Inverse Problems and Imaging
Volume 8, No. 2, 2014, 587–610
doi:10.3934/ipi.2014.8.587

A SEMI-BLIND SOURCE SEPARATION METHOD FOR DIFFERENTIAL OPTICAL ABSORPTION SPECTROSCOPY OF ATMOSPHERIC GAS MIXTURES

Yuanchang Sun
Department of Mathematics and Statistics
Florida International University
Miami, FL 33199, USA

Lisa M. Wingen, Barbara J. Finlayson-Pitts and Jack Xin
Department of Chemistry/Mathematics
University of California at Irvine
Irvine, CA 92697, USA

(Communicated by Hao-Min Zhou)

Abstract. Differential optical absorption spectroscopy (DOAS) is a powerful tool for detecting and quantifying trace gases in atmospheric chemistry [22]. DOAS spectra consist of a linear combination of complex multi-peak multi-scale structures. Most DOAS analysis routines in use today are based on least squares techniques, for example, the approach developed in the 1970s [18, 19, 20, 21] uses polynomial fits to remove a slowly varying background (broad spectral structures in the data), and known reference spectra to retrieve the identity and concentrations of reference gases [23]. An open problem [22] is that fitting residuals for complex atmospheric mixtures often still exhibit structure that indicates the presence of unknown absorbers.

In this work, we develop a novel three step semi-blind source separation method. The first step uses a multi-resolution analysis called empirical mode decomposition (EMD) to remove the slow-varying and fast-varying components in the DOAS spectral data matrix $\mathbf{X}$. This has the advantage of avoiding user bias in fitting the slow varying signal. The second step decomposes the preprocessed data $\hat{\mathbf{X}}$ in the first step into a linear combination of the reference spectra plus a remainder, or $\hat{\mathbf{X}} = \mathbf{A}\,\mathbf{S} + \mathbf{R}$, where columns of matrix $\mathbf{A}$ are known reference spectra, and the matrix $\mathbf{S}$ contains the unknown non-negative coefficients that are proportional to concentration. The second step is realized by a convex minimization problem $\mathbf{S} = \arg\min \operatorname{norm}(\hat{\mathbf{X}} - \mathbf{A}\,\mathbf{S})$, where the norm is a hybrid ℓ_1/ℓ_2 norm (Huber estimator) that helps to maintain the non-negativity of $\mathbf{S}$. Non-negative coefficients are necessary in order for the derived proportional concentrations to make physical sense. The third step performs a blind independent component analysis of the remainder matrix $\mathbf{R}$ to extract remnant gas components. This step demonstrates the ability of the new fitting method to extract orthogonal components without the use of reference spectra.

We illustrate utility of the proposed method in processing a set of DOAS experimental data by a satisfactory blind extraction of an a-priori unknown trace gas (ozone) from the remainder matrix. Numerical results also show that the method can identify trace gases from the residuals.

2010 *Mathematics Subject Classification.* Primary: 65F99, 62J05; Secondary: 92J99.

Key words and phrases. Blind source separation, trace gases, Huber estimator, independent component analysis.

1. **Introduction.** Trace gases play an important role in climate change and air quality of the Earth's atmosphere. Spectroscopic techniques are widely used today for measurements of many trace species, and have evolved over the past century from the first use of the sun as a light source to identify atmospheric trace gases. Many different light sources (e.g., infrared and UV-visible lamps, lasers, and natural sources such as the sun) are now conventionally used to identify light-absorbing species as well as determine their concentrations using Lambert-Beer's law,

$$I(\lambda) = I_0(\lambda) \cdot \exp(-\sigma(\lambda) \cdot \rho \cdot L) \,, \tag{1}$$

where $I_0(\lambda)$ is the initial intensity of light, $I(\lambda)$ is its intensity after traveling through a sample of path length, L, with concentration, ρ. Each species has its characteristic absorption cross section, $\sigma(\lambda)$, a measure of its ability to absorb light that varies with wavelength. The use of (1) is convenient for multi-component samples in laboratory spectrometers, but it is more difficult to determine the value of $I_0(\lambda)$ in the atmosphere over a large wavelength range.

A new method, differential optical absorption spectroscopy (DOAS), was introduced in the 1970s [22, 18, 19, 20, 21] to analyze atmospheric trace gas concentrations. DOAS analysis separates the trace gas absorptions, which typically vary quickly with wavelength, from features that vary slowly with wavelength, e.g., light scattering processes by molecules and aerosols. Differential cross sections are then defined relative to this new broad background in place of the true $I_0(\lambda)$. Several important trace gases were measured for the first time with DOAS, e.g., HONO, NO_3, BrO, ClO in the troposphere, and OClO and BrO in the stratosphere. A large number of other molecules absorb in the UV and the visible wavelength region and most aromatic hydrocarbons can also be detected. An advantage of DOAS is the ability to measure absolute trace gas concentrations *in situ*. DOAS is therefore especially useful for measuring highly reactive species such as the free radicals OH, NO_3, or BrO, and it provides a powerful tool for studying emissions, transformation and transport of chemicals throughout the troposphere and stratosphere. It can also help to understand the influence of atmospheric chemistry on climate and air quality. A detailed description of DOAS can be found elsewhere [22].

In general, DOAS spectra contain overlapping absorption structures which consist of complex multiple scales and peaks. They must be separated by the analysis routine to retrieve the concentrations of the trace gases. Least squares techniques are most often used for analysis of DOAS spectra, with the use of high pass filters to fit or separate out the slowly varying components. For example, the approach described in [22] applies a polynomial fit to remove the broad (slow-varying) spectral features, and known reference spectra to retrieve the concentrations of reference gases. However, the existing DOAS approaches have two limitations: 1) the condition of least squares (*that errors are normally distributed*) is often violated. This suggests that a different norm other than ℓ_2 (least squares) norm should be used; 2) the fitting residuals for atmospheric samples are in most cases not pure noise due to imperfect references, atmospheric turbulence, instrument effects, and unknown trace gases. Among other interesting problems, the identification of gas structures in the fitting residuals is of great importance. The method in this paper has been developed to address these issues, in the hope of providing a tool for atmospheric chemists to analyze the residuals for possible hidden trace gases. The method is designed to deal with the following three major challenges. First, DOAS spectra are complex multi-scale multi-peak structural data containing slow-varying features, structured signals due to the trace gases, and instrumental noise. Hence a multi-resolution analysis tool is needed for scale decomposition. Second, the identification of gases from the residuals is actually a problem of blind source separation (BSS) as both the trace gases (including their numbers) and mixing process are not known. A major problem is to find a working assumption on the source (hidden trace gas) signals and effective BSS algorithms. Third, the new objective function for data fitting should not only overcome the limitations of least squares fitting, but also help to maintain the non-negativity of mixing coefficients so that the trace gas concentrations derived from them are positive.

To tackle these problems, we have made an initial attempt of developing a semi-blind approach which contains three steps. The first step uses multi-resolution analysis to remove the very slow (e.g. scattering) and very fast components (noise) in the DOAS spectral data matrix $\mathbf{X}$. The second step decomposes the preprocessed data $\hat{\mathbf{X}}$ in the first step into a linear combination of the reference spectra plus a remainder, or $\hat{\mathbf{X}} = \mathbf{A}\,\mathbf{S}+\mathbf{R}$, where columns of matrix $\mathbf{A}$ are known reference spectra, and the matrix $\mathbf{S}$ contains the unknown non-negative coefficients. The second step is carried out by solving a convex minimization problem $\mathbf{S} = \arg\min \mathrm{norm}\,(\hat{\mathbf{X}} - \mathbf{A}\,\mathbf{S})$, where the norm is a hybrid ℓ_1/ℓ_2 norm that helps to maintain the non-negativity of $\mathbf{S}$. The third step performs a blind independent component analysis of the remainder matrix $\mathbf{R}$ to extract remnant gas components. The advantages of the new method are multifold: (1) separating the slow components is data-driven and avoids user input, (2) fitting the mixture using reference spectra results in non-negative coefficients that are required for trace gas concentrations, and (3) analyzing the mixture is also possible without the use of reference spectra to identify components in the residuals or in the original mixture. The ability to identify components without using their absorption cross sections as reference inputs offers remarkable improvements to the analysis of complex atmospheric mixtures.

The paper is organized as follows. In section 2, we review the essentials of DOAS and the existing approach, then introduce our method. In section 3, we illustrate the proposed method in processing a set of DOAS experimental data, and show satisfactory numerical results. Concluding remarks are in section 4.

2. DOAS and Signal processing methods.

2.1. DOAS and fitting methods.

A typical experimental setup for a DOAS instrument consists of a continuous light source, e.g., a Xe-arc lamp, a light-absorbing sample (the atmosphere or gases in a chamber), a grating spectrometer, and an optical detector. It is also possible to use the light from the sun or moon, or scattered sun light as light sources [22, 18, 19]. The typical length of the light path in the atmosphere ranges from several hundred meters to many kilometers, and < 100 m in laboratory DOAS experiments. The light of intensity $I_0(\lambda)$ passes through the sample, is typically dispersed by a grating spectrometer and is measured by a detector. During its way through the sample the light undergoes extinction due to absorption processes by trace gases and scattering by air molecules and aerosol particles. The intensity, $I(\lambda)$, at the end of the light path is given by the Lambert-Beer law,

$$(2)\quad I(\lambda) = I_0(\lambda)\exp\left[-\int_0^L \sum_j \sigma_j^{\mathrm{ABS}}(\lambda) \times \rho_j(l) + \varepsilon_{\mathbf{R}}(\lambda, l) + \varepsilon_{\mathrm{M}}(\lambda, l)\,\mathrm{d}l\right] + N(\lambda)\ ,$$

where σ_j^{ABS} is the absorption cross section of a trace gas j, ρ_j is its number density, L is the length of the light path, and $N(\lambda)$ is the measurement noise. The Rayleigh extinction by gases and Mie extinction by aerosols are described by $\varepsilon_{\mathbf{R}}$ and ε_{M} , but are not the parameters of interest here. The basic idea of DOAS is the separation of the cross section $\sigma_j^{\mathrm{ABS}} = \sigma_j^{\mathrm{B}} + \sigma_j'$ in which σ_j^{B} represents broad spectral features and the differential cross section σ_j' represents narrow spectral structures that are of interest for identification and quantification of the trace gases. If one considers only σ_j' as is done with DOAS, interferences with Rayleigh and Mie extinction are avoided. In logarithm form, the above equation can be simplified to having following terms

$$\mathbf{X} = \mathbf{A}\,\mathbf{S} + \mathbf{B} + \mathbf{N}\,, \tag{3}$$

where the details can be found in the Appendix. Each column of $\mathbf{X}$ is a mixed signal of several trace gases; the columns of matrix $\mathbf{A}$ correspond to the reference spectra of the known trace gases; the matrix $\mathbf{S}$ contains non-negative coefficients, which are proportional to the product of each trace gas concentration and path length; the matrix $\mathbf{B}$ includes the slow-varying components, and the matrix $\mathbf{N}$ contains the noise components. A more thorough mathematical description of the DOAS approach is given in the Appendix.

