LEAST DEPENDENT COMPONENT ANALYSIS FOR TRACE GASES RETRIEVAL FROM SATELLITE DATA

Pia Addabbo, Maurizio di Bisceglie, Carmela Galdi

Università degli Studi del Sannio Piazza Roma 21, I-82100, Benevento, Italy

1. INTRODUCTION

Recent research has proved that hyperspectral satellite observations can be successfully used to map atmospheric trace gases throughout the planet and that an appropriate processing of the retrieved information is, in turn, essential for understanding the global environmental changes. Several techniques have been developed for the retrieval of the major atmospheric and pollution constituents; among them, Differential Optical Absorption Spectroscopy (DOAS) is the most widely used approach. Although several atmospheric components (e.g. O_3, BrO, NO_2, SO_2) can be detected with this technique, DOAS can be used only when a clear molecular spectral structure is present in the sensed range of wavelengths.

Essentially, DOAS is based on precise measurements at some specific wavelengths, where absorption peaks and valleys are more evident. The contribution of this work moves from the consideration that a more robust and precise analysis can be carried out if the observed spectral waveform is more thoroughly exploited. The approach we propose is based on Independent Component Analysis (ICA).

ICA is a statistical method for extracting from a set of observed data, a new set of waveforms that are statistically independent from each other. In ICA, the unknown components are assumed completely independent, but in most cases, however, this hypothesis is not true. For example different chemical species do not necessarily have completely independent spectra, because the similarities in their chemical structure generate some dependencies. In such case the unmixing problem can be a hard task and any residual statistical dependence in the recovered sources might either signal a failure of the method, or reflect the fact that the goal of achieving independent spectra was inconsistent.

Least dependent Component Analysis (LCA) relaxes the hypothesis that signal sources can be linearly decomposed into exactly independent sources, and admits the introduction of a more general cost function to incorporate additional information about the sources. When some knowledge about the source waveforms is available, the semi–blind unmixing can be much more robust against artifacts and noise.

The proposed algorithm for trace gases retrieval consists of two fundamental steps:

1) a preprocessing filter for eliminating as much as possible any residual dependencies from the data;

2) the minimization of a cost function that minimizes among–source dependence and incorporates *a priori* additional information.

2. PROBLEM STATEMENT

According to the Beer-Lambert law, the negative logarithm of reflectance spectra can be expressed as a linear combination of unknown components, representing the scattering/absorption cross sections in their concentrations, that is [1]

$$-\log(R(\lambda)) = -\log\left(\frac{\pi I(\lambda)}{\mu_0 E(\lambda)}\right) = l\sum_i n_i \sigma_i(\lambda) \tag{1}$$

where $R(\lambda)$ is the reflectance, $I(\lambda)$ is the Earth radiance, $E(\lambda)$ the solar irradiance, μ_0 the cosine of the solar zenith angle, l the path length [cm], n_i the gas concentrations along the path [mol/cm^3], $\sigma_i(\lambda)$ their scattering/absorption cross-sections [cm^2/mol]. We explicitly note that the scattering/absorption cross sections include multiple Rayleigh scattering, Mie (aerosol) scattering and absorption, and surface albedo that can be filtered by a proper preprocessing.
Thus eq.(1) is a single observation of the model

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \tag{2}$$

where $\mathbf{x} = [x_1(\lambda), x_2(\lambda), ..., x_m(\lambda)]^T$ is the negative logarithm of m IFOV reflectance spectra, $\mathbf{s} = [s_1(\lambda), s_2(\lambda), ..., s_n(\lambda)]$ are n scattering/absorption cross sections, and $\mathbf{A} = l\{n_{ij}\}_{i=1,2,...,n, j=1,2,...,m}$ is the $n \times m$ mixing matrix of concentrations multiplied with the path length.

For simplicity consider the case of a single trace gas to be separated from a background component. We have two main difficulties when we think to the application of model (2).

1. The presence of all sources in any IFOV cannot be always assumed; thus we may have an ill-posed problem because not all observations contain the same number of components. To overcome this difficulty, for each field of view, two observations can be selected defining the second one as a contamination of the first observation with a known trace gas cross section waveform:

$$\begin{aligned} x_1(\lambda) &= -\log(R(\lambda)), \\ x_2(\lambda) &= -\log(R(\lambda)) + c_0\sigma(\lambda), \end{aligned} \tag{3}$$

where c_0 is the contamination factor. For the case at hand a reasonable value for c_0 is $2.69 \cdot 10^{16}$ [mol/cm^2], corresponding to the concentration of 1 DU.

