

Machine Learning Approaches and Pattern Recognition for Spectral Data

Thomas Villmann¹, Ersébet Merényi², and Udo Seiffert³

1- University Leipzig - Clinic for Psychotherapy
Semmelweisstr. 10, D-04103 Leipzig - Germany

2- Rice University, Electrical and Computer Engineering
6100 Main Street, Houston, TX, USA

3- Scottish Crop Research Institute (SCRI) - Mathematical Biology
Invergowrie, Dundee, DD2 5DA, Scotland, UK

Abstract. The adaptive and automated analysis of spectral data plays an important role in many areas of research such as physics, astronomy and geophysics, chemistry, bioinformatics, biochemistry, engineering, and others. The amount of data may range from several billion samples in geophysics to only a few in medical applications. Further, a vectorial representation of spectra typically leads to huge-dimensional problems. This scenario gives the background for particular requirements of respective machine learning approaches which will be the focus of this overview.

1 Introduction

Spectral data occur in many areas of theoretical and applied research like physics, astronomy and geophysics, chemistry, bioinformatics, biochemistry, engineering, and others. One key characteristic of such data is that their vectorial representation typically leads to huge-dimensional problems. However, spectral vectors are functional, i.e., the vector dimensions are not independent but reflect a functional relation. In the simplest case it is a data *vector* representing a one-dimensional function, vectorial functions may be described by *matrices*. Thereby, the amount of data may range from several billion samples in geophysics to only a few in medical applications.

The characteristic difference of functional data in comparison to usual vectorial data is the above mentioned dependency between the vectors dimensions, i.e. the vector components are functionally correlated. Thus, the inherent dimensionality of functional data vectors is usually much smaller than the vector dimension. This knowledge can be used to make feasible sparse high-dimensional data sets of functional data whereas non-functional data of a similar complexity may not be analyzed adequately. The locations, widths, skew, kurtosis, etc. and shapes of characteristic peaks or valleys (absorptions), as well as their co-occurrences are important for data analyses. These properties should be used for specific machine learning approaches designed for functional data analysis. In case of parametric models the promoters are usually chosen to be descriptors for shape and density and the machine learning task is to find their true value given the functional data examples. For example, the normal distribution is sufficiently described by mean and variance. Non-parametric models offer a greater

variability. However, the complexity has usually to be adapted during the machine learning process. Further, functional data frequently come from natural or technical processes known to be following mathematical laws like ordinary or partial differential equations. For these processes it is sufficient to estimate the parameters of the known functional form from the data stream.

In this paper we will focus on a special type of functional data: *spectral* data. In spectral data correlations can be two-fold: on the one hand, the correlation in vectorial representation may be in neighboring dimensions according to the shape of peaks. On the other hand, the occurrence and co-occurrence of peaks depends on the underlying physical, chemical, biochemical or technical process. Thus, long-range interactions may contribute to correlations which reduce the degree of freedom and, hence, the inner complexity. These characteristic properties can be used to handle spectral data effectively. Possibilities for this are particular metrics or similarity measures or special data transformations which make use of these characteristics. Different types of spectra may be distinguished like spectra with broad absorption bands, for example in remote sensing, line-spectra of isolated sharp peaks in chromatography or mass-spectrometry, etc. Each type has to be handled in different manner depending on the task and the underlying process.

In the following we will give a few general remarks highlighting some key principles for functional data analysis. After that, we give examples from three different areas of spectral data applications and their respective machine learning data analysis approaches: astronomy and geophysics, computational biology, and biochemical spectral data. These areas reflect typical issues of spectral and functional data analysis applications in machine learning: the underlying process of the data stream is not completely known therefore parametric approaches of the underlying model cannot be used. This is in contrast to many engineering problems where functional data analysis can frequently be reduced to parameter estimation of the respective theoretical functional model.

2 Some General Aspects of Functional Data Analysis

Functional data analysis is fundamentally based on the concept of similarity between functions which can often be described by functional norms. If a Hilbert space is assumed norms are related to inner products [1]. Well known examples are the family of \mathcal{L}_p -norms [2], divergence measures for density functions [3], or kernel approaches [4].

The \mathcal{L}_p -norms can be extended to take into account the spatial shape of the functions using the derivatives in case of differentiable functions. The respective norms are the Sobolev-norms, which can also be related to inner products [5]. Sobolev-norms can be used for spline approximation adapted to functional data as it is demonstrated in [6]. Other distance measures, which cannot be derived from norms but which are suitable for function shapes, may also be successfully applied in machine learning approaches [7]. Yet, the choice of an adequate similarity measure may crucially influence the performance of a method [8]. An adequate metric can reduce the complexity of the problem.