If minimal knowledge about the DOAS mixtures is available, for example, in the situation that no spectral references are known, we face a so called blind source separation problem where we need to separate the mixtures into a list of source signals without knowing how they are mixed. A widely used and powerful tool called multivariate curve resolution (MCR) [14], often used in the chemometrics community, could be used to solve for the spectra of the individual trace gases and their concentrations. For the readers' convenience, we provide some details of MCR in section 3.2. The MCR method solves a non-convex minimization whose solution is often a local minimum and may converge to a different one for a different initial guess. This limitation prevents MCR from being used in DOAS data analysis where some partial knowledge of the mixtures is actually available.

When we have full knowledge about what trace gases are in the mixtures, the conventional least squares (CLS) methods are often used to calculate $\mathbf{S}$ due to its computational simplicity. In the simplest case, linear least squares, reference spectra are used in a linear combination to form a model spectrum that most closely resembles the narrow features of the mixture spectrum. To deal with the problem that the data contain both broad and narrow spectral features, a high pass filter is needed to remove the broad spectral features. It is common to use polynomials to model and filter out the slowly varying parts from the narrow trace gas absorption. Given the order of polynomial and the known reference spectra, (3) can be solved with a least squares method. The polynomial fitting however has the following drawbacks: (1) the order of the polynomial is determined empirically and different orders might be used for different data; (2) the non-negativity of the concentration is not guaranteed during the fitting. An additional open problem after the fitting is how to identify and extract trace gases from the fitting residuals besides the noise. In particular, DOAS spectra of complex atmospheric samples typically have structured residuals after fitting, suggesting that additional absorbers are present. We propose a three step method in the next section that addresses each of these issues.

2.2. **Proposed semi-blind source separation method.** DOAS data can cover a range of scales and contain high frequency (<1 nm) artifact structures, for example due to pixel-to-pixel variability in the detector, while the reference spectra of the trace gases contain fewer peaks and peak widths on the order of several nm. Hence it is helpful to remove the fast varying artifacts from the spectra data by multi-resolution analysis. In addition, the broad features (slow-varying parts) in the data need to be eliminated in order to fit the reference spectra of the known trace gases. There are many methods available for high and low pass filtering of data and some that have been used for analysis of DOAS data are discussed in Platt and Stutz [22]. The filter ultimately chosen will preferably have little need for tuning and low influence by the user. Therefore, we chose to use empirical mode decomposition (EMD) to extract these components.

2.2.1. *Multi-resolution analysis: EMD.* EMD is a data analysis tool which decomposes a data set into a finite number of components called Intrinsic Mode Functions (IMF) via a so called sifting process. Every IMF satisfies two conditions: 1) In an IMF, the number of the extrema and the number of zero crossings must be equal or differ at most by one; 2) At any point of an IMF, the mean value of envelopes defined by those extrema are equal to zero. Several sifting steps might be required to produce an IMF, and the number of steps are determined by a stop criterion. Once an IMF is extracted, it will be subtracted from the signal, and the sifting process will then be applied to the new signal. The sifting process stops finally when the residue becomes a monotonic function. The residual of an EMD is also called the trend of the data. The detailed description of EMD can be found in [11]. EMD is fully adaptive to data even when it is non-stationary and contains nonlinear structures. Although EMD has been widely used as a successful tool for data analysis in science and engineering areas, it is an empirical approach defying rigorous analysis. Recent alternative mathematical methods in the spirit of EMD can be found in iterative filtering decomposition [16, 26], synchrosqueezing wavelet transformation [7, 24]. In the case of the DOAS data decomposition discussed here, EMD [11] works effectively.

The EMD method decomposes the DOAS data into a finite number of components of different frequencies. The advantage of EMD is that it is completely data-driven (no need to specify a parameter such as the order of a fitting polynomial as is normally done in DOAS analysis), fast and automatic. For DOAS data, the first IMF (the one with the highest frequency) will be taken as the noise and removed. As discussed above, the slowly-varying spectral features in DOAS data are not useful for trace gas quantification and therefore also need to be removed from the signal. Therefore, we need to extract the trend and subtract it from the data. In EMD decomposition, the residual after all IMFs are extracted is defined as the trend of the signal. Fig. 1 shows a typical DOAS spectrum of a trace gas mixture containing HONO and NO_2, whose reference spectra are shown in the middle and bottom panels. Fig. 2 shows results using EMD to extract the trend and, for comparison, use of the synchrosqueezing wavelet transform on the same original DOAS spectrum. The middle curve in each case is the residual, which should be slowly varying. This is the case for EMD, while for synchrosqueezing there are features from the trace gas, HONO, present in the residual. Removal of this non-ideal residual from the original spectrum would have the effect of producing a falsely low concentration for this trace gas upon decomposition of components in the next analysis step.

Additional testing of the performance of EMD was carried out in the presence of noise and perturbations in the data. White Gaussian noise of SNR = 60 dB is added to the DOAS signal and several spikes have been introduced to mimic damaged channels of the detector due to aging. Fig. 3 shows results using EMD to extract the trend from this modified DOAS signal. The decomposition residual does not deviate much from that of the un-modified signal when noise and spikes are introduced.

Fig. 4 shows the successful EMD decomposition of the original DOAS spectrum into its different frequencies. The resulting spectrum after removing fast (noise) and slowly varying trend (background) is now ready for trace gas decomposition. The EMD preprocessed (high-passed) data $\hat{\mathbf{X}}$ satisfies the following model

$$\hat{\mathbf{X}} = \mathbf{A}\,\mathbf{S} + \mathbf{R}, \tag{4}$$

where the columns of matrix $\mathbf{A}$ are the reference spectra of the known trace gases, and those of $\mathbf{S}$ matrix contains their concentrations, and $\mathbf{R}$ is the fitting residual which might contain the instrument noise, hidden trace gas structures, etc. For the estimation of the concentration matrix $\mathbf{S}$, we minimize the following constrained objective function:

$$\min_{\mathbf{S}} \operatorname{norm}(\hat{\mathbf{X}} - \mathbf{A}\,\mathbf{S}), \quad \text{s.t.} \;\; \mathbf{S}_{ij} \geq 0, \tag{5}$$

for a proper choice of the norm.

2.2.2. *Huber estimator and robust data fitting.* There are many kinds of norms available, e.g., ℓ_2 (least squares), ℓ_1 (least absolute deviations). The regular least squares method (ignoring the non-negative constraint on $\mathbf{S}$) is the conventional choice, if the unknown noise N is assumed to be Gaussian. However, it is very sensitive to the outliers in the data, even one outlier can destroy the estimation results. The least absolute deviations (ℓ_1 norm) is more robust to outliers in the data, however it is less effective if the peaks in N are not isolated (or sparse). We find that a hybrid ℓ_2 and ℓ_1 norm (ℓ_2 on small peaks and ℓ_1 on large peaks), or a Huber estimator [12], is able to resist the influence of outliers, and maintain non-negativity of $\mathbf{S}$ for our data fitting task. The Huber norm is both regular and convex. The corresponding non-linear function $H = H(x)$ is a parabola (ℓ_2) in the vicinity of zero, and increases linearly (ℓ_1) at $|x| > k$ for any positive constant k. More precisely,

$$H(x) = \begin{cases} \frac{1}{2}x^2, & |x| \leq k \\ k|x| - \frac{1}{2}k^2, & |x| \geq k. \end{cases} \tag{6}$$

k is called a tuning constant; smaller values of k produce more resistance to outliers, but at the expense of lower efficiency when the errors are normally distributed. The non-negativity of $\mathbf{S}$ under the Huber norm indicates that the choice fits the empirical distribution of the noise N arising from detection and photon statistics [22]. Fig. 5 is an example showing the superiority of Huber's estimator over conventional least squares: it is resistant to outliers in the data, while the least squares result deviated significantly from the exact line due to the two outliers. The least squares commonly assigns equal weighting to each observation; the weights for the Huber estimator decline when $|x| > k$ (see the weight functions in Fig. 5). The tuning constant is generally selected to yield high efficiency in the normal case; in particular, $k = 1.345\sigma$ (where σ is the standard deviation of the errors). The minimization of Huber's objective can be achieved by the method of iterative re-weighted least squares. The error is assumed to have a standard normal distribution ($\sigma = 1$). The

