Chemometrics and Intelligent Laboratory Systems 117 (2012) 61–69

Contents lists available at SciVerse ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Pretreatments by means of orthogonal projections

Jean-Claude Boulet [a,*], Jean-Michel Roger [b,1]

[a] INRA, UMR1083 Sciences Pour l'Oenologie, 2 place Viala, F-34060 Montpellier, France
[b] IRSTEA, UMR ITAP Information Technologies Analyse environnementale Procédés agricoles, F-34196 Montpellier, France

ARTICLE INFO

Article history:
Received 29 July 2011
Received in revised form 10 December 2011
Accepted 6 February 2012
Available online 17 February 2012

Keywords:
Orthogonal projection
Pretreatment
Preprocessing
Subspace
Linear model

ABSTRACT

This article describes several linear pretreatments based on orthogonal projections. The main differences of these pretreatments lie in the way the information to be removed are identified, using calibration dataset, pure spectra, experimental designs or mathematical models. Removing all the undesired spectral information yields spectra proportional to the net analyte signal, so it is important to collect the most complete information possible, using the complementarities of different approaches. The correction should then be processed with a single Euclidian orthogonal projection that gathers all the information, rather than with successive operations. By embedding Euclidian orthogonal projections into the calibration, it is not necessary to reapply them to new datasets.

1. Introduction

Spectroscopy has spread throughout many industries as an on-line process control tool because calibration models are able to extract quantitative information about a compound of interest from the spectra. Among the models proposed, regressions or inverse calibrations such as partial least square regression (PLSR) [1], extract the relevant spectral information by means of a calibration dataset. The term "partial" recalls that just a few dimensions or latent variables are used, and thus the information related to the other variables is dropped. Pretreatments or preprocessings are positioned prior to calibration. Their purpose is to identify and to remove spectral information that interferes with the desired prediction. Pretreatments and regressions share a same objective, so pretreatments allow regressions to perform better. Many different pretreatments are available. This paper focuses on describing the ones based upon orthogonal projections and complements a recent review by Rinnan et al. [2]. After introducing the notations, we describe several pretreatments focused on orthogonal projections, and then discuss their properties. We propose a clarified view of several pretreatments by putting forward their resemblances and complementarities and suggesting the best methods for their use. The detailed relationships of these pretreatments with other pretreatments (e.g. Savitsky–Golay (SG), standard normal variate) and with other calibration methods are outside the scope of this article.

1.1. Notations

Vectors are noted in bold lowercase, matrices in bold uppercase, and scalars in uppercase characters. Vectors are arranged in columns, except in matrices $\mathbf{X}$ and $\mathbf{X}_G$ where the lines represent the spectra. The transposed forms of vector $\mathbf{a}$ and matrix $\mathbf{A}$ are noted $\mathbf{a}'$ and $\mathbf{A}'$, respectively. The main notations are gathered in Table 1, and a glossary is also available in Table 2.

2. Pretreatments based on orthogonal projections

Pretreatments based on orthogonal projections deal with the correction of additive effects. For example, suppose that for sample i the observed spectrum $\mathbf{x}_{i,\,obs}$ is the sum of the expected spectrum $\mathbf{x}_i$ plus an unwanted contribution $\mathbf{h}_i$:

$$\mathbf{x}_{i,obs} = \mathbf{x}_i + \mathbf{h}_i \tag{1}$$

If a good estimation of $\mathbf{h}_i$ is available, the first possibility would be to perform a subtraction, and so $\mathbf{x}_i$ is estimated as:

$$\hat{x}_i = \mathbf{x}_{i,obs} - \hat{\mathbf{h}}_i$$

Unfortunately, because $\mathbf{h}_i$ is not well estimated for each spectrum i, this configuration is very uncommon in spectrometry. Nevertheless, it is possible to obtain a good estimation of the subspace ε^D spanned by the different vectors $\{\mathbf{h}_i\}$. Thus it becomes possible to build a projector orthogonal to this subspace. Let $\mathbf{P}$ be a matrix of dimensions

* Corresponding author. Tel.: +33 499613148; fax: +33 499612857.
E-mail addresses: bouletjc@supagro.inra.fr (J.-C. Boulet), jean-michel.roger@irstea.fr (J.-M. Roger).
[1] Tel.: +33 467046383; fax: +33 467166440.

