JOURNAL OF CHEMOMETRICS
J. Chemometrics 2007; **21**: 24–34
Published online 8 May 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/cem.1027

Estimating the number of pure chemical components in a mixture by maximum likelihood

E. Levina[1]*, A.S. Wagaman[1], A.F. Callender[2], G.S. Mandair[2] and M.D. Morris[2]

[1]Department of Statistics, The University of Michigan, Ann Arbor, MI 48109, USA
[2]Department of Chemistry, The University of Michigan, Ann Arbor, MI 48109, USA

Received 31 August 2006; Revised 4 December 2006; Accepted 6 December 2006

This paper addresses the problem of determining the number of pure chemical components in a mixture by applying the maximum likelihood estimator (MLE) of intrinsic dimension. The application here is to Raman spectroscopy data, although the method is general and can be applied to any type of data from a chemical mixture. We show that the MLE produces superior results compared to other methods on both simulated and real chemical mixtures, and is accurate even when minor components are present. Even if the signal-to-noise (SN) ratio is very low, accurate estimates can still be obtained by smoothing the data before applying the estimator, this approach is illustrated on two real datasets with high noise levels. Since the MLE is computed locally at every data point, we also show how the local estimates can be used for other applications, such as segmenting the specimen into homogeneous regions. Copyright © 2007 John Wiley & Sons, Ltd.

KEYWORDS: factor analysis; principal component analysis; pure components; maximum likelihood; Raman spectroscopy

1. INTRODUCTION

Self-modeling curve resolution (SMCR), a family of twenty or more multivariate algorithms extensively used in chemometrics, is designed to extract meaningful pure component spectra from unresolved multi-component mixture spectra [1,2]. SMCR does not require a priori knowledge of the pure component spectra, which is particularly useful in characterizing unknown chemical species [3,4]. In general, homogeneity, concentration, and the number of pure components within the chemical or biological system is seldom known in advance [5,6].

SMCR algorithms have a variety of applications, though our main focus will be on Raman spectroscopy. Morris and coworkers have extensively used factor analysis-based SMCR algorithms on Raman images to determine the mineral and matrix components contained within bone tissue [7–10]. In this context, factor analysis is used interactively to allow the user to visualize different linear combinations of potentially useful eigenvectors, which are in turn extracted by principal component analysis (PCA) [11]. This approach can be used to distinguish between healthy and diseased bone tissues [8–10], as well as to highlight chemical differences between trabecular and cortical bone structures at the micro-structural level [7]. Factor analysis is also useful for removal of non-informative eigenvectors associated with background tissue fluorescence and bone tissue embedding reagents, such as poly(methyl methacrylate) (PMMA) [7,8].

Although factor analysis is a useful and popular technique, it frequently runs into difficulties with over- or under-determination because it assumes that an appropriate number of eigenvectors has been selected [12,13]. In practice, the number of meaningful components (eigenvectors) in factor analysis is usually determined by the user through visual inspection, sometimes using prior knowledge on the chemical composition of the specimen. Alternatives to factor analysis include Simplisma and band-target entropy minimization (BTEM) [13]. BTEM allows the user to use forty or more additional eigenvalues and corresponding factors by performing exhaustive band targeting at different wavelengths. It has been shown to outperform Simplisma in terms of recovering minor components [13]. While the algorithm uses a large number of eigenvalues, the final decision on the number of extracted components still has to be made by the user and is based on visual inspection of the eigenvectors. It also places considerable demands on the user in terms of computational time and human interaction (manual rotation, exhaustive band targeting) [13]. This is especially undesirable if the user needs to analyze a large collection of Raman images or spectra.

All variants of SMCR in the literature start from extracting the eigenvectors by PCA, which is a linear dimension reduction method. There are good reasons for the use of linear methods, such as the Beer's law for mixture spectra [14]. However, the data are very complex, signal-to-noise

*Correspondence to: E. Levina, Department of Statistics, The University of Michigan, Ann Arbor, MI 48109, USA.
E-mail: elevina@umich.edu

(SN) ratios in real-time *in situ* Raman experiments are often low, and the structure of the noise may, in general, be non-linear. This situation may be further compounded by local variations in (SN) ratios as the Raman scattering properties of the irradiated specimen depend on its surface morphology and chemical composition; this may make the use of a global linear method like PCA inappropriate.

In the statistics and computer science literatures, recent focus has been on non-linear dimension reduction methods, such as the Locally Linear Embedding [15] and the Isomap [16]. These methods are designed for data on a 'manifold'—a non-linear smooth (low-dimensional) subspace of a bigger (high-dimensional) space. While these methods can easily handle non-linear dimension reduction, their major limitation is that they do not extract the principal components themselves—instead, they return projections of all the data points onto the reduced-dimension space. This makes them inappropriate for use in chemometrics, where the principal components are needed to obtain meaningful pure component spectra. However, their popularity has led to development of methods for estimating *intrinsic dimension* of non-linear manifolds from data, see for example [17]. These methods can handle non-linear data and produce accurate estimates of the manifold dimension, which, in the chemometrics context, corresponds to the number of pure components contained in the mixture. Having an accurate estimate of the number of components reduces the need for visual inspection and other user interventions, makes the analysis less subjective, and saves time in component extraction; it is also of independent interest when little a priori information is available about the specimen. Moreover, this estimate can be computed locally (i.e., at every data point); this local information can then be used for finding interesting regions in the image or for testing mixture homogeneity.

In this paper, we apply a non-linear dimension estimator, the maximum likelihood estimator (MLE) of intrinsic dimension, to the problem of determining the number of pure components in a mixture from Raman spectroscopy data, though the method can be applied to any spectral data and even more generally. We show how the intrinsic dimension corresponds to the number of pure components and introduce the MLE, as well as review existing methods in Section 2. In Section 3, we discuss the selection of a tuning parameter, and show on simulated mixtures that the MLE produces superior results compared to other methods, and is accurate even when minor components are present. In Section 4, we show how to handle low (SN) ratios in the data to obtain accurate estimates, and illustrate by applying the MLE to two real datasets with high noise levels in Section 5. In Section 6, we show how computing local estimates at every image pixel can be used to automatically divide the image into homogeneous regions. Section 7 concludes with discussion, and experimental details are given in the Appendix.

2. METHODS OF ESTIMATING THE NUMBER OF PURE COMPONENTS

Many SMCR methods require an estimate of the number of the components to be extracted, but few in fact make use of formal estimates. Examining a scree plot or plots of extracted eigenvectors by eye remains the prevalent method of analysis, and while the human eye can often be more accurate than an automated method, visual procedures are subjective, inconsistent across users, and time-consuming. In this section, we first review the methods that are currently available (even if not necessarily used) for estimating the number of components, and then present the new maximum likelihood method.