Further, classical mathematical methods like multivariate analysis can be

transferred to functional data analysis adequately for special data types: To give a prominent example, functional principal component analysis (FPCA) can be reduced to the usual principal component analysis (PCA) using approximation theory [9]. For this purpose it is assumed that the real functions f, g over $X \subseteq \mathbb{R}$ can be represented by orthogonal basis functions ϕ_k which form a basis of the functional space containing f and g . Thereby, orthogonality is defined by the (Euclidean) inner product

$$\langle \phi_k, \phi_j \rangle_{\mathbb{E}} = \int_X f(x) g(x) dx \quad (1)$$

$$= \delta_{k,j}. \quad (2)$$

The basis may contain an infinite number of basis functions. Prominent examples are the set of monomials $1, x, x^2, \dots, x^k, \dots$ or the Fourier-system of $\sin(k\omega x), \cos(k\omega x)$ with $k = 0, 1, 2, \dots$ in case of periodic functions. Using a basis system of K linearly independent functions, an arbitrary (continuous) function h can be approximated by

$$h(x) = \sum_{k=1}^K \alpha_k \phi_k(x) \quad (3)$$

which can be seen as a discrete Euclidean inner product $\langle \boldsymbol{\alpha}, \boldsymbol{\phi}(x) \rangle_{\mathbb{E}}$ of the coordinate vector $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_k)^T$ with the function vector $\boldsymbol{\phi} = (\phi_1(x), \dots, \phi_k(x))^T$. We denote by \mathcal{A} the function space spanned by all basis functions ϕ_k :

$$\mathcal{A} = \text{span}(\phi_1, \dots, \phi_k). \quad (4)$$

Following the suggestions in [10] and [11] to transfer the ideas of usual multivariate PCA to FPCA, we obtain for the Euclidean inner product (1) and function approximations according to (3)

$$\langle f, g \rangle_{\mathbb{E}} = \sum_{k=1}^K \sum_{j=1}^K \alpha_k \beta_j \int_X \phi_k(x) \phi_j(x) dx \quad (5)$$

$$= \sum_{k=1}^K \sum_{j=1}^K \alpha_k \beta_j \langle \phi_k, \phi_j \rangle_{\mathbb{E}} \quad (6)$$

whereby in the second line the Fubini-lemma was used to exchange the integral and the sums. Let $\boldsymbol{\Phi}$ be the symmetric matrix spanned by $\Phi_{k,j} = \langle \phi_k, \phi_j \rangle_{\mathbb{E}}$ using the symmetry of an inner product. Using this definition, the last equation can be rewritten as $\langle f, g \rangle_{\mathbb{E}} = \langle f, g \rangle_{\boldsymbol{\Phi}}$ with the new inner product

$$\langle f, g \rangle_{\boldsymbol{\Phi}} = \boldsymbol{\alpha}^T \boldsymbol{\Phi} \boldsymbol{\beta} \quad (7)$$

We remark that $\boldsymbol{\Phi}$ is independent of both f and g . If the basis is orthogonal, $\boldsymbol{\Phi}$ is diagonal with entries $\Phi_{k,k} = 1$. Thus, the inner product of functions is reduced to the inner product of the coordinate vectors

$$\langle f, g \rangle_{\mathbb{E}} = \langle \boldsymbol{\alpha}, \boldsymbol{\beta} \rangle_{\mathbb{E}} \quad (8)$$

and, hence, FPCA may be reduced to usual PCA of the coordinate space. For handling non-orthogonal basis systems we refer to [11].

Yet, there exists a great variety of other linear transformations of functional data, which can be used for complexity reduction and model simplification. A linear projection of spectral data based on noise variance estimation is demonstrated in [12]. A linear mapping for optimized learning vector quantization, dependent on class separation, is proposed in [13].

As we can see from the above example and remarks, the utilization of the knowledge of the data structure, here the functional behavior of the vector components, may be used for adequate handling of functional data. For further reading about general functional approaches we refer to the monograph [10].

3 Machine Learning of Spectral Data in Astronomy and Geosciences

Earth and space science have perhaps the longest history of using spectral data. Line spectra are used, at Angström resolution, to probe elemental composition, spectral measurements in the visible and near-infrared (VNIR), and thermal infrared (TIR) regions of the electromagnetic spectrum, sampled at a few to a few hundred nanometers, are used to infer mineralogical composition of various targets. In the VNIR and TIR, the many measured values (reflectances, transmittances, emitted heat, etc. at various wavelengths) are typically considered as one data "item" — a sampled spectrum — and the spectrum is used as a whole for species identification. The underlying physical process that determines the spectral shape is the preferential interaction of light with different materials at different wavelengths. In the VNIR, this manifests in absorption (transmission, emission) features (bands), whose depth, width and other properties are specific to a given material and wavelength. Depending on the sampling rate, we distinguish *multi-spectral* data (few spectral channels with wide bandpasses) and *hyperspectral* data (hundreds of narrowly spaced bandpasses, as in Figure 1).

