MIXTURE QUANTIFICATION IN THE PRESENCE OF UNKNOWN INTERFERENCES

Zelun Wang, Tiening Jin, Paotai Lin, Ricardo Gutierrez-Osuna
Texas A&M University, College Station, Texas 77840, USA
{wang14359, jtnjtnjtn, paolin, rgutier}@tamu.edu

ABSTRACT

Conventional methods for multicomponent analysis such as partial least squares perform well if the constituents of the chemical mixture are known. However, these methods degrade significantly when unknown interferents are present in the mixture. We describe a sparse active-sensing approach that addresses the interference problem in chemical mixture quantification. The approach assumes that a large library is available containing hundreds (or thousands) of potential interferences. This strategy is infeasible with conventional calibration models since it leads to ill-conditioning (i.e., many solutions exist that match the measurement vector). We show that combining active sensing with a sparsity constraint allows us to recover the concentration of the target compounds, even in the presence of high concentrations of unknown backgrounds. We evaluate the performance of the method against partial least squares using simulated mixtures from a database of Fourier transform infrared (FTIR) spectra.

Index Terms— Active sensing algorithm, mixture quantification, infrared spectroscopy, interference chemicals

1. INTRODUCTION

Infrared (IR) absorption spectra can be used for quantitative analysis of chemical mixtures. Given a spectral library of individual chemicals and the measured spectrum for a mixture, multicomponent analysis (MCA) techniques can be used to estimate the concentration of the individual components. Conventional MCA methods include classical and inverse least squares, principal component regression (PCR) [1], and partial least squares (PLS) regression [2]. These traditional methods degrade when unknown interference chemicals are present in the mixture. A potential solution is to increase the size of the spectra library to include all possible interference chemicals. However, this leads to numerical problems since the system becomes under-determined. In addition, when interference chemicals overlap significantly with the target chemicals, identifying individual components becomes even harder.

This paper proposes an active sensing (AS) algorithm [3] that addresses the limitations of conventional MCA methods. Our AS approach is designed to determine which chemicals are present in an unknown mixture from a large list of potential chemicals in a spectra library. It selects wavelengths actively and applies a sparsity constraint on the solution, thus recovering the true mixture constituents. We show that AS can be used to solve chemical quantification problems when unknown interference chemicals are present, even at high concentrations, and that the approach scales up to large spectral libraries.

2. METHODS

2.1. Mixture quantification

Let matrix A denote a library of IR spectra, each column corresponding to the absorption spectrum of one chemical. Further, let vector c denote the concentration of individual components in the mixture, and vector b denote the measured spectrum of the chemical mixture. The mixture quantification problem consists of estimating c given A and b. MCA methods can solve this problem when mixture constituents are known, since we have $A \times c = b$ when individual components are at low concentrations, according to Beer's law. However, when interferences are present in the mixture, estimating the concentrations of the individual components is no longer trivial. Our paper focuses on mixture quantification problem under such situations.

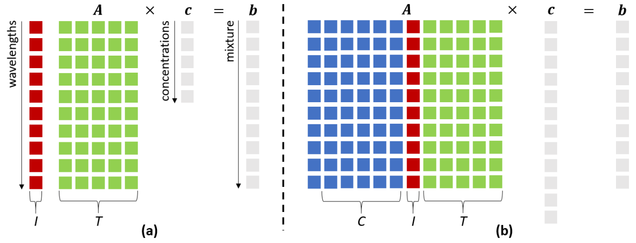

Fig 1. Each column in the library matrix A represents the spectrum of a chemical. Green columns represent targets; red columns represent interferences; blue columns represent confounders. In (a), A contains only the targets. In (b), A contains all targets, interferences, and confounders.

Let us define three sets of chemicals: targets T, interferences I, and confounders C. Targets T are chemicals that are known to be part of the mixture and whose concentration we seek to estimate. Interferences I are additional chemicals (generally undesired) that are present in the mixture, whereas confounders C are chemicals that are not present in the mixture, but whose spectra is included in the library.

978-1-5090-2392-9/17/$31.00 ©2017 IEEE

Assume we have a chemical mixture with no interferences, and that the library A contains only a small number of target chemicals (no confounders). In this case, the concentration vector c can be estimated using conventional MCA techniques, such as the pseudo-inverse solution [4]. The problem becomes more difficult when interference chemicals are present in the mixture; see Fig 1 (a). In this case, the influence of interferences on the measured spectra must be explained by other chemicals in the library (e.g., combinations chemicals with similar spectra), which in turn degrades the prediction accuracy.

As illustrated in Fig 1 (b), our approach consists of increasing the library size substantially to cover all possible interference chemicals. Unfortunately, increasing the number of columns in the library matrix A makes the problem harder since it introduces many potential confounders C and, more importantly, makes the system under-determined. It is here where active sensing becomes essential, as we describe next.

