Processing Spectral Data

L. Görlitz and F. A. Hamprecht*
Heidelberg Collaboratory for Image Processing (HCI),
Interdisciplinary Center for Scientific Computing (IWR),
University of Heidelberg, Germany

Abstract

Spectral images offer more information on complex probes than either conventional imagery (yielding only one or few measurements per voxel) or conventional spectroscopy (resulting in only one or a few spectra per probe) can. Spectral imaging is thus starting to become ubiquitous in areas ranging from materials science and process control to remote sensing and medicine. However, the information concealed in the massive amount of data generated by spectral imaging is not as easily accessible as in conventional (gray value or color) images and as in conventional spectroscopy, hence calling for new methods to analyze and visualize spectral images.
Broadly speaking, analysis methods can be classified in terms of the information used in the training phase, and by the extent to which they use spatial context in the analysis. This article gives a brief overview of "unsupervised" methods that require sample spectral images during the training stage as well as specification of a degree of complexity of the sample, either in terms of number of cluster centers or intrinsic dimensionality of the data (principal component analysis (PCA)). "Supervised" methods, on the other hand, require a training set in which a class membership, or label, is available for each pixel. Our discussion includes methods that deal well with the high dimensionality and high degree of correlation among individual features, such as partial least squares (PLS) and linear discriminant analysis (LDA).
All of the above methods ignore the spatial context when applied to individual spectra. Often, the label map can be assumed to exhibit some spatial coherence, a fact that can help in classifying low quality spectra and that can be exploited using conditional or Markov random fields. The methods are illustrated using examples from automated process control and quantitative medical diagnostics.

I. Introduction

Modern data acquisition procedures make spectral data abundant in many areas as diverse as tumor diagnosis, materials science and process control. The sheer amount of acquired spectra often prohibits complete analysis of the data by a human expert, leading to a strong demand for automatic analysis methods. Accessing the information contained in these spectra with conventional data analysis methods like linear regression or logistic regression may fail due to two reasons: the high collinearity and high dimensionality of the data.
Spectra can be represented as vectors with each channel of the spectrum corresponding to one component of the vector, such that a collection of different spectra corresponds to a data cloud in a high-dimensional feature space. The shape of a cloud centered around the origin can be approximated by an ellipsoid with estimated data covariance matrix X^TX (where matrix X holds one mean-removed spectrum per row). Many data analysis approaches can only provide reliable predictions if this matrix is well conditioned. Due to physical reasons or due to the point spread function of the acquisition device, a spectral signature is restricted to one single channel, instead nearby channels tend to have similar values. This high collinearity of spectral data leads to a badly conditioned estimated data covariance matrix X^TX which in turn leads to unreliable model parameter estimates and poorly performing models, which often do not even identify the major sources of influence for the given problem. If two channels of a spectrum were perfectly collinear (and one channel were thus superfluous) the estimated data covariance matrix would even be rank deficient and in linear regression the model parameters could no longer be inferred from the data and the model would likely identify both channels as important. The resulting model would only give unreliable predictions.
The high dimensionality of spectral data makes automatic analysis methods susceptible to the "curse of dimensionality".[1] It owes its reputation to the empirical observation that in order to reliably estimate

*corresponding author

a function from data, the required number of examples grows exponentially with the true, or intrinsic, dimensionality of the data. This is related to the fact that a fixed number of points populates the unit ball less densely with increasing dimensionality and that these points lie ever further apart, i.e. one has ever less information on the sampled space. This is well visualized by a small thought experiment: Repeatedly draw a fixed number of points p randomly from the N-dimensional unit sphere, and compute the median distance from the closest data point to the origin. This median $\mathcal{M}(p, N)$ for p points in N–dimensional space is known[2] to be

$$\mathcal{M}(p, N) = \left(1 - \frac{1}{2}^{\frac{1}{p}}\right)^{\frac{1}{N}}. \tag{1}$$

In the one-dimensional case the points tend to stay close together and the median of the distances is 0.0023 for $p = 300$. Moderately increasing the dimensionality to three already gives a median of 0.1322 for $p = 300$, almost 60 times as large as in the one-dimensional case. Already 600.000.000 sample points are required in three dimensions to reach the median distance of the closest point of $p = 300$ in one dimension, i.e. $\mathcal{M}(3, 6 \cdot 10^{-9}) = \mathcal{M}(1, 300)$! For hundreds of dimensions the resulting median distance is approximately one. In order to densely sample high dimensional spaces, an (exponentially!) increasing number of data points is necessary. As a consequence, an impractical amount of data needs to be labeled in order to reliably estimate all parameters of a linear model when naively applied to spectral data. Note, in this context, that the nominal dimension of spectra is often of the order of a few hundred but due to the correlation among channels the effective dimension may be significantly smaller. Thus one can hope to find a suitable low-dimensional representation containing most relevant discriminatory information.
The second, more direct implication of the high dimensional nature of spectral data is the impossibility to intuitively visualize their distribution in feature space for a human operator.

