DROP-D: dimension reduction by orthogonal projection for discrimination

Xavier Hadoux, Douglas Rutledge, Gilles Rabatel, Jean-Michel Roger

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2015, N°146, p. 221-231.
The original publication is available at http://www.sciencedirect.com
Doi: 10.1016/j.chemolab.2015.05.021

DROP-D: Dimension Reduction by Orthogonal Projection for Discrimination

Xavier Hadoux[a], Douglas N. Rutledge[b], Gilles Rabatel[a], Jean-Michel Roger[a]

[a]*Irstea UMR ITAP, 361 rue J-F Breton BP 5095, 34196 Montpellier Cedex 5, France*
[b]*AgroParisTech, UMR1145 Ingénierie Procédés Aliments, 16 rue Claude Bernard, 75005 Paris, France*

Abstract

The objective of this paper is two-fold. First, some theoretical aspects of dimension reduction in the context of supervised classification or discrimination are given. The emphasis is put on the different subspaces that can be defined in this context and what information is contained in each of them. Then, based on these theoretical aspects, we propose a novel method for supervised dimension reduction that is dedicated to discrimination purposes. The method, called Dimension Reduction by Orthogonal Projection for Discrimination (DROP-D) is particularly well suited to the high dimensionality and high intercorrelation of spectral variables. As with Fisher discriminant analysis, DROP-D aims at finding a lower dimensional subspace in which the classes are well separated. To do so, DROP-D cleans the observation matrix of variability sources that do not help with the classification task. For this purpose, the matrix is projected orthogonally to the within-class axes which prevent a good class separability. In cases where some between-class axes are collinear with the within-class axes, DROP-D can preserve these axes in order not to destroy the class separability. DROP-D discriminant axes are orthogonal to one another and thus offer a simplified interpretablility. The main advantage of DROP-D is that because it is based on removing unnecessary information, there is no need of a validation set to tune the model parameters. In contrast to modelling techniques, DROP-D thus cannot find class separability when there is none. In terms of results, DROP-D offers similar performances to the usual linear classification methods.

Keywords: , Discrimination, Dimension reduction, Orthogonal projection, Supervised Classification, Spectral discrimination, Discriminant analysis

1. Introduction

Due to their ability to perform accurate and non-destructive measurements, spectral data have been increasingly used in many scientific and industrial fields over the last decades. The acquired spectral data, measured as a function of wavelength, are often composed of more than a hundred narrow bands which means that the classical classification techniques fail. In practice, because spectral variables are also highly correlated their dimension can be reduced without loosing important information [1, 2, 3].

Email address: xavier@hadoux.com (Xavier Hadoux)

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2015, N°146, p. 221-231.
The original publication is available at http://www.sciencedirect.com
Doi: 10.1016/j.chemolab.2015.05.021

In the case of discriminant models, the variable to be predicted is qualitative. The basic discrimination approach as developed by Fisher in [4] calculates the discriminant subspace as the one spanned by the eigenvectors of $\mathbf{W}^{-1}\mathbf{B}$, where $\mathbf{W}$ and $\mathbf{B}$ are the within- and between-class dispersion matrices, respectively. Two problems arise when dealing with spectral variables — the first problem, related to the dimensionality (size) of the spectral data set ($\mathbf{X}$), is due to the fact that the calibration set generally contains many more variables (wavelengths) than individuals (spectra); the second problem, related to matrix conditioning, is due to the very high intercorrelation of the measured spectral variables. Actually, both problems are mathematically similar, i.e., they are both related to a lack of reliable dimensions of the spectral space, which inhibits the inversion of dispersion matrices [5, 6]. The direct consequence for discrimination of spectral data is that the within class dispersion matrix $\mathbf{W}$ is usually highly singular and so cannot be inverted.

Factorial discriminant methods solve the issues of dimensionality and conditioning in the same way as factorial regression [7], i.e., a classical discriminant analysis (DA) is performed on latent variables obtained by a factorial analysis algorithm. For instance, a Principal Components Analysis (PCA) of the spectral matrix followed by a DA is a PCA-DA and a Partial Least Squares (PLS) [8] between the spectral and the class membership matrices followed by a DA is a PLS-DA [9]. As far as regression is concerned, PLSR is generally more powerful than PCR, since the calculation of the PLS latent variables takes into account the relationship between the spectral variables and the responses and not only variability within the spectral data [10]. For the same reason, PLS-DA is generally more efficient than PCA-DA in the case of discrimination, i.e., by providing reduced variables with a smaller within-class spread and a larger distance between classes [11, 12, 13]. PLS-DA is however naturally prone to overfitting the training set and can find a class structure in the data even where there is none. With such a method, the parameters must therefore be tuned using either an independent validation set or using a cross-validation procedure [14], which is problematic when samples are expensive or difficult to obtain. Other criticisms often made towards PLS is that it is difficult to interpret loadings and scores because the procedure not only takes into account variations related to the response to be predicted but must also correct for interfering variations (see O-PLS [15] for a complete discussion on this topic).

Other approaches have been attempted to modify Fisher's discriminant paradigm directly without prior dimension reduction using mathematical tricks to overcome the within-class matrix inversion problem, e.g., using pseudo inverses or by inversion of the total dispersion matrix instead of the within-class matrix [16, 17, 18]. One of the best performing methods, Nullspace LDA (NLDA) uses as discriminant vectors the between-class principal directions computed on data where all within-class scatter is removed. In doing so, the weighting by $\mathbf{W}^{-1}$ originally proposed by Fisher to find the optimal discriminant vector is pushed to a limit where no within-class directions are left. However, in this nullspace, because the projections are orthogonal to all the within-class directions, the remaining information is very small and this leads to noisy discriminant vectors.

In this paper, we propose an approach in which, in contrast to NLDA, the removal of the within-class variability is controlled so that the most important discriminant information is preserved. This approach uses orthogonal projection in order to clean some of the within-class variability from the observed variable matrix while preserving the most important between-class directions. A PCA is then applied on the cleaned variables and the resulting scores are used for discrimination. This method is a general framework in which:
- PCA-DA is a special case, i.e., when no cleaning is performed and,

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2015, N°146, p. 221-231.
The original publication is available at http://www.sciencedirect.com
Doi: 10.1016/j.chemolab.2015.05.021

- NLDA is at the other extreme, i.e., when all within-class directions are removed.

We also show that the optimal solution offered by DROP-D generally lies between those of PCA-DA and NLDA.

2. Revisiting discrimination

The most important notations used in this paper are summarized in table 1.

Table 1: Notations

Symbol	Description
$\mathbf{A}$	Matrix of size $n \times m$.
$\mathbf{E}_k(\mathbf{A})$	Matrix $(m \times k)$ containing the k eigenvectors of $\mathbf{A}$ associated to its k largest eigenvalues.
$\mathcal{R}(\mathbf{A})$	Range of $\mathbf{A}$: Subspace spanned by the columns of $\mathbf{A}$.
$P_{\mathbf{P}}(\mathbf{A})$	Projection of the column vectors of $\mathbf{A}$ onto a subspace spanned by the matrix $\mathbf{P}$.
	In $\mathbb{R}^n$, $P_{\mathbf{P}}(\mathbf{A}) = \mathbf{P}(\mathbf{P}^\top\mathbf{P})^{-1}\mathbf{P}^\top\mathbf{A}$.
	In $\mathbb{R}^m$, $P_{\mathbf{P}}(\mathbf{A}) = \mathbf{A}\mathbf{P}(\mathbf{P}^\top\mathbf{P})^{-1}\mathbf{P}^\top$.
$P^\perp_{\mathbf{P}}(\mathbf{A})$	Projection of the column vectors of $\mathbf{A}$ orthogonally to a subspace spanned by the matrix $\mathbf{P}$.
	In $\mathbb{R}^n$, $P^\perp_{\mathbf{P}}(\mathbf{A}) = (\mathbf{I}_n - \mathbf{P}(\mathbf{P}^\top\mathbf{P})^{-1}\mathbf{P}^\top)\mathbf{A}$.
	In $\mathbb{R}^m$, $P^\perp_{\mathbf{P}}(\mathbf{A}) = \mathbf{A}(\mathbf{I}_m - \mathbf{P}(\mathbf{P}^\top\mathbf{P})^{-1}\mathbf{P}^\top)$.
$P_{\mathbf{P},k}(\mathbf{A})$	Projection of $\mathbf{A}$ onto the subspace spanned by the k main directions of $\mathcal{R}(\mathbf{P})$.
$P^\perp_{\mathbf{P},k}(\mathbf{A})$	Projection of $\mathbf{A}$ orthogonal to the subspace spanned by the k main directions of $\mathcal{R}(\mathbf{P})$.
$\mathbf{X}$	Matrix $(N \times P)$ containing N observations of a P-variable vector.
$\mathbf{Y}$	Dummy matrix $(N \times C)$ coding the class membership of each observation ($[0, 0, 1, 0]$ codes for the third class among four).
$\mathbf{T}(\mathbf{X})$	Total dispersion of $\mathbf{X}$; $\mathbf{T}(\mathbf{X}) = \mathbf{X}^\top\mathbf{X}$.
$\mathbf{B}(\mathbf{X}, \mathbf{Y})$	Between-class dispersion $\mathbf{B}(\mathbf{X}, \mathbf{Y}) = \mathbf{X}^\top\mathbf{Y}(\mathbf{Y}^\top\mathbf{Y})^{-1}\mathbf{Y}^\top\mathbf{X}$.
$\mathbf{W}(\mathbf{X}, \mathbf{Y})$	Within-class dispersion $\mathbf{W}(\mathbf{X}, \mathbf{Y}) = \mathbf{X}^\top\mathbf{X} - \mathbf{X}^\top\mathbf{Y}(\mathbf{Y}^\top\mathbf{Y})^{-1}\mathbf{Y}^\top\mathbf{X}$.
	When there is no risk of confusion, the symbols $\mathbf{T}$, $\mathbf{B}$ and $\mathbf{W}$ are used without arguments.
$\mathcal{E} = \mathcal{R}(\mathbf{X})$	Individual space ($\mathcal{E} \subseteq \mathbb{R}^N$).
$\mathcal{F} = \mathcal{R}(\mathbf{X}^\top)$	Feature (Variable) space ($\mathcal{F} \subseteq \mathbb{R}^P$).

