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
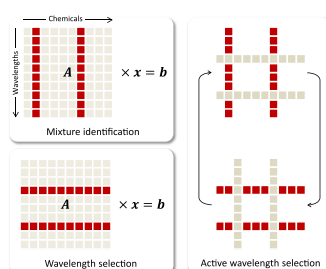
ABSTRACT

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.

© 2016 Elsevier B.V. All rights reserved.

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).

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:

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

where column matrix A is a reference library containing the spectrum for each possible chemical constituent, column vector b denotes the measured spectrum of a mixture, and 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 A, i.e., determine the non-zero elements in the solution vector x. By contrast, the goal of wavelength selection is to find a small number of rows in matrix 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 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 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 $\lambda_m = \{\lambda_1, \lambda_2, ..., \lambda_m\}$ and obtained the corresponding observations $b_{\lambda_m} = \{b_{\lambda_1}, b_{\lambda_2}, ..., b_{\lambda_m}\}$. The goal of GPR is to reconstruct the full spectrum $b_{GP} = \{b_{\lambda_1}, b_{\lambda_2}, ..., b_{\lambda_M}\}$ with $M \gg m$, and estimate its variance $S_{GP}^2 = \{S_{\lambda_1}^2, S_{\lambda_2}^2, ..., S_{\lambda_M}^2\}$. Gaussian processes model an arbitrary function as a random vector that follows a multivariate normal distribution $b_{GP} \sim \mu + \mathcal{N}(0, R_0)$ where μ is a scalar, and R_0 is a covariance matrix $cov(\lambda_M, \lambda_M)$. The output of GPR is a multivariate distribution $\mathcal{N}(b_{GP}, \Sigma_{GP})$ where Σ_{GP} is a full covariance matrix
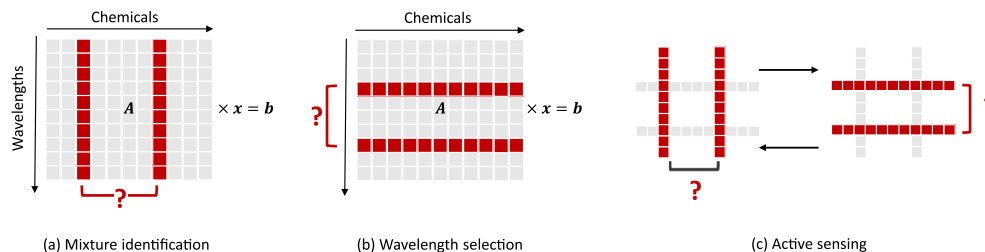
Fig. 1. The underlying mathematical problems of (a) mixture identification and (b) wavelength selection. (c) Active wavelength selection alternates between the two subproblems.
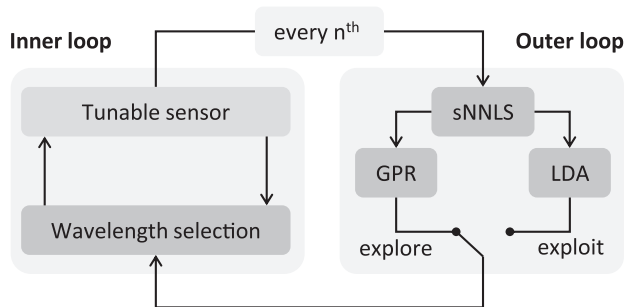

Fig. 2. Building blocks of our active sensing framework for mixture identification. The inner-loop on the left selects one wavelength at a time; the outer loop on the right estimates the concentration of the chemical mixture. Depending on the concentration vector, the algorithm switches between exploration mode (using Gaussian Process Regression) and exploitation mode (using Linear Discriminants Analysis).
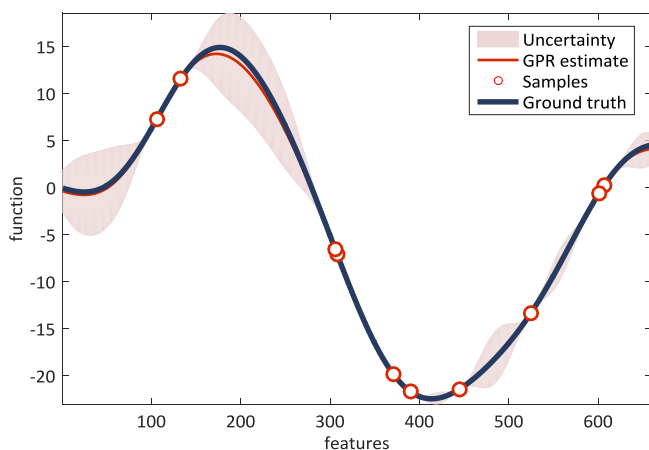