2.1. Current methods for estimating the number of pure components

Principal components analysis (PCA) is a dimension reduction technique which finds linear combinations of the original variables that best explain the variability in the data. Let $X_1, X_2, \ldots, X_n$ be the p-dimensional data vectors, where, in the context of Raman spectroscopy data, p is the total number of wavenumber values and n the number of pixels in the image. PCA consists of computing the eigenvalues and eigenvectors of the data covariance matrix $\Sigma = 1/n \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})'$, where A' denotes matrix transpose, and $\bar{X} = 1/n \sum_{i=1}^{n} X_i$. Alternatively, the principal components can be computed from the singular value decomposition of the data matrix X. The eigenvalues of Σ, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$, represent the amount of variation in the data explained by the corresponding principal component. A scree plot of these eigenvalues can be used to estimate the true dimension by eye. A somewhat more principled approach is to estimate the dimension of the data by the number of principal components that explain a pre-specified (large) fraction of the variance in the data:

$$\hat{s} = \arg\min_{s} \left\{ s : \sum_{j=1}^{s} \lambda_j \geq (1-\varepsilon) \sum_{j=1}^{p} \lambda_j \right\} \tag{1}$$

Choosing an appropriate fraction $1 - \varepsilon$ generally depends on the amount of noise in the data, which is not known in advance. This is one difficulty with using PCA for estimating dimension. Typically the fraction chosen is at least 90%; we use $\varepsilon = 0.01$ in the results shown below.

The Malinowski's F-test [18] was introduced in the chemometrics literature to differentiate between significant and noise eigenvectors in PCA. The sum of the eigenvalues $\sum_{j=1}^{p} \lambda_j$ can be decomposed into pieces representing significant and noise eigenvalues, with the number of significant eigenvalues providing an estimate of the number of pure components. The test starts from the smallest eigenvalue λ_p and goes through the eigenvalues in increasing order until it finds the first significant one. Once one eigenvalue has been determined to be significant, all larger eigenvalues are also considered significant. The Malinowski's F-statistic for testing the significance of the s-th eigenvalue is given by

$$F_s = \frac{\lambda_s}{\sum_{j=s+1}^{p} \lambda_j / (p-s)} \tag{2}$$

DOI: 10.1002/cem

Under the null hypothesis that the s-th eigenvector is noise, Malinowski argued that F_s has an F distribution with 1 and $p-s$ degrees of freedom. The estimated number of components based on the F-test can be computed as

$$\hat{s} = \min_s \{F_s > f_{1,p-s}(1-\alpha)\} \tag{3}$$

where $f_{1,p-s}(1-\alpha)$ is the $(1-\alpha)$ critical value for the $F(1, p-s)$ distribution. Again, the choice of α, like the choice of ε in Equation (1), is at the user's discretion; we will use $\alpha = 0.01$ throughout the paper. In any case, since the test is repeated until the first significant eigenvalue is found, this creates a multiple testing problem (see, e.g., [19]), and the actual overall significance level will be higher than α. A comparison of similar techniques and an adapted F-test can be found in [20].

2.2. Maximum likelihood estimation of dimension

The maximum likelihood estimator (MLE) of intrinsic dimension [17] was originally proposed for estimating the intrinsic dimension s of data $X_1, X_2, \ldots, X_n$ which are measured as p-dimensional vectors, but in fact lie on a 'manifold'—that is, in an s-dimensional subspace of the p-dimensional space, with s typically much smaller than p. Our goal here is to show how the MLE of intrinsic dimension can be applied to spectroscopy data and to resolve practical issues that arise in the process, such as choosing the tuning parameter and dealing with high levels of noise in the data.

The idea of the MLE is to fix an arbitrary point x, and assume the density $f(x)$ of the observations is constant in a small sphere of radius R around x. Then the points falling into this sphere form a Poisson process, the likelihood of which can be written explicitly (see [17] for details). This leads to a local estimator of dimension around a point x

$$\hat{s}_R(x) = \left[\frac{1}{N(R,x)} \sum_{j=1}^{N(R,x)} \log \frac{R}{T_j(x)} \right]^{-1} \tag{4}$$

where $N(R, x)$ is the number of points in the sphere of radius R around x, and $T_j(x)$ is the Euclidean distance from x to its j-th nearest neighbor in the sample.

Alternatively, instead of fixing the radius R of the sphere, one may fix the number of points falling into the sphere, which is often more intuitive and hence easier to pick. The estimate can then be rewritten as

$$\hat{s}_k(x) = \left[\frac{1}{k-2} \sum_{j=1}^{k-1} \log \frac{T_k(x)}{T_j(x)} \right]^{-1} \tag{5}$$

The factor $k-2$ appears in the denominator to make the estimator unbiased. We will use the fixed k version expressed in Equation (5) throughout this paper.

Equation (5) allows us to obtain an estimate of the intrinsic dimension at every data point X_i. Then the global estimate (or an estimate over a particular region) can be computed by averaging the local estimates over the entire data set (or the region in question). The global MLE for the whole dataset is given by

$$\hat{s}_k = \frac{1}{n} \sum_{i=1}^{n} \hat{s}_k(X_i) \tag{6}$$

We discuss the choice of k and the sensitivity of the estimator to k in Section 3; usually k is chosen to be a relatively small number, and the estimator is robust to the choice of k. Note also that in general neither Equation (5) nor Equation (6) give an integer estimate of dimension; in practice Equation (6) is rounded to the nearest integer.

The intrinsic dimension of the space generated by the spectra is not in itself an estimate of the number of pure components present in the mixture. Consider the following: a 'mixture' of two points generates a line, which has intrinsic dimension one, and three points generate a plane, which has intrinsic dimension two. Since the MLE estimates the dimension of the manifold generated by the mixture (the line or the plane), we add one to the MLE of intrinsic dimension in order to obtain the MLE for the number of pure components present in the mixture.

An advantage of the MLE is that it automatically generates an estimate of the number of components at every data point (pixel). While some variability in local estimates is expected even in homogeneous mixtures due to noise, a lot of variability indicates that the mixture under examination is likely not homogeneous. Using the local MLEs for testing for mixture homogeneity is currently under investigation; an application of local MLEs to segmenting the Raman image into homogeneous regions is presented in Section 6.

3. SIMULATION RESULTS

In this section we investigate the performance of all three estimators (PCA, F-test, and the MLE) on simulated mixtures obtained from real spectra. Each pure material was scanned separately, and their individual spectra are combined into a mixture as follows. The pure component spectra are combined into a single $s \times p$ matrix A, where s is the number of pure component spectra and p the number of different wavenumber values at which the spectra were measured. To generate n mixture spectra $X_1, \ldots, X_n$, which we combine into a single $n \times p$ matrix X, we first generated a random $n \times s$ matrix of component weights W. The distribution of the weights in W is described in detail below. The mixture spectra are generated according to Beer's law, which states that pure component spectra are linearly combined to form mixture spectra [14]. Since the application we focus on here is Raman spectroscopy, the linear Beer's law is the most obvious and appropriate way to produce simulated data, which makes it an easier problem for PCA. In other situations, where the pure components are combined in a non-linear way, the PCA would suffer much more, whereas the MLE would not be affected. Once the mixture spectra matrix is obtained from Beer's law, a $n \times p$ matrix of i.i.d. Gaussian noise, ϵ, is added to the generated mixture spectra, yielding the following data generation model:

$$X = WA + \epsilon \tag{7}$$

DOI: 10.1002/cem

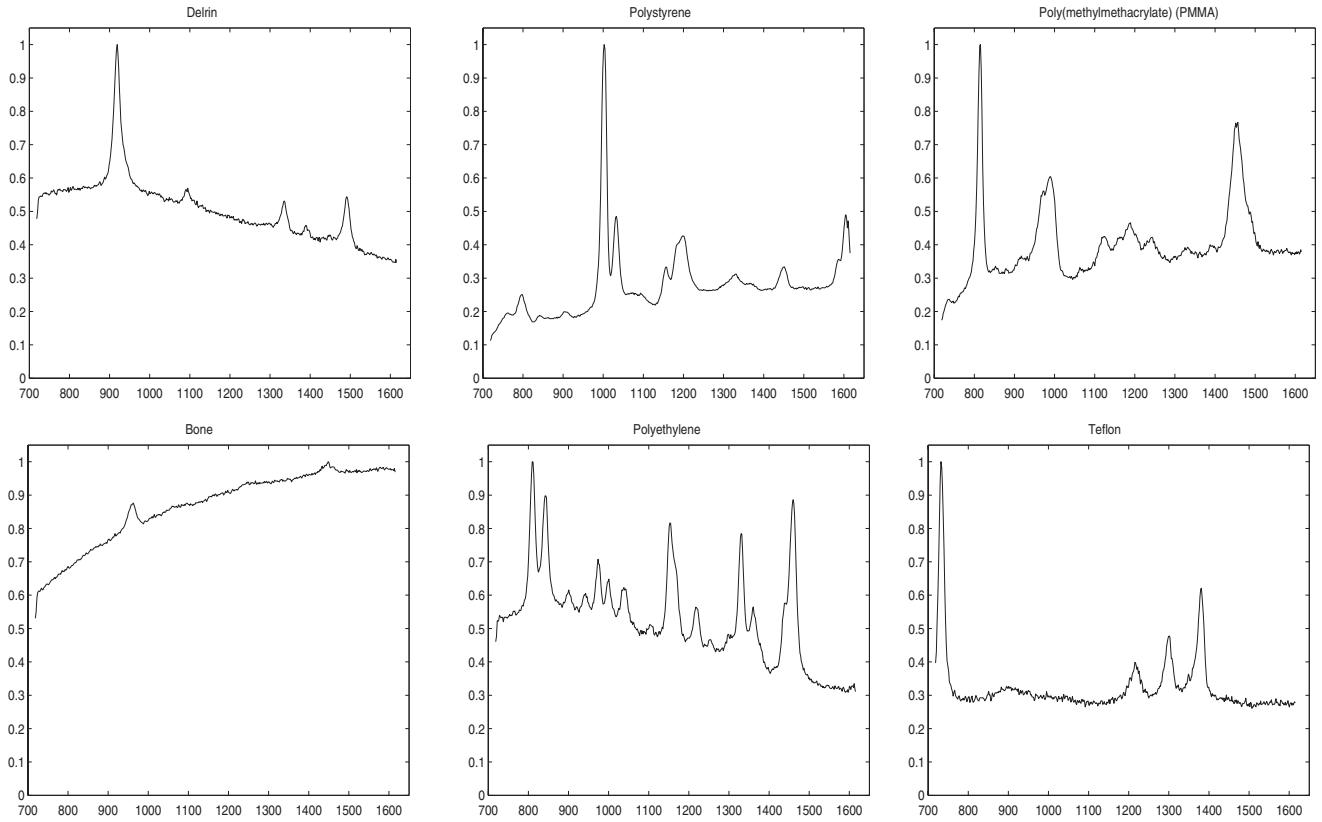

Figure 1. Test set 1 (dissimilar spectra). The pure component spectra of plastics and bovine bone are rescaled to maximum intensity 1, horizontal axis shows Raman shift (cm^{-1}).

3.1. Data description

For generating simulated mixtures, we used two separate test sets of pure component spectra. The first set consists of five plastics and one bovine bone spectra collected on the visible Raman system, which are quite dissimilar from each other and should be easy to discriminate. The pure components in this set are polyethylene, Delrin, polystyrene, poly(methyl methacrylate) (PMMA), bovine bone, and Teflon, measured at $p = 512$ wavenumber values in the range 700–1600 cm^{-1}. For details on the visible Raman system and experimental conditions, see Appendix A. The pure spectra rescaled to have maximum intensity 1 (to compensate for different amount of material present in each scan) are shown in Figure 1, with distinct spectral features of each component clearly visible by naked eye.

The second set of spectra contains five spectra of a fractured mouse tibia bone and one plastic (PMMA) collected on the NIR system (see Appendix A for details) and measured at 815 different spectral values. The PMMA is used to embed the fractured bone, and the five bone spectra are measured at different distances along the mouse bone, gradually moving away from the fracture. The five bone spectra vary with the distance away from the fracture but these differences in the spectra are minute (see Figure 2). The further away from the fracture, the less the spectra differ; in fact, the last two bone spectra, measured at 900 μm and 1100 μm away from the fracture, are identical. Hence, this set of 6 spectra contains only 5 distinct pure components.

We examined many combinations of weight matrices and noise levels in our simulations. For each set of six spectra we always select four major components (Delrin, polystyrene, PMMA, and bone for set 1, and PMMA and the three bone spectra closest to the fracture for set 2). The remaining two spectra in each set were used as minor components, to test whether the methods are able to pick components present in small amounts. For the setting with just four major components, each component's weight was drawn uniformly from the interval (0.15, 0.30). When minor components were added at 10% and 5% levels, major component weights were drawn uniformly from (0.15, 0.25), and minor weights from (0.05, 0.15) and (0.03, 0.07), respectively. In Section 4, we push the minor components level down to 1%, in which case we draw the major weights from (0.20, 0.30), and the minor weights from (0.00, 0.02). In each case, weights randomly drawn from a uniform distribution on each interval were rescaled to sum to one. Gaussian noise was added at a q% level, which means that the noise has mean 0 and variance $(0.01q)^2$.

3.2. Results

The results presented for each spectra set (Tables I and II) are representative of all simulations we performed. In each case, the number of pixels is $n = 3600$ (60 by 60 image), and the estimated numbers of components are averaged over 100 replications. We compare the MLE at $k = 20$ (the choice of tuning parameter is discussed below), PCA at 99% variance explained, and Malinowski's F-test at 1% significance level. If the estimate is given as an integer for example integer 4, it means there was no variation in the estimate across the 100 replications. If there was variation but the average came out to be 4, it is given as 4.0.