2.1. Prerequisites

In the following we call the *individual space*, the N-dimensional space (one axis per observation) in which we can represent the variables (wavelengths) as vectors. Conversely, the *variable space* is the P-dimensional space (one axis per variable) in which we can represent the observations as vectors.

2.1.1. Reminders on orthogonal projection [19, 20]

For any column vector $\mathbf{v}$, and for any subspace defined by its basis $\mathbf{P}$, the orthogonal projection of $\mathbf{v}$ on $\mathbf{P}$ is given by

$$P^\perp_{\mathbf{P}}(\mathbf{v}) = \mathbf{P}(\mathbf{P}^\top\mathbf{P})^{-1}\mathbf{P}^\top\mathbf{v}^\top \tag{1}$$

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2015, N°146, p. 221-231.
The original publication is available at http://www.sciencedirect.com
Doi: 10.1016/j.chemolab.2015.05.021

2.2. Subspace decomposition : problem statement

Using a data matrix $\mathbf{X}$ and a class matrix $\mathbf{Y}$ based on training samples, supervised classification consists in finding a model that is capable of predicting the class of any observation $\mathbf{x}$ using its P descriptors. With spectral data, classification is often done in two steps [13] –1) projection of the observation in a lower-dimensional subspace; 2) attribution of the individual to a class. The efficacy of the second step is highly influenced by that of the first one. Hence, we are looking for a subspace in which class centres are well separated and class spread around their centers is small.

From a mathematical point of view, this corresponds to finding Q discriminant axes $\mathbf{D}$ ($P \times Q$) such that the coordinates of $\mathbf{X}$ in the subspace defined by $\mathbf{D}$, given by $\mathbf{XD}$ maximize the between-class dispersion given by

$$\mathbf{B}(\mathbf{XD}, \mathbf{Y}) = (\mathbf{XD})^{\top}\mathbf{Y}(\mathbf{Y}^{\top}\mathbf{Y})^{-1}\mathbf{Y}^{\top}(\mathbf{XD}) \tag{2}$$

and minimizes the within-class dispersion given by

$$\mathbf{W}(\mathbf{XD}, \mathbf{Y}) = (\mathbf{XD})^{\top}(\mathbf{XD}) - (\mathbf{XD})^{\top}\mathbf{Y}(\mathbf{Y}^{\top}\mathbf{Y})^{-1}\mathbf{Y}^{\top}(\mathbf{XD}). \tag{3}$$

In addition, we are looking for a subspace of reduced dimensions, i.e., Q minimal. These three constraints $\min(Q)$, $\max\Big(\text{trace}(\mathbf{B})\Big)$ and $\min\Big(\text{trace}(\mathbf{W})\Big)$ on the way to build the set of axes are illustrated in Figure 1. This can be achieved by minimizing the ratio of within- to total-class dispersion given by the Wilks' Lambda

$$\Lambda_{Wilks} = \frac{|\,\mathbf{W}\,|}{|\,\mathbf{T}\,|} = \frac{|\,\mathbf{W}\,|}{|\,\mathbf{B}+\mathbf{W}\,|} \tag{4}$$

where $|\,\mathbf{A}\,|$ corresponds to the determinant of the square matrix $\mathbf{A}$.

In cases where the data are well conditioned, a solution is given by the Fisher Linear Discriminant Analysis (LDA) paradigm, which can be expressed as:

$$\mathbf{D} = \arg\max_{\mathbf{D}}\Big(\text{trace}(\mathbf{D}^{\top}\mathbf{W}^{-1}\mathbf{BD})\Big) = \mathbf{E}_Q(\mathbf{W}^{-1}\mathbf{B}) \tag{5}$$

where the notation $\mathbf{E}_Q(\mathbf{A})$ corresponds to the Q eigenvectors associated to the Q largest eigenvalues of the matrix $\mathbf{A}$. However, for ill-conditioned data, the inversion of $\mathbf{W}$ is problematic. Thus, LDA is known to be unable to deal with spectral data and several solutions have been proposed in the literature to overcome this problem [21].

Nevertheless, the construction of a classification model corresponds to finding a subspace of the variable space that 'copies' the class structure observed in the individual space of the sample set. The Fisher LDA does this by contracting the subspace corresponding to the within-class dispersion and by focusing on the subspace carried by between-class dispersion.

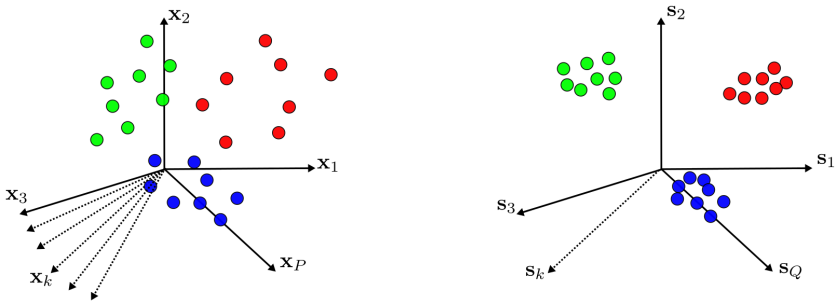
The method proposed in this paper offers another way to realize this 'copy'. This idea is to use the between- and within-class dispersion to decompose the variable space into different subspaces so that one of them corresponds to a large part of between-class dispersion and a small part of within-class dispersion.

2.3. Variability decomposition in $\mathbb{R}^N$ and $\mathbb{R}^P$

Suppose we have a matrix $\mathbf{X}$ containing N observations of P variables from C classes coded using dummy variables and stored in a matrix $\mathbf{Y}$. We can then define a mean observation per class using the matrix operation

$$\mathbf{X}_B = \mathbf{Y}(\mathbf{Y}^{\top}\mathbf{Y})^{-1}\mathbf{Y}^{\top}\mathbf{X}. \tag{6}$$

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2015, N°146, p. 221-231.
The original publication is available at http://www.sciencedirect.com
Doi: 10.1016/j.chemolab.2015.05.021



(a) $\mathbb{R}^p$ (b) $\mathbb{R}^q$

Figure 1: Illustration of the dimension reduction for classification purposes, i.e., fewer axes ($Q \leq P$), a small within-class dispersion and a large distance between class centroids

Thanks to the dummy variable coding, each of $\mathbf{X}_B$ row contains, instead of the original observation, the mean observation of its own class. The operation $\mathbf{X} \mapsto \mathbf{Y}(\mathbf{Y}^\top\mathbf{Y})^{-1}\mathbf{Y}^\top\mathbf{X}$ thus defines a new sample in which each observation is replaced by the mean of its class (centroid).

We can similarly define $\mathbf{X}_W$ as

$$\begin{aligned}\mathbf{X}_W &= \mathbf{X} - \mathbf{X}_B \\ &= (\mathbf{I}_N - \mathbf{Y}(\mathbf{Y}^\top\mathbf{Y})^{-1}\mathbf{Y}^\top)\mathbf{X}. \end{aligned} \tag{7}$$

The matrix $\mathbf{X}_W$ thus contains the observations centred on their class centroid.

What happens in the individual space ($\mathbb{R}^N$)?

In this space, we can represent the $\mathbf{Y}$ matrix, i.e., each column of $\mathbf{Y}$ corresponds to one vertex of the unit N-dimensional hypercube.

In this space, the operation $\mathbf{X} \mapsto \mathbf{Y}(\mathbf{Y}^\top\mathbf{Y})^{-1}\mathbf{Y}^\top\mathbf{X} = \mathbf{X}_B$ projects the columns of $\mathbf{X}$ (individuals) on the subspace defined by $\mathbf{Y}$. The removed part corresponds to $\mathbf{X}_W$ which is also an orthogonal projection, but on the orthogonal complement of $\mathbf{Y}$.

$$\mathbf{X}_B = \mathcal{P}_{\mathbf{Y}}(\mathbf{X}) = \mathbf{Y}(\mathbf{Y}^\top\mathbf{Y})^{-1}\mathbf{Y}^\top\mathbf{X} \tag{8}$$

$$\mathbf{X}_W = \mathcal{P}^{\perp}_{\mathbf{Y}}(\mathbf{X}) = (\mathbf{I}_N - \mathbf{Y}(\mathbf{Y}^\top\mathbf{Y})^{-1}\mathbf{Y}^\top)\mathbf{X} \tag{9}$$

The subspaces spanned by these matrices, $\mathcal{E}_B = \mathcal{R}(\mathbf{X}_B)$ and $\mathcal{E}_W = \mathcal{R}(\mathbf{X}_W)$, are orthogonal and complementary subspaces of $\mathcal{E}_T$ in $\mathbb{R}^N$ (equation 10).

$$\mathcal{E}_T = \mathcal{E}_B \oplus \mathcal{E}_W \subseteq \mathbb{R}^N \tag{10}$$

In this space ($\mathbb{R}^N$), owing to the orthogonality, we can thus completely eliminate $\mathbf{X}_W$ without affecting $\mathbf{X}_B$.