2. In most applications of interest, the spectral sources are not realistically independent. This happens when chemical compounds in a mixture share common or similar structural groups [2]. In addition, scattering from air molecules, aerosols and clouds as well as absorption from the ground often dominate the extinction of sunlight. Extinction from scattering and absorption on the ground usually varies smoothly with wavelength and represents a source of dependence among spectra. An easy way to remove the slower component is to use second derivative of spectroscopic signals as an high-pass filter for pre-processing. The derivative removes slowly varying components and extracts a spectral information which is more independent from source to source.
A simple way to make such filtering is through the finite difference

$$\left.\frac{d^2\mathbf{x}(\lambda)}{d\lambda^2}\right|_{\lambda_i} = \mathbf{x}(\lambda_{i-1}) - 2\mathbf{x}(\lambda_i) + \mathbf{x}(\lambda_{i+1}). \tag{4}$$

Thereby eq.(2) becomes

$$\mathbf{x}'' = \mathbf{A}\mathbf{s}'', \tag{5}$$

denoting with $\mathbf{x}''$ and $\mathbf{s}''$ the second order derivatives of $\mathbf{x}$ and $\mathbf{s}$ respectively. Because the second order derivative doesn't change the mixing matrix, the LDA algorithm can be applied directly on the signal $\mathbf{x}''$.

We implement the LDA in two stages.

a) A transformation $\mathbf{W}$ of $\mathbf{x}''$

$$\mathbf{y} = \mathbf{W}\mathbf{x}''. \tag{6}$$

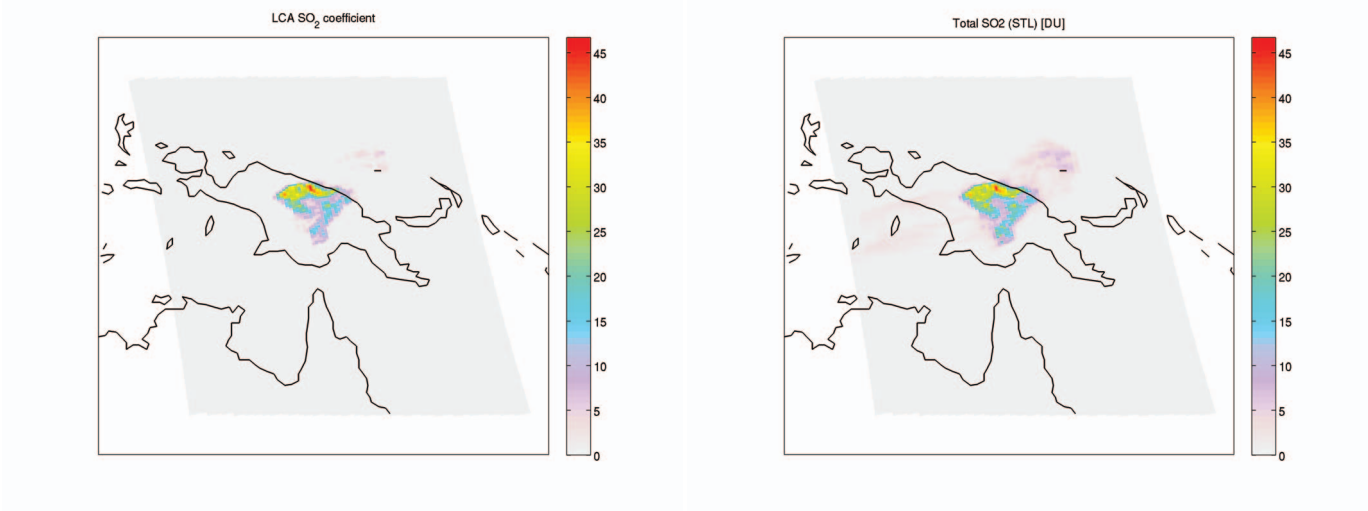

Fig. 1. SO_2 concentration from Manam volcano eruption (January 28, 2005) using: Left) Least Dependent Analysis, Right) OMI team Linear Fit algorithm.