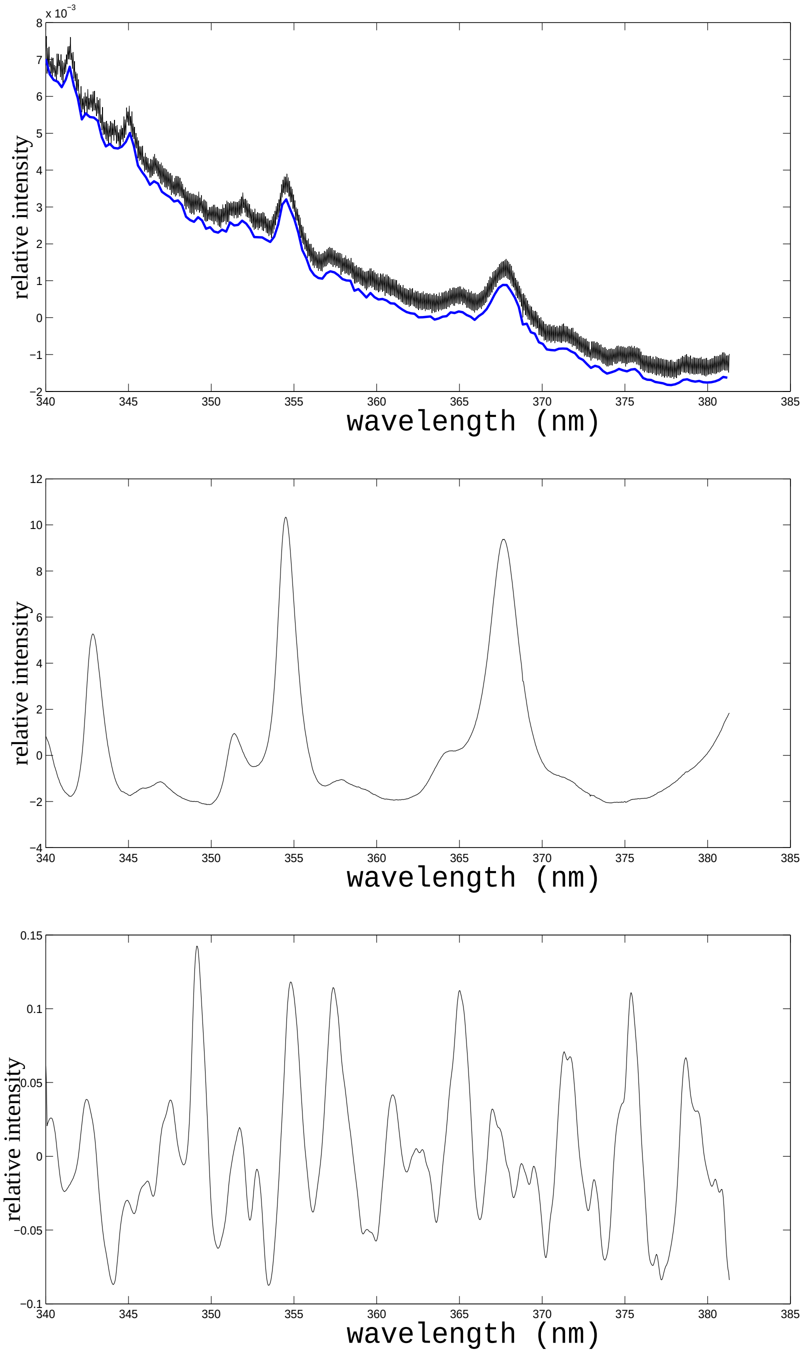

Figure 1. Top panel is a mixed DOAS spectrum (black) from the experiment and its smoothed counterpart (blue). Middle and bottom panels are the spectral absorption references of trace gas HONO, and NO_2, respectively.

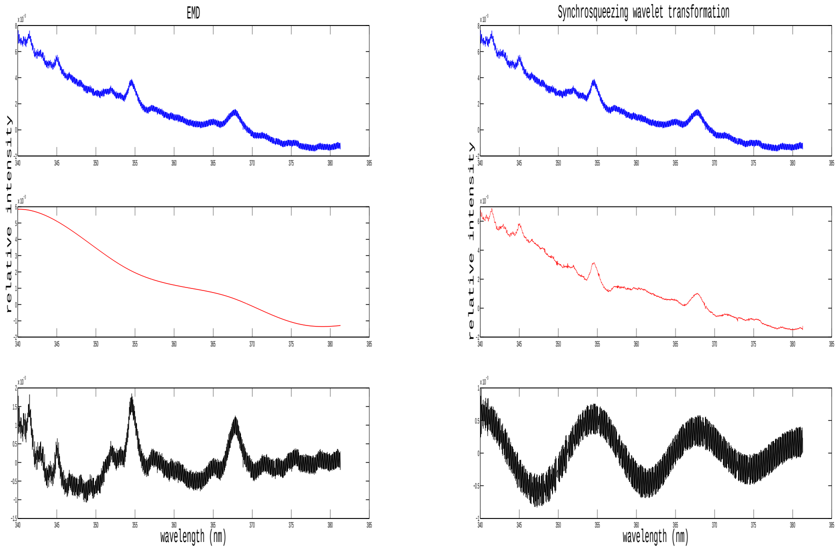

Figure 2. **Left panel**: results of EMD; first row is the signal, the middle is the trend, and bottom row is the resulting signal after removing the trend. **Right panel**: the corresponding results of synchrosqueezing wavelet transformation.

95% asymptotic efficiency on the standard normal distribution is obtained with the tuning constant $k = 1.345$, which provides a good starting point for a newcomer. Although $k = 1.345$ works nicely with the DOAS data we tested, for other data sets difficulties might arise due to its discontinuous second derivative. In those cases, a tuning constant with different value might be employed, and the non-negativity constraint on the concentration matrix $\mathbf{S}$ might serve as a selection criterion for k. In the numerical tests, we have varied k in a range of $(0.5, 1.5)$ to find that similar nonnegative results for $\mathbf{S}$ are obtained without explicit enforcement of non-negativity constraint, which suggests that k could take any value from that range with reasonably good results.

It should be noted there are other non-least squares techniques for robust data fitting called M-estimators in statistics, which are resistant to outliers. For example, if the minimized function is chosen to be $0.5c^2 \log(1+x^2/c^2)$, with a preset constant c, we have the Cauchy estimator. A limitation of Cauchy estimator is that a unique solution is not always guaranteed, although this does not occur with DOAS data we have tested. In fact, in numerical tests we found that Cauchy estimator can deliver similar results to Huber's when c is chosen from $(0.6, 2.05)$, while a recommended value $c = 2.3849$ for most data sets produces negative values in the concentration coefficients. Fig. 7 and 8 provide comparisons of the fitting results using the Huber estimator and the Cauchy estimator to the CLS method. Both estimators provide similar, favorable results. It is difficult to select a function for general use without testing and parameter selection based on many data sets. However, we have chosen the Huber estimator for its resistance to outliers, its capability for fitting both large and small peaks, and its tested and proven ability with typical DOAS data here to produce unique solutions.

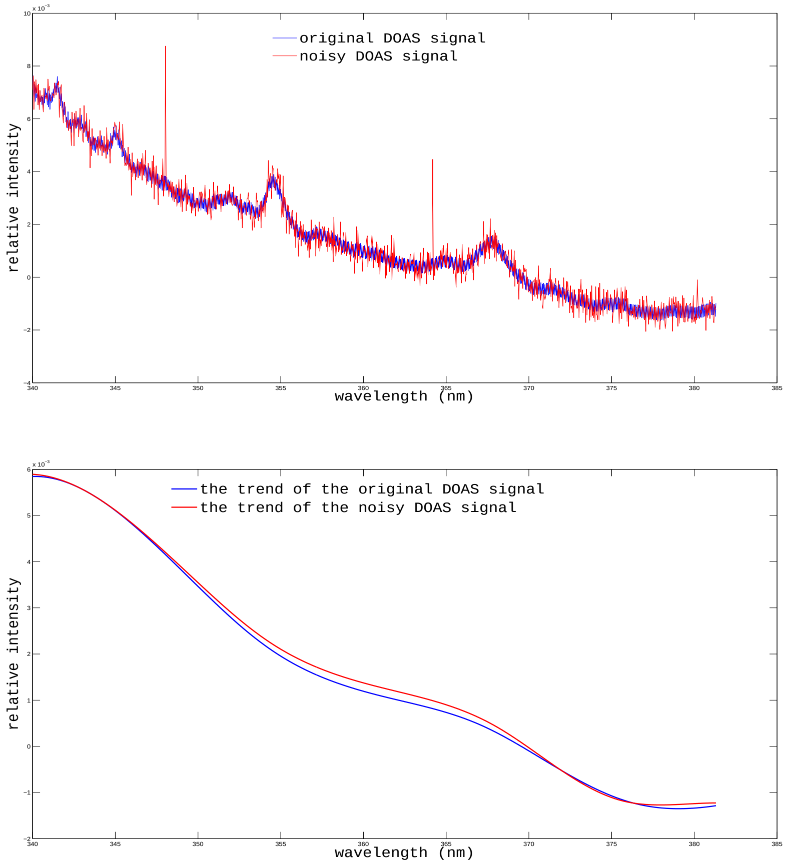

Figure 3. Top: A DOAS signal with noise being added and spikes being introduced; Bottom: the residuals of EMD of the clean signal (blue) and noisy signal (red).

In the residual of Huber estimation, there might be spectral structures (one or many) of trace gases buried in noise, or just random noise. In either case, we decompose the residuals in a blind fashion due to the lack of the knowledge of the hidden trace gases. The source signal assumption required for the decomposition is that the spectra of different trace gases are statistically independent (orthogonal). This appears to be a reasonable working assumption for many trace gases. Independent component analysis (ICA) can now be readily applied.

