potassium sulphate. There are no lower rated spectra. Therefore, approximately 10.3% of the data set contains regions where we see sodium sulphate and potassium sulphate together. This again refers to the centre component of the mapping data, where the compounds are mixed.

**Trend 3: Figure 5d**

Potassium sulphate is the highest rated spectrum, calcium carbonate is the lowest rated spectrum, whilst in between is a ‘mix’ spectrum having peaks from both potassium sulphate and calcium carbonate and another two background spectra (one shown previously in Trend 1, and another generic background—both of these can be attributed to fluorescence). Trend 3 is due to the experimental environment. When the potassium sulphate and calcium carbonate were mixed together, they did not form a completely homogeneous mixture; patches of just potassium sulphate and patches of just calcium carbonate remain. This explains the highest rated and lowest rated spectra. By contrast, points in the sample where the two compounds were found in equal measure were rarer, as represented by the lower rating of the mix spectrum than for potassium sulphate alone. Trend 3 is uncommon, with a percentage variance of only 6.8%.

**Trend 4: Figure 5e**

The highest rated spectra here is a mixed spectrum, containing peaks from both potassium sulphate and calcium carbonate. These are highly rated, in comparison to the low ratings of the spectra corresponding to pure sodium sulphate, potassium sulphate and the previously mentioned fluorescence spectrum. This demonstrates the spatial independence of the mixed compounds from both the fluorescence-causing contamination and from regions of pure chemicals. The low percentage variance (3.3%) for this trend means that it is found infrequently in the data set.

### 4.1.2 Discussion

All the predicted trends for the Raman spectroscopy experimental data were observed in the actual PCA components discussed above. The only trend observed, but not predicted, was that of the fluorescence (i.e., contamination) observed in Trends 1 and 3. However, this is explainable by the 99% purity of the powders, which still allows for possible contamination. In addition, the Raman spectroscopy was not performed in a clean room, so other sources of contamination are also possible. The orthogonal nature of the results should be noted, as without it Trends 2, 3 and 4 seem contradictory. It is not be possible to have data that have two or more compounds appearing together, but also not appearing together. However, here is where the independence of the trends is reiterated. The depicted trends are orthogonal and therefore refer to different parts of the data set and/or different regions of the mapping area.

Identification of these trends would not be possible from a conventional application of either NMF or PCA. The ability to relate the appearance of one chemical species to the appearance—or disappearance—of another is a valuable insight into the chemical reaction history of a sample.

### 4.2 Investigation 2

A second Raman spectroscopy data set is presented as a demonstration of this technique applied to a sample where a number of the surface compounds are entirely unknown. The sample used was a cutoff from a turbine blade from a modern Rolls-Royce engine. The sample was subjected to an artificial corrosion environment to simulate the chemical environment that can cause in-service corrosion (though the samples were not mechanically stressed and were therefore not subject to mechanical degradation such as creep and fatigue damage). The sample material was CMSX-4, a nickel-based superalloy. The sample was salt-sprayed with sodium chloride (NaCl) and then thermally cycled at 700°C for 200 h in a sulphurous gas environment.

The same Raman spectroscopy system and conditions were the same as in Investigation 1, with the difference being a 50X Long Working Distance objective was used (rather than a 20x objective). A map was taken with 176 spectra, forming a rectangular $22 \times 8$ array, with 3-μm spacing. The map location is shown in Figure 6, with the white box showing the analysed area. The average spectrum is shown in Figure 7.

NMF was performed on the collected data, using 10 derived spectra. PCA was then performed on the spectral weightings matrix to extract trends, which explained certain percentages of the variance in the data set. These trends are shown in Figure 8a–d, along with the respective percentage variances explained by each trend. Table 1 is the legend for Figure 8, showing the derived spectra and the current state of the spectral identification.
As with Investigation 1, only the derived spectra that contributed significantly are plotted. This is to show the highest contributing spectra for each trend, and remove extraneous data. In addition, trends displaying less than 3% of the total variance have been excluded for brevity. For reference, all 10 of the derived spectra are displayed in the supporting information (Figure S10).

