Chemometrics and Intelligent Laboratory Systems 138 (2014) 48–56

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

A simple, projection-based geometric model for several linear pretreatment and calibration methods

Jean-Claude Boulet [a,*], Steven D. Brown [b,1], Jean-Michel Roger [c,2]

[a] INRA, UMR1083, 2 Place Viala, F34060 Montpellier, France
[b] University of Delaware, Newark, DE 19716, USA
[c] IRSTEA, UMR ITAP, 361 Rue J.F. Breton, F34191 Montpellier, France

ARTICLE INFO

Article history:
Received 15 April 2014
Received in revised form 7 July 2014
Accepted 15 July 2014
Available online 21 July 2014

Keywords:
Model
Projection
Calibration
Direct
Regression
Pretreatment

ABSTRACT

A general model is proposed for the explanation of several linear methods used for quantification. A first projection approach concerns pretreatments and calibrations. It consists of an oblique projection of the spectral data onto a subspace containing useful information for calibrations or detrimental information for pretreatments. Corrected spectra and scores are obtained for pretreatments and calibrations, respectively. A second projection approach concerns only calibrations. The regression vector is deduced after an orthogonal projection of the reference values onto the scores previously obtained. Several pretreatments, and direct and indirect (inverse) calibrations also called regressions are reviewed according to this model. The methods described are focused on spectroscopic applications. Some are very specific to spectroscopy; however, most of them also can be applied in other situations.

1. Introduction

A calibration is an operation that…in a first step, establishes a relation between the quantity values…provided by measurement standards and corresponding indications…and, in a second step, uses this information to establish a relation for obtaining a measurement result from an indication [1]. The metrological terms number of measurements, quantity values and indications are hereafter called number of observations, quantities of interest and variables respectively, terms that are more common in the literature. The quantities of interest correspond to the measured ($\mathbf{y}$) or estimated ($\hat{\mathbf{y}}$) values for a set of N observations of a compound of interest $\mathcal{Y}$ (*e.g.* sugar). Let $\mathbf{X}$ be the set of values of Q variables for the same observations. Then a calibration is a function f which yields $\hat{\mathbf{y}}$ from $\mathbf{X}$: $\hat{\mathbf{y}} = f(\mathbf{X})$. Finding f is a real issue. Countless solutions have been proposed; we do not intend to cover all of them. Non-linear methods, such as those based on artificial neural networks, are beyond the scope of this review. This paper is restricted to linear calibration methods. Still, there is a need for classification.

Three families of linear methods have been proposed: indirect calibrations (also called inverse calibrations or regressions), direct calibrations, and pretreatments [2]. For an observation i, a set of Q measurements is represented by the vector $\mathbf{x}_i$. Linear calibrations provide an estimate $\hat{y}_i$ of the quantity of interest by determining a vector of regression coefficients $\mathbf{b}$ of dimensions $(Q \times 1)$ such that: $\hat{y}_i = \mathbf{x}_i'\mathbf{b}$ with the constraint that $\mathbf{b}$ minimizes $|y_i - \hat{y}_i|$. Indirect calibrations build $\mathbf{b}$ with a calibration dataset, while direct calibrations build $\mathbf{b}$ with component responses (often spectra). On the other hand, pretreatments yield a corrected vector $\mathbf{x}_{i,corr}$. Separate reviews have shown links between the methods. Among the regressions, multiple linear regression (MLR), principal regression (PCR) and projection to latent structures regression (PLSR) have been compared according to the way in which scores and loadings are produced [2], their relevant components [3], the H-principle [4,5], the geometry [6], and tuning parameters [7]. Indirect methods have also been compared, *e.g.* [2,8,9] or with PLSR [10]. In the discussion that follows, pretreatments are considered separately from calibrations [11] with two exceptions: orthogonal signal correction (OSC), a preprocessing method whose concept originates from PLSR [12] and orthogonal subspace projection (OSP) [13], another preprocessing method which also involves a direct calibration method [14]. Thus, there are links between pretreatments, direct and inverse calibrations. The question is then: what is the general framework that underlies all these linear methods? A term commonly used is projections.