This is a prewhitening and, afterwards, we will restrict the minimization of the cost function to pure coordinates rotations [3, 4]. Thus, the matrix $\mathbf{W}$ is decomposed in two factors as

$$\mathbf{W} = \mathbf{RV}, \tag{7}$$

where the prewhitening matrix $\mathbf{V}$ transforms the covariance matrix $\mathbf{C}$ into an identity matrix $\mathbf{I} = \mathbf{VCV}^T$, and $\mathbf{R}$ is a pure rotation.

b) Some knowledge available on the sources can be treated as an *a priori* constraint in the LDA to guide the separation of the desired components [5]. In our case we know that one desired source is the trace gas absorption cross-section. So a laboratory measured cross-section can be used as reference spectrum in our algorithm. This additional knowledge can be incorporated in a cost function through the mean square error between the retrieved least dependent source and the measured absorption cross-section. Since the data are prewhitened the maximization of statistical independence is obtained simply minimizing the sum of the Shannon entropies of $y_1(\lambda)$ and $y_2(\lambda)$ [3], and the cost function is

$$\alpha \left[\widehat{H}[y_1(\lambda_1), y_1(\lambda_2), ..., y_1(\lambda_p)] + \widehat{H}[y_2(\lambda_1), y_2(\lambda_2), ..., y_2(\lambda_p)] \right] + (1-\alpha)\frac{1}{p}\sum_{i=1}^{p}\left[y_2(\lambda_i) - \sigma(\lambda_i)\right]^2, \tag{8}$$

with $\widehat{H}$ the estimated Shannon entropy, α a weighting costant, and p the number of observed wavelengths.
After the minimization of the cost function, the estimated mixing matrix can be calculated as

$$\mathbf{A}^* = (\mathbf{W}^*)^{-1} = (\mathbf{R}^*\mathbf{V})^{-1}, \tag{9}$$

and the estimated sources can be recovered by applying the demixing transformation $\mathbf{W}^*$ on the original measured mixture signals $\mathbf{x}$

$$\mathbf{s}^* = \mathbf{W}^*\mathbf{x}. \tag{10}$$

Finally the estimated mixing matrix coefficient related to the absorption cross section is proportional to the trace gas concentration.

3. RESULTS

The procedure has been applied to the retrieval of sulphur dioxide SO_2 volcano emission using data from the NASA Ozone Monitoring Instrument (OMI) and the SCIAMACHY preflight model SO_2 absorption cross section [6] as reference spectrum. The test scenario is Manam volcano eruption on January 2005. When the Manam volcano erupted explosively in the night on January 27, 2005, it sent a cloud of ash and sulfur dioxide over New Guinea. About 12 hours after the eruption, OMI flew over on NASA Aura satellite. Figure 1 shows the SO_2 plume produced over Papua New Guinea on January 28, 2005. The image on the left is obtained using the LDA algorithm. On the right it is shown the upper tropospheric and stratospheric SO_2 column in Dobson Units produced by the OMI team. Red pixels represent areas of high SO_2 concentration, while the lowest concentrations are indicated with pink pixels. By comparing the two images, we observe that the LDA algorithm is able to correctly detect the SO_2 plume.
In view of retrievals of different atmospheric component results seem to be promising but refinements are surely necessary in many contexts as the calibration of the concentrations in Dobson units.

4. REFERENCES

[1] *OMI Algorithm Theoretical Basis Document Volume IV*, August 2002.

[2] J. Chen and X. Z. Wang, "A new approach to near-infrared spectral data analysis using independent component," *Journal of Chemical Information and Computer Science*, vol. 41, 2001.

[3] E. G. Learned-Miller and J. W. Fisher, "Ica using spacing estimates of entropy," *Journal of Machine Learning Research*, 2003.

[4] S. A. Astakhov, H. Stogbauer, A. Kraskov, and P. Grassberger, "Spectral mixture decomposition by least dependent component analysis," *Physical Review E*, February 2004.

[5] W. Lu ang J. C. Rajapakse, "Approach and applications of constrained ica," *IEEE Transactions on Neural Networks*, , no. 1, January 2005.

[6] K. Bogumil, J. Orphal, T. Homann, S. Voigt, P. Spietz, O.C. Fleischmann, A. Vogel, M. Hartmann, H. Kromminga, H. Bovensmann, J. Frerick, and J.P. Burrows, "Measurements of molecular absorption spectra with SCIAMACHY preflight model: instrument characterization and reference data for atmospheric remote-sensing in the 230-2380 nm region," *Journal of Photochemistry and Photobiology*, 2003.