2.2.3. *Independent component analysis.* ICA is a useful and generic tool for solving blind source separation problems (BSS), which arise when one attempts to recover source signals from their mixtures without knowing the mixing process [5, 6]. ICA finds the independent components in the mixtures by maximizing the statistical independence (minimizing mutual information) of the estimated components. Mathematically, given the mixture matrix $\mathbf{R} \in \mathbb{R}^{p\times n}$ and the number of independent source components d, ICA finds a full rank matrix $\mathbf{W} \in \mathbb{R}^{d\times n}$ such that the output

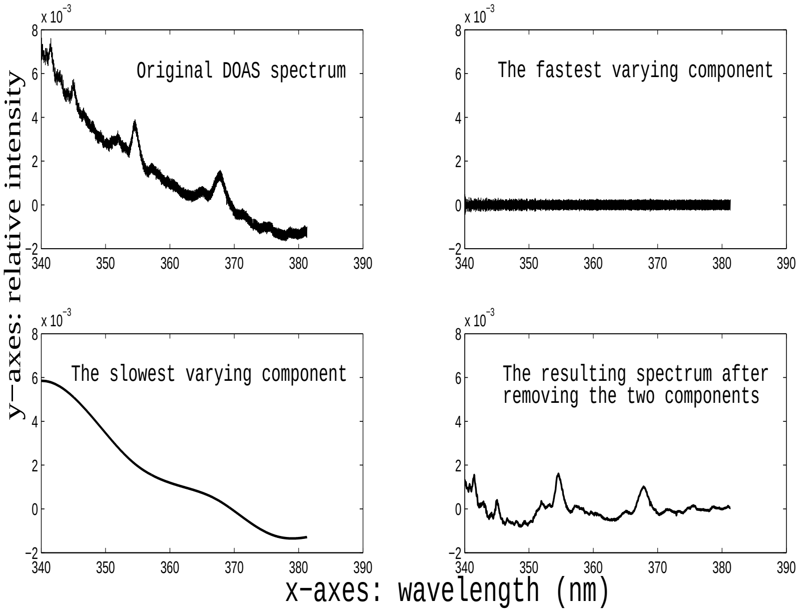

Figure 4. The preprocessed DOAS data from Fig. 1 after removing the trend (broad feature) and the fastest varying components.

matrix $\mathbf{U} \in \mathbb{R}^{d\times p}$ given by

$$\mathbf{U} = \mathbf{W}\,\mathbf{R}' \tag{7}$$

contains columns (recovered source signals) as independent from each other as possible. Here n is the number of residuals from the data fitting, p is the number of wavelength pixels. The columns of $\mathbf{U}$ correspond to the decomposed independent source signals. We may choose one of many ways to approximate independence, and this choice governs the form of the ICA algorithm. The two broadest definitions of independence for ICA are: (1) minimization of mutual information; (2) maximization of non-Gaussianity. The non-Gaussianity family of ICA algorithms use kurtosis and negentropy. The minimization of mutual information family of ICA algorithms use the Kullback-Leibler divergence and maximum-entropy, however, the knowledge of source signal probability distribution function (PDF) is needed. Algorithms for ICA include infomax [1], FastICA [13], and JADE [4]. We opt for JADE because JADE is based on cumulants (2nd and 4th order statistics) and the approximate joint diagonalization of cumulant matrices (hence does not rely on PDF information of source signals). For moderate number of sources, it is more direct and stable than iterative methods such as infomax [1] and FastICA [13]. It was recently found [17] that the infomax method [1] may even diverge and that it only converges in a weak sense under proper rescaling and soft dynamic control of the iterations. The most attractive aspect of JADE is that it does not require parameter tuning (e.g. choosing the learning parameter in the iterative methods). In general, ICA algorithms cannot identify the actual number of source signals, so this number needs to be found by other means, for example by human evaluation of the end results. In our decomposition of Huber residuals, we tested a range for this number, and

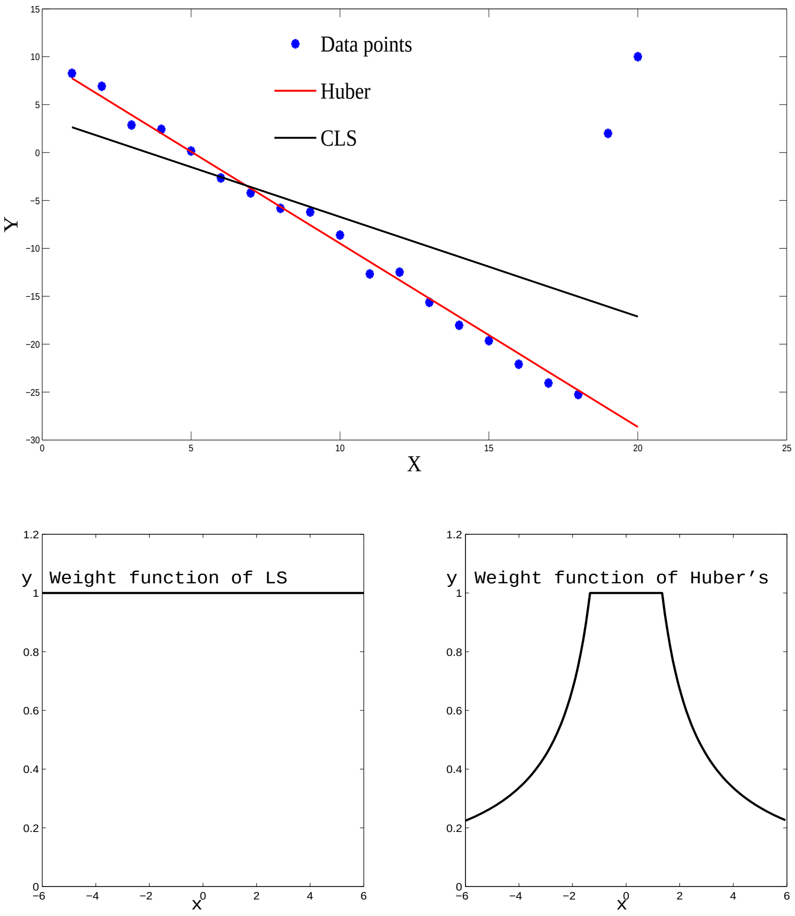

Figure 5. Comparison of Huber's estimator and CLS (their weight functions are shown in the bottom plots). The data points are generated by $y = -2x + 10$ plus noise. Huber: $y = -1.9794x + 9.9318$; CLS: $y = -1.0504x + 3.5819$.

pinpointed the one with the most reliable and meaningful outcomes when calibrated with the knowledge of the existing trace gas spectral properties.

3. Experiments and computational results.

3.1. Experimental setup.

Spectra of chemical mixtures were collected using an environmental chamber [8] for which DOAS is one of the analytical techniques used to measure species during experiments. Fig. 6 shows a simplified schematic of the chamber and optical arrangement for DOAS. The chamber is 561 L in volume and can be evacuated to a pressure of $\sim 10^{-2}$ Torr for collection of true $I_0(\lambda)$ spectra. Spectra can also be collected before and after addition of ultrapure air and each gaseous analyte of interest through various ports.

The DOAS instrumentation consists of a high pressure Xe arc lamp (Oriel, Model 6263) as the UV-visible light source. The light beam enters the chamber through a quartz window and undergoes multiple reflections using White cell mirrors [27] through the gas mixture in the chamber. The multiple reflections increase the path length of the light beam through the sample to a total path length of $L = 52$ m. The light beam exits the chamber through the quartz window and is focused on the entrance slit of a monochromator (Jobin Yvon-Spex, Model HR460) with a diode array detector (Princeton Instruments, model PDA-1024 ST121). The grating (1200 grooves mm^{-1} blazed at 330 nm) gives a dispersion of ~ 0.043 nm /pixel and the detector has 1024 channels giving each spectrum a total wavelength range of ~ 44 nm. Spectra can be collected in different wavelength ranges by moving the grating motor. Changes in grating position as well as temperature lead to changes in dispersion of the light beam on the detector. This is taken into account in the least squares analysis by allowing for shifting or linear compression/expansion in one or more reference spectra along the wavelength axis to obtain the best fit. The use of such techniques is standard and user controlled to correlate wavelengths with channels of the detector. Absolute dispersion and wavelength were calibrated using a mercury lamp spectrum that was recorded daily at the beginning of each experiment.

The analytes added to the chamber were NO_2 and O_3 at a total pressure of ~ 1 atm at room temperature in dry ultrapure air (Scott-Marrin, Riverside, CA). The wavelength range typically used to measure NO_2 by DOAS is 340 - 380 nm. Although the air was dry (relative humidity $< 0.8\%$), even small amounts of water react with NO_2 to form HONO [9]. As a result HONO is almost always present in detectable quantities with NO_2. HONO is also typically measured using the 340 - 380 nm wavelength range, thus the mixture of HONO and NO_2 was used as a convenient test case for the new DOAS analysis technique. The addition of O_3 leads to formation of NO_3 radicals ($NO_2 + O_3 \rightarrow NO_2 + O_2$). Analysis for O_3 is typically carried out in a different wavelength range, 290 - 330 nm. It should be noted that the wavelength range chosen for analysis is usually that in which the cross sections are highest for that analyte in order to optimize the detection limits. O_3 continues to absorb at wavelengths > 330 nm, albeit with absorption cross sections that are lower by a factor of 100 or more [25] compared to those at shorter wavelengths. Another test of the technique was to determine if it could identify the presence of this third component, O_3, in the 340 - 380 nm range where its detection is not optimal. NO_3 analysis was carried out in a different range, 600 - 640 and 640 - 680 nm, and is not discussed here.

In addition to the new DOAS analysis technique introduced in this work, the typical linear least squares analysis was carried out on HONO and NO_2 using MFC [10] for which reference spectra are needed. A reference spectrum for NO_2 was generated by adding a known quantity of NO_2 to the chamber and collecting DOAS spectra with the instrumentation described above. Pure samples of HONO are difficult to generate without the presence of NO_2, thus HONO reference spectra were generated from published cross sections [3, 2] which were convoluted to the dispersion and resolution of our spectrometer. Reference spectra for O_3 were generated from published cross sections [25] and also converted to the dispersion and resolution of the spectrometer.

Chemicals used in these experiments are as follows: Gaseous NO_2 was synthesized by reaction of gaseous NO (Matheson, 99%) which was first passed through a cold

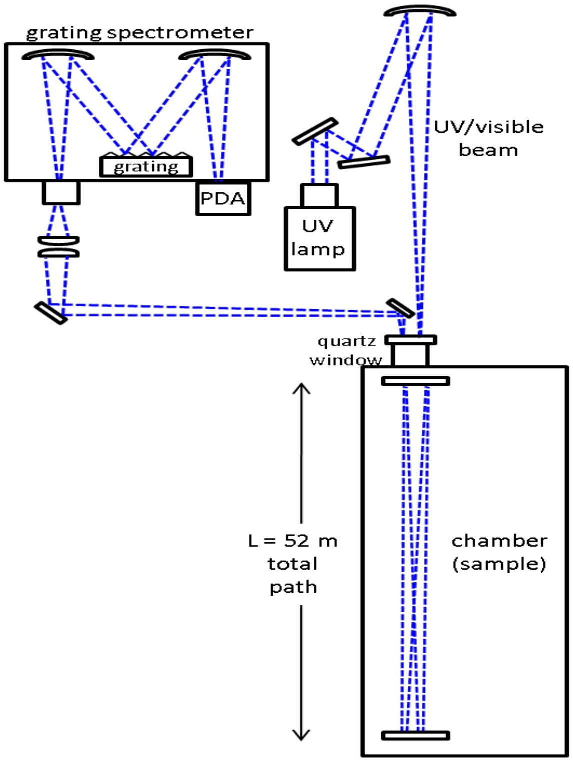

Figure 6. Schematic of chamber, instrumentation, and optical setup used to make DOAS measurements of gaseous mixtures.

trap at 195 K to remove impurities such as HNO_3, with an excess of O_2 (Oxygen Services Co., 99.993%). The mixture was allowed to react for 2 hrs. and then purified by condensing the NO_2 at 195 K to pump away excess O_2. Gaseous O_3 was generated as a mixture in O_2 using a commercial ozonizer (Polymetrics, Model T-816).