DOI: 10.1002/cem

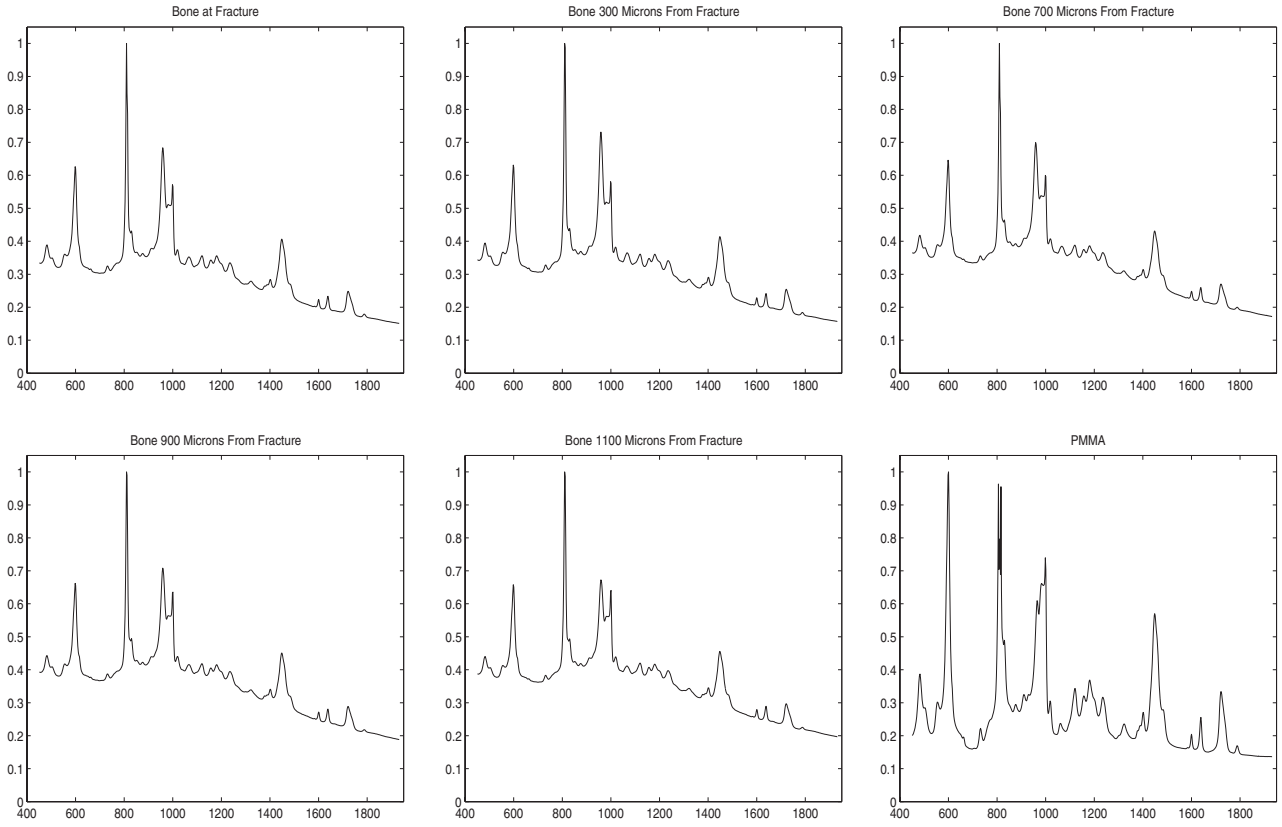

Figure 2. Test set 2 (similar spectra). The pure component spectra of mouse bone and PMMA are rescaled to maximum intensity 1, horizontal axis shows Raman shift (cm^{-1}).

The levels of noise are chosen to show where the MLE starts picking up noise components. For both test sets, it is clear that at low noise levels the MLE estimate performs as well as or better than the other two. Note that PCA fails to obtain the correct number of components even with no noise if minor components are present. The *F*-test can obtain the correct estimates with no noise, but fails once noise is added. In fact, as noise levels increase, all methods begin to suffer, but the MLE is more sensitive to noise than PCA. As expected, it takes less noise for the estimates to break down when the spectra are very similar than when they are different. Otherwise, the same pattern holds for both test sets. The issue of dealing with high levels of noise is addressed in Section 4.

3.3. Choice of the tuning parameter *k* for the MLE

The MLE estimate requires choosing a value of k, the number of nearest neighbors around each point on which the local estimator is based. The impact of k on the estimate is examined via simulation. Figure 3 shows the MLE estimate plus and minus one standard deviation versus k over 100 replications for three different settings. Keeping in mind that

Table I. Test set 1 (dissimilar spectra): estimated number of pure components

No. of components	4	4	4	4	6	6	6	6	6	6	6	6
Minor level (%)	0	0	0	0	10	10	10	10	5	5	5	5
Noise level (%)	0	0.05	0.1	0.3	0	0.05	0.1	0.3	0	0.05	0.1	0.3
MLE	4.0	4.2	4.9	10.9	5.6	5.8	6.4	12.2	5.5	5.8	6.6	15
PCA (99%)	4	4.0	5	5	5	5	5.2	6	5	5	5.1	5.1
F-test (1%)	4	4	4	1	6	6	1	1	6	6	1	1

Table II. Test set 2 (similar spectra): estimated number of pure components

No. of components	4	4	4	4	5	5	5	5	5	5	5	5
Minor level (%)	0	0	0	0	10	10	10	10	5	5	5	5
Noise level (%)	0	0.005	0.01	0.03	0	0.005	0.01	0.03	0	0.005	0.01	0.03
MLE	3.9	4.1	4.7	10.2	4.7	4.9	5.5	11.7	4.5	4.5	5.7	14
PCA (99%)	3	4	4.0	5	4	5	5	6	4	5	5.1	6
F-test (1%)	4	4	3	2	6	3	3	2	6	3	2	2

DOI: 10.1002/cem

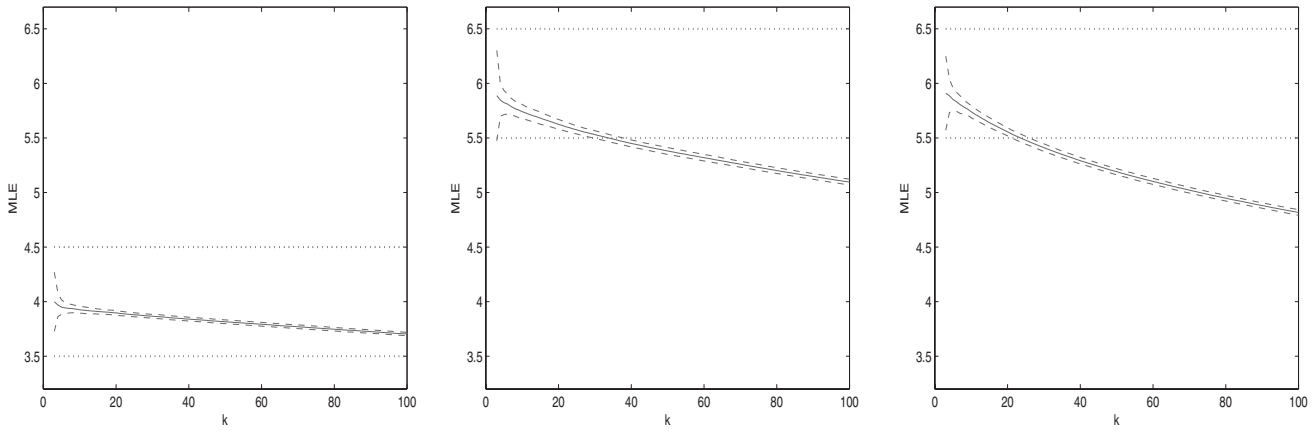

Figure 3. Sensitivity to k for (a) 4 major components, (b) 4 major and 2 minor components at 10% and (c) 4 major and 2 minor components at 5% for sample size, n = 1000, and results averaged over 100 replications. Dashed lines show the range where the estimate is rounded to the correct value.

the estimate is rounded to the nearest integer, we see that the MLE estimates vary very little across replications (small standard deviation), and are fairly robust to the choice of k. The derivation of the MLE involves an approximation that requires k to be small relative to n, and smaller values of k reduce the amount of computation; on the other hand, very small values of k may lead to too much variability in local estimates. On the balance, we chose $k = 20$ and kept it constant for all simulations and real data applications.