Sample VNIR spectra in Figure 1 illustrate the variety of features that exists even among similar species. The functional relations among the spectral channels manifest in multiple correlations with any index differences. Materials can have multiple absorption features, each of which may be very narrow or quite wide. For example, the clays all have a sharp feature near $2.1 \mu\text{m}$, and also at 1.4 and $1.9 \mu\text{m}$. However, the depth and width of those features varies across the individual species. The overall spectral shape is also important in material identification.

In Earth and space science spectra are obtained mostly by remote sensing, from telescopes, aircraft or spacecraft, and by robots such as the Mars Exploration Rovers. In the VNIR and TIR range, imaging spectroscopy (acquiring *high-resolution* spectra in image context, as opposed to spatially sparse measurements of single spots) became the standard for many applications. Mapping the geology on remote planets; precision agriculture; monitoring environmental contamination are but a few. Since most often the whole spectral shape is used for identification of materials pattern recognition, with either or both supervised classification and unsupervised clustering, is a primary task. Machine learning (ML) has become increasingly attractive for spectral data because it effectively

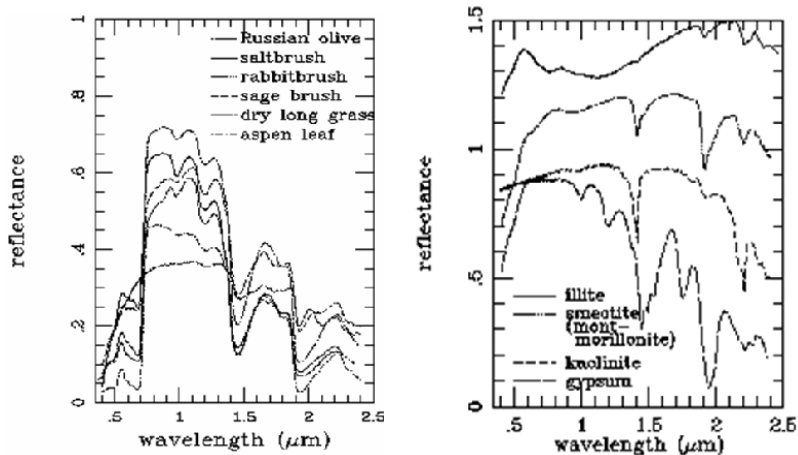


Fig. 1: VNIR spectra of plants and geologic materials (clay minerals). Both illustrate the range of variations in absorption features, unique to the particular species, and the degree of (dis)similarities within the same family of materials.

handles the associated pattern recognition challenges. Some of these are:

1. The spectral shapes are extremely hard to model from first principles.
2. Data vectors can be high dimensional (hundreds to thousands of channels).
3. Imaging spectroscopy maps large areas, therefore a large number of spectral species (classes, clusters) is expected to be found.
4. Subtle but important differences (such as those between some of the plant spectra in Figure 1) are expected to be recognized.
5. High spectral dimensionality may be aggravated by a scarcity of data points (e.g., spectra taken of distant asteroids or planetary surfaces such as Pluto, one at a time, through telescopes, using hours of integration time).

VNIR spectroscopy has also spread outside the fields of astronomy and planetary science. Examples are quality control in food industry, drug manufacturing, and gemology (mostly using spot measurements), and imaging spectroscopy in medical diagnostics. These data have similar *general* characteristics, thus much of this discussion also applies to them. An important difference is that remote sensing spectra typically exhibit more complicated structure. The reader is invited to compare the plots in Figure 1 with, for example, spectra of food in [12].

Machine learning of multi- and hyperspectral data started in the 1980-s and early 1990-s, respectively, mostly applying Back Propagation (BP) nets, and reporting improvement over more traditional methods for classification of terrestrial [14, 15] and simulated Martian spectra [16], for moderate number of

classes. The difficulty of training BP nets for many inputs and classes, however, turned attention to other ML schemes. SVMs are favored by many [17, 18], partly because of the justified use of small training sample size. Hybrid architectures that consist of an SOM hidden layer coupled with a categorization output layer, alleviate the training difficulties of BP nets and can produce precise classification of high-dimensional spectra into many classes [19].