It is undisputed that spectra and spectral images contain much information; however, the latter cannot easily be accessed by standard methods, and appropriate concepts need to be applied or suitable preprocessing steps have to be taken to overcome the aforementioned problems of collinearity and high dimensionality for high quality automatic processing of spectra. This paper presents general principles of supervised and unsupervised dimensionality reduction and preprocessing as well as classification approaches with and without spatial interaction. The concepts are illustrated using examples from medical diagnostics and materials science.

II. Data

Three spectral data sets from two application areas are used to illustrate the methods discussed in this paper. Magnetic resonance (MR) spectroscopy is a non-invasive diagnostic method used to determine relative concentrations of specific metabolites at arbitrary locations *in vivo*, and ^{1}H NMR spectra (MRS) and spectroscopic images (MRSI) offer information in tumor diagnostics. Characteristic changes in the spectral pattern can be linked to specific changes of the metabolism, providing means for the grading and localization of tumors, e.g. in the brain, breast and prostate. Magnetic resonance spectroscopic imaging allows to acquire such spectra on two- or three-dimensional spatial grids. When searching for tumorous changes of the spectrum, pattern recognition methods can be applied to evaluate the data in a highly automated fashion and to guide the radiologist to the relevant regions of the spectroscopic image. The spectra shown in this paper were recorded on a clinical 1.5T full body MR scanner at the German Cancer Research Center (dkfz, Heidelberg). Each spectrum consists of a total of 512 data points in the range from 1000 – 1250 Hz. In the examples shown in section III, MR spectra of the brain and prostate are used. The information in a long echo-time brain MR spectrum is mostly restricted to the three metabolites choline, creatine and N-acetyl-aspartate (NAA), with a decrease of NAA and an increase of choline in the tumorous spectrum.[3] Prostate spectra reflect the different metabolism of tumorous and healthy tissue

in a similar way, with citrate replacing NAA.[4]

The second data set consists of energy-dispersive X-ray (EDX) spectra recorded with a scanning electron microscope operated with 20 keV acceleration voltage and operated in variable pressure mode. Bombarding a material with electrons can provoke the excitation of atoms by knocking free inner shell electrons. The atom relaxes from this instable state by filling the vacancy with an outer shell electron and releases the energy difference by emitting an X-ray quantum (or an Auger electron). The energy of these quanta is characteristic for different elements. The data set used here consists of spectral EDX images recorded on a 26×29 grid and contains five different elements. Each single spectrum was recorded with 90 ms acquisition time.

III. Dimensionality and collinearity reduction

In order to automatically process and analyze spectra and to extract reliable model predictions, it is of crucial importance to overcome the curse of dimensionality and deal with collinearity. The dimension reduction approaches introduced in this paper can be classified into i) model driven, ii) unsupervised and iii) supervised approaches. In model driven approaches, theoretical models for the resonance lines are fit to the spectrum to obtain low dimensional representations. If no such model exists, methods which rely on the data and, if available, on class assignments of spectra have to be applied.

Model Based / Parametric Approach

A straightforward way to reduce the dimensionality of spectral data is to represent a spectrum in terms of a set of resonance lines with a model for each line. These model based methods can thus be applied only if prior knowledge on the process or on the informative resonance lines such as expected position, shape, or width is available. These parametric approaches explain the given spectrum as a (linear) superposition of basis functions which are either determined empirically or given as parametric functions, e.g. Lorentzian, Gaussian or Voigt shaped. The contribution of each basis function is found by estimating the optimal mixture parameters via (non-)linear least squares or regularized versions thereof such as the lasso[5] or ridge regression.[6] The low dimensional representation is then simply given by all model parameters (such as abundance, peak height, width, position, etc.). If the basis functions are chosen properly (reflecting all collinearities), the resulting representation of the spectra is free of collinearities . The resulting dimensionality is dependent on the used number of basis functions and the number of parameters per basis function. Often it suffices to use few basis functions to capture all discriminatory information between different classes, thus significantly reducing the dimensionality. Approaches of this kind are widely used in standard-free quantification of elements in EDX-spectra[7] and diagnostic processing of MR spectra. For the latter this approach is performed by quantification methods like QUEST[8] or AMARES.[9] An example for prostate MR spectra is shown in figure 1.

If the relevant resonance lines are unknown, or no models are available for these, one must turn to multivariate methods for dimension reduction. These methods rely on the observed spectra and, if possible, on class assignments of several spectra.

Unsupervised Approach

The best known purely data driven approach that does not use any prior knowledge is principal component analysis (PCA) – often called “Karhunen-Loève” or “whitening” transform – which tries to find the linear subspace that best approximates the data in a least squares sense. An orthogonal basis $\{e_k\}$ of this subspace is given by the eigenvectors of the largest eigenvalues of the (estimated) data covariance matrix $C = X^T X$ of the spectral data matrix X and can be found as solution to

$$e_k = \underset{\substack{\|e\|=1 \\ \mathrm{Corr}(e^T C, e_j^T C)=0,\ j<k}}{\operatorname{argmax}} \mathrm{Var}(e^T C)$$