4.2.1 | Results

Trend 0 (Figure 8a) explains 52.1% of the variance in this data set and therefore is the most commonly found trend. The top two ranked spectra are Derived Spectra 2 and 6, respectively (top here meaning closest to the maximum score of 1). By comparison with literature spectra (using Renishaw’s spectral ID function in WIRE), Derived Spectrum 2 spectrum corresponds well with nickel oxide (NiO), whereas Derived Spectrum 6 is as yet unidentified, though it contains some noisy peaks from the Raman spectrum of chromite (Fe$^{2+}$Cr$_2$O$_4$, a black spinel). From Trend 0, we can infer that both compounds commonly appear together. However, from Trend 1, we also see Derived Spectrum 6 (or rather, the compound responsible for it) appears mostly by itself without nickel oxide. This could be explained as a chemical relationship such as the compound responsible for Derived Spectrum 6 being intrinsically linked to the production, or destruction, of nickel oxide.

Trend 2 (Figure 8c, explaining 10.8% variance) shows just one spectrum, Derived Spectrum 7 at a height of +1. This spectrum is a very good match for chromite (Fe$^{2+}$Cr$_2$O$_4$), identified via RRUFF $^\text{TM}[21]$. This is a spinel, which is dark grey to black in colour (further confirmed by the white light image), and it is formed in air at temperatures between 600°C and 700°C, as per the experimental conditions for this sample. From this trend, where it is the most contributing component to the trend—and by the absence of any other plotted spectra—it is the only contributing compound in the data set. Other spectra may be present and not plotted, due to the filtering applied to the NMF-PCA scores, but these bear no relation to the presence or absence of chromite. Therefore, on just over 10% of occasions, we see this compound mostly by itself.

Trend 3 explains the smallest amount of variance with 6.7% (shown in Figure 8d). The top ranking spectrum, Derived Spectrum 4 at a height of +1, is a very good match for sodium sulphate, Na$_2$SO$_4$. No other
**FIGURE 8**  (a–d) Derived non-negative matrix factorisation-principal component analysis (NMF-PCA) Trends 0–3 for Investigation 2, showing the four major trends for the mapped data in order of the amount of variance in the data set they explain [Colour figure can be viewed at wileyonlinelibrary.com]

**TABLE 1**  A table detailing the 10 derived NMF components with colours corresponding to Figure 8 and the state of the spectral identification process

| PLOT COLOUR | DERIVED SPECTRUM No. | SPECTRUM IDENTIFICATION |
|-------------|----------------------|--------------------------|
|             | 0                    | Unknown                  |
|             | 1                    | SO$_4$ peak at 984 cm$^{-1}$, Cosmic Ray at 1660 cm$^{-1}$ |
|             | 2                    | Nickel Oxide and Cobalt(II,III) Oxide (Co$_2$O$_3$) |
|             | 3                    | Unknown, SO$_4$ peak (?) present at 983 cm$^{-1}$ |
|             | 4                    | Unknown, SO$_4$ peak present at 983 cm$^{-1}$, potentially a poor copy of Sodium Sulphate |
|             | 5                    | Unknown, contains peaks from Chromite (Fe$^{2+}$Cr$_2$O$_4$) and Nickel Oxide |
|             | 6                    | Unknown, also contains very noisy Chromite peaks |
|             | 7                    | Chromite (Fe$^{2+}$Cr$_2$O$_4$) |
|             | 8                    | Sulphur based compound (SO$_4$ peak at 984 cm$^{-1}$). Other peaks unknown. |
|             | 9                    | Unknown, similar to Nickel Oxide |

Abbreviation: NMF, non-negative matrix factorisation.
compounds appear, and therefore, sodium sulphate is found by itself on 6.7% of occasions in this data set.