3.4. Impact of image size

The behavior of all estimators can in general be affected by the amount of data available. In general, larger images are better since they contain more information about the mixture. To study this effect, we performed simulations with image sizes of $n = 400, 1000$, and 3600 (detailed results omitted). The standard errors of the MLE estimate decrease as the image size increases, as expected. When the noise level is high, the MLE estimates for $n = 3600$ are much higher than the $n = 400$ estimates. This is expected: the noise is overwhelming the signal, and as the estimate becomes more accurate for larger n, it picks up more noise components. This issue is addressed in detail in Section 4. Finally, we note that for large image sizes, the computational complexity associated with the SVD makes the MLE, which only requires finding the nearest neighbors, a much more attractive choice than both the PCA and the F-test.

4. DEALING WITH HIGH LEVELS OF NOISE

Simulations showed the MLE method is sensitive to noise, which is common in real data. When high levels of noise are present, smoothing the data before applying the procedure can enhance the performance of the estimator. There are two types of smoothing one can consider: smoothing along each spectrum, and smoothing spatially across the image. Individual spectra can be smoothed, for example, with a Blackman-Harris filter, a signal processing tool available in many software libraries [21]. We found that smoothing the spectra helps somewhat, but is less efficient than spatial smoothing. When both methods of smoothing are combined, the effect is the same as that of spatial smoothing alone. Therefore we choose not to smooth the individual spectra at all, which allows us to better preserve the peaks and other spectral features.

Spatial smoothing can be achieved via a convolution of the image X with a filter matrix Q. At each spectral wavenumber l and pixel location (x, y), we compute

$$X_Q(l, x, y) = \sum_{u=-\infty}^{\infty} \sum_{v=-\infty}^{\infty} X(l, u, v) F(x - u, y - v) \qquad (8)$$

where values of matrices outside of the valid index range are defined to be zero.

The filter matrix, generally speaking, averages the values around (x, y), and many choices are possible (simple averaging, weighted averaging over a fixed window, exponentially decaying weights over the whole image, etc). We found a simple spatial moving average (MA) filter to perform very well in this context. The MA filter replaces the value at each pixel with the average of pixel values in a $w \times w$ window around it. The window size is taken to be odd for convenience, $w = 2m + 1$, and the convolution formula reduces to

$$X_Q(l, x, y) = \frac{1}{w^2} \sum_{u=x-m}^{x+m} \sum_{v=y-m}^{y+m} X(l, u, v) \qquad (9)$$

We only compute this for x and y that are at least m pixels away from the edges of the image and discard the rest.

Finally, we investigated smoothing across neighbors in terms of spectral similarity rather than spatial location. This technique is often used for data on a manifold, for smoothing over manifold neighbors. We have investigated a moving average smoother over 'spectral' neighbors and the iterative locally linear smoothing technique [22]. The results were found to be inferior to spatial smoothing. The spatial moving average is therefore our final choice and the only technique we present results for, due to space limitations.

DOI: 10.1002/cem

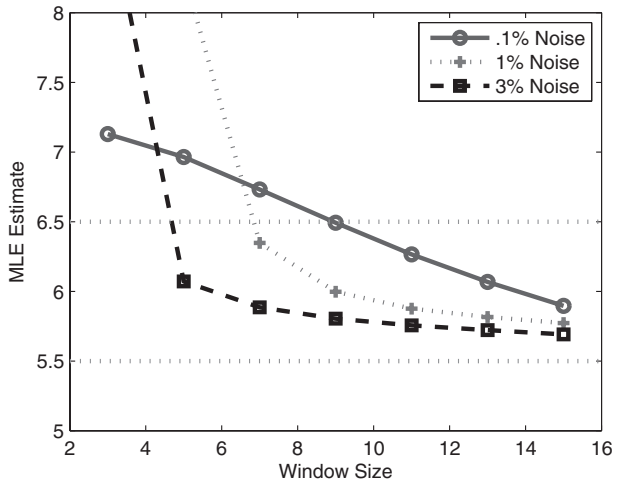

Figure 4. MLE estimates on smoothed data as a function of MA window size at various noise levels for dissimilar spectra with 4 major + 2 minor components at 10%, $n = 3600$, $k = 20$. The horizontal lines show the range where the estimator is rounded to the correct value.

4.1. Choice of window size for smoothing

The window size w for the MA smoother is another tuning parameter to be selected. We investigated many window sizes in extensive simulations; Figure 4 shows a representative plot of MLE estimates applied to smoothed data for several noise levels as a function of window size. The setting is dissimilar spectra with four major components and two minor components at 10%.

An important conclusion from Figure 4 is that it is better to over-smooth than to under-smooth. For consistency throughout the paper, we selected the first window size for which the MLE estimates at all noise levels obtained a correct estimate, which is $w = 9$ (a 9×9 window), which we will use in all results below. In general, we recommend this as a rule-of-thumb starting value, since over-smoothing does not appear to present a problem. Plots of the MLE as a function of window size like the ones in Figure 4 can also be investigated and w selected as the point where the plot 'levels off'.

4.2. Simulation results with high levels of noise

To examine the impact of smoothing on all three estimators, we performed simulations with increased noise levels. All settings were the same as before, except Gaussian noise was added at higher levels (0.1%, 1%, or 3%). The results are shown in Table III (dissimilar spectra, all six are distinct) and Table IV (similar spectra, five out of six are distinct). A 9×9 window was used for spatial smoothing of the 60×60 image.

The results show that, while the MLE estimate is the most sensitive to high levels of noise, it is also the only one that is able to obtain correct (on average) estimates on smoothed data. Both PCA and the *F*-test are not coming close to the truth whether the data are smoothed or not, which suggests that generally their results cannot be trusted for data with high noise levels. The MLE, on the other hand, performs well on smoothed data even when the noise level (3%) is higher than the amount of minor components present (1%).

5. APPLICATIONS TO REAL DATA

The simulation results in the previous section were based on using real spectra which were artificially combined into a simulated mixture. In contrast, here we apply the proposed methodology to Raman images of real specimen (see Section A for experimental details). Based on results in Sections 3 and 4, we set the MLE tuning parameter to $k = 20$ and apply a spatial moving average smoother with a window size of nine as a preprocessing step. For a fair comparison, we report

Table III. Dissimilar Spectra smoothing results: $n = 3600$, $k = 20$, $w = 9$, 4 major + 2 minor components, results averaged over 100 replications

Minor level (%)	10	10	10	5	5	5	1	1	1
Noise level (%)	0.1	1	3	0.1	1	3	0.1	1	3
MLE	6.7	58.7	101.2	7.1	65.7	104.6	8.2	74.5	110.7
Smoothed MLE	6.0	5.4	5.3	6.1	5.5	5.2	6.4	5.5	5.2
PCA	5	31	38	4	31	37	4	31	36.1
Smoothed PCA	1	5	20	1	4.1	19	1	4.6	17.9
F-test	1	0	0	1	0	0	1	0	0
Smoothed *F*-test	1	1	1	1	1	1	1	1	1

Table IV. Similar Spectra Smoothing Results: $n = 3600$, $k = 20$, $w = 9$, 4 major + 2 minor components (five distinct), results averaged over 100 replications

Minor level (%)	10	10	10	5	5	5	1	1	1
Noise level (%)	0.1	1	3	0.1	1	3	0.1	1	3
MLE	64.7	132.9	153.1	59.9	128.5	155.6	82.1	137.6	158.4
Smoothed MLE	6.5	5.4	5.4	6.0	5.4	5.6	5.7	5.5	5.6
PCA	2	33	37	2	34	38	2	34.1	40
Smoothed PCA	1	14	24.9	1	14	26	1	14.1	26.9
F-test	1	0	0	1	0	0	1	0	0
Smoothed *F*-test	1	1	1	1	1	1	1	1	1

DOI: 10.1002/cem

Table V. Real data results for the three estimators applied to raw and smoothed (denoted by Sm) images

Data	MLE	MLE(Sm)	PCA	PCA(Sm)	*F*-test	*F*-test(Sm)
PMMA with 2 curing times	50	4	6	4	1	1
Bone embedded in PMMA	47	5	5	4	9	9

the results for the other two estimators for both raw and smoothed data.