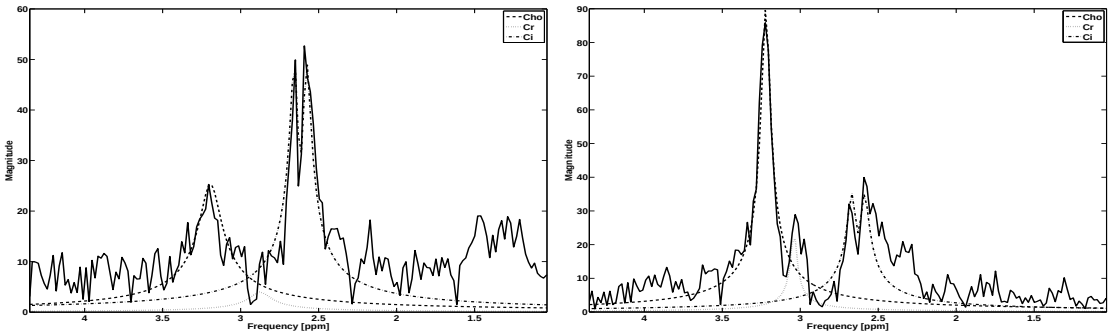

Fig. 1. Example of model based ("parametric") dimensionality and collinearity reduction for prostate MR spectra from healthy (left) and tumorous (right) tissue.[4] Each spectrum with hundreds of channels (dimensions) is represented in terms of three basis functions only.

The coordinates $s_k = e_k^T X_i$ of a spectrum X_i relative to the new basis vectors $\{e_k\}$ are called "scores" and are often used to produce scatter plots summarizing the data. These basis vectors can no longer be interpreted as spectra. This is particularly obvious for count data or absorption spectra, where negative components would be meaningless; and yet these necessarily arise when the standard orthogonal basis is replaced by a rotated version.
After projecting the data onto the PCA eigenbasis, the covariance matrix is diagonal. As most of the variation is contained in the first few eigenvectors, it (often) suffices to use only these few basis functions. Consequently, a lossy compression in terms of a low dimensional representation of spectra has been found, which has additionally removed collinearities (they are condensed into the new basis vectors) and is thus much better suited for visualization or further processing with classification or regression algorithms.

The first PCA component calculated on brain MR spectral data from Menze et al.[3] shown in figure 2 demonstrates that the first PCA component is not able to identify all discriminatory information between healthy and tumorous MR spectra. It nicely captures variation in the NAA metabolite, which is known to change from healthy to tumorous metabolism, but misses the variation in the choline peak. This is due to the fact that PCA cannot distinguish between informative and uninformative (in the discriminatory sense) variation in the data, but only tries to approximate the cloud of all observations in feature space. Prior knowledge on a problem can be introduced by removing irrelevant regions (and thus irrelevant variation) from all spectra prior to calculation of the PCA vectors, and can significantly improve classification results.

Supervised Approach I

The removal of irrelevant spectral regions depends on the availability of prior problem-specific knowledge, and no general guidelines can be given. In practical applications, such prior knowledge is often not available, which may be the very reason why the experiment was conducted. In such cases, it may still be possible to assign some spectra to a class, such as healthy vs. diseased, signal vs. background, interesting vs. non-interesting, etc. This kind of information can then be used to distinguish discriminatory data variation from noise, and to find subspaces which allow for good class separation.
The calculation of the PCA directions only uses observed spectra in the selection of a low dimensional representation, but no class labels. For dimensionality reduction, the additional information inherent in class labels can be used to find directions which not only approximate the spectra well, but also allow for good class separation within the low dimensional space. A common method in chemometrics is partial least squares (PLS) regression,[10] where a linear subspace is sought similar to PCA, except this time the basis vectors are found by maximizing

$$e_k = \underset{\substack{\|e\|=1 \\ \mathrm{Corr}(e^T C, e_j^T C)=0,\ j<k}}{\operatorname{argmax}} \mathrm{Corr}^2(e^T C, y)\mathrm{Var}(e^T C)$$

By incorporating the correlation of the low dimensional projections $s_k = e_k^T C$ with the labels y, this approach explicitly considers the separability of the classes in the projection space. Often these basis vectors nicely reflect the differences between two classes as incorporating label information into the criterion for finding the linear subspace helps in distinguishing between discriminatory variation and noise. Again, it is no longer guaranteed that these basis vectors can be interpreted as spectra themselves.