3.2. **Computational results.**

We report here the computational results for the proposed method. In the first example, we fit the known reference spectra of NO_2 and HONO to the DOAS data. The results are shown in a series of plots, Fig. 4, Fig. 7, and Fig. 9. We use 11 sets of data corresponding to different reaction times and hence different gas concentrations ($\mathbf{X}$ has 11 columns) from the experiment. Fig. 4 illustrates the data preprocessing (EMD) described earlier which removes the fastest and slowest varying components. The Huber fitting results are presented in Fig. 7 which shows the coefficients of NO_2 and HONO in the 11 mixtures in comparison with the coefficients and concentrations determined using the CLS fitting technique and their two standard deviation errors ($2s$) calculated based on the difference between the real mixture and the model spectrum generated by the fit. The coefficients determined using the hybrid ℓ_1/ℓ_2 fitting technique are all non-negative as well as in very good quantitative agreement with values from CLS fitting. The fitting residual is in the third plot of Fig. 7.

Though some structure can be seen in the residuals, it is not clear if there are other spectral structures embedded in the fitting residuals. Then further identification was done by JADE. For the 11 residuals, we vary the number of independent components in the JADE computation. We observed that the structure of the first plot in Fig. 9 remains approximately invariant as the number varies. This invariance implies that it should be a hidden trace gas in the fitting residual. It can be seen that the identified structure resembles O_3 in several peak locations as indicated in Fig. 9, especially the region 340–355 nm. This further analysis of the residuals is currently not possible with CLS methods used today for DOAS analysis. The identification step by JADE is a major advantage of the new method and can provide extremely useful chemical information in the analysis of DOAS spectra. It should be noted that the ℓ_1/ℓ_2 fitting technique currently does not incorporate shifting and squeezing of spectra to optimize fitting, but this can be implemented in the future. Spectral shifts and squeezes are often used in DOAS analysis routines to account for changes in grating dispersion due to temperature fluctuations and grating positioning accuracy [22, 23]. The example shows that the new method can identify an *a-priori* unknown trace gas, ozone (O_3), from the fitting residuals, which offers a powerful improvement to DOAS analysis.

The second example uses the same set of data (11 mixtures), however only the reference spectrum of NO_2 is used to fit the data. Ideally, we should recover O_3 and HONO from the residuals. The two identified hidden spectral signals are in Fig. 10 and Fig. 11. The recovered fits are recognizable as HONO and O_3 upon comparison with reference spectra, demonstrating the ability of the technique to identify absorption features *without* the use of reference spectra during the fitting procedure. While the ℓ_1/ℓ_2 technique is demonstrated here for laboratory DOAS data with three components, its utility lies in the analysis of atmospheric DOAS spectra, which are more complex. The least squares method works best when reference spectra for all known absorbing species are used to carry out the fitting, i.e., when the fit residuals are unstructured and do not vary considerably with wavelength. Given that this condition is rarely satisfied for complex atmospheric measurements, the method described here is complementary in that it can identify species that are either not known to be present or do not yet have available published cross sections. In addition, the results in Fig. 7 show its value as an alternative standalone technique for analyzing DOAS spectra with the use of appropriate reference spectra.

To see how close the recovered structures are to their spectral references in Fig. 10 and 11, we have computed the angles between them. For Fig. 7 (HONO), the cosine value of the angle is 0.0633; while the value for Fig. 8 (O_3) is 0.6653. While these values suggest that the recovered structures are not similar to their spectral references, these measures often lack efficiency as they include many insignificant features and regions such as oscillatory noisy peaks. For example, the absorption cross section of ozone is almost flat at the incident wavelengths $\geq$ 353 nm (See O_3 spectrum in Fig. 11), hence focus should be put on the region of [340,353] nm for the computation. Likewise, the spectral reference of HONO contains just a few of major and significant peaks, hence more weight should be added to pixels around those peaks. To better measure the similarity, we shall differentiate weights through the regions, i.e., we shall use a weighted-distance (angle) similarity measure. More weight will be placed around the significant features (e.g., major peaks in the spectral references), while small values should be assigned to other regions. To

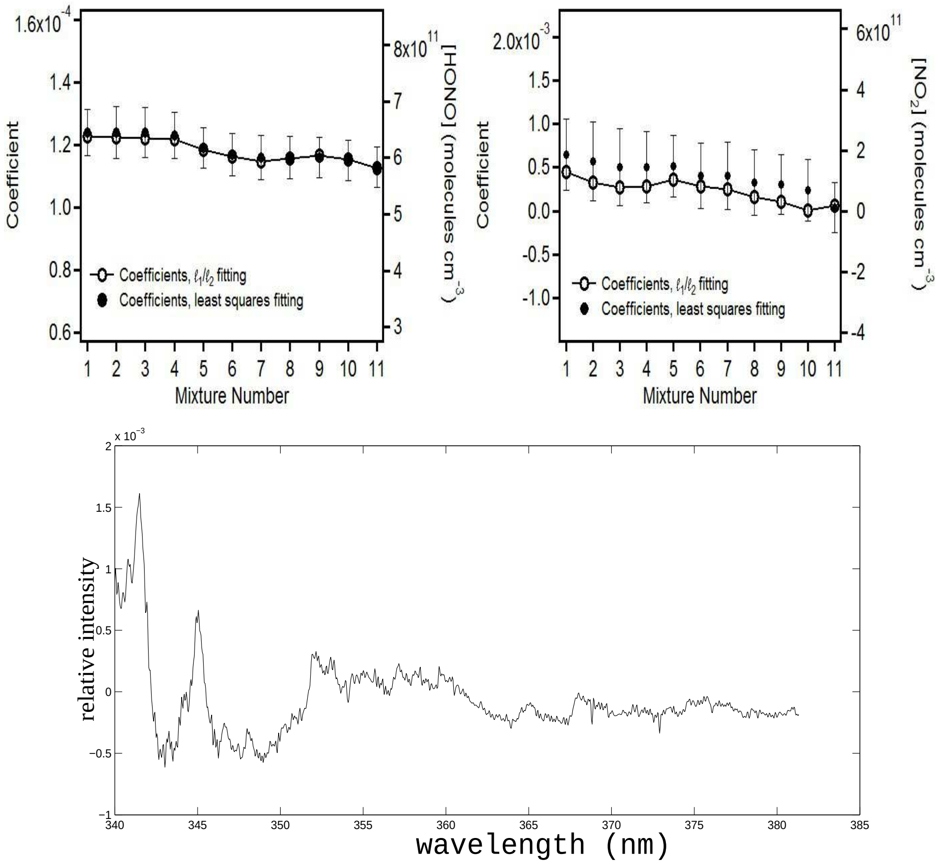

Figure 7. (Top row) comparison of the Huber estimator fitting and CLS techniques; HONO coefficients with 2s errors (left), NO_2 coefficients with 2s errors (right), showing good quantitative agreement for eleven mixture spectra collected sequentially over 11 minutes. Corresponding concentrations in molecules cm^{-3} are provided on the right axis for each plot. (Bottom row) one fitting residual from robust data fitting.

match the recovered spectral structure with the HONO reference, we assign large weights (0.95 is used in the paper) on the peak regions (denoted by arrows in the plot). The feature peaks can be located by either a visual inspection or a peak finder method[1]. The weight on other parts is 0.05. For ozone in Fig. 8, it is more reasonable to consider the spectra of 340-353 nm which includes most of the significant peaks. Using the weighted distance similarity measures, the cosine values are now 0.5815 for HONO and 0.7473 for O_3. Given the fact that shifts and squeezing (or other distortion) occurs in the absorption structures of HONO and O_3 during the experiments, these numbers actually show a fairly good agreement to

[1] for example, we can detect peaks by looking for downward zero-crossings in the first derivative that exceeds a given slope threshold.

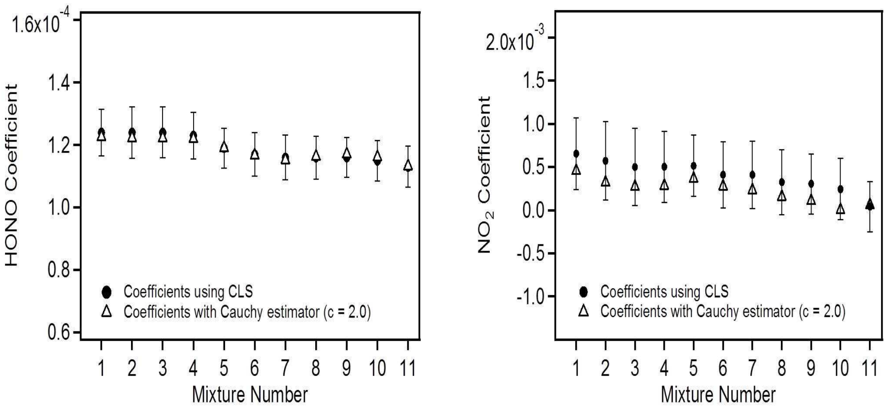

Figure 8. Comparison of the Cauchy estimator fitting and CLS techniques on the HONO coefficients with 2s errors (left) and NO_2 coefficients with 2s errors (right).