In unsupervised tasks, SOMs proved their discovery power for a variety of situations: low-dimensional large data sets of Earth and Mars [19, 20, 21], small number of high-dimensional astronomical spectra with many clusters [22] and massive hyperspectral imagery with very large number of clusters [23]. A successful alternative to SOMs are ART maps [24] for clustering and novelty detection. With Associative Memories [25, 26] improved on traditional spectral unmixing (a frequently used analysis tool for spectral images), by *automatic* identification of endmembers and by the use of a *large* number of endmembers, both of which have limitations in traditional methods.

Estimation of physical parameters from complex spectral shapes is an important task, whose difficulties can be addressed by ML, as in [27], this session.

Another related issue is feature extraction in the spectral dimension. Existing methods are inapplicable because most operate in the spatial domain, which misaligns the spectra. Methods simultaneously handling all spectral bands are not yet generally available. ML efforts in this area started in the early 1990s but remained scarce. In [28] an interesting Decision Boundary Extractor is shown to improve classification accuracy, in addition to making the reduced hyperspectral data suitable for BP learning. The invention of the Generalized Relevance Learning Vector Quantization [29] opened new powerful principled possibilities, by jointly optimizing classification performance and feature extraction. This was further engineered by [30] for hyperspectral data. In this session, [13] offers additional developments of GRLVQ, while [12] proposes a different way by identifying latent variables for nonlinear models.

Various transformations have also been proposed for a preprocessing step, which — indirectly — effect a more advantageous metric in the transform space. In all cases, sampling of continuous functions is involved. The question of classification consistency for sampled functions is addressed theoretically in [6].

We encourage the reader to explore the cited articles for more details.

4 Machine Learning Techniques for the Analysis of Functional Data in Computational Biology

The amount of data in typical *computational biology* (bioinformatics) applications [31] tends to be quite large but is on a manageable scale. In contrast, astrophysical applications have huge amount of data, and medical research often only has a rather limited number of samples. The challenges in bioinformatics seem to be:

- Diversity and inconsistency of biological data,
- Unresolved functional relationships within the data,

- Variability of different underlying biological applications/problems.

As in many other areas, this requires the utilization of adaptive and implicit methods, as provided by machine learning [32, 33]. Due to the above mentioned wide scope of potential bioinformatics applications, we have restricted this review to a number of key issues with a focus on spectral data.

Protein function, interaction, and localization is definitely one of the key research areas in bioinformatics where machine learning techniques can beneficially be applied. Protein localization data, no matter whether on tissue, cell or even subcellular level, are essential to understand specific functions and regulation mechanisms in a quantitative manner. The data can be obtained, for example, by fluorescence measurements of appropriately labelled proteins. Now the challenge is to recognize different proteins, and classes of them, respectively, which usually leads to either an unsupervised clustering problem or, in case available a-priori information is to be considered, a supervised classification task. Here a number of different neural networks have been used [34, 35, 36, 37, 38, 39]. Due to the underlying measurement technique, often artifacts are observed and have to be eliminated. Since the definition of these artifacts is not straightforward, here too, trainable methods are used. In this context, for the separation of artifact vs. all other data, support vector machines have successfully been applied as well [40].

Further major applications areas comprise the analysis of genomic data on transcript and metabolic level [41, 42]. The particular field of spectral data will be covered by the following section.

4.1 Spectral Data in Bioinformatics

The analysis of biochemical data is a common task in many life science disciplines as well as in chemistry and physics, food industry etc. [32],[43]. Frequently used measurement techniques providing such data are mass spectrometry (MS) and nuclear magnetic resonance spectroscopy (NMR). Typical fields, where such techniques are applied in biochemistry and medicine, are the analysis of small molecules, e.g., metabolite studies, or studies of medium or larger molecules, e.g., peptides and small proteins in case of mass spectrometry. One major objective is the search for potential biomarkers in complex body fluids like serum, plasma, urine, saliva, or cerebral spinal fluid in case of MS or search for characteristic metabolites as a result of metabolism in cells (NMR).

Spectral data in this field have in common that the raw functional data vectors, representing the spectra, are very high-dimensional, usually containing many thousands of dimensions depending on the resolution of the measurement instruments and/or the specific task [44]. Moreover, the raw spectra are usually contaminated with high-frequency noise and systematic baseline disturbances. Thus, before any data analysis may be done, advanced pre-processing has to be applied. Here application specific knowledge can be involved. For example, for comparison of spectra an alignment, i.e., a frequency shifting, is necessary to remove the inaccuracy of the instruments [45], [46]. A second step usually follows the alignment to reduce the noise, Figure 2. Here machine learning methods including neural networks offer alternatives to traditional methods like