Projections applied to the variable space (see the definition given hereafter) reveal a general framework that unifies several pretreatment and calibration methods. Projections can be introduced by their linear

* Corresponding author. Tel.: +33 499613148; fax: +33 499612857.
E-mail addresses: bouletjc@supagro.inra.fr (J.-C. Boulet), sdb@udel.edu (S.D. Brown), jean-michel.roger@cemagref.fr (J.-M. Roger).
[1] Tel.: +1 3028316861; fax: +1 3028316335.
[2] Tel.: +33 467046383; fax: +33 467166440.

http://dx.doi.org/10.1016/j.chemolab.2014.07.012
0169-7439/

algebra properties, it is the aim of the next paragraph. Projections can also be introduced by their geometrical properties, as in a tutorial recently submitted [15]. Thus to begin with and to provide a quick view of their interest, orthogonal projections are illustrated with a short example based on spectroscopy. In the Chimiometrie 2007 Challenge, twelve spectra were acquired for a sample # 1 with different experimental conditions (temperature, grinding, and moisture). The operation was repeated for a sample # 2. The spectra obtained for sample # 1 are all different, due to the experimental conditions (Fig. 1(a)). Differences are also observed for sample # 2; removing them is the issue addressed by pretreatments. One possibility is to center the set of spectra # 1 in order to keep only the spectral information due to the experimental conditions (Fig. 1(b)). This centered set of spectra is used to build an orthogonal projector which is applied to the sample # 2 spectra, yielding the spectra in Fig. 1(c). Just one curve is visible. It means that the twelve spectra overlap almost perfectly. One can notice that the result is a reduction of the differences due to the experimental conditions, which is expected. But the original shape of the spectra has also been lost.

This paper begins with a brief overview of linear algebra. Then a general model is described. The following sections review direct calibrations, indirect calibrations and related pretreatment methods in accordance with the proposed framework. Calibrations are considered for the prediction of a single quantity of interest. Each method is briefly described with a focus on its geometry; more information is available in the original papers. A section is also dedicated to methods which are closely related, yet not part of the proposed framework. General properties of the methods are then discussed. Along with a classification of methods, this discussion develops the notion of useful and detrimental information and how the different methods are able to identify these kinds of information, then process them.

2. A brief overview of linear algebra

Concepts of orthogonal and oblique projections, spaces and metrics are used throughout the paper. The main acronyms are explained in Table 1. The notation used throughout the paper is conventional. Scalars are rendered in normal uppercase font (with italic lowercase for indices), vectors are rendered in bold lowercase font, and matrices are rendered in bold uppercase font. Subspaces are represented by the script fonts $\mathcal{E}$ and $\mathcal{R}$. The transposed form of any vector $\mathbf{u}$ and matrix $\mathbf{U}$ are $\mathbf{u}'$ and $\mathbf{U}'$, respectively. The inverse of a square matrix $\mathbf{U}$ is noted $\mathbf{U}^{-1}$. The Moore–Penrose pseudo-inverse of any matrix $\mathbf{U}$ is $\mathbf{U}^{+}$. Finally, the same symbol can be used in different methods because the value is different, but the meaning is the same. This is especially the case for the matrices $\mathbf{P}$, $\mathbf{M}$ and $\mathbf{D}$ which represent loadings, a metric, and a matrix of detrimental information respectively. When more precision is needed, a subscript is used: *e.g.* $\mathbf{P}_{PLSR}$.

- Vectorial spaces and subspaces, bases:
  The vector $\mathbf{x}_i$ is defined in the vector space $\mathbb{R}^Q$, also called the variable space. Subspaces of $\mathbb{R}^Q$ are defined by a basis. For example, the set of vectors $\{\mathbf{p}_1, \mathbf{p}_2, \ldots \mathbf{p}_A\}$, the column vectors of $\mathbf{P}$, defines a basis of a unique subspace $\mathcal{R}(\mathbf{P})$ [16]. However, different bases, yielding different matrices $\mathbf{P}$, may span the same subspace. The row space and column spaces of $\mathbf{X}$ are $\mathcal{R}(\mathbf{X}')$ and $\mathcal{R}(\mathbf{X})$, subspaces of $\mathbb{R}^Q$ and $\mathbb{R}^N$, respectively.
- Metric: A metric $\mathbf{M}$ for vector space $\mathbb{R}^Q$ is a symmetrical matrix of dimension $(Q \times Q)$ with specific properties [17]. $\mathbf{M}$ is used for the calculation of the generalized inner product between vectors $\mathbf{u}$ and $\mathbf{v}$: $< \mathbf{u}, \mathbf{v} > = \mathbf{u}'\mathbf{M}\mathbf{v}$. Metrics are often set to the identity matrix, but sometimes they are not, *e.g.* for calculating a Mahalanobis distance.
- Orthogonal and oblique projections:
  A matrix Ψ of dimensions $(Q \times Q)$ is a projector if $\Psi^2 = \Psi$. Projectors are built using two matrices associated with the range-space and null-space [16], or alternatively with a metric and a basis [17,18]. Only the second of these two approaches is considered here. The oblique projector Ψ onto $\mathcal{R}(\mathbf{P})$ using the metric $\mathbf{M}$, also called the $\mathbf{M}$-projector onto $\mathcal{R}(\mathbf{P})$, is:

$$\Psi = \mathbf{M}\mathbf{P}\left(\mathbf{P}'\mathbf{M}\mathbf{P}\right)^{-1}\mathbf{P}'. \quad (1)$$

The M-antiprojector to $\mathcal{R}(\mathbf{P})$ is: $\Psi^{\perp} = \mathbf{I}_Q - \Psi$. These projectors have been designed to be applied to the right of matrices or vectors.

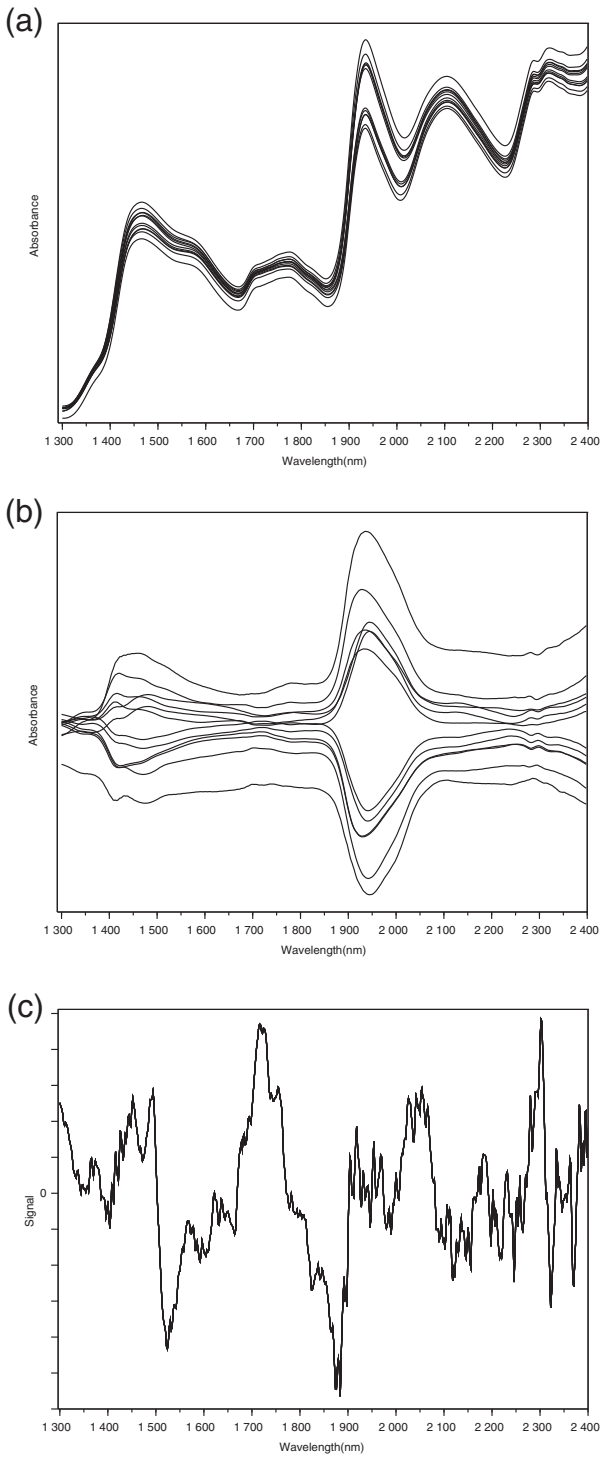

Fig. 1. An orthogonal projection illustrated with near infrared spectra of wheat flour. (a) Twelve spectra of sample # 1 acquired at 3 temperatures, 2 moistures and 2 grindings. (b) Previous spectra after centering. (c) The twelve spectra of sample # 2, after an orthogonal projection to all but one of the centered spectra of sample # 1.