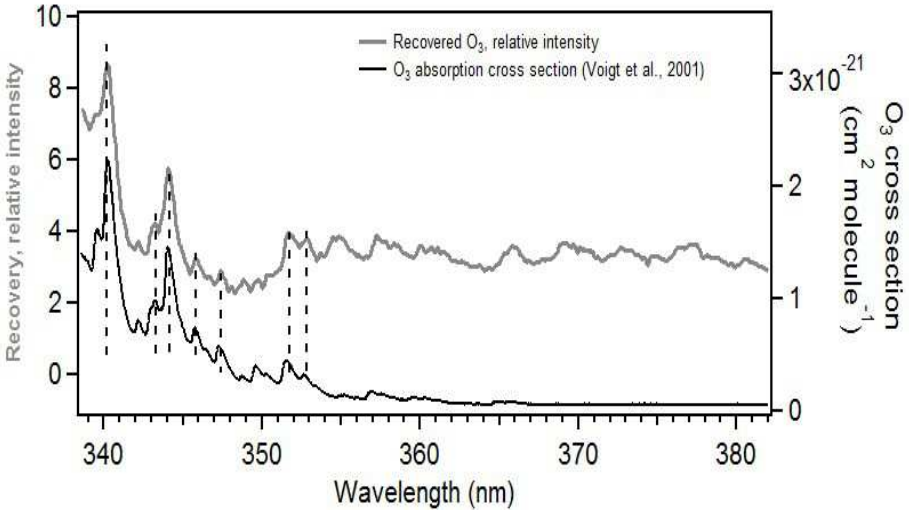

Figure 9. Recovered O_3 and its absorption cross section [25] for comparison

their reference spectra. As a matter of fact, even without computing these similarity measures, Fig. 11 is easily recognizable as O_3 to an atmospheric chemist.

In the situation of no spectral references available (or no knowledge of the mixtures), a blind source separation approach such as multivariate curve resolution (MCR) [14] could be employed to extract the spectra of the individual trace gases and their estimated concentrations. Due to the lack of the knowledge of the trace gases other than nonnegativity constraint on the concentrations, MCR method decomposes the mixture $\mathbf{X}$ into $\mathbf{A}$ and $\mathbf{S}$ by solving a non-convex optimization problem

$$\min_{\mathbf{A},\mathbf{S}} \|\mathbf{X} - \mathbf{A}\ \mathbf{S}\|_2^2, \text{ subject to } \mathbf{S}_{ij} \geq 0\ . \tag{8}$$

The MCR algorithm used in the paper is MCR alternating least squares algorithm (MCR-ALS) with nonnegativity constraint on the concentration matrix **S**. The MCR-ALS, an iterative approach [15], requires an initial guess and input of the number of source components. Note that MCR-ALS is a non-convex optimization method, and different initial guesses generally give different solutions. As a comparison, we have tested the DOAS data using MCR-ALS method, and the results are shown in Fig. 12. The top plot is the recovered spectral structures (columns of **A**) and the bottom plot is their coefficients (matrix **S**). Given a-priori that there are three species present in the mixture, MCR analysis yields that all three components are the same species. As shown in Fig. 12 (top), the results are recognizable as HONO (major peaks at 343, 354 and 368 nm), but MCR fails to extract NO_2 and O_3, which are separate, orthogonal species. In addition, because it is a non-convex method, MCR can yield different results upon each analysis. The method proposed in the paper does not exhibit this variability and can be used with or without reference spectra for quantitative or qualitative analysis, making it superior for DOAS data analysis.

Each step of the process is designed to accomplish a particular task, and the overall goal is to find a good estimate of the concentration of known trace gases and to identify possibly hidden gaseous structures from the residual. Although each step of our method can be replaced by a different approach, reasonable and meaningful results really depend on the integration of all three steps. Note that our method aims to achieve three objectives: 1) automate the detection of broad background and remove it from the original data sets; 2) deliver better estimates of concentration of the known trace gases compared to the CLS in the presence of outliers in the data; 3) help the recognition of spectral structures from the fitting residuals. The Huber estimator is chosen to prevent the outliers from destroying the results. Note that the Huber estimator will deliver similar results as CLS if the data do not have outliers. To demonstrate this, we have replaced the second step of our method (Huber estimator) with CLS and identified the residual by ICA. The results are plotted in Fig. 13. Overall, the two spectra look similar as might be expected based on the absence of outliers. For the DOAS data in our numerical experiments, we found that the Huber's estimator produced non-negative solutions **S** without explicit enforcement of non-negativity constraint for certain values of the tuning parameter k. However, negative values may exist in CLS solution for **S** if non-negativity is not enforced.

As multichannel detectors age, some of the pixels can become damaged or dust particles can settle on them rendering them unresponsive. Material defects can also give rise to unresponsive channels. In the example below, we added two spikes (outliers) to the DOAS data to mimic this case ranging from 5 to 7 pixels in width. Fig. 14 shows one mixture spectrum with damaged pixels. We decompose the damaged data to a linear combination of reference spectra of HONO and NO_2 via CLS and Huber's estimator. The coefficients of HONO are shown in Fig. 15. It can be seen that a few damaged pixels in the data destroy the CLS results, while the Huber estimator remains robust with outliers. The results for NO_2 were similar with Huber's estimator being more robust in the presence of outliers. This

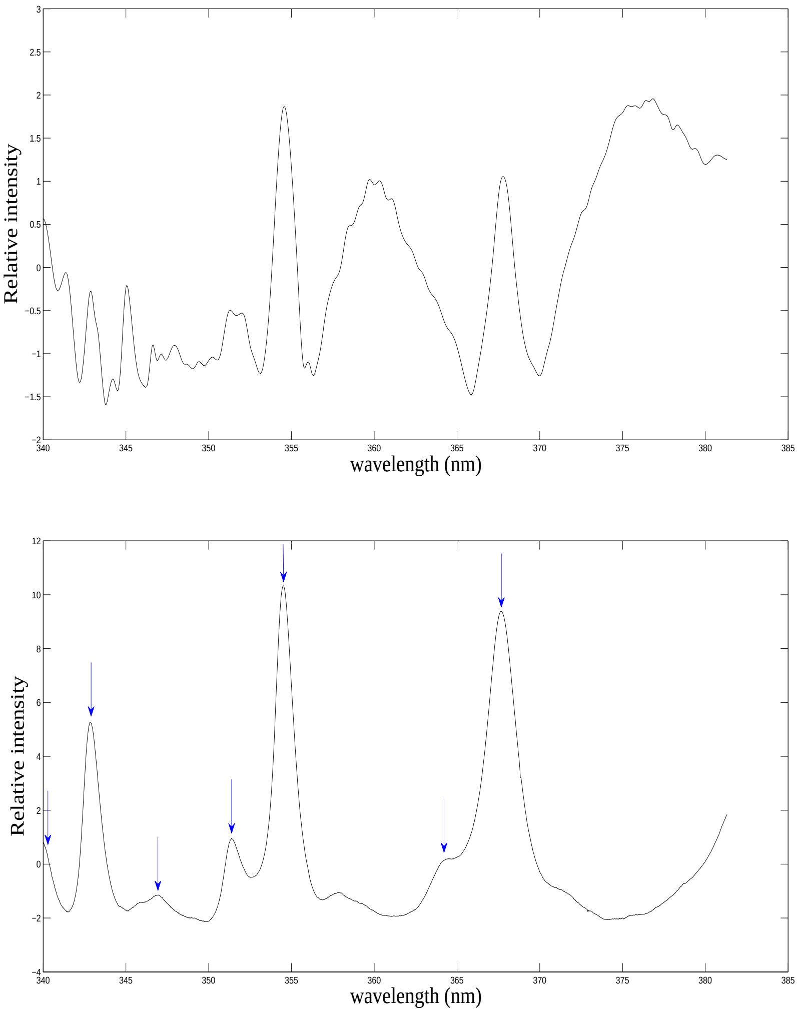

Figure 10. Top plot is the identified spectral structure 1 compared to the spectral reference of HONO (bottom).

example demonstrates the superiority of the Huber estimator to CLS when outliers are present.

4. **Concluding remarks.** We developed a semi-blind source separation method for retrieving the concentrations and performing identifications of trace gases from DOAS spectra. The method is designed to identify potentially hidden trace gases after fitting the known trace gases to the data, which is a challenging problem. Our method can be useful for separating unknown source signals from the residuals after any known reference spectra have been first deployed to fit the data. Analysis of DOAS data using another potential technique, multivariate curve resolution (MCR), to extract hidden signals yielded highly variable results and did not identify some components. The first novelty of the new method is to employ the

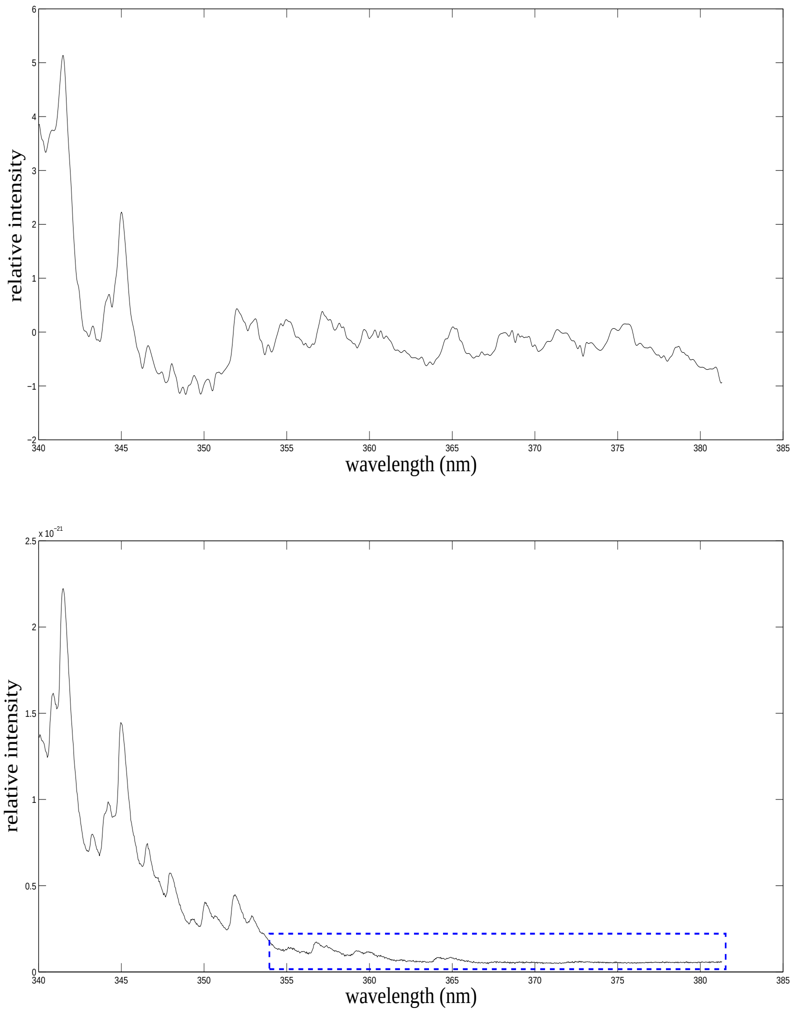

Figure 11. Top plot is the the identified spectral structure 2 compared to reference spectrum for O_3 (bottom). Note that very few feature peaks are in the rectangular region.