0169-7439/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2012.02.002

Table 1
Main notations.

$\mathbf{X}$	Matrix $N\times P$, N samples and P spectral variables
$\mathbf{y}$	Vector $N\times 1$, the reference values
$\mathbf{X}_{1:\,i}$	Projection of $\mathbf{X}$ orthogonally to $\{\mathbf{t}_1, \mathbf{t}_2, \ldots\mathbf{t}_i\}$
$\mathbf{T}$	Matrix $N\times A$, scores for $\mathbf{X}$
$\mathbf{P}$	Matrix $P\times A$, loadings for $\mathbf{X}$
$\mathbf{W}$	Matrix $P\times A$, weights for $\mathbf{X}$
Σ	Moore–Penrose pseudo-inverse of $\left(\mathbf{X}'\mathbf{X}\right)$; $\Sigma = \left(\mathbf{X}'\mathbf{X}\right)^{+}$
$\mathbf{I}_N$, $\mathbf{I}_P$	Identity matrices for $\mathbb{R}^N$ and $\mathbb{R}^P$ spaces
$\mathcal{P}_P$	Euclidian orthogonal projector $P\times P$ onto $\mathbf{P}$; $\mathcal{P}_P = \mathbf{P}\left(\mathbf{P}'\mathbf{P}\right)^{-1}\mathbf{P}'$
$\mathcal{P}_P^{\perp}$	Euclidian projector $P\times P$ orthogonally to $\mathbf{P}$; $\mathcal{P}_P^{\perp} = \mathbf{I}-\mathcal{P}_P$
$\mathbf{t}_i$	i^{eme} column vector of $\mathbf{T}$
$\mathbf{p}_i$	i^{eme} column vector of $\mathbf{P}$
$\mathbf{w}_i$	i^{eme} column vector of $\mathbf{W}$
ε_X	Subspace of $\mathbb{R}^P$ spanned by the line vectors of $\mathbf{X}$
ε^U, ε^D	Useful/detrimental subspaces of $\mathbb{R}^P$

$P\times A$ whose column-vectors $\{\mathbf{p}_1, \mathbf{p}_2, \ldots\mathbf{p}_A\}$ form a basis of ε^D. Let $\mathbf{I}_P$ be the identity matrix of dimensions $P\times P$. The Euclidian orthogonal projector to $\mathbf{P}$ is:

$$\mathcal{P}_P^{\perp} = \mathbf{I}_P - \mathbf{P}\left(\mathbf{P}'\mathbf{P}\right)^{-1}\mathbf{P}'$$

A spectrum $\mathbf{x}_{i,\,corr}$ corrected from the information due to any $\mathbf{h}_i$ is obtained after a projection of $\mathbf{x}_{i,\,obs}$ orthogonally to $\mathbf{P}$:

$$\mathbf{x}_{i,corr} = \mathcal{P}_P^{\perp}\mathbf{x}_{i,obs} = \mathcal{P}_P^{\perp}\mathbf{x}_i \tag{2}$$

Note that the vector $\mathbf{x}_{i,\,corr}$ that is obtained after an orthogonal projection is very different from $\hat{x}_i$ obtained when a subtraction is possible. However in both cases, the influence of $\mathbf{h}_i$ has been reduced to nought. The orthogonal projector $\mathcal{P}_P^{\perp}$ is symmetrical: $\mathcal{P}'^{\perp}_{P} = \mathcal{P}^{\perp}_{P}$, so for N spectra forming the matrix $\mathbf{X}$ of dimensions $N\times P$:

$$\mathbf{X}_{corr} = \mathbf{X}\mathcal{P}_P^{\perp} = \mathbf{X}\left(\mathbf{I}_P - \mathbf{P}\left(\mathbf{P}'\mathbf{P}\right)^{-1}\mathbf{P}'\right) \tag{3}$$

Table 2
Glossary.