2.2. Active sensing

The proposed active sensing algorithm [3] was designed to identify mixtures when its constituents are unknown[1], even when the system is ill-conditioned. The algorithm iterates between selecting wavelengths as new measurements and identifying chemicals as predictions. It measures a subset of wavelengths in a sequential manner, each new wavelength being selected based on the previous measurements. After each new measurement, the algorithm alternates between an explorative stage guided by Gaussian process regression (GPR), which reconstructs the spectrum of the unknown mixture, and an exploitative state guided by linear discriminant analysis (LDA), which eliminates irrelevant mixture components. After each iteration, the algorithm uses non-negative least squares (NNLS) to estimate the concentrations of individual components. It then uses the Bayesian information criterion (BIC) to guide a shrinkage process, which sequentially eliminates the least significant component from the solution vector to guarantee that the solution is sparse. The estimated concentrations are then used to identify the components and refine the wavelength selection process. See reference [3] for details.

3. EXPERIMENTS

We validated the approach on a dataset of synthetic IR spectra from NIST's Webbook [5]. The dataset contains Fourier Transform Infrared (FTIR) spectra from 500 chemicals, each spectrum containing 660 wavelengths in the range 3-11.5 μm. We synthesize the mixture spectrum as a weighted sum of individual spectra, each weighted by its respective concentration; this is a valid first-order approximation since IR radiation absorbs in linear proportion to concentration. We used the AS algorithm to estimate the concentrations of the targets given the mixture spectrum. For comparison, we used PLS regression as a baseline method.

To ensure consistency across experimental runs, we selected 5 chemicals as the target set T. These 5 chemicals were selected to be the most distinctive (least correlated) among the 500 chemicals in the database using a greedy search: we randomly select one chemical from the dataset and added it to T; then, we select the chemical that has the least *similarity* with the current target set T, and repeat the previous step until T contains 5 chemicals. We define the *similarity* between a chemical x and those in the set T as $sim_{x,T} = \max_{k \in \mathrm{T}} corr(a_k, a_x)$, where $corr$ is the Pearson correlation, a_x is the IR spectrum of chemical x, and a_k is the spectrum of the *k-th* chemical in T.

For evaluation, we compute the root mean square error (RMS) of the estimated concentrations as follows:

$$RMS = \sqrt{\frac{1}{|T|} \sum_{i=1}^{|T|} |(c_i - \widehat{c_i})/c_i|^2} \qquad (1)$$

where c_i is the ground truth concentration of the ith target, $\widehat{c_i}$ is its estimated concentration, and $|T|$ is the number of targets (i.e., $|T| = 5$ in our study). We conducted three experiments and reported the RMS given by AS and PLS.

3.1. Experiment I

In a first experiment, we tested AS and PLS on mixtures that contained only the selected 5 target chemicals, without interferences or confounders. First, we generated 110 random mixtures, each mixture containing the 5 target chemicals at random concentrations in the range from 1 to 1000 (arbitrary units). We used 10 of these mixtures as the test set T_1, and the remaining 100 mixtures as the training set. Then, we used mixture spectra in the training set as the observations (inputs) and their concentrations as the responses (outputs) to train the PLS model. For the AS algorithm, no training is required; we simply run AS on a library matrix that has only the 5 target chemicals. We tested both PLS and AS on the test set T_1. Experimental results show that both PLS and AS achieved low average RMS: 1.33×10^{-14} for PLS and 1.92×10^{-6} for AS.

3.2. Experiment II

In a second experiment, we evaluated the PLS model when the mixtures contained not only chemical targets but also interferences. How much an interference affects the prediction model depends on how similar the interference is to the target chemicals in T. For this reason, we generated a new set of mixtures consisting of 6 chemicals: the 5 targets plus 1 interference, all at equal concentration of 500 units. The result is a test set T_2 with 495 mixtures, one for each of the remaining chemicals in the library (excluding the 5

[1] According to our terminology, the targets in the mixture are known but the interferences are unknown.

targets).

Fig 2 (a) shows the RMS of the PLS model trained in Experiment I, as a function of the similarity between the interference and the 5 target chemicals. As expected, the PLS error increases with the similarity score (the correlation coefficient between RMS and $sim_{x,T}$is 0.62). The average RMS among the 495 mixtures is 0.29, meaning that the estimated concentrations deviates from the ground truth concentrations by 29%.