multi-resolution analysis (EMD) to remove the slowest varying component from the data. The removal of such components relies on a polynomial fit in the existing methods. Different polynomials may produce different results, and the degree of the fitting polynomial is often empirically defined. The multi-resolution approach avoids specifying the order of polynomial, and it extracts the slow component in an automatic fashion. The second novelty is to use a hybrid ℓ_1/ℓ_2 interpolated norm (Huber function) to fit the data, which reduces the effects of outliers and keeps the concentrations non-negative. The ability of Huber's estimator to succeed in the presence of a small number of typical outliers is demonstrated, whereas the CLS methods fall short under these circumstances. Lastly, a multi-channel signal decomposition method (JADE) produced encouraging results on extracting hidden source

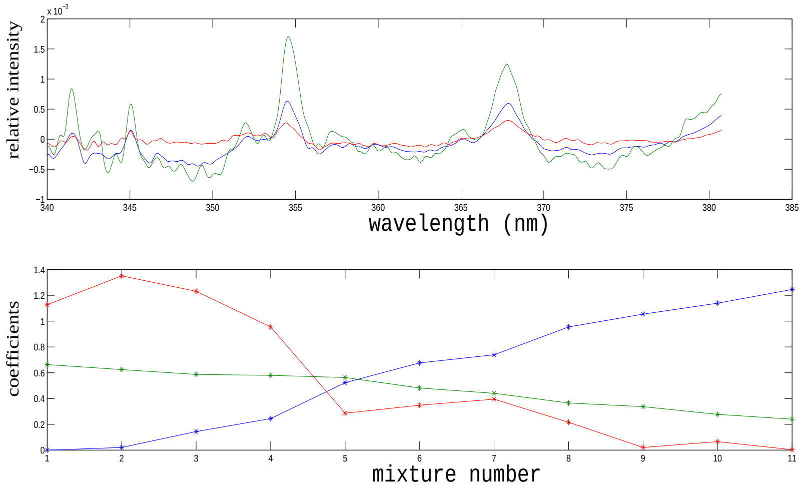

Figure 12. Top panel is the three spectral structures computed by MCR, where the three colored curves correspond to the spectral structures of three species. The bottom row is their coefficients.

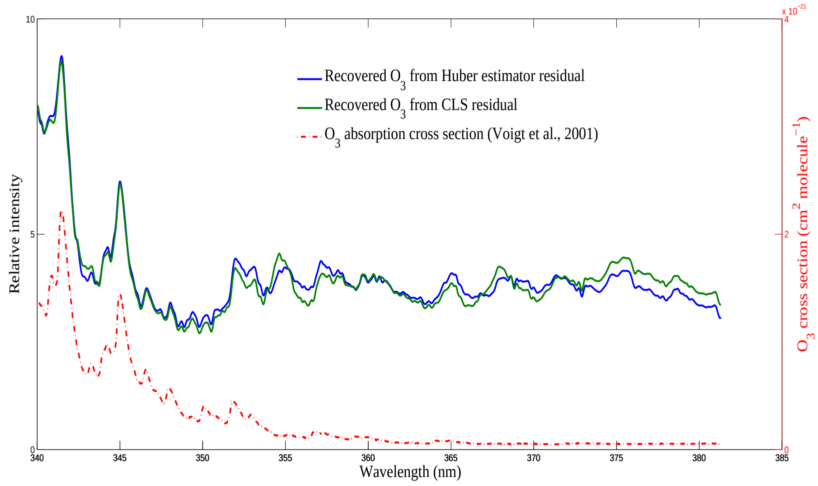

Figure 13. Comparison between the CLS and Huber estimator.

signals from the fitting residuals. While use of the least squares fitting procedure for atmospheric data can quantify several trace species simultaneously, typical fit residuals often suggest there are remaining absorbers. In some cases, species can be inferred based on known atmospheric chemistry, e.g., HONO is often present in NO_2 mixtures. The major strength of the technique described here is its ability to be used either with existing published reference spectra for quantification or without references for identification of new absorbers. Numerical results on DOAS

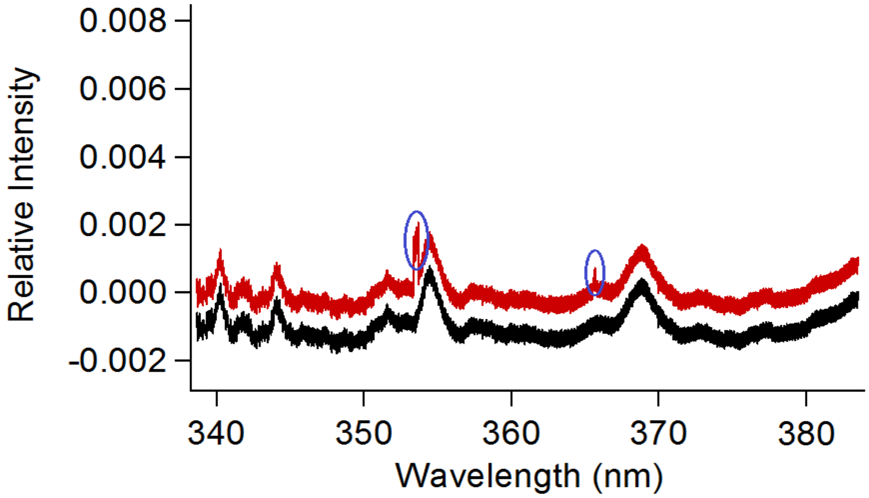

Figure 14. Original mixture spectrum (black) and same spectrum with typical outliers (red). Damaged pixels are shown in the circled regions.

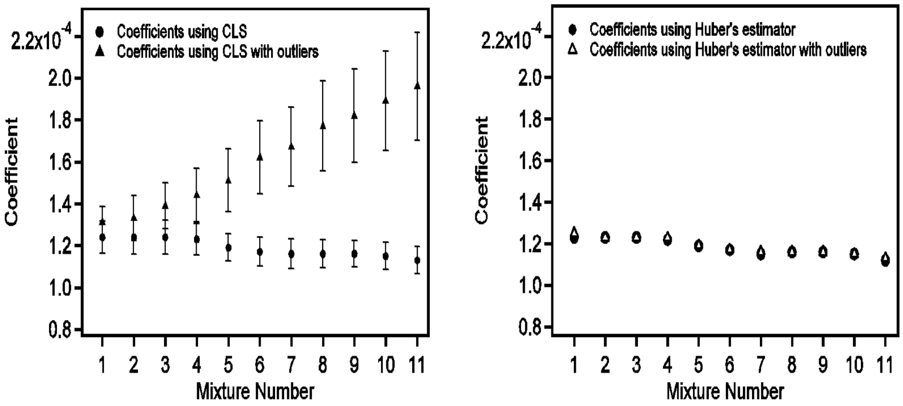

Figure 15. Computational results of coefficients of HONO in the mixtures with and without damaged spikes using CLS (left) and the Huber estimator (right).

data show the promising potential of our method for both trace gas recovery and quantification.

Acknowledgments. This work was partially supported by NSF-ADT grants DMS-0911277, DMS-1222507 and NSF-CHE 0909227. The authors thank Professor Jochen Stutz for helpful discussions and the anonymous referees for their constructive comments.

REFERENCES

[1] A. Bell and T. Sejnowski, An information-maximization approach to blind separation and blind deconvolution, *Neural Computation,* **7** (1995), 1129–1159.

[2] A. Bongartz, J. Kames, U. Schurath, CH. George, PH. Mirabel and J. L. Ponche, Experimental determination of hono mass accommodation coefficients using two different techniques, *J. Atmos. Chem.,* **18** (1994), 149–169.

[3] A. Bongartz, J. Kames, F. Welter and U. Schurath, Near-UV absorption cross sections and trans/cis equilibrium of nitrous acid, *J. Phys. Chem.,* **95** (1991), 1076–1082.

[4] J.-F. Cardoso, High-order contrasts for independent component analysis, *Neural Computation,* **11** (1999), 157–192.

[5] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications,* John Wiley and Sons, New York, 2005.

[6] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications,* Academic Press, 2010.

[7] I. Daubechies, J. Lu and H.-T. Wu, Synchrosqueezed wavelet transforms: An empirical mode decomposition-like tool, *Applied and Computational Harmonic Analysis,* **30** (2011), 243–261.

[8] D. DeHaan, T. Brauers, K. Oum, J. Stutz, T. Nordmeyer and BJ. Finlayson-Pitts, Heterogeneous chemistry in the troposphere: Experimental approaches and applications to the chemistry of sea salt particles, *Intern. Rev. Phys. Chem.,* **18** (1999), 343–385.

[9] BJ. Finlayson-Pitts, LM. Wingen, AL. Sumner, D. Syomin and KA. Ramazan, The Heterogeneous Hydrolysis of NO_2 in Laboratory Systems and in Outdoor and Indoor Atmospheres: An Integrated Mechanism, *Phys. Chem. Chem. Phys.,* **5** (2003), 223–242.

[10] T. Gomer, T. Brauers, T. Heintz, J. Stutz and U. Platt, MFC Version 1.99, 1995.

[11] N. Huang, Z. Shen, S. Long, M. Wu, H. Shih, Q. Zheng, NC. Yen, C. Tung and H. Liu, The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis, *Proc. R. Soc. Lond. A.,* **454** (1998), 903–995.

[12] J. Huber and E. Ronchetti, *Robust Statistics*, Wiley, 2009.

[13] A. Hyvärinen, Fast and robust fixed-point algorithms for independent component analysis, *IEEE Trans. on Neural Networks,* **10** (1999), 626–634.