4.2.2 | Discussion

The work presented in this section still requires further work in the form of a full and rigorous identification of each derived spectrum (as shown by the incomplete spectral analysis shown in Table 1). Traditionally, this would have precluded any further interpretation of the data.

However, this technique opens an avenue to ongoing analysis. Physically realistic spectra have been extracted from the data set and presented within orthogonal trends, thus aiding future analysis. Moreover, once chemical species are identified, knowing how they are related is highly useful. For example, the Trends 0 and 1 show that nickel oxide is found overwhelmingly when sulphur-based compounds are not present—as is described in literature explanations of corrosion mechanisms.\(^{[12]}\) Even without knowing the exact stoichiometric makeup of the sulphurous compounds, this is powerful information when investigating corrosion mechanisms in complex environments. In practice, the products that form on a turbine blade during service can appear and disappear as corrosion progresses. As a consequence, the link between each compound and the others is invaluable knowledge for an operator.

5 | CONCLUSIONS

The analysis of multivariate data is highly important across many areas of scientific research. In particular, Raman spectroscopy data can be measured from many different areas of science, engineering, medicine and technology.

The proposed technique, non-negative assisted PCA, combines NMF and PCA to produce a novel method of data analysis, which will advance the interrogation of multivariate data, in particular when applied to large Raman data-sets.

A data analysis technique that provides realistic spectra, displayed across orthogonal (and therefore independent) trends, is extremely useful when analysing all samples, not just those where the sample composition is unknown. Being able to infer trends from a chemical sample is important for the broad adoption of Raman spectroscopy as an analytical tool in industrial settings.

Future work will include the incorporation of a spectral ID function into the NMF step of the method, such that spectra can be automatically assigned to chemical compounds, thus allowing the interpretation of trends to be further improved.

ACKNOWLEDGEMENTS

This work was funded by EPSRC iCASE Award No. 17000030 and by Rolls-Royce plc. The method described is protected under patent GB2004459.0 (patent pending). We would like to thank Dr Adriana Encinas-Oropesa for performing the artificial corrosion of the CMSX-4 samples at Cranfield University.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Astrid L. Blee https://orcid.org/0000-0001-5316-7718