Fig. 3. An example of Gaussian Process Regression (GPR). The goal is to reconstruct the underlying function (ground truth) using a small number of measurements (samples), which GPR accomplishes by taking into account the smoothness of the function.

whose diagonal elements are $\boldsymbol{S}_{GP}^2$. Given input measurements $\boldsymbol{b}_{\lambda_m}$, the best linear unbiased predictor to reconstruct the spectrum can be calculated as [10]:

$$\boldsymbol{b}_{GP} = \mu + \boldsymbol{r}^T \boldsymbol{R}^{-1} (\boldsymbol{b}_{\lambda_m} - \mathbf{1}\mu) \tag{2}$$

where $\boldsymbol{R}$ denotes the $m \times m$ (auto)covariance matrix of the sampled wavelengths λ_m, i.e., $cov(\lambda_m, \lambda_m)$; and $\boldsymbol{r}$ denotes the $m \times M$ covariance matrix between sampled wavelengths λ_m and all wavelengths λ_M, i.e., $cov(\lambda_m, \lambda_M)$. Though possible, calculating the full covariance matrix Σ_{GP} is not necessary in our case because only its diagonal elements are needed for wavelength selection. Accordingly, and

following [10,11], we calculate these diagonal elements as:

$$\boldsymbol{S}_{GP}^2 = \sigma^2 \left(1 - \boldsymbol{r}'\boldsymbol{R}^{-1}\boldsymbol{r} + \frac{(1 - \mathbf{1}'\boldsymbol{R}^{-1}\boldsymbol{r})^2}{\mathbf{1}'\boldsymbol{R}^{-1}\boldsymbol{r}} \right). \tag{3}$$

where $\boldsymbol{R}$, $\boldsymbol{r}$ are the covariance matrices defined above.

Constructing theses matrices is non-trivial because covariance matrices must be positive semi-definite. For this purpose, we define a covariance function[1] $cov(\lambda_i, \lambda_j)$ that calculates the covariance between pairs of wavelengths (λ_i, λ_j) as a weighted sum of three covariance functions, each of which is also positive semi-definite:

$$cov(\lambda_i, \lambda_j) = a_{SE} \exp\left(-(\lambda_i - \lambda_j)^2 / \rho \right) + a_{prod} \boldsymbol{b}_{NNLS}(\lambda_i) \cdot \boldsymbol{b}_{NNLS}(\lambda_j)$$
$$+ \delta_{ij}\sigma^2 \tag{4}$$

where $\boldsymbol{b}_{NNLS}$ is the projection from the concentration $\boldsymbol{x}$ estimated via shrinkage non-negative least squares (sNNLS): $\boldsymbol{b}_{NNLS} = \boldsymbol{Ax}$. Section 2.2 describes sNNLS in detail.

These three terms allow us to impose constraints on the covariance matrix based on the smoothness of the spectrum, previous measurements, and sensor noise, respectively:

- The first component, $exp(-(\lambda_i - \lambda_j)^2/\rho)$, is known as the squared exponential covariance function [12]. It captures the global smoothness of the spectrum, which is determined by the effective resolution of the sensor and the optics. As such, we assume that this term and its smoothing parameter (ρ) are analyte independent.
- The second component, $\boldsymbol{b}_{NNLS}(\lambda_i) \cdot_{NNLS}(\lambda_j)$, is known as the product covariance function [13]. It allows us to incorporates information from all previous measurements, as captured by the current estimate of the spectrum $\boldsymbol{b}_{NNLS} = \boldsymbol{Ax}$.
- The third component, $\delta_{ij}\sigma^2$, captures sensor noise. The higher the noise, the less credible each observation is, and the less strongly the model responds to each new observation. We measure sensor noise level beforehand, assuming that it is analyte independent. We add this term to the diagonal of the covariance matrix with $\delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$.