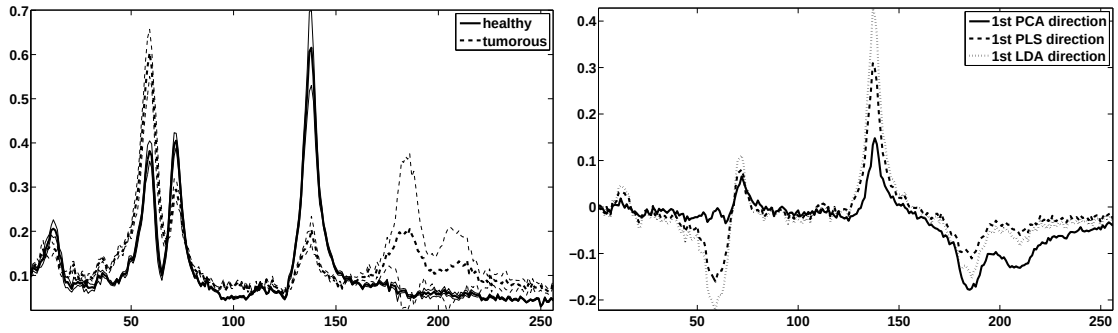

Fig. 2. LEFT: Mean healthy and tumorous in vivo magnetic resonance brain spectra along with an indication of the spread in each spectral channel. RIGHT:: First PCA, PLS and LDA directions. The PCA direction almost misses the important variation in the choline peak (around channel 60), whereas the differences between the two classes are well represented in the PLS direction and even better in the LDA direction.

Applying PLS to the previous example shows that the first PLS direction concentrates much more on the interesting, discriminatory information between healthy and tumorous spectra. Using the labels, PLS is able to identify this kind of variation between classes and, in contrast to the first PCA direction, it identifies variation in choline as highly relevant for classification of spectra (Fig. 2).

Supervised Approach II

Both PCA and PLS (which is known to be dominated by the variance contribution) will fail if the difference between classes is dwarfed by variation of the data along another direction in feature space. In these situations, Fisher's linear discriminant analysis[11] can be used for dimensionality (and again collinearity) reduction. This approach tries to find those directions in the space of all spectra that separate the different classes well. In LDA, the basis of a linear subspace is found according to

$$e_k = \underset{\|e\|=1}{\operatorname{argmax}} \operatorname{Corr}^2(e^T C, y),$$

Thus it searches for directions which are capable of distinguishing between different classes and completely ignores the variation within the data. The differences between PCA, PLS and LDA can be nicely demonstrated using the toy example shown in figure 3. 500 points are randomly drawn from two Gaussian distributions with identical covariance matrices and different means. From left to right the distributions are rotated by 45 degrees. In the left example the within class variation is larger than the spread between the two classes and consequently PCA is not able to identify a direction which allows for good class separation. PLS and LDA again find good directions by exploiting the provided labels. In the right example the main direction of the within class variation is closer to the direction of the between class variation. PLS and LDA again find direction allowing for good class separation. The direction identified by PCA is a mixture of the within and between class variation. Projecting the classes onto this direction still leads to severe overlap. This example demonstrates that projection onto PCA components will only allow for good class separation if the directions of within class and between class variation are almost parallel.

LDA assumes that the spectra from either class come from Gaussian distributions with identical covariance matrices. LDA can fail if the distributions of both classes are in fact very different. In these cases it is advisable to use quadratic discriminant analysis (QDA) which explicitly allows for different covariance matrices per class, or other nonlinear classifiers. The prize for the higher flexibility of QDA is that more

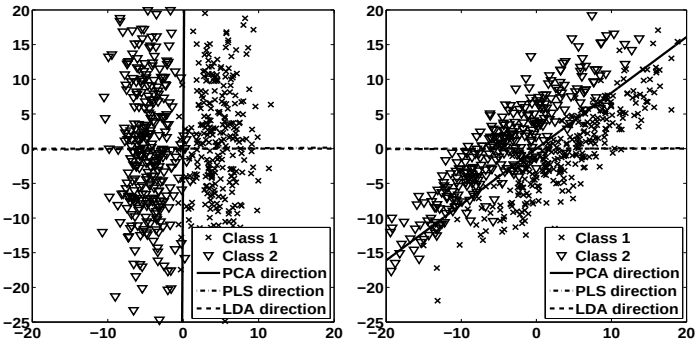

Fig. 3. 500 points are randomly sampled from each of two 2-D Gaussian distributions with common covariance matrix and different means. From left to right the distributions are rotated by 45 degrees with fixed mean LEFT: Projection onto the first PLS and LDA direction allows for good class separation. Projecting onto the first PCA direction leads to a complete class overlap as it only captures the large within class variation and ignores the smaller between class variation; RIGHT: PLS and LDA again find highly discriminatory directions. The variation captured by the first PCA direction is a mixture of within class and between class variation. If projected onto this direction the classes overlap severely.

labeled examples are required to estimate the covariance matrices per class in contrast to LDA which uses all labeled spectra to estimate one covariance matrix. In general, there are rarely enough labeled spectra to estimate a full covariance matrix for LDA and therefore regularized methods[12] are used which significantly reduce the number of parameters and thus the required number of labeled examples.

