Analytica Chimica Acta 734 (2012) 45–53

Contents lists available at SciVerse ScienceDirect

Analytica Chimica Acta

journal homepage: www.elsevier.com/locate/aca

Application of direct calibration in multivariate image analysis of heterogeneous materials

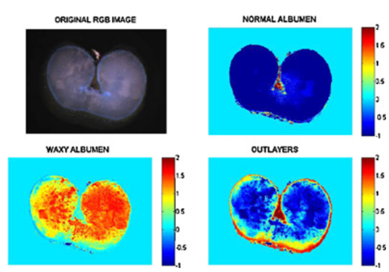
Benoît Jaillais[a,*], Jean-Claude Boulet[b], Jean-Michel Roger[c], François Balfourier[d], Pierre Berbezy[e], Dominique Bertrand[f]

[a] INRA, UR 1268 Biopolymères Interactions Assemblages, F-44316 Nantes, France
[b] INRA, UMR 1083 Sciences Pour l'Oenologie, F-34060 Montpellier, France
[c] Irstea, UMR ITAP, F-34196 Montpellier, France
[d] INRA, UMR 1095 Génétique Diversité et Ecophysiologie des Céréales, F-63100 Clermont-Ferrand, France
[e] Ulice, F-63204 Riom, France
[f] Data_frame, F-44300 Nantes, France

HIGHLIGHTS

- ► Screening of wheat accessions based on multivariate images.
- ► Processing of multivariate images by Direct Calibration.
- ► Useful and harmful spaces determination and choice of dimensions in PCA improved.
- ► Possible application to very large images, such as hyperspectral images.
- ► False RGB images in which an RGB channel corresponds to a specific tissue.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 22 September 2011
Received in revised form 6 March 2012
Accepted 1 May 2012
Available online 17 May 2012

Keywords:
Direct Calibration
Multivariate imaging
Orthogonal subspace projection
Wheat
Endosperm

ABSTRACT

Many scientific instruments produce multivariate images characterized by three-way tables, an element of which represents the intensity value at a spatial location for a given spectral channel. A problem frequently encountered is to attempt estimating the contributions of some compounds at each location of these images. Usual regression methods of calibration, such as PLS, require having a matrix of calibration $\boldsymbol{X}$ $(n \times p)$ and the corresponding vector $\boldsymbol{y}$ of the dependent variable $(n \times 1)$. $\boldsymbol{X}$ can be built up by sampling pixel-vectors in the images, but $\boldsymbol{y}$ is sometimes difficult to obtain, if the surface of the samples is formed by chemically heterogeneous regions. In this case, the quantitative analyses related to $\boldsymbol{y}$ may be difficult, if the pixels represent very small areas (for example on microscopic images) or very large ones (satellite images). This is for example the case when dealing with biological solid samples representing different tissues. Direct Calibration (DC), sometimes referred to as "spectral unmixing", do not require having such a calibration set. However, it is indeed needed to have both a matrix of "perturbing" pixel-vectors (noted $\boldsymbol{K}$) and a vector of the "pure" component spectrum to be analyzed ($\boldsymbol{p}$), which are more easily obtainable. For estimating the contribution, the unknown pixel vector $\mathbf{x}$ and the pure spectrum $\boldsymbol{p}$ are first projected orthogonally onto $\boldsymbol{K}$ giving the vectors $\boldsymbol{x}_{\perp}$ onto $\boldsymbol{p}_{\perp}$, respectively. The contribution is then estimated by a second projection of $\boldsymbol{x}_{\perp}$ onto $\boldsymbol{p}_{\perp}$. A method, based on principal component analysis, for determining the optimal dimensions of $\boldsymbol{K}$ is proposed. DC was applied on a collection of multivariate images of kernel of wheat to estimate the proportion of three tissues, namely *out-layers*, "*waxy*" *endosperm* and

* Corresponding author at: INRA UR1268 Biopolymères Interactions Assemblages, rue de la géraudière, BP 71627, 44316 Nantes, France. Tel.: +33 240675085.
E-mail address: benoit.jaillais@nantes.inra.fr (B. Jaillais).

0003-2670/$ – see front matter
http://dx.doi.org/10.1016/j.aca.2012.05.002

normal endosperm. The eventual results are presented as images of wheat kernels in false colors associated to the estimated proportions of the tissues. It is shown that DC is appropriate for estimating contributions in situations in which the more usual methods of calibration cannot be applied.

1. Introduction

Multivariate imaging instruments are now currently available in research laboratories or in industries. Such instruments basically produce images which can be represented by 3-way tables, with two dimensions associated with the spatial location and one related to the spectral measurements such as absorbance values. The spectroscopic techniques which can be applied in multivariate imaging are very diverse, and almost all of them have been adapted to imaging. The vibrational regions of visible, near infrared and mid infrared are now rather commonly exploited in imaging. Many other techniques such as nuclear magnetic resonance or mass spectroscopy are also applicable in multivariate imaging. Basic information on multispectral imaging can be found in [1].

The data processing of such images is not straightforward. Collections of multivariate images are generally very large, which implies that their processing must remain sufficiently simple and fast.