BB	Best basis
CPSA	Constrained principal spectra analysis
DFT	Discrete Fourier transform
DO	Direct orthogonalization
DOP	Dynamic orthogonal projection
DOSC	Direct orthogonal signal correction
DWT	Discrete wavelet transform
EMSC	Extended multiplicative signal correction
EPO	External parameter orthogonalization
EROS	Error removal by orthogonal subtraction
IIR	Independent interference reduction
MSC	Multiplicative signal correction
MDL	Minimum description length
NAP	Net analyte preprocessing
NAS	Net analyte signal
NSV	Net sensitivity vector
OPLS	Orthogonal projection to latent structures
OSC	Orthogonal signal correction
OSP	Orthogonal subspace projection
PCA	Principal component analysis
PCDA	Principal component discriminant analysis
PLSR	Projection to latent structures regression
RMSEP	Root mean square error of prediction
SG	Savitsky–Golay
SNV	Standard normal variate
SVD	Singular value decomposition
TOP	Transfer by orthogonal projection
WPT	Wavelet packet transform

The performances of the different pretreatments are directly explained by their ability to obtain a good approximation of a basis of ε^D. Different approaches are possible: using pure spectra, information extracted from experimental design, models, and calibration datasets. For each method, matrix $\mathbf{X}$ represents centered or uncentered data, depending on the centering option chosen. In order to simplify the presentation, all pretreatments presented here are for correcting spectra in which just one compound of interest is to be quantified. However, some pretreatments can also be written for the correction of several compounds of interest.

2.1. Pretreatment using pure spectra

A basis of the space spanned by chemical components is given by their pure spectra. A method derived from hyperspectral imaging uses this property.

2.1.1. Orthogonal subspace projection

The orthogonal subspace projection (OSP) uses pure spectra, called undesired signatures, which are associated with all the chemical influences present except the one of the compound of interest. These undesired signatures form matrix $\mathbf{K}$. They can be determined after a clustering process [3,4], in which homogeneous groups of spectra are obtained, followed by selection of a spectrum representative of each group. However in Harsanyi et al. [5] they were chosen within the image. The OSP correction is a projection that is orthogonal to $\mathbf{K}$, in accordance with Eq. (3):

$$\mathbf{X}_{OSP} = \mathbf{X}\left(\mathbf{I}_P - \mathbf{K}\left(\mathbf{K}'\mathbf{K}\right)^{-1}\mathbf{K}'\right)$$

In Harsanyi et al. [5], the OSP method was applied to a hyperspectral image from an airborne VIS-IR spectrometer using the radiance spectra directly. Several end members were identified and alternatively chosen as the compound of interest. An OSP was performed for each endmember. For each OSP, the corrected spectra were used to classify the pixels, and the results were in accordance with the measured values, or ground truth. However, the limits of this method are such that all the pure spectra must be known in advance and they cannot be collinear. In addition, influences such as temperature are not taken into account.

2.2. Pretreatments using spectra issued from an experimental design

If pure spectra are not available due to the chemical complexity of the samples or because the influence to be removed is physical and no pure spectrum exists, OSP cannot be applied. It is possible to construct experimental designs to obtain a matrix $\mathbf{X}_G$ whose spectra contain targeted spectral perturbations without any useful information. A singular value decomposition (SVD) or a principal component analysis (PCA) applied to $\mathbf{X}_G$ gives a matrix of eigenvectors $\mathbf{P}$ of dimensions $(P\times A)$ whose columns represent an orthonormal basis of the subspace to be removed. The matrix $\mathbf{X}$ is corrected to $\mathbf{X}_{corr}$ by a projection orthogonal to $\mathbf{P}$. The following formula is the same as Eq. (3) and can be simplified to:

$$\mathbf{X}_{corr} = \mathbf{X}\left(\mathbf{I}_P - \mathbf{P}\left(\mathbf{P}'\mathbf{P}\right)^{-1}\mathbf{P}'\right) = \mathbf{X}\left(\mathbf{I}_P - \mathbf{P}\mathbf{P}'\right)$$

Several methods have been based on this principle, but they differ in the way $\mathbf{X}_G$ is obtained, and in how the dimension of the SVD or the PCA applied to $\mathbf{X}_G$ is determined.

2.2.1. Independent interference reduction

The independent interference reduction method (IIR), Hansen [6] uses spectra from samples where the compound of interest is null and