Analytica Chimica Acta 937 (2016) 11–20

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Active wavelength selection for mixture identification with tunable mid-infrared detectors

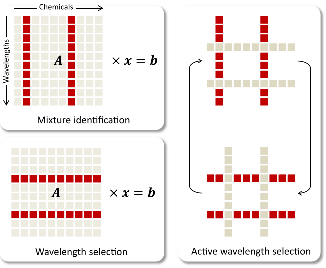
Jin Huang, Ricardo Gutierrez-Osuna*

Department of Computer Science and Engineering, Texas A&M University, United States

HIGHLIGHTS

- An active wavelength-selection algorithm is proposed for mixture identification.
- The algorithm runs in real-time, interleaving wavelength selection with sensing.
- Wavelength selection is analyte-specific and based on previous measurements.
- Active wavelength selection operates in two stages: exploration and exploitation.
- The approach is compared against a passive strategy based on successive projection.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 21 May 2016
Received in revised form
19 July 2016
Accepted 4 August 2016
Available online 10 August 2016

Keywords:
Active wavelength selection
Mixture identification
Fabry-Perot interferometry
Gaussian process regression
Shrinkage non-negative least squares
Linear discriminant analysis

ABSTRACT

This article presents a wavelength selection framework for mixture identification problems. In contrast with multivariate calibration, where the mixture constituents are known and the goal is to estimate their concentration, in mixture identification the goal is to determine which of a large number of chemicals is present. Due to the combinatorial nature of this problem, traditional wavelength selection algorithms are unsuitable because the optimal set of wavelengths is mixture dependent. To address this issue, our framework interleaves wavelength selection with the sensing process, such that each subsequent wavelength is determined on-the-fly based on previous measurements. To avoid early convergence, our approach starts with an exploratory criterion that samples the spectrum broadly, then switches to an exploitative criterion that selects increasingly more relevant wavelengths as the solution approaches the true constituents of the mixture. We compare this "active" wavelength selection algorithm against a state-of-the-art passive algorithm (successive projection algorithm), both experimentally using a tunable spectrometer and in simulation using a large spectral library of chemicals. Our results show that our active method can converge to the true solution more frequently and with fewer measurements than the passive algorithm. The active method also leads to more compact solutions with fewer false positives.

1. Introduction

Infrared (IR) spectroscopy is a powerful tool for qualitative and quantitative analysis of chemical mixtures. Mixture analysis typically requires using multivariate techniques since the IR spectra of

---

* Corresponding author.
E-mail addresses: tonmey@cse.tamu.edu, rgutier@cse.tamu.edu (R. Gutierrez-Osuna).

http://dx.doi.org/10.1016/j.aca.2016.08.005

individual chemicals can overlap significantly. However, not all wavelengths in the IR spectrum are useful. As an example, in the context of multivariate calibration, it has been shown −both theoretically [1] and experimentally [2]− that accuracy improves if a subset of the wavelengths is selected before conducting multivariate analysis. Accordingly, a number of wavelength selection algorithms have been proposed in the chemometrics literature, including exhaustive search (e.g., branch-and-bound [3]); randomized search (e.g., genetic algorithms [4], simulated annealing [5], ant colony optimization [6]); and greedy search (e.g., successive projection algorithms [7], uninformative variable elimination [8]). These algorithms work well for multicomponent calibration, when the target constituents are known and only their concentrations need to be estimated. In such cases, the linear system is known, so a globally optimal subset (containing a few wavelengths) exists. A more challenging problem is selecting wavelengths when the mixture constituents are unknown. This is the case for mixture identification problems, where the goal is to determine the constituents of a mixture from among a large number of chemicals. While mixture identification problems use the same linear model of multicomponent calibration, and therefore can also benefit from wavelength selection, the linear model becomes ill-defined because the spectral signature of the mixture can vary significantly (i.e., for a library with N constituents there are 2^N possible mixtures). Thus, different wavelength selection strategies are needed depending on how much is known (or can be assumed) about the constituents of the mixture.

To address this issue, we propose an active sensing strategy that interleaves wavelength selection with the sensing process, so that the next wavelength to be sampled is a function of previous measurements. This approach is fundamentally different from those outlined above because it does not generate a fixed “universal” subset of wavelengths but a unique sequence of wavelengths for each analyte. In other words, our approach assumes that the optimum subset of wavelengths is analyte dependent. Starting with an exploratory criterion that samples the spectrum broadly, our active sensing algorithm selects increasingly more relevant (i.e., exploitative) wavelengths as the sensing process continues and its estimates approach the true constituents of the mixture.