However, we are in the individual space, which means that this operation can be applied to vectors expressed as combinations of individuals and only modify the variable values of the N observations of the training set, i.e., it is not applicable to any incoming spectrum.

What happens in the variable space ($\mathbb{R}^P$)?

In this space, the operation $\mathbf{X} \mapsto \mathbf{X}_B$ defined by a matrix $N \times N$ in the individual space cannot be directly applied to a unique observation.

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2015, N°146, p. 221-231.
The original publication is available at http://www.sciencedirect.com
Doi: 10.1016/j.chemolab.2015.05.021

However, in this space, we can use the subspaces spanned by $\mathbf{X}_B$ and $\mathbf{X}_W$. We can show that the between- and within-class dispersion matrices define an orthogonal basis for these subspaces and can therefore be used to 'copy' the class structure from $\mathbb{R}^N$ to $\mathbb{R}^P$. Indeed, it can be easily shown that formulas $\mathbf{B}(\mathbf{X}, \mathbf{Y})$ and $\mathbf{W}(\mathbf{X}, \mathbf{Y})$ can be retrieved by computing

$$\mathbf{T}(\mathbf{X}_B) = \mathbf{B}(\mathbf{X}, \mathbf{Y}) \tag{11}$$

$$\mathbf{T}(\mathbf{X}_W) = \mathbf{W}(\mathbf{X}, \mathbf{Y}) \tag{12}$$

, i.e., the total dispersion of $\mathbf{X}_B$ and $\mathbf{X}_W$ corresponds to $\mathbf{B}$ and $\mathbf{W}$ respectively.

Proof. Let us call $\widetilde{\mathbf{Y}} = \mathbf{Y}(\mathbf{Y}^\intercal\mathbf{Y})^{-1}\mathbf{Y}^\intercal$ the projector. Noting that this projector is symmetric ($\widetilde{\mathbf{Y}}^\intercal = \widetilde{\mathbf{Y}}$) and idempotent ($\widetilde{\mathbf{Y}}^2 = \tilde{\mathbf{Y}}$) we have:

$$\mathbf{T}(\mathbf{X}_B) \triangleq (\widetilde{\mathbf{Y}}\mathbf{X})^\intercal(\widetilde{\mathbf{Y}}\mathbf{X}) = \mathbf{X}^\intercal\widetilde{\mathbf{Y}}\mathbf{X} \triangleq \mathbf{B}(\mathbf{X}, \mathbf{Y}) \tag{13}$$

$$\begin{aligned}\mathbf{T}(\mathbf{X}_W) &\triangleq \Big((\mathbf{I}_N - \widetilde{\mathbf{Y}})\mathbf{X}\Big)^\intercal\Big((\mathbf{I}_N - \widetilde{\mathbf{Y}})\mathbf{X}\Big)\\ &= (\mathbf{X}^\intercal - \mathbf{X}^\intercal\widetilde{\mathbf{Y}})(\mathbf{X} - \widetilde{\mathbf{Y}}\mathbf{X})\\ &= \mathbf{X}^\intercal\mathbf{X} - \mathbf{X}^\intercal\widetilde{\mathbf{Y}}\mathbf{X}\\ &\triangleq \mathbf{W}(\mathbf{X}, \mathbf{Y})\end{aligned} \tag{14}$$

□

Hence, the subspace spanned by $\mathbf{X}_B$ contains the between-class dispersion while the subspace spanned by $\mathbf{X}_W$ contains the within-class dispersion.

In the variable space $\mathbb{R}^P$, let us define these two subspaces $\mathcal{F}_B = \mathcal{R}(\mathbf{X}_B^\intercal)$ and $\mathcal{F}_W = \mathcal{R}(\mathbf{X}_W^\intercal)$. Using the range property $\mathcal{R}(\mathbf{A}^\intercal) = \mathcal{R}(\mathbf{A}^\intercal\mathbf{A})$, these subspaces are expressed as

$$\mathcal{F}_T = \mathcal{R}(\mathbf{X}^\intercal) = \mathcal{R}(\mathbf{X}^\intercal\mathbf{X}) = \mathcal{R}(\mathbf{T}) \tag{15}$$

$$\mathcal{F}_B = \mathcal{R}(\mathbf{X}_B^\intercal) = \mathcal{R}(\mathbf{X}_B^\intercal\mathbf{X}_B) = \mathcal{R}(\mathbf{B}) \tag{16}$$

$$\mathcal{F}_W = \mathcal{R}(\mathbf{X}_W^\intercal) = \mathcal{R}(\mathbf{X}_W^\intercal\mathbf{X}_W) = \mathcal{R}(\mathbf{W}) \tag{17}$$

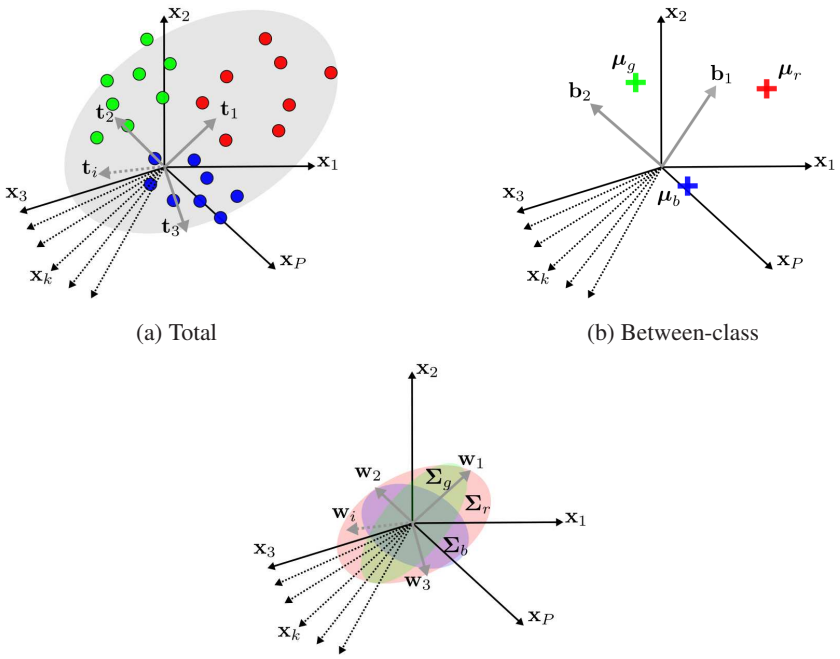
where we have by construction $\mathbf{T} = \mathbf{B} + \mathbf{W}$. Figure 2 illustrates this decomposition in the feature space:

- The total subspace $\mathcal{F}_\mathbf{T}$, whose dimension is bounded by $\dim(\mathcal{F}_\mathbf{T}) \leq \min(N, P)$, represents the overall data variability in the variable space without considering classes.
- The between-class subspace $\mathcal{F}_\mathbf{B}$ is defined by the spread of the class centroids in the variable space. Its dimension is thus bounded by $\dim(\mathcal{F}_\mathbf{B}) \leq \min(C - 1, P)$.
- The within-class subspace $\mathcal{F}_\mathbf{W}$ corresponds to the overall spread of the data around the class centroids. Its dimension is bounded by $\dim(\mathcal{F}_\mathbf{W}) \leq \min(N, P)$.

Then, because the subspace dimension and matrix rank are linked by the fundamental relation

$$dim\,\mathcal{R}(\mathbf{A}) = rank(\mathbf{A}) = rank(\mathbf{A}^\intercal) = dim\,\mathcal{R}(\mathbf{A}^\intercal) \tag{18}$$

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2015, N°146, p. 221-231.
The original publication is available at http://www.sciencedirect.com
Doi: 10.1016/j.chemolab.2015.05.021



(a) Total

(b) Between-class

(c) Within-class

Figure 2: Decomposition in the feature space $\mathbf{T} = \mathbf{B} + \mathbf{W}$. Note the possible collinearity between $\mathbf{b}_i$ and $\mathbf{w}_j$.

and since, $dim\ \mathcal{F}_\mathbf{T} = dim\ \mathcal{E}_\mathbf{T}$, $dim\ \mathcal{F}_\mathbf{B} = dim\ \mathcal{E}_\mathbf{B}$ and $dim\ \mathcal{F}_\mathbf{W} = dim\ \mathcal{E}_\mathbf{W}$, therefore $\mathcal{F}_\mathbf{B}$ and $\mathcal{F}_\mathbf{W}$ define two subspaces of $\mathcal{F}_\mathbf{T}$ in the variable space such that

$$\mathcal{F}_\mathbf{T} = \mathcal{F}_\mathbf{B} + \mathcal{F}_\mathbf{W} \subseteq \mathbb{R}^p \tag{19}$$

These subspaces $\mathcal{F}_\mathbf{B}$ and $\mathcal{F}_\mathbf{W}$ are however not orthogonal in $\mathbb{R}^P$ and their intersection is not necessarily empty. The separation of the between- and within-class dispersion is therefore less obvious than in the individual space. Hence, depending on the class configuration, removing within-class variability does not necessarily improves the separability as illustrated with Figure 3.

