ISPRS Journal of Photogrammetry and Remote Sensing 132 (2017) 88–101

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

# Hyperspectral dimensionality reduction for biophysical variable statistical retrieval

Juan Pablo Rivera-Caicedo [a,b], Jochem Verrelst [a,*], Jordi Muñoz-Marí [a], Gustau Camps-Valls [a], José Moreno [a]

[a] Image Processing Laboratory (IPL), Parc Científic, Universitat de València, 46980 Paterna, Spain
[b] Secretary of Research and Postgraduate, CONACYT-UAN, 63155 Tepic, Nayarit, Mexico

ARTICLE INFO

*Article history:*
Received 28 February 2017
Received in revised form 21 August 2017
Accepted 25 August 2017

*Keywords:*
Spectral dimensionality reduction methods
Machine learning regression algorithms
Biophysical parameter retrieval
ARTMO
Hyperspectral
Vegetation properties

ABSTRACT

Current and upcoming airborne and spaceborne imaging spectrometers lead to vast hyperspectral data streams. This scenario calls for automated and optimized spectral dimensionality reduction techniques to enable fast and efficient hyperspectral data processing, such as inferring vegetation properties. In preparation of next generation biophysical variable retrieval methods applicable to hyperspectral data, we present the evaluation of 11 dimensionality reduction (DR) methods in combination with advanced machine learning regression algorithms (MLRAs) for statistical variable retrieval. Two unique hyperspectral datasets were analyzed on the predictive power of DR + MLRA methods to retrieve leaf area index (LAI): (1) a simulated PROSAIL reflectance data (2101 bands), and (2) a field dataset from airborne HyMap data (125 bands). For the majority of MLRAs, applying first a DR method leads to superior retrieval accuracies and substantial gains in processing speed as opposed to using all bands into the regression algorithm. This was especially noticeable for the PROSAIL dataset: in the most extreme case, using the classical linear regression (LR), validation results $R^2_{CV}$ ($RMSE_{CV}$) improved from 0.06 (12.23) without a DR method to 0.93 (0.53) when combining it with a best performing DR method (i.e., CCA or OPLS). However, these DR methods no longer excelled when applied to noisy or real sensor data such as HyMap. Then the combination of kernel CCA (KCCA) with LR, or a classical PCA and PLS with a MLRA showed more robust performances ($R^2_{CV}$ of 0.93). Gaussian processes regression (GPR) uncertainty estimates revealed that LAI maps as trained in combination with a DR method can lead to lower uncertainties, as opposed to using all HyMap bands. The obtained results demonstrated that, in general, biophysical variable retrieval from hyperspectral data can largely benefit from dimensionality reduction in both accuracy and computational efficiency.


## 1. Introduction

Spatio-temporally explicit, quantitative retrieval methods for Earth surface are a requirement in a variety of Earth system applications. Optical Earth observing satellites, endowed with a high spectral resolution, enable the retrieval and hence monitoring of continuous bio-geophysical variables (Schaepman et al., 2009). With forthcoming operational imaging spectrometers, such as EnMAP (Guanter et al., 2015), HyspIRI (Roberts et al., 2012), PRISMA (Labate et al., 2009) and ESA's 8th Earth Explorer FLEX mission (Drusch et al., 2016), an unprecedented data stream for land monitoring will soon become available to a diverse user community. These massive data streams will require enhanced processing techniques that are accurate, robust and fast. One of the major challenges with these data streams is the large amount of spectral data that has to be processed.

Over the last few decades, a wide diversity of bio-geophysical retrieval methods have been developed, but only a few of them made it into operational processing chains and many of them are still in its infancy and not fully adapted to hyperspectral data (Verrelst et al., 2015a). Essentially, we may find four main approaches for the inverse problem of estimating biophysical variables from spectra: statistical, i.e. (1) parametric and (2) nonparametric regression; (3) physically-based; and (4) hybrid regression methods. Hybrid methods combine elements of non-parametric regression and physically-based methods. These methods exploit the generic properties of physically-based models combined with the flexibility and computational efficiency of non-parametric, non-linear regression models (Verrelst et al., 2015a). They proved

* Corresponding author.
*E-mail address:* jochem.verrelst@uv.es (J. Verrelst).

http://dx.doi.org/10.1016/j.isprsjprs.2017.08.012

to be particularly successful in operational generation of land products such as leaf area index (LAI). However, current hybrid methods rely exclusively on neural networks (NN), typically trained by a very large amount of simulated data as generated by radiative transfer models (RTMs) (e.g., Baret et al., 2007, 2013; Verger et al., 2008). For instance, when it comes to LAI retrieval then commonly the PROSAIL model (PROSPECT + SAIL) is used to generate training data (Jacquemoud et al., 2009). This approach works fine to multispectral data but becomes challenging when applied to hyperspectral data due to the computational cost in training a NN with many bands.

Beyond NN, various alternative nonparametric methods in the field of machine learning regression algorithms (MLRAs) have been recently introduced, many of them with interesting properties. Especially bagging/boosting of regression trees (RT), random forests (RF) and kernel-based methods such as kernel ridge regression (KRR) have proven to be simpler and faster to train, providing competitive accuracies. Some of these kernel-based MLRAs such as Gaussian processes regression (GPR) even provide associated uncertainties in a Bayesian framework (Verrelst et al., 2012b, 2015b). A drawback of these advanced statistical regression algorithms (including NN) for retrieving biophysical variables, however, is that they also come with a computational cost, especially when large datasets are involved in the training phase, such as when simulated data are used typically in hybrid schemes. Consequently, reduction of the training data space while retaining as much information as possible would enable to alleviate these computational drawbacks.