Our work builds on a previous algorithm for active wavelength selection [9] based on multi-modal solvers. In that early work, a multi-modal solver was used to generate multiple candidate spectra that fit the measurements well, and the wavelength with maximum variance across the candidate spectra was chosen as the next measurement. However, the multi-modal solver does not scale up to higher-order mixtures since its computational complexity grows with the number of chemicals in the spectral library. To overcome these computational issues, the work present here guides the wavelength selection process with two methods whose complexity grows with the number of wavelengths in the spectrum: Gaussian process regression (GPR) and linear discriminant analysis (LDA). Namely, we use GPR to reconstruct the spectrum of the unknown mixture −one wavelength at a time, and LDA to select wavelengths that allow us to eliminate irrelevant mixture components from the solution.

2. Methods

The problem of mixture identification can be formulated as:

$$\boldsymbol{Ax} = \boldsymbol{b} \quad \text{s.t.}\boldsymbol{x} \geq \boldsymbol{0} \tag{1}$$

where column matrix $\boldsymbol{A}$ is a reference library containing the spectrum for each possible chemical constituent, column vector $\boldsymbol{b}$ denotes the measured spectrum of a mixture, and $\boldsymbol{x}$ represents the concentration (non-negative) of the mixture. As illustrated in Fig. 1(a), the goal of mixture identification is to select the correct columns in matrix $\boldsymbol{A}$, i.e., determine the non-zero elements in the solution vector $\boldsymbol{x}$. By contrast, the goal of wavelength selection is to find a small number of rows in matrix $\boldsymbol{A}$ that offer good accuracy −Figure 1(b). Wavelength-selection algorithms assume that the identity of the chemicals is known (i.e., the correct columns in matrix $\boldsymbol{A}$ have been preselected), in which case supervised learning can be used to find a subset of wavelengths that maximize the effective rank for the known components. However, if the identity of the components is unknown, the problem becomes ill-defined because then not only the rows but also the columns in matrix $\boldsymbol{A}$ must be selected. This is a paradoxical problem because selecting optimal wavelengths requires knowledge of the mixture components, and identifying those components requires a set of wavelengths to be measured. We address this problem by using an iterative process that alternates between selecting rows (wavelengths) and columns (chemicals), as illustrated in Fig. 1(c). This requires an active sensing strategy that performs wavelength selection on-the-fly, interleaved with the sensing process.

Fig. 2 shows the building blocks of our algorithm for active wavelength selection. At a high level, the algorithm consists of two loops: an inner-loop that performs wavelength selection, and an outer-loop that performs mixture identification. The inner-loop (wavelength selection) operates in two distinctive stages: an initial exploratory stage that aims to reconstruct the entire spectrum, and a later exploitative stage that targets at distinctive regions in the spectrum. The outer-loop (mixture identification) uses a sparse solver to estimate the concentration of the analytes in the mixture. The estimated concentration is then used to identify the analyte and refine the wavelength selection process as it moves from exploration to exploitation. The outer loop is the more computationally intensive of the two, so it is executed once every n-th wavelength measurements. Parameter n provides a balance between computational costs and adaptiveness. We observed empirically that varying n from 1 to 20 had minimal impact on convergence, so we chose the median n=10 for the work reported here.

2.1. Explorative stage

The explorative stage of wavelength selection is guided by Gaussian Process Regression (GPR). Also known as kriging in geostatistics, GPR is an interpolation method that can be used to approximate a smooth arbitrary function using a set of sparse samplings. Fig. 3 illustrates GPR on a toy one-dimensional function. In this example, the goal is to reconstruct the function using a small number of samples. Because of the inherent smoothness of the function, GPR only requires 10 samples to recover it accurately. Also illustrated in Fig. 3 (shaded areas), GPR provides an estimate of the variance of the reconstruction, which indicates how uncertain the estimation is across all wavelengths. As we will see, this variance serves as the utility function for explorative wavelength selection.

Consider the case where we have selected m wavelengths $\boldsymbol{\lambda}_m = \{\lambda_1,\lambda_2,\ldots,\lambda_m\}$ and obtained the corresponding observations $\boldsymbol{b}_{\lambda_m} = \{b_{\lambda_1}, b_{\lambda_2}, \ \ldots, b_{\lambda_m}\}$. The goal of GPR is to reconstruct the full spectrum $\boldsymbol{b}_{GP} = \{b_{\lambda_1}, b_{\lambda_2}, \ldots, \ b_{\lambda_M}\}$ with $M \gg m$, and estimate its variance $\boldsymbol{S}^2_{GP} = \{S^2_{\lambda_1}, S^2_{\lambda_2}, \ldots, S^2_{\lambda_M}\}$. Gaussian processes model an arbitrary function as a random vector that follows a multivariate normal distribution$\boldsymbol{b}_{GP} \sim \mu + \mathscr{N}(0, \boldsymbol{R}_0)$ where μ is a scalar, and $\boldsymbol{R}_0$ is a covariance matrix $cov(\boldsymbol{\lambda}_M,\boldsymbol{\lambda}_M)$. The output of GPR is a multivariate distribution $\mathscr{N}(\boldsymbol{b}_{GP}, \boldsymbol{\Sigma}_{GP})$ where $\boldsymbol{\Sigma}_{GP}$ is a full covariance matrix