REFERENCES

[1] D. J. Gardiner, in Practical Raman Spectroscopy, Springer, Berlin Heidelberg 1989, 1.
[2] C. V. Raman, Indian J. Phys 1928, 2, 387.
[3] T. Vankeirsbilck, A. Vercauteren, W. Baeyens, G. Van der Weken, F. Verpoort, G. Vergote, J. P. Remon, Trends Anal. Chem. 2002, 21(12), 869.
[4] A. Kudelski, Talanta 2008, 76(1), 1.
[5] L.-P. Choo-Smith, H. G. M. Edwards, H. P. Endtz, J. M. Kros, F. Heule, H. Barr, J. S. Robinson Jr, H. A. Bruining, G. J. Puppels, Biopolymers 2002, 67(1), 1.
[6] D. Liu, Spectrosc. Prop. Inorg. Organomet. Compd. 2014, 45, 141.
[7] D. Liu, P. E. J. Flewitt, Acta Mater. 2013, 61(1), 12.
[8] P. Vandenabeele, H. G. M. Edwards, J. Jehlicˇka, Chem. Soc. Rev. 2014, 43(8), 2628.
[9] P. Buzzini, E. Suzuki, J. Raman Spectrosc. 2016, 47(1), 16.
[10] R. C. Reed, in The superalloys, Cambridge University Press 2006, 1.
[11] R. A. Rapp, Corros. Sci. 2002, 44(2), 209.
[12] J. Stringer, Ann. Rev. Mater. Sci 1977, 7, 477.
[13] J. A. Goebel, Metall. Trans. 1973, 4(1), 261.
[14] E. Kosieniak, K. Biesiada, J. Kaczorowski, M. Innocenti, J. Failure Anal. Prev. 2012, 12, 330.
[15] H. Shimzawa, K. Awa, W. Kanematsu, Y. Ozaki, J. Raman Spectrosc. 2009, 40(12), 1720.
[16] M. J. Zavaleta, S. S. Gambhir, Proc. Natl. Acad. Sci. 2009, 106(32).
[17] A. Pallaro, M. R. Hoonejani, G. B. Braun, C. D. Meinhart, M. Moskovits, ACS Nano 2015, 9(4).
[18] H. J. Byrne, P. Knief, M. E. Keating, F. Bonnier, Chem. Soc. Rev. 2016, 45(7), 1865.
[19] A. De Juan, R. Tauler, Crit. Rev. Anal. Chem. 2006, 36(3-4), 163.
[20] H. Li, T. Adal, W. Wang, D. Emge, A. Cichocki, J. VLSI Sig. Proc. 2007, 48(1-2), 83.
[21] B. Lafuente, R. T. Downs, H. Yang, N. Stone, in Highlights in mineralogical crystallography, Walter de Gruyter GmbH 2016, 1.
[22] O. Alter, P. O. Brown, D. Botstein, Proc. Natl. Acad. Sci. 2000, 97(18), 10101.
[23] G. W. Stewart, SIAM Rev. 1993, 35(4), 551.
[24] S. Wold, K. Esbensen, P. Geladi, Chemom. Intell. Lab. Syst. 1987, 2, 37.
[25] H. Hotelling, J. Ed. Psychol. 1933, 24, 417.
[26] H. Abdi, L. J. Williams, Wiley Interdiscip. Rev.: Comp. Stat. 2010, 2(4), 433.
[27] R. Bro, A. K. Smilde, Anal. Methods 2014, 6(9), 2812.
[28] I. T. Jollife, Philos. Trans. R. Soc., A 2016, 374(2065).
[29] M. Ringnér, Nat. Biotechnol. 2008, 26(3), 303.
[30] S. Karamizadeh, S. M. Abdullah, A. A. Manaf, M. Zamani, A. Hooman, J. Signal Inf. Proc. 2013, 4, 173.
[31] C. Ding, X. He, H. D. Simon, in Proceedings of the 2005 siam international conference on data mining, SIAM, Newport Beach, CA, USA 2005, 606.
[32] H. Han, in The eighth asia pacific bioinformatics conference, II (Suppl I). BMC Bioinformatics 2010.
[33] R. Zass, A. Shashua, in Advances in neural inf. proc. systems, Citeseer 2007, 1561.
[34] H. Han, in 22nd international conference on genome informatics, BMC Systems Biology, Busan, Korea 2011.
[35] C. D. Sigg, J. M. Buhmann, in Proc. 25th int. conf. machine learning., Helsinki, Finland 2008.
[36] D. D. Lee, H. S. Seung, Nature 1999, 401, 788.
[37] N. Gillis, Regularization, Optimization, Kernels, and Support Vector Machines, 12, (257), 2014.
[38] V. P. Pauca, J. Piper, R. J. Plemmons, Linear algebra and its applications 2006, 416(1), 29.
[39] D. D. Lee, H. S. Seung, in Adv. neural inf. proc. syst. 2001, 556.
[40] M. W. Berry, A. N. Langville, V. P. Pauca, R. J. Plemmons, Comput. Stat. Data Anal. 2007, 52(1), 155.
[41] X. Geng, L. Ji, K. Sun, Front. of Inf. Technol. Electron. Eng. 2016, 17(5), 403.
[42] Anaconda Software Distribution, Anaconda, Vers. 2-2.4.0, 2016.
[43] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Mach. Learn. Res. 2011, 12, 2825.

SUPPORTING INFORMATION
Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Blee AL, Day JCC, Flewitt PEJ, Jeketo A, Megson-Smith D. Non-negative assisted principal component analysis: A novel method of data analysis for raman spectroscopy. J Raman Spectrosc. 2021;52: 1135–1147. https://doi.org/10.1002/jrs.6112