In the following (section 3.1), we propose a method, called DROP-D, that enables a controlled removal of the within-class dispersion, i.e., by preserving its axes collinear with $\mathcal{F}_\mathbf{B}$.

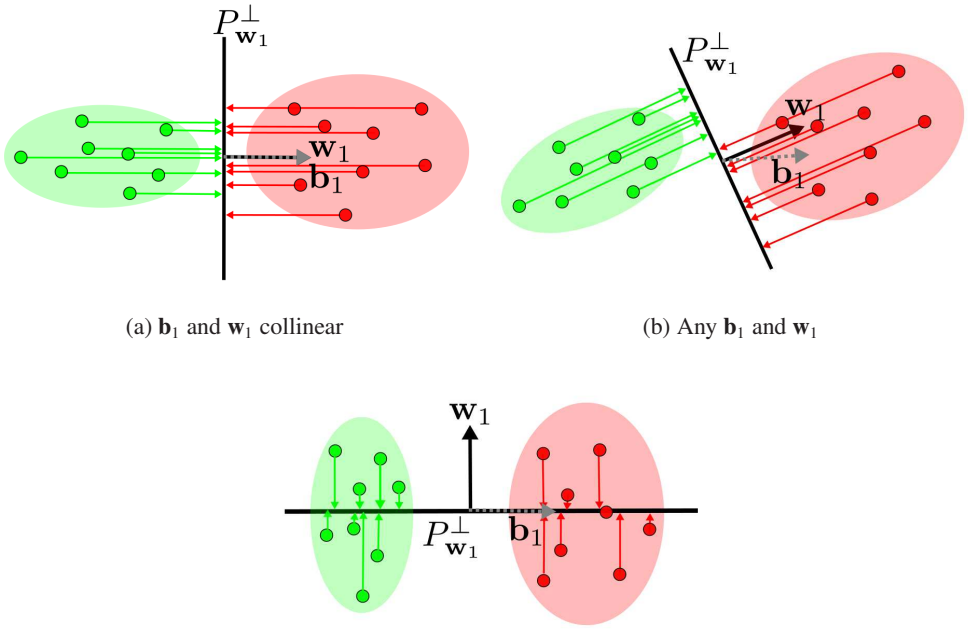
3. Proposed approach

3.1. DROP-D

Let us define the notations $P_{\mathbf{P},k}(\mathbf{A})$ (resp. $P^{\perp}_{\mathbf{P},k}(\mathbf{A})$) to correspond to the projection of the matrix $\mathbf{A}$ onto (resp. orthogonal) to the subspace spanned by the k main directions of $\mathcal{R}(\mathbf{P})$, where $\mathcal{R}(\mathbf{P})$, is the subspace spanned by the columns of $\mathbf{P}$. The matrix containing the k eigenvectors of $\mathbf{A}$ associated to its k largest eigenvalues is noted $\mathbf{E}_k(\mathbf{A})$. These notations are summarized in Table 1.

Dimension Reduction by Orthogonal Projection for Discrimination method (DROP-D) is in three steps.

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2015, N°146, p. 221-231.
The original publication is available at http://www.sciencedirect.com
Doi: 10.1016/j.chemolab.2015.05.021



(a) $\mathbf{b}_1$ and $\mathbf{w}_1$ collinear

(b) Any $\mathbf{b}_1$ and $\mathbf{w}_1$

(c) $\mathbf{b}_1$ and $\mathbf{w}_1$ orthogonal

Figure 3: Effect of removing the first within-class axis with different class configurations in $\mathbb{R}^P$

The first step consists in (temporarily) removing from $\mathbf{X}$ the b principal directions of the between-class scatter, as expressed in equation 20.

$$\mathbf{X}_b^{\perp} = P^{\perp}_{\mathbf{B},b}(\mathbf{X}) \tag{20}$$

In a second step, the within-class dispersion matrix is computed with $(\mathbf{X}_b^{\perp}, \mathbf{Y})$. Then, the w principal directions linked to this within-class dispersion ($\mathbf{W}^*$) are eliminated according to the equation 21.

$$\mathbf{X}_{clean} = P^{\perp}_{\mathbf{W}^*\left(\mathbf{X}_b^{\perp},\mathbf{Y}\right),w}(\mathbf{X}) \tag{21}$$

The third step is to extract the Q principal directions of $\mathbf{X}_{clean}$ which are given by the principal directions of the total dispersion matrix computed with $\mathbf{X}_{clean}$:

$$\mathbf{D} = \mathbf{E}_Q\Big(\mathbf{T}(\mathbf{X}_{clean})\Big). \tag{22}$$

To summarize, DROP-D defines three subspaces $\mathcal{F}_B$, $\mathcal{F}_{W^*}$ and $\mathcal{F}_D$ of $\mathbb{R}^P$, such that:

- $\mathcal{F}_B$ is linked to the b principal directions of the between-class dispersion;
- $\mathcal{F}_{W^*}$ contains the w principal directions of the within-class variance that are orthogonal to $\mathcal{F}_B$;
- $\mathcal{F}_D$ contains the Q directions that include the b principal directions of the between-class dispersion and the $Q-b$ principal directions that are orthogonal to the within-class dispersion.

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2015, N°146, p. 221-231.
The original publication is available at http://www.sciencedirect.com
Doi: 10.1016/j.chemolab.2015.05.021

In doing so, DROP-D eliminates the principal directions of the within-class dispersion while preserving the most important directions of the between-class dispersion. An unconstrained projection orthogonal to $\mathbf{W}$ would introduce the risk of removing important axes of $\mathbf{B}$, because $\mathcal{F}_B$ and $\mathcal{F}_W$ can have a collinear part. In that sense, the step 1 of DROP-D guarantees preserving at least the b most important axes of $\mathcal{F}_B$. In addition, axes of $\mathcal{F}_B$ that were not included in step 1 preservation, but that are orthogonal to $\mathcal{F}_W$, are preserved as well.

3.2. Algorithm

If the data are not already centred: compute the training set mean spectrum ($\mathbf{xm} \leftarrow \mathit{mean}(\mathbf{X})$) and subtract it from every row of the matrix ($\mathbf{X} \leftarrow \mathit{center}(\mathbf{X}, \mathbf{xm})$).

DROP-D algorithm is as follows:

1. $\mathbf{B} \leftarrow \mathbf{X}^\top\mathbf{Y}(\mathbf{Y}^\top\mathbf{Y})^{-1}\mathbf{Y}^\top\mathbf{X}$ *Compute the between-class dispersion*
2. $\mathbf{B}_b \leftarrow \mathbf{E}_b(\mathbf{B})$ *Extract the b principal eigenvectors of* $\mathbf{B}$ *(via SVD*$(\mathbf{B})$*)*
3. $\mathbf{X}_b^\perp \leftarrow \mathbf{X}(\mathbf{I}_P - \mathbf{B}_b(\mathbf{B}_b^\top\mathbf{B}_b)^{-1}\mathbf{B}_b^\top)$ *Remove from* $\mathbf{X}$ *these b directions*
4. $\mathbf{W}^* \leftarrow \mathbf{X}_b^{\perp\top}\mathbf{X}_b^\perp - \mathbf{X}_b^{\perp\top}\mathbf{Y}(\mathbf{Y}^\top\mathbf{Y})^{-1}\mathbf{Y}^\top\mathbf{X}_b^\perp$ *Compute the within-class dispersion with* $\mathbf{X}_b^{\perp\top}$ *and* $\mathbf{Y}$
5. $\mathbf{W}_w^* \leftarrow \mathbf{E}_w(\mathbf{W}^*)$ *Extract the w principal eigenvectors of* $\mathbf{W}^*$ *(via SVD*$(\mathbf{W}^*)$*). These w directions are assured to be at least orthogonal to the b previously removed directions*
6. $\mathbf{X}_{clean} \leftarrow \mathbf{X}(\mathbf{I}_P - \mathbf{W}_w^*(\mathbf{W}_w^{*\top}\mathbf{W}_w^*)^{-1}\mathbf{W}_w^{*\top})$ *Remove from THE ORIGINAL* $\mathbf{X}$ *these w directions*
   $\mathbf{T}^* \leftarrow \mathbf{X}_{clean}^\top\mathbf{X}_{clean}$ *Compute the total dispersion of* $\mathbf{X}_{clean}$
7. $\mathbf{D} \leftarrow \mathbf{E}_Q(\mathbf{T})$ *Extract the Q principal eigenvectors of* $\mathbf{T}$ *via SVD*$(\mathbf{T})$
8. Optimize b, w and Q

Any new vector $\mathbf{x}$ is projected on this new basis by computing $\mathbf{s} = (\mathbf{x} - \mathbf{xm})\mathbf{D}^\top$.

4. Experimental results and discussions

In the following, discrimination results are obtained using the Mahalanobis distance on the scores of each of DROP-D, NLDA, PLS and PCA [22].