Reduction of the training dataset can essentially take place in two domains: (1) in the sampling domain, i.e. by selecting only the most informative samples, e.g. through active learning techniques (MacKay, 1992; Tuia et al., 2011; Crawford et al., 2013; Verrelst et al., 2016a), and (2) in the spectral domain, i.e. by making use of feature (band) selection and feature extraction or dimensionality reduction (DR) techniques (Van Der Maaten et al., 2009). While the first type of methods aim to minimize the amount of samples while preserving high accuracies, the second type of methods aim to bypass the so-called "curse of dimensionality" (Hughes phenomenon) (Hughes, 1968) that is commonly observed in hyperspectral data. Adjacent hyperspectral bands carry highly correlated information which may result in redundant data and possible noise and potentially suboptimal performances. In feature (band) selection, the aim is to define a subset of the original bands that maintains the useful information to apply regression with highly correlated and redundant bands excluded from the regression analysis. In parametric regression, this is typically done by systematically calculating all possible two-band combinations in vegetation indices formulations, (e.g., le Maire et al., 2008; Rivera et al., 2014b). More elegant methods exist by making use of band ranking properties provided by regression methods, such as in GPR or random forests, e.g. (Van Wittenberghe et al., 2014; Feilhauer et al., 2015). For instance, (Verrelst et al., 2016b) recently developed an automated sequential band removal procedure to identify most sensitive bands based on GPR band ranking.

Alternatively, in DR methods the original spectral data is transformed in some way that allows the definition of a small set of new features (components) in a lower-dimensional space which contain the vast majority of the original data set's information (Liu and Motoda, 1998; Lee and Verleysen, 2007). As such, there is no need to search for most relevant spectral bands, and thus simplifies the retrieval problem. Especially in data classification a plethora of feature extraction and DR methods are available in the literature (e.g., Arenas-Garcia et al., 2013; Damodaran and Nidamanuri, 2014). Surprisingly less progress in DR methods has been presented when it comes to biophysical variable retrieval (regression). If a DR method at all is applied, then it is by the classical principal component analysis (PCA) (Jolliffe, 1986; Liu et al., 2016). Although PCA has proven its use in a broad diversity of applications, and continues to be the first choice in vegetation properties mapping based on hyperspectral data, situations may occur where PCA is not the best choice and alternatives have to be sought. As an extension of PCA, partial least squares (PLS) introduces some refinements by looking for projections that maximize the covariance and correlations between spectral information and input variables (Wold, 1966). PLS regression (PLSR) became a popular regression method in chemometrics and remote sensing applications (e.g. see Verrelst et al. (2015a) for review), however, the regression part of PLSR and principal component regression (PCR) has always been restricted to multiple linear regression. It remains to be questioned how well PLS combines with more advanced, nonlinear regression methods. Beyond PCA and PLS, only a few DR-regression studies have been presented, including a semi-supervised DR where the data distribution resides on a low-dimensional manifold has been proposed (Uto et al., 2014). But this method was only applied to linear regression. Apart from Laparra et al. (2015) and Arenas-Garcia et al. (2013) where a few alternative DR methods were proposed, the combined use of DR with advanced regression methods for biophysical variable estimation has been largely left unexplored. Nonetheless, there is no doubt DR methods may become prevalent within the context of introducing advanced regression methods into new generation hybrid retrieval processing chains. This especially holds for LAI retrieval; LAI is characterized by a broad sensitive spectral range (e.g. see global sensitivity analysis Verrelst et al. (2015c)) and thus perfectly suited for a DR conversion step.

In this respect, apart from PCA and PLS, in this work we evaluate 9 alternative DR methods into regression, including canonical correlation analysis (CCA), orthonormalized PLS (OPLS) and minimum noise fraction (MNF), as well as their nonlinear extensions derived by means of the theory of reproducing kernel Hilbert spaces. All these methods have been put together into an in-house developed MATLAB library called SIMFEAT (Arenas-Garcia et al., 2013), which has been now included in a free graphical user interface (GUI) retrieval toolbox.

This brings us to the following objectives: (1) to implement multiple DR methods into a software framework that enables semi-automatic development and validation of (hybrid) statistical retrieval strategies, and (2) to evaluate the efficacy of the SIMFEAT DR methods in combination with advanced regression methods in optimizing statistical LAI retrieval from hyperspectral data. Two experiments are presented. First, a hybrid scheme where the regression algorithms are trained by simulated data coming from PROSAIL. Second, an experimental dataset where the regression algorithms are trained by data coming from ESA's SPARC campaign (Barrax, Spain).

In the following, we will explain the implemented DR methods and used regression techniques (Section 2). This is followed by a description of the developed software and experimental setup (Section 3) and a presentation of the results (Section 4). The work closes with a discussion (Section 5) and a conclusion (Section 6).

## 2. Function approximation as a multivariate data analysis problem

The problem of regression and variable retrieval aims to *learn* a function $f(\cdot)$ that, based on input hyperspectral data $x \in \mathcal{X}$ can predict an output target variable or biophysical parameter $y \in \mathcal{Y}$. The problem can be approached *directly* with nonlinear regression methods implementing $f(\cdot)$, e.g. with neural networks, random forests, or kernel machines. Despite its efficiency, this approach leads to hidden representations that are hard to analyze, understand and visualize. Alternatively, one can approach the problem by learning