Fig. 2 shows that LDA works well for the MR example. The first LDA direction perfectly identifies the variation in the NAA (around channel 140) and choline peak (around channel 60) as discriminatory information and almost completely discards the noisy regions (due to lipids and lactate) to the right of the NAA peak. Consequently it allows for a very good separation between healthy and tumorous spectra. Only concentrating on those directions in the space of all spectra which separate different classes (as LDA does) is often too short-sighted as shown in figure 4. On the spatial EDX example, PLS has proven to give a sound trade-off between data and label variation, leading to well-performing models.
From the above, it is unclear how to choose an appropriate number of PCA, PLS or LDA directions in order to ensure that enough discriminatory information is contained in the resulting low dimensional representation. Choosing this number too low leads to overly naive models, while choosing it too high includes uninformative noise variation and makes the classification step susceptible to overfitting. For PCA, a manual analysis of the eigenvalues can be performed. For PLS and LDA, this number can be inferred from training data by methods like cross-validation,[2] Akaike Information criterion[17] or Bayesian information criterion.[18]

Nonlinear Approaches

In this paper only linear methods for dimensionality reduction are discussed. If the spectra form a highly nonlinear manifold in feature space, projection onto a linear subspace can introduce significant distortions of the relative positions of the spectra, resulting in improper variance estimates. Dimensionality reduction methods such as the Laplacian Eigenmap[13] or local linear embedding[14] can be used in these situations. They rely on a distance measure between two spectra to construct a weighted graph to create a discrete representation of the manifold structure and find an embedding of the manifold of all spectra in a low dimensional linear space which preserves relative distances to the extent possible. Identification of major sources of influence can then be achieved with the embedded spectra.
All previous approaches represented a spectrum as a linear superposition of basis functions. If higher order, nonlinear interactions of the basis functions (e.g. products of two basis functions) have to be incorporated, direct modeling can often be performed but the resulting model will require significantly more labeled spectra and might again fall for the “curse of dimensionality” due to the increased number of parameters.

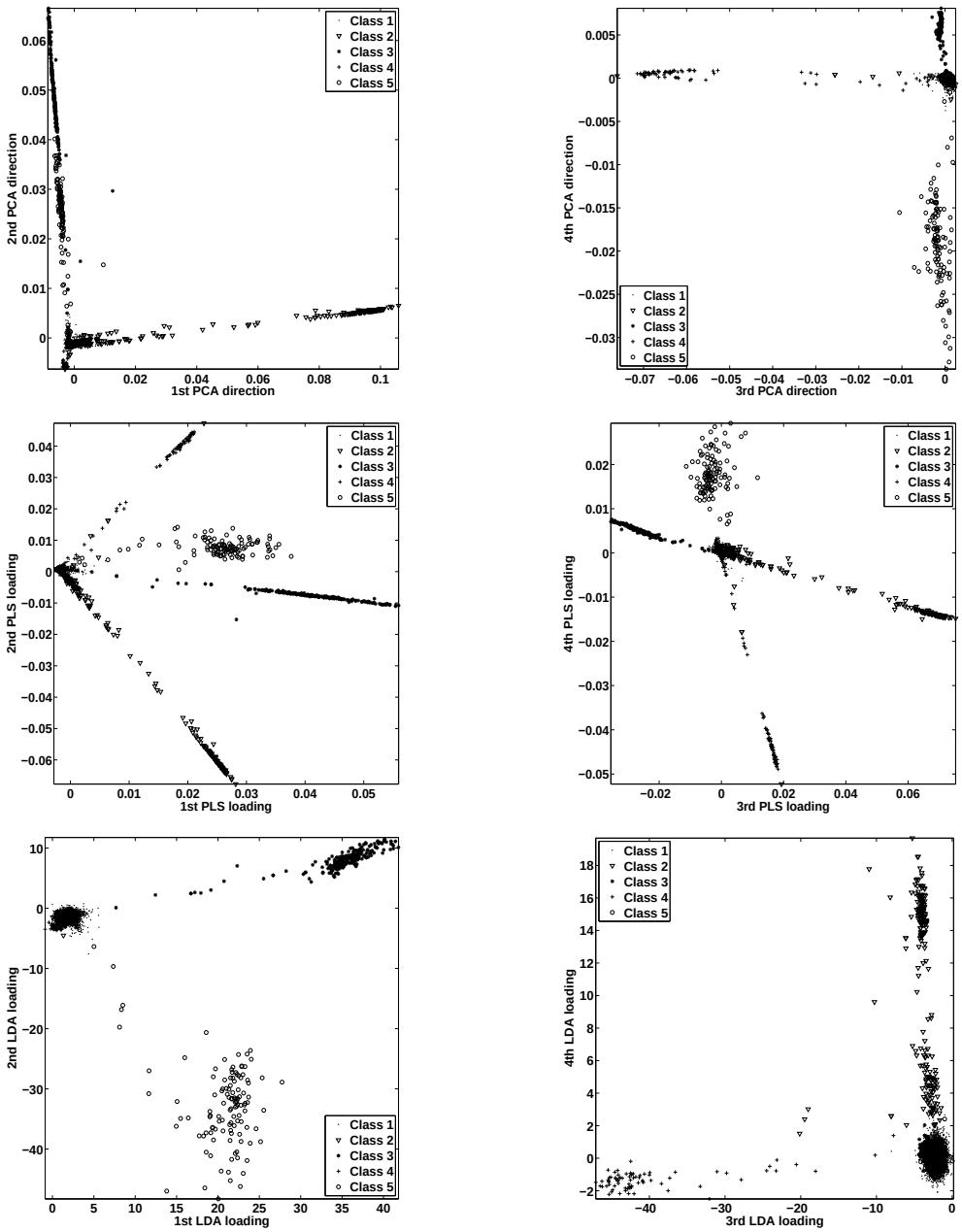

Fig. 4. Scatter plots of EDX spectra of different materials projected onto the first four PCA, PLS and LDA directions. The first two PCA and LDA directions are insufficient for further processing. The first two LDA directions separate classes 3 and 5 whereas directions 3 and 4 allow for classification of 2 and 4. With PCA the third and fourth direction do not help in reducing the overlap between class 1 and 2. In this example, PLS is the best preprocessing step as it allows for very good classification even with only two directions. The two extra directions do not provide significant additional discriminatory power.