In many situations, the user would expect to have a relevant summary of a given multivariate image in the form of a single image in false colors which emphasizes the surfaces presenting different spectral characteristics. For this purpose, unsupervised or supervised clustering approaches have been followed. In a previous study [2] we have applied an unsupervised approach, based on the K-nearest neighbors clustering, for detecting pathological regions in infrared multispectral images of human aortic tissues. With this approach, it was possible to build up labeled images emphasizing the healthy or pathological regions, in false color the hue of which indicates the possible degree of the disease.

Many methods of unsupervised or supervised image classification are based on orthogonal subspace projection (OSP) [3]. The basic concept is to project each pixel vector onto a subspace which is orthogonal to some undesired signatures. Once the interfering signatures have been nullified, projecting the residual onto the signature of interest maximizes the signal-to-noise ratio and results in a single component image that represents a classification for the signature of interest. Ren [4] has carried out unsupervised multispectral image classifications, using a technique of OSP. He has proposed an algorithm enabling the detection of prototype pixel vectors called “targets”. OSP was applied on the multivariate images, taking the targets as desired or undesired pixel vectors. In projection methods, it is needed that the number of targets were smaller than the number of channels (or “bands”). In the work [4] this requirement was not respected. For this reason, the authors created a new set of additional bands that are generated nonlinearly from original multispectral bands. They have called “generalized OSP” (GOSP) this variant of the method.

The supervised approaches have often relied on discriminant analyses. In most cases, the qualitative groups are visually identified from the examination of some reference images. Signals associated with each qualitative group can be selected to constitute the training set of discriminant analysis. Here again, the final result can be presented as a false image showing the identified qualitative groups. As in spectroscopic techniques, the method for discrimination must be adapted to the usual situation of near-colinearity of the variables. Methods such as partial-least square discriminant analyses (PLSDA) are often appropriate in such a situation [5]. Wang et al. [6] have classified images of human brain obtained in magnetic resonance using GOSP. The qualitative groups represented three cerebral tissues (gray matter, white mater and cerebral spinal fluid). Their spectral signatures were extracted directly from the images by experienced radiologists.

The problem of estimating the contribution of a given material at each location of a multivariate image remains difficult. Often In some cases, the *indirect calibration methods* such as partial least squares, principal component or ridge regression cannot be directly applied, because the assessment of the regression model requires having a calibration data set including a matrix of predictive variables ($\boldsymbol{X}$) and the corresponding dependant variable ($\boldsymbol{y}$). If the sample has a heterogeneous surface, it may be impossible to measure the concentration at the surface of a sample, and $\boldsymbol{y}$ vector may be unavailable [1]. This situation commonly appears when the study deals with very large pixels, as obtained with aerial or satellite images, or on the contrary with small biological materials: it will indeed be very difficult to sample and analyze a great number of voxels (volume pixel elements) representing very large or very small amount of materials.

Thus, some cases, as the one studied here, deserve trying using *direct calibration methods* (DC). DC do not need a vector y of contribution, but they always require having a data matrix representing the “perturbing” spectra ($\boldsymbol{H}$) and a vector giving the “pure” ($\boldsymbol{p}$) spectrum. $\boldsymbol{H}$ represents spectra which contain information able to disturb the prediction of the measured compound. The “pure” spectrum $\boldsymbol{p}$ is the spectrum associated with the measured compound, without contaminant. It appears that, in many practical situations, $\boldsymbol{H}$ and $\boldsymbol{p}$ can be made available from multivariate images. For example, when dealing with section of biological material, one can be interested in labeling images according to the tissues which appear at the surface of the studied object. In such cases, each studied tissue can be seen as a pure compound. The specialist is generally able to identify regions in the images in which the studied tissue is very abundant, or, on the contrary, almost absent. It is thus rather easy to select spectra for building up $\boldsymbol{H}$. $\boldsymbol{p}$ can be obtained by acquiring spectra of the isolated tissue. The same kind of selection can be carried out in images of heterogeneous scenes in which the different compounds are, at least at some locations of the images, clearly separated.

Multivariate images of sections of wheat kernels are taken as an illustrative example. These images have been acquired in the frame of a genetic study of wheat in relation with their nutritional interest in the human diet (NOMAC, New tools for managing the fate of cereal nutrients in the gut, project of *Agence Nationale de la Recherche*, France). The wheat starch is usually composed of two α-glucans built mainly upon α-(1, 4) linkages: amylose, an essentially linear polymer, and amylopectin, a branched polymer containing 5–6% of α-(1, 6) linkages. In most common types of starch, the amount of amylose ranges between 18 and 28%, and those of amylopectin between 72 and 82%. Some properties of starch, such as swelling power, solubility, *in vitro* glycemic index and viscosity, are determined by the macromolecular characteristics and the conformation in solution of both constitutive polymers [7], and by the amylose content [8]. It was also shown that the amylose/amylopectin ratio may influence the rate of glucose absorption in the intestine [9].

In western countries, diseases related to foods represent a major issue in public policy. Overweight and obesity are increasing at an alarming rate in the world and in Europe more particularly. Nutrition is a major health determinant and is one of the key priorities in public health policy, especially in Europe. The consumption of cereals-based foods with low glycemic indexes are