4.1. Data sets

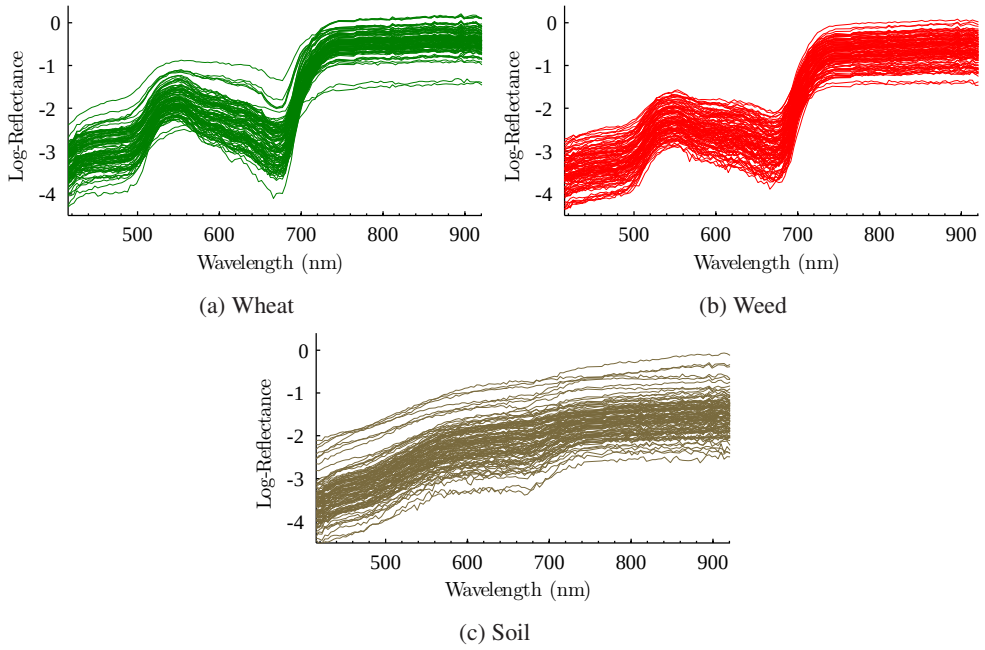
For this study, we used spectra provided by an hyperspectral camera Hyspex V-NIR 1600 (Norks Elektro Optikk, Norway).

4.1.1. Data set A

Data set A contains spectra extracted from an hyperspectral image of vegetation scene acquired under natural lighting at short range ($\sim$ 1 m). More details on the acquisition setup are described in [23]. The acquired spectra were composed of 160 spectral bands ranging from 415.11 nm to 993.54 nm. The radiance image was transformed into a reflectance image using a ceramic plate calibrated in the laboratory. Due to the low sensitivity of the camera sensor in the near infrared (NIR), the 20 spectral bands above 920.78 nm were discarded. Then, because of the high atmospheric absorption of oxygen at around 750 nm which affects natural lighting spectrum, bands 93 to 96 were discarded as well. In order to linearise the effect of leaf inclination on spectrum amplitude, a logarithmic transformation was applied to each reflectance spectrum [24].

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2015, N°146, p. 221-231.
The original publication is available at http://www.sciencedirect.com
Doi: 10.1016/j.chemolab.2015.05.021

For the discrimination, around 3000 spectra were manually labelled in the hyperspectral image into one of the three classes: wheat, weed or soil. For training the models, 100 spectra per class (300 in total) were randomly extracted from the available ground truth, the rest being left for model validation. The log-transformed training spectra for each class are presented in Figure 4.



(a) Wheat

(b) Weed

(c) Soil

Figure 4: Data set A presentation: Log-transformed training spectra for each class.

4.1.2. Data set B

Data set B contains spectra extracted from a remotely-sensed hyperspectral image acquired with a camera embedded in a light aircraft (Piper Seneca II PA 34) flying at ~ 650 m. This image was extracted from a field measurement campaign carried out over the Quiberon peninsulas (France) by Actimar (Actimar, Brest, France.) within the exploratory research and innovation project named HypLitt (more information on the measurement campaign are given in [25]). The acquired spectra were composed of 160 spectral bands ranging from 409.6 nm to 986.8 nm. Reflectance images were obtained through the atmospheric model ATCOR and then adjusted using spectroradiometric measurement on the ground using reference surfaces [25]. Because of low signal values below 442.3 nm, the first ten spectral bands were discarded. Then, because of saturation of vegetation spectra above 841.6 nm, spectral bands from 121 to 160 were discarded as well. Finally, the oxygen absorption bands at around 750 nm, which corresponds to spectral bands 84 to 90 were also removed. The topographic effects on spectrum amplitude were linearised using a logarithmic transformation on each reflectance spectrum.

A ground truth map that discriminates four classes, i.e., grass, deciduous, conifer and sand was manually created. For training the models, 100 spectra per class (400 in total) were randomly extracted from the available ground truth map, which correspond to approximately 4.2% of the

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2015, N°146, p. 221-231.
The original publication is available at http://www.sciencedirect.com
Doi: 10.1016/j.chemolab.2015.05.021

available data, the rest being left for model validation. The log-transformed training spectra for each class are presented in Figure 5.

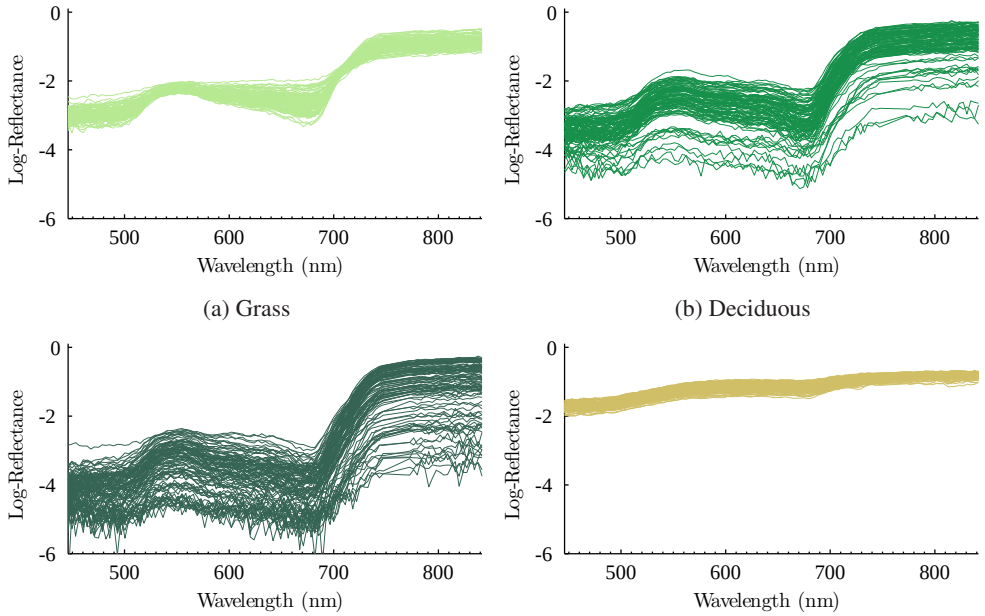


Figure 5: Data set B presentation: Log-transformed training spectra for each class.

4.1.3. Procedures

In order to provide numerical results, we used the overall classification error, which corresponds to the ratio of pixels incorrectly classified over the total number of pixels, expressed in percent. For cross-validation, where applicable (i.e., number of latent variable for PLS, number of principal axes for PCA), we used a 10-fold procedure on the training data [14].

In the following, we present different aspects of the dimension reduction method DROP-D using our two data sets. With data set A, we first illustrate the collinearity of the dispersion matrices' eigenvectors in the variable space. Secondly, we show DROP-D in action step by step and illustrate the effect of removing the within-class variability on the class separability. Thirdly, we show that the number of within-class axes to remove can be tuned without cross-validation by comparing calibration and cross validation results. Using data set B, we illustrate how a full calibration procedure of DROP-D can be accomplished using only the training set. We also illustrate, using an artificial data set, that since DROP-D operates by removing information, it cannot model a class structure when there is none. Finally, with both data sets A and B, we compare DROP-D classification performances with PCA-DA, Nullspace LDA (NLDA) and PLS-DA.