Another way of extending linear models to capture higher order correlations is to apply the so called "kernel trick". A detailed explanation of this approach is beyond the scope of this paper and the interested reader is referred to the literature on kernel machines.[15]

IV. Classification

The classification of spectra can be performed using any of the aforementioned low dimensional representations of the spectra and any parametric (linear or quadratic discriminant analysis, logistic regression,[2] ...) or nonparametric (k-nearest neighbor, support vector machines,[15] random forest,[16] Bayes classifier with

kernel density estimation, ...) classifier. Well-known, versatile classifiers are support vector machines, which after employing the "kernel-trick" find a hyperplane in a high-dimensional kernel space which classifies the spectra as well as possible (according to the assigned labels). Projecting this decision boundary back into the original feature space can lead to highly nonlinear surfaces, depending on the kernel used. Another good classifier is random forest, which is a nonlinear classifier combining the response of many decision trees grown on a subset of the whole data set.

Comparing data driven and model based approaches it is tempting to assume that the latter can be better classified as prior knowledge is used during preprocessing. Extensive work with MRS data has shown that purely data driven approaches consistently perform at least as well as model driven approaches, often even better[3] if sufficiently large training sets are available; at the same time, they are computationally significantly faster when classifying new spectra.

The successful application of classification methods to brain tumor MR spectra is shown in figure 5. For dimensionality reduction the MR spectra were projected onto a PLS basis calculated on hold-out data. The resulting PLS-scores were classified using logistic regression. The classifier's output ranges from 0 to 1, with "0" indicating a highly probable healthy and "1" a highly probable tumorous MR spectrum.

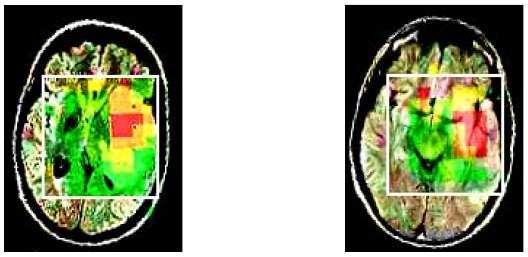

Fig. 5. Inside the area indicated by the white rectangle, MR spectra were recorded on a 16×16 grid. Each spectrum was projected onto the first two PLS directions estimated from hold-out data. Finally, a logistic regression model trained on separate training data was applied to these PLS-scores. The output of the model ranges from 0 to 1 for each spectrum and is color-coded with green indicating a highly confident healthy prediction for the spectrum and red a highly confident tumorous diagnosis. Intermediate colors represent less confident classifications of the spectra.

V. Including Spatial Information

Often the quantity of interest changes smoothly in a (spectral) image, an information which has not been exploited up to now as all previously mentioned approaches processed single spectra and totally disregarded their spatial context. Randomly permuting the spectra in space would not change the classification result, contradicting the spatial smoothness assumption. For example in the spectral EDX image shown in figure 6, the spatial sampling distance of the red points in the left image is significantly smaller than the particle, and it is much more likely for neighboring spectra to belong to the same class "particle" than to the different classes "particle" and "organic background". Using this information on the spatial neighborhood helps in classifying spectra with unclear characteristics or low signal-to-noise ratios as the method "knows" what class the neighboring spectra belong to. To model this kind of interaction, it suffices to allow short-range interaction (often only between neighboring spectra) as this indirectly yields long-range correlation of spectral labels and allows for efficient classification schemes. Conventional approaches to incorporate spatial information in the classification of spectral images include relaxation labeling[19] or a mere smoothing of the estimated labels from the classification step. Alternatively it is possible to model the probability distribution of all spectra of the image and their labels via Markov random fields[20] or via conditional random fields (CRF)[21] (which then model a probability distribution of all possible labelings given the spectral image). The resulting structure for a CRF used to classify spectral EDX images is represented as a graphical model in figure 7, with the pair-potential terms $\mathrm{pp}(Y_i, Y_j | X_i, X_j)$ indicating the degree of interaction between the spectra at positions i and j, and the single site potentials $\mathrm{ssp}(Y_i | X_i)$ suggesting a class assignment depending only on the spectrum X_i. The latter can, for example, be derived from any

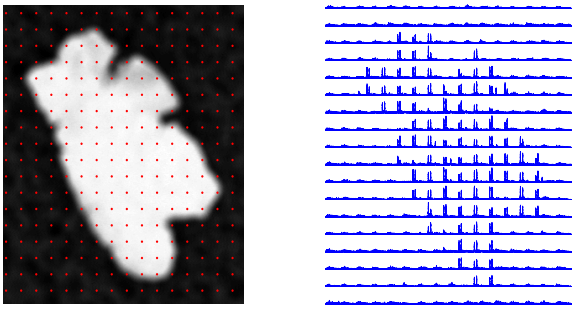

Fig. 6. LEFT: Scanning electron microscope back-scatter detector image, red points indicate points of spectrum acquisition; RIGHT: Energy dispersive X–ray spectral image; information on the elemental distribution from the SEM image and the spectral image coincide. Although the spectral image has a lower resolution than the SEM image, it carries the information on the elemental distribution much more reliably and allows for identification of the particle's material.