[14] A. Juan and R. Tauler, Multivariate Curve Resolution (MCR) from 2000: Progress in Concepts and Applications, *Critical Reviews in Analytical Chemistry,* **36** (2006), 163–176.

[15] E. J. Karjalainen, The Spectrum Reconstruction Problem: Use of Alternating Regression for Unexpected Spectral Components in two-dimensional Spectroscopies, *Chemometrics and Intelligent Laboratory Systems,* **7** (1989), 31–38.

[16] L. Lin, Y. Wang and H. Zhou, Iterative filtering as an alternative algorithm for empirical mode decomposition, *Adv. Adapt. Data Anal.,* **1** (2009), 543–560.

[17] J. Liu, J. Xin and Y. Y. Qi, A Soft-Constrained Dynamic Iterative Method of Blind Source Separation, *SIAM J. Multiscale Modeling Simulation,* **7** (2009), 1795–1810.

[18] J. Noxon, Nitrogen dioxide in the stratophere and troposphere measured by ground-based absorption spectroscopy, *Science,* **189** (1975), 547–549.

[19] J. Noxon, E. Whipple and R. Hyde, Stratospheric NO_2. 1. Observational method and behavior at midlatitudes, *J. Geophys. Res.,* **84** (1979), 5047–5076.

[20] D. Perner, D. Ehhalt, H. Patz, U. Platt, E. Roth and A. Volz, OH-radicals in the lower troposphere, *Geophys. Res. Lett.,* **3** (1976), 466–468.

[21] U. Platt, D. Perner and H. Pätz, Simultaneous measurements of atmospheric CH_2O, O_3 and NO_2 by differential optical absorption, *J. Geophys. Res.,* **84** (1979), 6329–6335.

[22] U. Platt and J. Stutz, *Differential Optical Absorption Spectroscopy: Principles and Applications*, Springer, 2008.

[23] J. Stutz and U. Platt, Numerical analysis and estimation of the statistical error of differential optical absorption spectroscopy measurements with least-squares methods, *Appl. Optics.,* **35** (1996), 6041–6053.

[24] G. Thakur, E. Brevdo, N. S. Fučkar and H.-T. Wu, The Synchrosqueezing algorithm for time-varying spectral analysis: Robustness properties and new paleoclimate applications, *Signal Processing,* **93** (2013), 1079–1094.

[25] S. Voigt, J. Orphal, K. Bogumil and J. P. Burrows, The temperature dependence (203–293 K) of the absorption cross sections of O_3 in the 230 - 850 nm region measured by Fourier-transform spectroscopy, *J. Photochem. Photobiol. A: Chemistry,* **143** (2001), 1–9.

[26] Y. Wang, G.-W. Wei and S. Yang, Iterative filtering decomposition based on local spectral evolution kernel, *J. of Sci. Comp.,* **50** (2012), 629–664.

[27] J. White, Long Optical Paths of Large Aperture, *J. Opt. Soc. Amer.,* **32** (1942), 285–288.

[28] ZH. Wu and N. Huang, Ensemble empirical mode decomposition: A noise-assisted data analysis method, *Advances in Adaptive Data Analysis,* **1** (2009), 1–41.

Appendix: Mathematical details of the DOAS approach. The DOAS approach is based on Lambert-Beer's law relating light intensity through a sample to the concentration of the light absorbing species, the length of the light path through the sample, and the absorption cross section or strength of absorption of each species. The intensity, $I(\lambda)$, at the end of the light path as shown in the text is,

$$(9)\quad I(\lambda) = I_0(\lambda)\exp\left[-\int_0^L \sum_j \sigma_j^{\mathrm{ABS}}(\lambda)\times\rho_j(l) + \varepsilon_{\mathbf{R}}(\lambda,l) + \varepsilon_{\mathrm{M}}(\lambda,l)\,\mathrm{d}l\right] + N(\lambda)\ ,$$

where σ_j^{ABS} is the absorption cross section of a trace gas j, ρ_j is its number density, L is the length of the light path, and $N(\lambda)$ is the measurement noise. The parameters $\varepsilon_{\mathbf{R}}$ and ε_{M} represent the Rayleigh extinction by gases and Mie extinction by aerosols, respectively. DOAS utilizes a separation of the cross section $\sigma_j^{\mathrm{ABS}} = \sigma_j^{\mathrm{B}} + \sigma_j'$ into broad spectral features, denoted by σ_j^{B}, and the narrow spectral features of the differential cross section, denoted by σ_j'. The narrow spectral features of the gas absorptions, σ_j', are treated in order to obtain trace gas concentrations. The mathematical description of this process is a convolution of $I(\lambda)$ with the instrument function H of the spectrometer,

$$\begin{aligned} I^*(\lambda) &= I(\lambda)*H = \int I(\lambda-\lambda')\,H(\lambda')\mathrm{d}\lambda' \\ &= \int_{-\Delta\lambda}^{\Delta\lambda} I_0(\lambda-\lambda')\exp\Bigg[-\int_0^L\sum_j \sigma_j^{\mathrm{ABS}}(\lambda-\lambda')\,\rho_j(l) + \\ &\qquad (\varepsilon_{\mathbf{R}}+\varepsilon_{\mathrm{M}})(\lambda-\lambda',l)\,\mathrm{d}l\Bigg]\,H(\lambda')\mathrm{d}\lambda' + N(\lambda)*H \\ &= \int_{-\Delta\lambda}^{\Delta\lambda} I_0'(\lambda-\lambda')\exp\left[-\int_0^L\sum_j \sigma_j'(\lambda-\lambda')\rho_j(l)\mathrm{d}l\right]\,H(\lambda')\mathrm{d}\lambda' + N(\lambda)*H \end{aligned}$$

where

$$I_0'(\lambda-\lambda') = I_0(\lambda-\lambda')\exp\left[-\int_0^L\sum_j \sigma_j^{\mathrm{B}}(\lambda-\lambda')\,\rho_j(l) + (\varepsilon_{\mathbf{R}}+\varepsilon_{\mathrm{M}})(\lambda-\lambda',l)\,\mathrm{d}l\right]$$

describes the broad spectral structures due to the characteristics of the light source I_0, the Rayleigh and Mie's extinction, and the broad absorption by trace gases. I_0' is a slow-varying function of wavelength, so $I^*(\lambda)$ can be approximated by

$$\begin{aligned} I^*(\lambda) &= I_0'(\lambda)\int_{-\Delta\lambda}^{\Delta\lambda}\exp\left[-\int_0^L\sum_j \sigma_j'(\lambda-\lambda')\times\rho_j(l)\mathrm{d}l\right]H(\lambda')\mathrm{d}\lambda' + N(\lambda)*H \\ &= I_0'(\lambda)\exp\left[\mathbf{S}_j\times\int_{-\Delta\lambda}^{\Delta\lambda}\sum_j -\sigma_j'(\lambda-\lambda')\,\mathrm{d}l\right]H(\lambda')\mathrm{d}\lambda' + N(\lambda)*H \end{aligned}$$

$$\begin{aligned}
&= I_0'(\lambda)\exp\left[\sum_j \mathbf{S}_j \times \mathbf{A}_j(\lambda)\right] + N(\lambda) * H \\
&= I_0'(\lambda)\exp\left[\sum_j \mathbf{S}_j \times \mathbf{A}_j(\lambda)\right] + N(\lambda) * H \\
&= I_0'(\lambda)\exp\left[\sum_j \mathbf{S}_j \times \mathbf{A}_j(\lambda)\right]\left\{1 + \frac{N(\lambda) * H}{I_0'(\lambda)\exp\left[\sum_j \mathbf{S}_j \times \mathbf{A}_j(\lambda)\right]}\right\}
\end{aligned}$$

where $\mathbf{S}_j = \int_0^L \rho_j(l)\,\mathrm{d}l$, and $\mathbf{A}_j(\lambda) = -\sigma_j'(\lambda) * H(\lambda)$ denote the narrow absorption structures of the trace gases measured with the same instrument. Suppose that there are m known trace gases in the data. The logarithm of the above equation becomes

$$x(\lambda) = \sum_{j=1}^{m} \mathbf{S}_j \times \mathbf{A}_j(\lambda) + B'(\lambda) + N'(\lambda)\ , \tag{10}$$

where $x(\lambda) = \ln I^*(\lambda)$ and $B'(\lambda) = \ln I_0'(\lambda)$ represents the broad spectral features. The noise $N'(\lambda) = \ln\left\{1+\frac{N(\lambda) * H}{I_0'(\lambda)\exp\left[\sum_j \mathbf{S}_j \times \mathbf{A}_j(\lambda)\right]}\right\} \approx \frac{N(\lambda) * H}{I_0'(\lambda)\exp\left[\sum_j \mathbf{S}_j \times \mathbf{A}_j(\lambda)\right]}$ given the fact that $N(\lambda) * H \ll I_0'(\lambda)\exp\left[\sum_j \mathbf{S}_j \times \mathbf{A}_j(\lambda)\right]$. In the experiment, the wavelength range is mapped onto p discrete pixels of the detector. The sampled data points form p-dimensional column vectors of a data matrix $\mathbf{X}$. Suppose there are n measurements, then $\mathbf{X} = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{p\times n}$ whose column vectors are recorded DOAS data points. Eq. 10 in matrix form is

$$\mathbf{X} = \mathbf{A}\,\mathbf{S} + \mathbf{B} + \mathbf{N}\ , \tag{11}$$

where the columns of matrix $\mathbf{A}$ correspond to the reference spectra of the known trace gases; the matrix $\mathbf{S}$ contains non-negative coefficients, which are proportional to the product of each trace gas concentration and pathlength; the matrix $\mathbf{B}$ includes the slow-varying components, and the matrix $\mathbf{N}$ contains the noise components. As discussed in the text, DOAS analysis methods typically apply least squares methods to calculate the value of $\mathbf{S}$ for each mixture spectrum. $\mathbf{S}$ is proportional to the product of the concentration and the path length through the use of the concentration of the reference spectrum utilized in the fitting procedure.

Received September 2012; revised December 2013.

E-mail address: yuasun@fiu.edu
E-mail address: wingenit@uci.edu
E-mail address: bjfinlay@uci.edu
E-mail address: jxin@math.uci.edu