4.2. Collinearity in $\mathbb{R}^P$

In order to assess the collinearity issue in the variable space, let us first have a look at the eigenstructure of the total, between- and within- class dispersion matrices for our data set. In

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2015, N°146, p. 221-231.
The original publication is available at http://www.sciencedirect.com
Doi: 10.1016/j.chemolab.2015.05.021

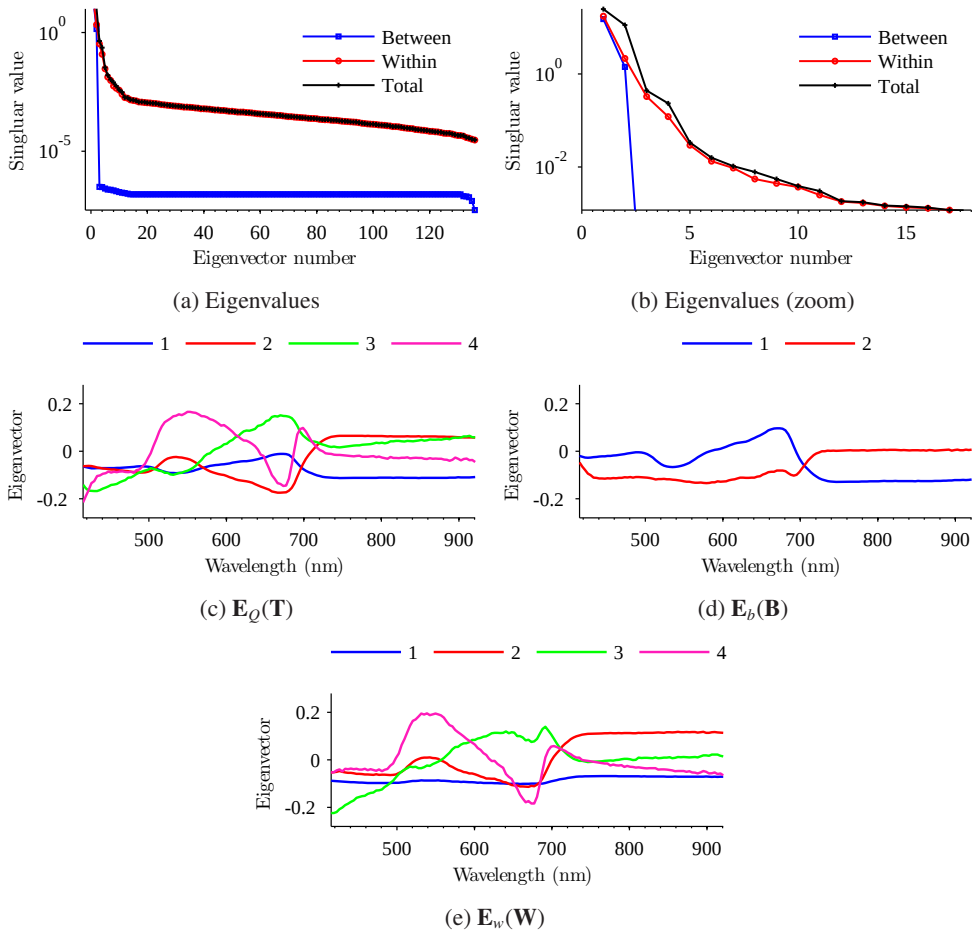


Figure 6: Eigenvalues and the principal eigenvectors of the dispersion matrices **T**, **B** and **W**

		$E_b(\mathbf{B})$		$E_Q(\mathbf{T})$					
		$b=1$	$b=2$	$Q=1$	$Q=2$	$Q=3$	$Q=4$	$Q=5$	$Q=6$
$E_w(\mathbf{W})$	$w=1$	63	**30**	**29**	62	88	89	90	90
	$w=2$	**34**	61	62	**30**	80	90	88	90
	$w=3$	90	90	85	83	**28**	64	88	89
	$w=4$	89	88	87	87	65	**27**	89	82
	$w=5$	89	89	89	88	88	90	**8**	89
	$w=6$	89	89	90	90	87	86	89	**23**
$E_b(\mathbf{B})$	$b=1$	**34**	56	88	89	90	90		
	$b=2$	58	**38**	77	77	89	88		

Table 2: Angle (in degrees) between the principal eigenvectors of the dispersion matrices **T**, **B** and **W**

Figure 6 are represented the eigenvalue plot and the main eigenvectors for this data set. The angle between each combination of these eigenvectors is also given is Table 2.

The eigenvalue plot illustrates that the maximum number of eigenvectors for **B** is $C-1$,

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2015, N°146, p. 221-231.
The original publication is available at http://www.sciencedirect.com
Doi: 10.1016/j.chemolab.2015.05.021

where C is the number of different classes. In the data set, in which three classes have to be discriminated, there are thus only 2 non-zero eigenvalues (Figure 6).

For $\mathbf{W}$ and $\mathbf{T}$, the number of non-zero eigenvalue is at most $\min(P-1, N)$, where P is the number of variables and N the number of observations in the training set. Note that in this data set, we have more observations than variables. Therefore, these matrices ranks are numerically full. However, they cannot be inverted because of their bad conditioning (ratio of the maximum to minimum eigenvalue). The actual rank is thus only around 10 to 15, the remaining is only due to observation noise.

Observing the shape of the eigenvectors plotted as curves gives a first indication of the non-orthogonality of the eigenvectors: some of them are nearly identical between the different sets. This intuition is confirmed by the angles reported in Table: 2.

Some eigenvectors are clearly non-orthogonal (angle < 90 degrees). In this particular case, most non-orthogonal eigenvectors have an angle of approximately 30 degrees. As we will see in the following, with this type of data, a compromise has to be made on removing or keeping these within-class eigenvectors using an orthogonal projection because of a potential loss in discrimination power.

Although this approach, which consists in examining the angle between eigenvectors is interesting to understand the class structure of the data set, tuning DROP-D parameters looking at these tables is unmanageable as the number of classes increases. Therefore, in the following, we adopt a classical parameter tuning approach.

Final remarks concern the 'shape' of the obtained eigenvectors. As was observed in the data set presentation (Figure 4), vegetation spectra have a strong reflectance feature at around 700 nm. This transition, situated at the edge of the red and infrared part of the electromagnetic spectrum is very distinctive for vegetation spectra and is often referred to as the red-edge. This red-edge is mostly due to a strong absorption by the chlorophyll within the vegetation and is thus characteristic of the type of plant. The structure of the red edge (position, slope) is thus naturally found as a discriminative feature by classifiers. At the end of the red-edge, another characteristic feature of the vegetation spectra is the near infrared plateau. Finally, the greenness of the vegetation is also often discriminative between vegetation types and specific features are thus often found in the 500 to 600 nm range.

4.3. Effect of removing W on the class separability

As we have seen in the previous section, the between- and within-class dispersion matrices can have some non-orthogonal principal directions. Also, in order to decrease the Wilks' Lambda and thus to increase class separability, a possible approach would be to suppress the within-class variability by removing the principal axis of the within-class dispersion. However, because of this non-orthogonality, removing these directions may affect the class separability as well. The idea of DROP-D, as presented in the previous section, is thus to prevent the suppression of too much between-class dispersion. In the following we present this effect step by step for different numbers (w) of within-class axes removed and by preserving different numbers (b) of between-class directions.

With $b = 0$, no between-class direction is a priori preserved. Figure 7 show the Wilks' Lambda of the training data plotted as a function of the number of within-class directions removed. We also represent the obtained dispersion plot for three specified values. At $w = 0$, the discriminant vectors correspond to those of the Principal Component Analysis (PCA). The two vegetation classes (wheat and weed) are poorly separated but distinct from the third class (soil).

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2015, N°146, p. 221-231.
The original publication is available at http://www.sciencedirect.com
Doi: 10.1016/j.chemolab.2015.05.021

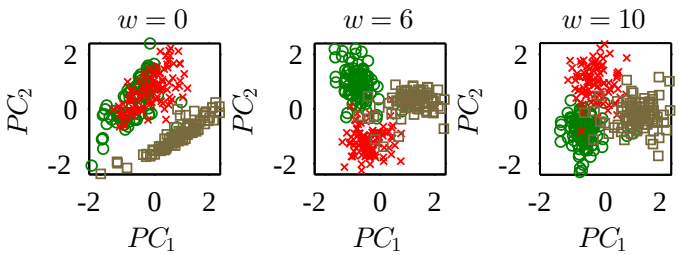
Then, until $w = 9$, the 'noisy' aspect of the obtained curve is due to the non-orthogonality of **W** and **T**. The discriminant vectors keep changing due to the removal of within-class directions. Then, at around 9 or 10 removed axes, a clearer minimum is obtained. Removing more axes only degrades the class separability (even for the training set).

With $b = 1$, the principal direction of the between-class dispersion is preserved. As observed in Figure 8, a similar 'noisy' pattern is obtained, but only until $w = 5$. In this case only the second axis of the between-class is affected by the successive cleaning of the within-class directions. The clear minimum obtained for $w = 5$ is stable until $w = 8$. Then, removing more directions starts affecting the class separability as well, e.g., see the Figure 8 at $w = 12$ where the wheat (green) and weed (red) starts to cluster together.

With $b = 2$, every between-class direction is preserved. With three classes, this corresponds to the limiting case of DROP-D in which any cleaning does not change the class separability unless more discriminant axes (Q) are kept (results not shown).

These results on class separability were all obtained with $Q = 2$ in order to provide these two-dimensional scatter plots. In the general case, Q, which corresponds to the final number of discriminant vectors to be used, is another parameter to tune. It actually corresponds to a PCA on the cleaned spectral matrix as explained in the previous section. Usually, with PCA, the correct number of components to retain is always subject to discussion because the error decreases only slowly and an optimal threshold is difficult to estimate. The rule of thumb in such cases is 'less is better'. Fortunately, we will see in the following figures that when cleaning the spectral matrix with DROP-D, this threshold appears to be easier to find.

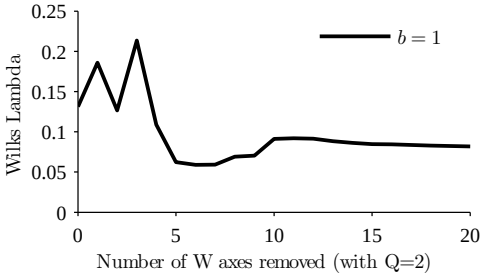
(a) Wilks' Lambda plotted as a function of the number of within-class principal axes removed.