5.1. Dataset 1: PMMA with two different curing times

Our first dataset is a 130 by 30 image of a polymer (PMMA), with Raman spectra measured at 512 spectral values. The specimen was obtained by combining two Koldmount mixtures at different stages of polymerization. Koldmount is commonly used to embed biological specimens. The solid component and the liquid are mixed; the reaction proceeds quickly to produce a translucent material. In this dataset, 'fresh' PMMA (three minutes after mixing) was layered onto partially cured PMMA (eight minutes after mixing). The image was taken at the interface. The details of the experiment are given in the Appendix.

The initial mixture contains four chemical components—PMMA particles, unreacted monomer, and two initiators (trace amounts). The mixture is not necessarily homogeneous. The volume fraction of unreacted monomer depends on the reaction rate and the time (post-mixing) at which any particular pixel was imaged. As a result, substantial variation in the proportions of the two major components is expected. This relatively simple system serves as an excellent test case for estimating the number of pure components.

The results are shown in Table V. If no smoothing is applied as a pre-processing step, all estimators give incorrect results. With smoothing, the MLE and PCA both pick up 4 components. The *F*-test only detects 1 component, with and without smoothing. The MLE on smoothed data was the same for a range of values of k and the moving average window size w.

5.2. Dataset 2: Bone

We also examined a 300 by 50 bone image consisting of Raman spectra measured at 512 spectral values. The bone was a murine femur, embedded in PMMA resin. A transverse section was chosen at the edge of the bone to include both bone and resin in the field of view. However, no significant concentration of resin was seen in the data; the reduced collection efficiency at the edges of the CCD left the section known to contain PMMA relatively dark. Based on previous experiments on similar specimens, the presence of PMMA distributed within the bone tissue is still expected. Thus, there are at least three major components expected in the data—PMMA, bone mineral, and bone matrix. There may also be additional bone components, depending on age and damage [13]. Here, MLE and PCA obtain five and four components respectively on the smoothed data, with the *F*-test obtaining nine.

Even though in real data, unlike in simulations, we do not know the correct answer exactly, the MLE appears to perform well on smoothed data. For these datasets, PCA and the MLE give comparable results; however, results in Tables III and IV suggest that in general the MLE of smoothed data is likely to be more reliable when high levels of noise are present.

6. USING LOCAL DIMENSION ESTIMATES FOR IMAGE SEGMENTATION

The MLE in Equation (5) is computed at every pixel, but so far we have been using the global average given in Equation (6) as the estimate for the number of components. We can also use the pointwise estimates for other tasks, such as finding regions with different numbers of components (areas with more components may be more chemically interesting), or evaluating homogeneity of the mixture. To illustrate the potential of local estimators, we demonstrate how they can be used to segment an image into regions with homogeneous numbers of components. While there are many segmentation procedures that could be applied (see, e.g., [23] and [24] for approaches based on Markov random fields, [25] for a contour-based segmentation, and [26] for an algorithm combining contour and texture information), the segmentation procedure we use here is normalized cuts [27], a general purpose graph clustering and image segmentation algorithm that has been shown to give good results in a variety of applications.

6.1. Image segmentation technique: normalized cuts

Normalized cuts, or Ncuts [27] is an image segmentation procedure that divides an image into regions by both maximizing similarity of points within each region and maximizing dissimilarity between regions. The procedure treats segmenting the data into regions as a graph partitioning problem. Pixels form the set of vertices V, and weights $w(x, y)$ on the edges between points $x \in V$ and $y \in V$ represent a measure of similarity between x and y. The partition of V into two non-overlapping sets A and B is then found by minimizing a function of the data called the normalized cut. The normalized cut between two regions A and B is defined to be

$$\mathrm{Ncut}(A, B) = \frac{\mathrm{cut}(A, B)}{\mathrm{assoc}(A, V)} + \frac{\mathrm{cut}(A, B)}{\mathrm{assoc}(B, V)} \tag{10}$$

where association and cut are defined as

$$\mathrm{assoc}(A, V) = \sum_{x \in A, y \in V} w(x, y); \quad \mathrm{cut}(A, B) = \sum_{x \in A, y \in B} w(x, y) \tag{11}$$

The idea is to find A and B that have the least similarities between them (minimize the cut) but penalize for segmenting out the regions that are not well connected within themselves—that is the purpose of normalizing by the association. If the normalization is omitted, the segmentation will tend to cut off single points.

DOI: 10.1002/cem

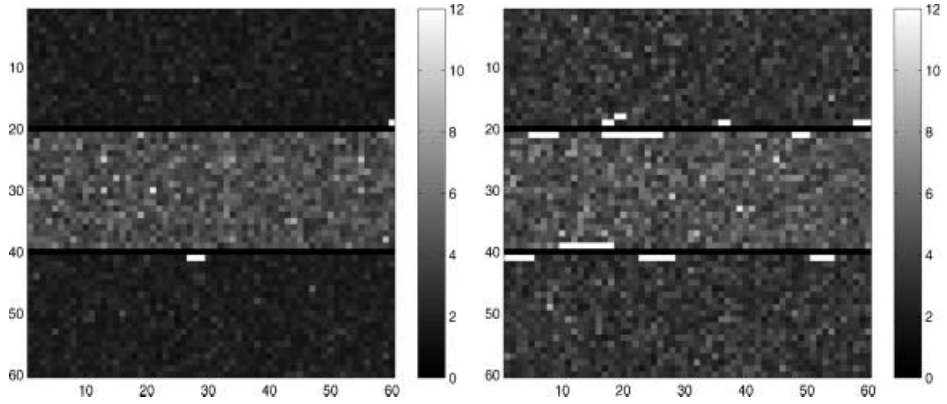

Figure 5. Segmentation results with 0.1% and 0.3% Gaussian noise; the pixel color values show local dimension estimates. Estimated boundaries are shown in white where they do not match the true boundaries in black.

In order to implement the normalized cuts, we need an appropriate measure of similarity between pixels. For regular image segmentation, Shi and Malik [27] used a similarity measure based on spatial distance between pixels and differences in their brightness values. For our application, we propose a similarity measure which reflects both the spatial distance between pixels and the differences in the estimated number of components at each pixel. For a pair of data points x located at (i_x, j_x) and y located at (i_y, j_y), with $\hat{s}_x$ and $\hat{s}_y$ the number of components estimated at each point, we define the weight on the edge between x and y to be

$$w(x, y) = \exp\left(-\frac{(\hat{s}_x - \hat{s}_y)^2}{\sigma_1} - \frac{(i_x - i_y)^2 + (j_x - j_y)^2}{\sigma_2}\right) \quad (12)$$

The scaling factors σ_1 and σ_2 can be used to vary how much importance is given to spatial proximity vs. the similarity in the number of components. They can also be used to resolve scaling issues if the two measures combined are not on the same scale. In this case, the two components in (12) are on the same scale already, and we set $\sigma_1 = \sigma_2 = 1$ in the results shown below.