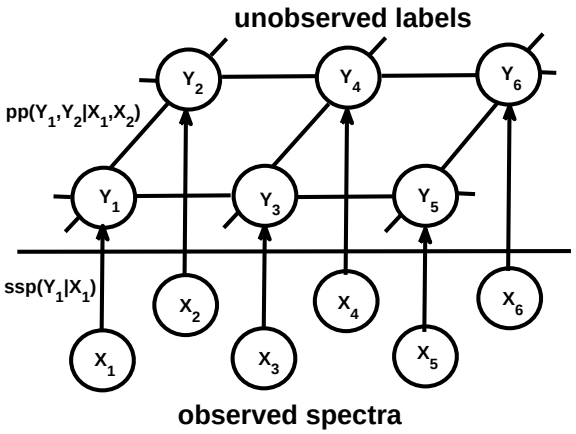

Fig. 7. Graphical model representing a discriminative random field used to classify spectral EDX images. Each observed spectrum corresponds to a node X_i in the graph.

single-spectrum modeling scheme discussed in section III. The resulting probability distribution of all labels $\mathbf{Y}$ given all spectra $\mathbf{X}$ of the image in this case is

$$p(\mathbf{Y}|\mathbf{X}) = \frac{1}{Z(\mathbf{X})} \cdot \exp\left(-\sum_i \mathrm{ssp}(Y_i|X_i) - \nu \cdot \sum_{i\sim j} \mathrm{pp}(Y_i, Y_j|X_i, X_j) \right). \tag{2}$$

Using these complex models, classification is not as easy as before. Two possible and the most widely used approaches are searching for the maximum of this multidimensional distribution (e.g. with GraphCut[22]), i.e. finding the most probable global state according to the distribution (2), or by maximizing the marginal at each node (e.g. with belief propagation[23]), i.e. finding locally the most probable classification for each spectrum given its neighboring spectra. In binary classification problems it is advisable to use the maximum a posteriori estimate as it can mostly be calculated exactly.[22] This is no longer true for multi-class problems and we advise to use the marginal posterior mode estimator (MPME) as classification of all spectra of an image. According to the MPME, each individual spectrum is assigned to that class that is most probable when averaging over all possible combinations of labels maps, with each combinations weighted by its probability. It is known to often yield better results in multi class problems with skewed posterior distribution.[24] In applications, using spatial information has shown to significantly improve a model's classification accuracy and robustness to noise. Applying a conditional random field to the task of assigning noisy EDX spectra to one of five material classes showed significantly improved results

(Fig. 8). The single spectrum approach had severe problems in assigning spectra with unclear elemental characteristics correctly. The incorporation of information on the class assignment of neighboring EDX spectra helps in classifying these spectra correctly .

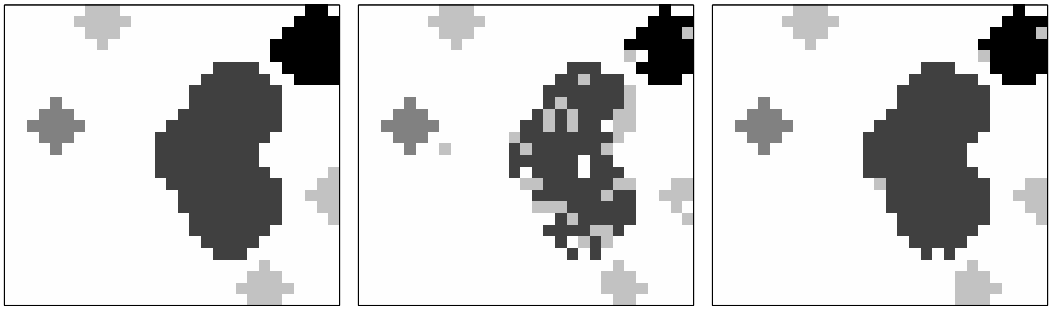

Fig. 8. Comparison of classification of EDX spectra with and without spatial interaction.LEFT: Ground truth for membership in five classes; MIDDLE: Classification based on individual spectra; RIGHT: Classification taking into account the spatial context by means of a conditional random field 2.