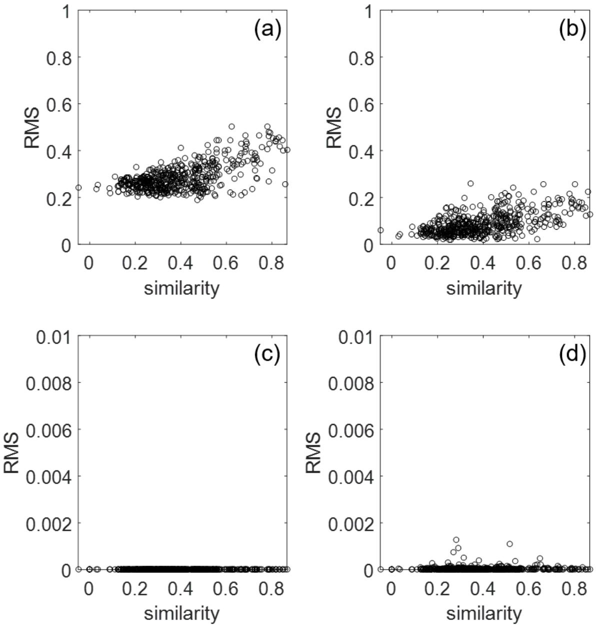

Fig 2. RMS of PLS (a,b) and AS (c,d) as a function of the similarity between the 5 chemicals in the mixture and the interference. (a) PLS trained on mixtures with 5 targets and tested on T_2. (b) PLS trained on mixtures with 5 targets + 1 interference, and tested on T_2. (c) AS tested on T_2. (d) AS tested on T_3.

The high RMS of the PLS model may be attributed to the fact that interference chemicals were not part of the training set. Thus, we retrained the PLS model on a new training set that included the possible interference chemicals. The training set was generated as follows. First, we selected one interference (out of 495), added it to the 5 targets, and then generated 100 mixtures, each containing these 6 individual components at random concentrations ranging from 1 to 1000. We repeated this process for each of the 495 interference chemicals, which resulted in a training set with 49,500 spectra. Next, we trained a new PLS model based on this training set and tested on T_2. Results are shown in Fig 2 (b). As before, the RMS increases with the similarity score (a correlation coefficient of 0.56), though the errors are significantly lower than those in Fig 2 (a). Thus, training the PLS model on a dataset that contains interferences makes it more robust. Notice, however, that the average RMS is still relatively high (0.09). In addition, this new PLS model can only handle mixtures with one interference.

3.3. Experiment III

In a third experiment, we introduced interferences to the mixtures and tested their effect on the AS model. In this case, AS does not require any training process; instead, it only requires that the library incorporates all possible interference chemicals, which result in a library matrix $A \in \mathbb{R}^{660\times 500}$.

First, we tested AS on dataset T_2. Results are shown in Fig 2 (c). AS performs accurately regardless of the similarity between the interference and the targets (the average RMS is 1.09×10^{-6}, 5 orders of magnitude lower than that in Fig 2 (a). Next, we generated a more challenging test set T_3 with 495 mixtures. Each mixture in T_3 contains one interference at high concentration. Specifically, we set the concentrations of the 5 target chemicals to be random numbers between 1 and 100, and the concentration of the interference to 1000 (this is unlike in T_2, where all individual concentrations are set to be 500). Fig 2 (d) shows the results of evaluating the active sensing algorithm on T_3. Even when the interference is at higher concentration than the targets, AS is able to obtain accurate estimate of the target concentrations (the average RMS is 4.30×10^{-5}), though with a slightly higher error than on T_2.

4. CONCLUSION

We have described an active sensing approach can be used to solve mixture quantification problems when unknown interference chemicals are present. To incorporate knowledge of all the possible interference chemicals, we increase the library size substantially, which introduces a large number of confounders and makes the problem ill-conditioned for traditional MCA methods. Our proposed AS algorithm performs remarkably better than PLS regression in the presence of unaccounted interferences. An added advantage of AS is that it requires no training data. Moreover, the proposed AS algorithm scales up to large library sizes, and is robust to overlapping spectral and interferences at high concentrations.

REFERENCES

[1] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems,* vol. 2, pp. 37-52, 1987.

[2] P. Geladi and B. R. Kowalski, "Partial least-squares regression: a tutorial," *Analytica chimica acta,* vol. 185, pp. 1-17, 1986.

[3] J. Huang and R. Gutierrez-Osuna, "Active wavelength selection for mixture identification with tunable mid-infrared detectors," *Analytica chimica acta,* vol. 937, pp. 11-20, 2016.

[4] R. Gutierrez-Osuna, "Pattern analysis for machine olfaction: a review," *IEEE Sensors journal,* vol. 2, pp. 189-202, 2002.

[5] P. J. Linstrom and W. Mallard, "NIST Chemistry webbook; NIST standard reference database No. 69," 2001.