The normalized cut problem itself is NP-hard, but a relaxation can be solved efficiently through a generalized eigenvalue problem. Here we briefly summarize the algorithm and refer the reader to [27] for details. Let $d(i) = \sum_j w(i, j)$ and let D be a diagonal matrix with diagonal d. Let $W = [w(i, j)]_{1 \le i, j \le n}$ be the symmetric matrix of edge weights. Finally, let v be an $n \times 1$ indicator vector with $v(i) = 1$ if the i-th data point is in A and -1 if it is in B. Now let

$$b = \sum_{x \in A} d(x) / \sum_{x \in B} d(x) \quad (13)$$

$$c = (1 + v) - b(1 - v) \quad (14)$$

In effect, c is a continuous approximation to v. Shi and Malik show that the solution to the normalized cuts problem can be found by solving the eigenvalue system

$$(D - W)c = \lambda Dc \quad (15)$$

The second smallest eigenvector of the system gives the first split, with the partition based on the signs of the entries in the eigenvector. The procedure can then be repeated to split the two regions A and B further.

6.2. Segmentation results

For our simulations, we divided a 60×60 image into three 20×60 horizontal strips with 3, 6, and 3 components, respectively. The three major components were bone, PMMA, and Delrin, with polystyrene added as a major component and Teflon and polyethylene as minor components at 10% in the middle region. Then we generated two images, with two levels of noise (0.1% and 0.3%). The local MLEs were computed at each pixel with $k = 20$. To keep the example straightforward, we set the levels of noise low enough so that smoothing was unnecessary. Similar results (not shown) were obtained with MA smoothing as a preprocessing step. The local estimates were used in Equation (12) to create the weights and normalized cuts were applied to segment out three regions. The resulting segmentations are shown in Figure 5. The procedure correctly finds the three regions we built into the data generation. For low noise, the average MLE in each region was 3.0, 5.4, and 3.0, respectively. For higher noise, the region averages were 4.4, 5.9, and 4.5. We can see that the MLE performs better at low noise levels in terms of obtaining the correct number of components (recall that no smoothing was applied to the data), but the segmentation is still correct at the higher noise level.

Since normalized cuts require the user to specify the number of regions to be segmented, we experimented with asking for more than three regions. In this case the procedure segments out additional very small areas, but the main three regions are still clearly visible. Hence this segmentation procedure can be used even if little is known a priori about the number of different regions in the image. Also note that we did not incorporate spectral similarity into the measure (12), and it is of course possible to have spectroscopically different regions with the same total number of pure components. This example was intended to illustrate the potential of local estimates; an in-depth investigation of their applications to segmentation and homogeneity testing is a subject of current research.

7. CONCLUSIONS

Determining the number of pure component spectra present in a mixture is an important step in SMCR, and having a

DOI: 10.1002/cem

reliable estimate of how many components to extract leads to more objective and accurate data analysis, as well as reduces the amount of human visual inspection and other manipulations. We have shown the maximum likelihood estimator of intrinsic dimension, designed for general data on non-linear manifolds, can be successfully applied to this problem, and tested its performance on both real and simulated mixtures of Raman spectra. The method is robust to the choice of tuning parameter and outperforms PCA and the Malinowski's *F*-test, particularly when minor components are present and/or the SN ratios are low. When the noise level is very high, additional preprocessing via spatial smoothing has been shown to produce good results.

The MLE of dimension is a general method and is likely to find applications in other areas of chemometrics. One advantage of the MLE is that it automatically generates an estimate of the number of components at every data point (in case of images, at every pixel). Here we illustrated the potential of these local estimates by using them to segment the specimen into homogeneous regions in terms of the number of components present. Local estimates can also be used to test for mixture homogeneity, which is an important pharmaceutical application; this application is a subject of future work.

APPENDIX A: EXPERIMENTAL MATERIALS AND METHODS

A.1. Raman instrumentation

Raman spectra were collected using two different systems: a Raman microprobe optimized for collection in the near-infrared (NIR) [13] and a purpose-built, visible Raman microscope [28]. Briefly, the NIR system consists of an epi-illumination microscope frame (Olympus, BH-2) and a 400 mW 785 nm laser (Invictus, Kaiser Optical Systems, Inc.). The laser light is line-focused through a Powell lens (Stocker Yale) and into a 20x/0.75 NA Fluar objective (Carl Zeiss, Inc). For the visible Raman system, a research grade microscope (Nikon E600) and a 2 W 532 nm laser (Spectra Physics, Millenia II) were used. The circular beam profile of the Millenia II laser is reshaped into a line using a Powell lens and focused through a 4x/0.20 NA infinity-corrected objective (Nikon). For visible Raman hyperspectral imaging, a single axis scanning mirror (64240H, Cambridge Technology, Inc., Cambridge, MA) was used [28]. A LabVIEW (National Instruments, Austin, TX) program controlled the mirror's position by adjusting the voltage sent to the mirror control board through a 12-bit digital-analog converter. The mirror could be positioned to approximately $\pm 0.2\,\mu$m with a setting of 1–3 ms. Raman scatters from both systems were collected using an $f/1.8$ axial transmissive spectrograph (Kaiser, HoloSpec). NIR and visible Raman scatter were detected using a back-thinned, deep depletion 1024×128 pixel CCD camera (Andor Technology) or an 512×512 pixel electron-multiplying CCD camera (iXon Andor Technology), respectively. The spectral axis was calibrated (pixel to wavenumbers) using emission lines from a neon or argon discharge lamp. Curvature corrections and data analysis were performed in Matlab 6.1 (The Mathworks Inc., Natick, MA) using built-in and locally-written scripts.

A.2. Chemical components (dissimilar spectra)

Raman spectra of bovine bone, polyethylene, polystyrene, Teflon, Delrin, and PMMA were acquired using the visible Raman system. All spectra were collected using an acquisition time of 10 seconds and within the 700–1600 cm^{-1} spectral range. Bovine bone specimens were obtained from a local abattoir and sectioned into $5 \times 10 \times 2$ mm blocks under constant irrigation using a diamond wheel saw. The sections were rinsed with calcium-buffered saline solution to remove any blood residues, and stored at -30°C until required.

A.3. Fractured mouse bone (similar spectra)

Raman spectra of the fractured bone specimen embedded in PMMA were collected using the NIR Raman microprobe. A series of spectra were taken in parallel with the fracture from the edge at 100 μm intervals. Spectra were collected using a seven minutes integration time to ensure good SN ratios. To prepare the fractured mouse bone specimens, a heavy rounded blade was dropped onto the tibia of a 10 months old wild-type mouse. Fractured mouse tibias were harvested according to a protocol approved by the University of Michigan Institutional Committee on Use and Care of Animals. The specimens were embedded in PMMA, sectioned, and polished to reveal the fractured ends of the bone.