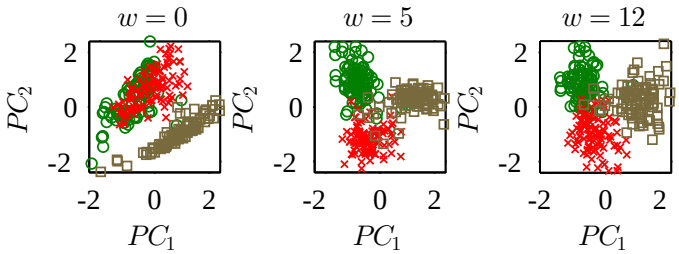
(b) Scores on principal components 1 and 2 for the training sets

Figure 7: Class separability as a function of the number of within-class principal directions removed. ($b = 0$)

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2015, N°146, p. 221-231.
The original publication is available at http://www.sciencedirect.com
Doi: 10.1016/j.chemolab.2015.05.021



(a) Wilks' Lambda plotted as a function of the number of within-class principal axes removed.



(b) Scores on principal components 1 and 2 for the training set

Figure 8: Class separability as a function of the number of within-class principal directions removed. (b = 1)

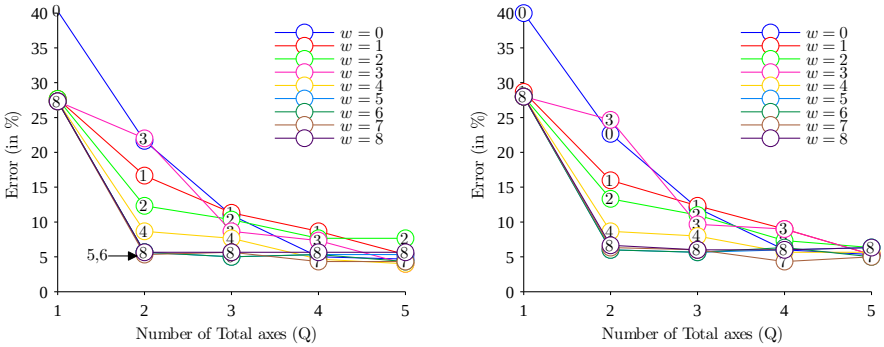
4.4. Model calibration

As seen in the previous section, with a careful selection of the between- and within-class principal axes to keep or to remove, various class separabilities can be obtained. Also, it was seen that owing to the DROP-D approach, which consists in removing information (contrary to PLS-DA, which learns the class structure by modelling **B**), overfitting can be detected directly on the training set by observing the class separability. In the following, we show that a similar behaviour is observed with the classification performance.

In figure 9, we show the classification error obtained by DROP-D with the training set (calibration error) and using a 10-fold cross-validation on the training set. This graph presents the classification error as a function of the number of final discriminant axes (Q) for different numbers of within-class axes removed (number inside the circle). With $w = 0$ (which corresponds to a classical PCA) both calibration and cross-validation errors decrease smoothly without any clear minimum. Then, from $w = 1$ to $w = 4$, we observe the similar noisy appearance, but in terms of classification performance. From $w = 5$ (optimal) to $w = 7$, the same classification error is obtained. In addition, a clear optimal value for Q emerges ($Q = 2$). Then, observed with the class separability, when removing one more axis, the error starts to increase. Therefore, $w = 5$ is chosen as an optimal value since it corresponds to the smallest value for which the optimal results are obtained. Also note a similar behaviour obtained for both calibration and cross-validation curves.

Let us assess the optimal parameters b, w and Q for the data set B using only the calibration error. Figure 10 shows the four sets of curves that correspond to every possible values for b. Without preserving the first two between-class axes, removing w always leads to worse results.

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2015, N°146, p. 221-231.
The original publication is available at http://www.sciencedirect.com
Doi: 10.1016/j.chemolab.2015.05.021



(a) Calibration error (b) Cross-validation error

Figure 9: Representation of the classification error of calibration (a) and cross-validation (b) for different values for w and Q with $b = 1$.

Note that in these cases, deciding for an optimal value for Q is not an easy task as explained before. When $b = 2$, a clear optimum is reached by removing only one within-class axis. In addition, the optimal value for Q also becomes more obvious to choose. In particular, removing more w or increasing Q both lead to worse results. Finally, when preserving the last possible between-class direction, classification results become slightly worse. The optimal parameters for this data set are thus $b = 2$, $w = 1$ and $Q = 3$. These values actually correspond to those obtained with the 10-fold cross-validation (not represented here).

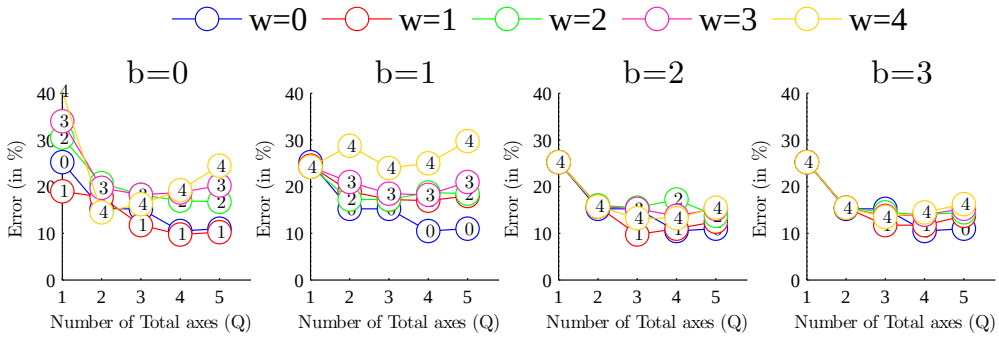


Figure 10: Data set B: Illustration of a full calibration procedure. The curves represent the training error of classification for different parameters b,w and Q.

Finally, to illustrate that DROP-D cannot learn a class structure when there is none, we show in Figure 11 the classification error obtained with the same data set in which the class matrix has been randomly permuted. For any number of removed within-class axes, no structure can be extracted and the classification results remain constant. On the other hand, with a PLS-DA model trained on the same data, a class structure can be learned and that method is thus prone to overfitting. This major difference comes from the fact that DROP-D removes $\mathbf{W}$ while PLS-DA learns a class structure by modelling $\mathbf{B}$ [9]. Therefore, because of the high dimensionality, a class structure can potentially be learned, especially with a small training set. On the contrary, when

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2015, N°146, p. 221-231.
The original publication is available at http://www.sciencedirect.com
Doi: 10.1016/j.chemolab.2015.05.021

removing information with DROP-D, if the information was useful for discrimination, even the training data is affected by the loss.

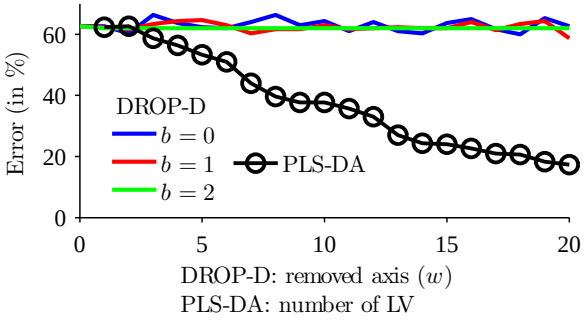


Figure 11: Comparison of DROP-D and PLS-DA classification errors on the training set with a random class matrix.

4.5. Classification performances

As we have seen, a major interest of DROP-D is to provide a method robust to overfitting. Let us now have a look at the classification performance that this method provides on our data sets. We also compare the results with the most commonly used dimension reduction methods, i.e., PCA-DA, NLDA (setting $w = 15$) and PLS-DA. For all methods, the class decision is made using the minimum of Mahalanobis distance between each class. Note that because of equal class prior in our case, this affectation corresponds to a Quadratic Discriminant Analysis (QDA) on the obtained scores [22].

An initial classification performance assessment is qualitative and is made by observing the shape of the obtained discriminant vectors for data set A (Figure 12). Indeed, as every dimension reduction method used is linear, the obtained discriminant vectors can be plotted as spectra and can be analysed in the same way. Only those of DROP-D and NLDA are orthogonal since they correspond to the eigenvectors of symmetric matrices. However, although it leads to high classification performances, the 'shapes' of NLDA discriminant vectors are not interpretable in practice, which probably explains the lack of interest from the chemometrics community. The DROP-D discriminant vectors that come from preserved between-class principal directions appear less noisy due to the averaging involved in the computation of these eigenvectors. However, the discriminant vectors obtained by removing within-class axes are noisier, but offer a different type of information to PLS-DA or PCA-DA.

Then, for practical uses, it is interesting to assess the classification performances for different numbers of training samples. For this purpose, with both data sets, we randomly selected among the training set from 10 to 60 spectra per class, in steps of 10. These results are presented in Figure 13.