VI. CONCLUSION

This paper gives an introduction to the large field of automated processing of spectral images with a focus on suitable low-dimensional representations of spectra. Many popular statistical modeling methods are not straightforwardly applicable to this type of data as they may perform badly due to the high dimensionality and collinearity of spectral data. This results in the necessity of preprocessing spectra to find a suitable low-dimensional representation that still contains enough discriminatory information for the solution of the task. This can be done in a purely model based approach by trying to represent the spectrum as a linear combination of resonance peaks, each of which is fit by a model. This approach is only viable if sufficient prior knowledge of the recorded spectra is available. In all other situations using a data (and label) driven approach like PCA, PLS or LDA can be used to find a low-dimensional space which condenses the information provided by the spectrum into few dimensions.

Which method performs best depends on whether labels are available and whether the total variation of the spectra carries the discriminatory information. If no labels are provided one is left with PCA, an approach which has often proven to give satisfactory results. If the variation of the data is the result of a significant amount of uninformative noise, excluding uninformative regions of the spectrum can guide PCA in finding a better subspace. If labels are available, LDA or PLS is the method of choice, as the labels help in distinguishing between discriminatory and uninformative variation. Comparative studies have shown that although data driven approaches use almost no prior knowledge, they often outperform model based methods if enough training data is available.

Incorporating neighborhood information significantly improves classification accuracies if the quantity of interest is known to vary smoothly across the spatial dimension, especially if the quality of the spectra is low. This information can be explicitly modeled by using Markov random field approaches, which allow interaction only between neighboring spectra. In these models, all spectra of a spectral image are classified at once (and not on a spectrum-by-spectrum basis) using a maximum a posteriori or marginal posterior mode estimator.

Spectral data are important already in scientific and industrial areas ranging from materials science and quality control to tumor detection, and this trend will intensify in the future. The resulting vast amount of spectral data can no longer be analyzed by human experts but requires fast automatic methods to process them. Duda et al. provide an accessible introduction to statistical learning,[25] and Otto demonstrates the applicability of many methods to interesting chemometrical problems.[26]

REFERENCES

1. Bellman, R. *Adaptive Control Processes: A Guided Tour*; Princeton University Press: Princeton, 1961.
2. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer: New York, 2001.
3. Menze, B. H.; Lichy, M. P.; Bachert, P.; Kelm, B. M.; Schlemmer, H.-P.; Hamprecht, F. A. *NMR Biomed* **2006**, *19*, pp. 599.
4. Kelm, B. M.; Menze, B. H.; Zechmann, C. M.; Baudendistel, K. T.; Hamprecht, F. A. *Magnet Reson Med* **2007**, *57*, pp. 150.
5. Tibshirani, R. *J Roy Stat Soc B* **1996**, *58*, pp. 267.
6. Hoerl, A. E.; Kennard, R. W. *Technometrics* **1970**, *12*, pp. 55.
7. Statham, P. *X–Ray Spectrom* **1976**, *5*, pp. 16.
8. Ratiney, H.; Sdika, M.; Coenradie, Y.; Carvassila, S.; van Ormondt, D.; Graveron-Demilly, D. *NMR Biomed* **2005**, *18*, pp. 1.
9. Vanhamme, L.; van den Boogaart, A.; van Huffel, S. *J Magn Reson* **1997**, *129*, pp. 35.
10. Geladi, P.; Kowalski, B. *Anal Chim Acta* **1986**, *185*, pp. 1.
11. Fisher, R. A. *A Eug* **1938**, *8*, pp. 376.
12. Friedman, J. H. *J Am Stat Assoc* **1989**, *84*, pp. 165.
13. Belkin, M.; Niyogi, P. *Neural Comput* **2003**, *15*, pp. 1373.
14. Roweis, S. T.; Saul, L. K. *Science* **2000**, *290*, p. 2323.
15. Schölkopf, B.; Smola, A. J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*; MIT Press: Cambridge, 2002.
16. Breiman, L. *Mach Learn* **2001**, *45*, pp. 5.
17. Akaike, H. *IEEE T Automat Contr* **1974**, *19*, pp. 716.
18. Schwarz, G. *Ann Stat* **1978**, *6*, pp. 461.
19. Rosenfeld, A.; Hummel, R.; Zucker, S. *IEEE T Syst Man Cyb* **1976**, *6*, pp. 420.
20. Besag, J. *J Roy Stat Soc* **1974**, *36*, pp. 192.
21. Lafferty, J.; McCallum, A.; Pereira, F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proceedings of the 18th International Conference on Machine Learning*, 2001; pp. 282.
22. Kolmogorov, V.; Zabih, R. *IEEE T Pattern Anal* **2004**, *26*, pp. 147.
23. Yedida, J.; Freeman, W.; Weiss, Y. *Understanding Belief Propagation and its Generalizations*; Technical Report TR–2001–22, 2002.
24. Fox, C.; Nicholls, G. K. Exact MAP states and expectations from perfect sampling: Greig, Porteous and Seheult revisited. *AIP Conference Proceedings 568: Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, 2001; pp. 252.
25. Duda, R.; Hart, P.; Stork, D. *Pattern Classification*, 2nd ed.; Wiley–Interscience, 2000.
26. Otto, M. *Chemometrics. Statistics and Computer Application in Analystical Chemistry*; Wiley: New York, 1998.