A.4. PMMA curing

Koldmount (Vernon & Bishoff, Albany, NY) is a two-part acrylic resin commonly used to embed biological specimens for microscopy and archival preservation. The solid component (poly(methyl methacrylate) plus benzoyl peroxide as an initiator) and the liquid (methyl methacrylate monomer plus *N*,*N*-dimethyl-*p*-toluidine as an initiator) are mixed; the reaction proceeds quickly to produce a translucent (highly scattering) material. Koldmount powder (2.3 g) and Koldmount liquid (1.5 mL) were mixed together using vendor-supplied protocols. The mixture was stirred briefly at ambient temperature and immediately poured into a polystyrene cuvette. A second batch of Koldmount was prepared after five minutes, mixed and poured into the same cuvette, on top of the partially cured material. The fresh mixture was allowed to cure for three minutes (the minimum at which it no longer flows as a liquid). The cuvette was then turned on its side and placed on the microscope stage for Raman imaging. The reaction continued during the imaging.

Transects (30 in all) for the Raman image were collected on the 532 nm system with 500 mW excitation power and four seconds acquisition time. The spectra—initially 512 spectral values by 390 spatial pixels, 30 exposures altogether—were binned spatially to improve the SN ratio. This gave a data set consisting of 3900 spectra (130×30), each with 512 values in the spectral dimension (800–1500 cm^{-1}).

A.5. Bone image

Visible hyperspectral imaging of mouse bone specimens embedded in PMMA were performed using the scanning mirror described previously [28], measured over the range 800–1500 cm^{-1} (512 spectral values). Spectra were acquired

DOI: 10.1002/cem

with an integration time of 4 seconds per line; the excitation power was 300 mW. The image comprised of 30 lines with a reduced pixel size of 300 × 50.

Acknowledgements

We thank Martin Strauss and Mark Iwen for helpful discussions. This project was supported by the University of Michigan LSA Jump Start Fund for interdisciplinary research.

REFERENCES

1. Jiang JH, Liang Y, Ozaki Y. Principles and methodologies in self-modeling curve resolution. *Chemom. Intell. Lab. Syst.* 2004; **71**(1): 1–12.
2. Jiang JH, Liang Y, Ozaki Y. Self-modeling curve resolution (SMCR): principles, techniques, and applications. *Appl. Spectrosc. Rev.* 2002; **37**(3): 321–345.
3. Kleimeyer JA, Harris JM. Monitoring the formation and decay of transient photosensitized intermediates using pump-probe UV resonance Raman spectroscopy. I: Self-modeling curve resolution. *Appl. Spectrosc.* 2003; **57**(4): 439–447.
4. Chew W, Widjaja E, Garland M. Band-target entropy minimization (BTEM): an advanced method for recovering unknown pure component spectra. application to the FTIR spectra of unstable organometallic mixtures. *Organometallics* 2002; **21**(9): 1982–1990.
5. Sasic S, Clark DA. Defining a strategy for chemical imaging of industrial pharmaceutical samples on Raman line-mapping and global illumination instruments. *Appl. Spectrosc.* 2006; **60**(5): 494–502.
6. Budevska BO, Sum ST, Jones TJ. Application of multivariate curve resolution for analysis of FT-IR microscospectroscopic images of in situ plant tissue. *Appl. Spectrosc.* 2003; **57**(2): 124–131.
7. Timlin JA, Carden A, Morris MD, Bonadio JF, Hoffer CE, Kozloff KM, Goldstein SA. Spatial distribution of phosphate species in mature and newly generated mammalian bone by hyperspectral Raman imaging. *J. Biomed. Optics* 1999; **4**(1): 28–34.
8. Morris MD, Crane NJ, Gomez LE, Ignelzi MA. Compatibility of staining protocols for bone tissue with Raman imaging. *Calcif. Tissue Int.* 2003; **74**(1): 86–94.
9. Tarnowski CP, Ignelzi MA, Wang W, Taboas JM, Goldstein SA, Morris MD. Earliest mineral and matrix changes in force-induced musculoskeletal disease as revealed by Raman microspectroscopic imaging. *J. Bone Miner. Res.* 2004; **19**(1): 64–71.
10. Timlin JA, Carden A, Morris MD, Rajachar RM, Kohn DH. Raman spectroscopic imaging markers for fatigue-related microdamage in bovine bone. *Anal. Chem.* 2000; **72**(10): 2229–2236.
11. Otto M. *Chemometrics*. Wiley-Vch Verlag: Germany, 1999.
12. Hair JF, Anderson RE, Tatham RL, Black WC. *Multivariate Data Analysis*. Prentice Hall: New Jersey, 1998.
13. Widjaja E, Crane N, Chen TC, Morris MD, Ignelzi MA, McCreadie BR. Band-target entropy minimization (BTEM) applied to hyperspectral Raman image data. *Appl. Spectrosc.* 2003; **57**(11): 1353–1362.
14. Widjaja E, Garland M. Pure component spectral reconstruction from mixture data using SVD, global entropy minimization, and simulated annealing. Numerical investigations of admissible objective functions using a synthetic 7-spectra data set. *J. Comput. Chem.* 2002; **23**(9): 911–919.
15. Roweis ST, Saul LK. Nonlinear dimensionality reduction by local linear embedding. *Science* 2000; **290**(5500): 2323–2326.
16. Tenenbaum JB, de Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. *Science* 2000; **290**(5500): 2319–2323.
17. Levina E, Bickel PJ. Maximum likelihood estimation of intrinsic dimension. *Advances in NIPS* 17 MIT Press: Cambridge, 2005.
18. Malinowski ER. Statistical F-tests for abstract factor analysis and target testing. *J. Chemom.* 1988; **3**(1): 49–60.
19. Storey JD, Tibshirani R. Statistical significance for genome-wide studies. *PNAS* 2003; **100**(16): 9440–9445.
20. Malinowski ER. Abstract factor analysis of data with multiple sources of error and a modified Faber-Kowalski F-test. *J. Chemom.* 1999; **13**(2): 69–81.
21. Harris FJ. On the use of windows for harmonic analysis with the discrete Fourier transform. *Proceedings of the IEEE* 1978; **66**(1): 51–83.
22. Zhang Z, Zha H. Local linear smoothing for nonlinear manifold learning. CSE-03-003, Technical Report, CSE, Penn State University, 2003.
23. Geman S, Geman D. Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* 1984; **6**: 721–741.
24. Comer ML, Delp EJ. The EM/MPM algorithm for segmentation of textured images: analysis and further experimental results. *IEEE Trans. Image Proc.* 2000; **9**(10): 1731–1744.
25. Jacobs D. Robust and efficient detection of salient convex groups. *IEEE Trans. Pattern Anal. Mach. Intell.* 1996; **18**(1): 23–37.
26. Malik J, Belongie S, Leung T, Shi J. Contour and texture analysis for image segmentation. *IJCV* 2001; **43**(1): 7–27.
27. Shi J, Malik J. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. and Mach. Intel.* 2000; **22**(8): 888–905.
28. Golcuk K, Mandair G, Callender A, Sahar N, Kohn D, Morris MD. Is photobleaching necessary for Raman imaging of bone tissue using a green laser? *Biochim. Biophys. Acta* 2006; **1758**(7): 868–873.

DOI: 10.1002/cem