The obtained results in terms of classification performance are very similar with these data sets. All these methods appear to be relatively non sensitive to the lack of training samples. In particular, above 30 samples per class, the classification stabilizes to its optimal value, which is of great interest for practical uses. The only noticeable difference was with data set A, when less than 30 training samples per class were used, comes from the estimation of the class covariance at the decision stage (Mahalanobis distance). DROP-D and NLDA subspaces were

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2015, N°146, p. 221-231.
The original publication is available at http://www.sciencedirect.com
Doi: 10.1016/j.chemolab.2015.05.021

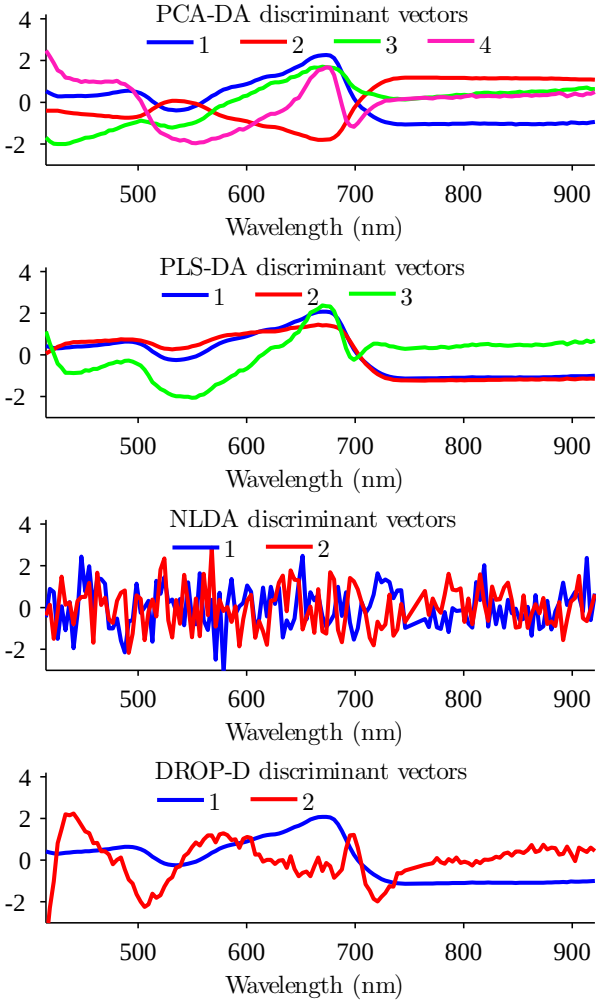
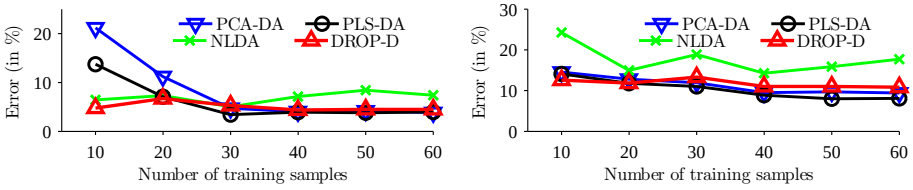


Figure 12: Discriminant vectors obtained for the four tested methods.



(a) Data set A (b) Data set B

Figure 13: Validation error for an increasing number of training samples.

Author-produced version of the article published in Chemometrics and Intelligent Laboratory Systems, 2015, N°146, p. 221-231.
The original publication is available at http://www.sciencedirect.com
Doi: 10.1016/j.chemolab.2015.05.021

two-dimensional, PLS-DA was three dimensional and PCA-DA was four dimensional, therefore, with such a small number of samples, the class covariance estimation was better in the lower dimensional spaces.

5. Conclusion

The aim of this paper was two-fold. First, to give a new vision of the discrimination paradigm using subspace decomposition. Then, based on these observations, we proposed a novel discrimination approach that uses orthogonal projection to clean the data before dimension reduction.

The proposed method is called DROP-D and performs a data cleaning using an orthogonal projection onto the within-class principal directions. In that sense, it mimics Fisher LDA, but instead of weighting the projection by the within-class inversion, it directly removes the information due to this within class-variations.

We also show that unless being very careful when removing the within-class information, the class separability can be lost because of non-orthogonality of the within- and between-class principal directions. Therefore, in DROP-D, a first step consists in preserving the most important between-class directions so that no cleaning can be performed on them. Once the data is cleaned, a classical Principal Component Analysis is performed in order to provide reduced data.

This method provides similar results to the more classical approaches (PCA-DA, PLS-DA and NLDA) in terms of classification performances. However, contrary to PLS-DA, due to the nature of the method, overfitting can be avoided without using the cross-validation procedure. Indeed, since it is cleaning the data instead of learning a class structure, DROP-D classification results are affected if useful information is removed even during the training phase.

Other dimension reduction strategies, which may prove to be even more powerful and subtle in their cleaning performance are currently under investigation. Another research direction is to link this work to existing orthogonal filtering techniques already used in regression analysis.

6. References

[1] P. Geladi, Chemometrics in spectroscopy. Part 1. Classical chemometrics, Spectrochimica Acta Part B: Atomic Spectroscopy 58 (5) (2003) 767–782.
[2] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemometrics and Intelligent Laboratory Systems 2 (1-3) (1987) 37–52.
[3] P. Eshghi, Dimensionality choice in principal components analysis via cross-validatory methods, Chemometrics and Intelligent Laboratory Systems 130 (2014) 6–13.
[4] R. A. Fisher, The use of multiple measurements in taxonomic problems, Annals of eugenics 7 (2) (1936) 179–188.
[5] N. Tormod, M. Bjorn-Helge, Understanding the collinearity problem in regression and discriminant, Journal of Chemometrics 426 (1998) (2001) 413–426.
[6] L. O. Jimenez, A. Member, D. A. Landgrebe, L. Fellow, Supervised Classification in High-Dimensional Space : Geometrical , Statistical , and Asymptotical Properties of Multivariate Data, IEEE Transactions on Systems, Man and Cybernetics 28 (1) (1998) 39–54.
[7] T. Naes, T. Isaksson, T. Fearn, T. Davies, A user friendly guide to multivariate calibration and classification., NIR publications Chichester, 2002.
[8] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemometrics and Intelligent Laboratory Systems 58 (2001) 109–130.
[9] M. Barker, W. Rayens, Partial least squares for discrimination, Journal of Chemometrics 17 (3) (2003) 166–173.
[10] J. L. Godoy, J. R. Vega, J. L. Marchetti, Relationships between PCA and PLS-regression, Chemometrics and Intelligent Laboratory Systems 130 (2014) 182–191. doi:10.1016/j.chemolab.2013.11.008.
[11] H. Nocairi, E. Mostafa Qannari, E. Vigneau, D. Bertrand, Discrimination on latent components with respect to patterns. Application to multicollinear data, Computational Statistics & Data Analysis 48 (1) (2005) 139–147.

[12] E. Kemsley, Discriminant analysis of high-dimensional data: a comparison of principal components analysis and partial least squares data reduction methods, Chemometrics and Intelligent Laboratory Systems 33 (1) (1996) 47–61.
[13] U. G. Indahl, N. S. Sahni, B. Kirkhus, T. Næ s, Multivariate strategies for classification based on NIR-spectra—with application to mayonnaise, Chemometrics and Intelligent Laboratory Systems 49 (1) (1999) 19–31.
[14] K. H. Esbensen, P. Geladi, Principles of Proper Validation: use and abuse of re-sampling for validation, Journal of Chemometrics 24 (3-4) (2010) 168–187.
[15] J. Trygg, S. Wold, Orthogonal projections to latent structures (O-PLS), Journal of Chemometrics 16 (3) (2002) 119–128.
[16] J. Ye, Q. Li, A two-stage linear discriminant analysis via QR-decomposition, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (6) (2005) 929–941. doi:10.1109/TPAMI.2005.110.
[17] Y.-F. Guo, L. Wu, H. Lu, Z. Feng, X. Xue, Null Foley–Sammon transform, Pattern Recognition 39 (11) (2006) 2248–2251.
[18] D. M. Witten, R. Tibshirani, Penalized classification using Fisher's linear discriminant., Journal of the Royal Statistical Society. Series B, Statistical methodology 73 (5) (2011) 753–772.
[19] C. D. Meyer, Matrix Analysis and Linear Algebra, 2000.
[20] J.-C. Boulet, J.-M. Roger, Pretreatments by means of orthogonal projections, Chemometrics and Intelligent Laboratory Systems 117 (2012) 61–69.
[21] X. Hadoux, Some contributions to supervised classification of hyperspectral data, Ph.D. thesis, Université de Montpellier 2, France (2014).
[22] R. De Maesschalck, D. Jouan-Rimbaud, D. L. Massart, The mahalanobis distance, Chemometrics and intelligent laboratory systems 50 (1) (2000) 1–18.
[23] N. Vigneau, M. Ecarnot, G. Rabatel, P. Roumet, Potential of field hyperspectral imaging as a non destructive method to assess leaf nitrogen content in Wheat, Field Crops Research 122 (1) (2011) 25–31.
[24] X. Hadoux, N. Gorretta, J.-M. Roger, R. Bendoula, G. Rabatel, Comparison of the efficacy of spectral pretreatments for wheat and weed discrimination in outdoor conditions, Computers and Electronics in Agriculture 108 (0) (2014) 242 – 249.
[25] S. Smet, G. Sicot, M. Lennon, Evaluation des capacités de le télédétection hyperspectrale et développement de méthodes innovantes de traitement d'images pour des applications défense en zone littorale (hyplitt), Tech. rep., , contrat de recherche DGA 2010 34 0014 (2010).