2.1.1. Wavelength selection criterion

We guide wavelength selection using the variance $\boldsymbol{S}_{GP}^2 = \{\sigma_{\lambda_1}^2, \sigma_{\lambda_2}^2, \ldots, \sigma_{\lambda_M}^2\}$ in equation (3). Namely, we select the wavelength that maximally reduces this variance. Writing the

---

[1] Note that all covariance matrices ($\boldsymbol{R_0}$, $\boldsymbol{R}$, and $\boldsymbol{r}$) are constructed using this covariance function $cov(\lambda_i, \lambda_j)$.

variance as $\mathbf{S}_{GP}^2(\lambda_m)$, we compute the total reduction in variance if wavelength λ_i were sampled as:

$$\Delta_{\sigma_i^2} = 1' \times \left(\mathbf{S}^2(\lambda_{\mathbf{m}}) - \mathbf{S}^2(\lambda_{\mathbf{m} \cup i}) \right), \tag{5}$$

where 1 is a column vector of ones. Using Δ_{σ}^2 as the utility[2] of each wavelength, we select the next wavelength randomly but with probability[3]:

$$p(\lambda_i) \sim \frac{\Delta_{\sigma^2}}{\sum_{i=1}^M \Delta_{\sigma_i^2}}, \tag{6}$$

This strategy allows the sampling process to adapt to previous observations and sample unexplored areas. Our implementation allows each wavelength to be sampled multiple times; this helps de-noise measurements at critical wavelengths by averaging multiple observations.

2.2. Shrinkage non-negative least squares

Following absorption measurements at each new set of n wavelengths, we use non-negative least squares (NNLS) to generate a new solution vector $\mathbf{x}$. Our NNLS solver is based on the original solver by Lawson [14], but uses a post-processing step to sparsify the NNLS solution.[4] Namely, we sequentially eliminate the least significant component in the solution vector (i.e., the one with the lowest concentration) until a measure of model complexity ceases to improve. Two measures of complexity are commonly used: the Akaike information criterion (AIC) [15] and the Bayesian information criterion (BIC) [16]. Both criteria encourage parsimony by penalizing model complexity, but BIC's penalty grows stronger as the number of measurements increases —BIC assumes that only one true model exists.[5] Since our goal is to recover the mixture constituents (which implicitly assumes a true model must exist), we chose BIC for the shrinkage criterion. The BIC score can be calculated as:

$$BIC = -2\log(\mathscr{L}) + n\log(m) \tag{7}$$

where n is the number of non-zero components in the solution $\mathbf{x}$ (a measure of model complexity); m is the number of measurements; and $\mathscr{L}$ is the likelihood of the model, which can be calculated as:

$$\mathscr{L} = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2\sigma^2}(\mathbf{b} - \mathbf{A}\mathbf{x})^T(\mathbf{b} - \mathbf{A}\mathbf{x}) \right\}. \tag{8}$$

where σ is the spread of the Gaussian noise, and $(\mathbf{b}-\mathbf{A}\mathbf{x})^T(\mathbf{b}-\mathbf{A}\mathbf{x})$ is the sum squared error. Table 1 shows the pseudo-code of the BIC guided shrinkage method.

2.3. Exploitative stage

The goal of the initial explorative stage (section 2.1) is to reconstruct the spectrum as closely as possible. As such, the explorative stage has a tendency to overfit the observations by adding new constituents (false-positives) to the solution. This problem occurs when the ground truth is sparse (i.e., the mixture consists of only a few components) and becomes worse as the size of the reference library grows. To address this issue, we switch to an exploitative strategy at the later stages of sensing that allows us to further sparsify the sNNLS solution. Our approach works as follows. Recall that sNNLS generates an overfitting solution $\mathbf{x}'$, and then sparsifies it to $\mathbf{x}$ via the shrinkage method in Table 1. Chemical constituents eliminated during this shrinkage step are potential confounders (i.e., false-positives). Accordingly, during the exploitation stage we select wavelengths according to their ability to discriminate between these false-positives and the remaining constituents in the solution, which we treat as an approximation of the ground truth. Fig. 4 illustrates this process. Let $\mathbf{x}'$ (p' non-zero entries) be the overfitting sNNLS solution, and $\mathbf{x}$ ($p < p'$ non-zero entries) be the sparsified sNNLS solution. Let $\mathbf{x}_{\mathbb{Z}}$ be the concentration of the eliminated entries $\mathbf{x}_{\mathbb{Z}} = \{x_{\mathbb{Z}_1}, x_{\mathbb{Z}_2}, ..., x_{\mathbb{Z}_z}\}$ with indices $\mathbb{Z} = \{\mathbb{Z}_1, ..., \mathbb{Z}_z\}$. To determine the next wavelength to be sampled, we project each eliminated component back to absorbance $\mathbf{b}_{\mathbb{Z}_i} = \mathbf{A}_{\mathbb{Z}_i}\mathbf{x}_{\mathbb{Z}_i}$ where $\mathbf{A}_{\mathbb{Z}_i}$ is the corresponding $\mathbb{Z}_i th$ column vector in the library matrix $\mathbf{A}$. Then, we calculate the Fisher's Linear Discriminant Analysis (LDA) solution for the binary discrimination problem ($class_1 = \{\mathbf{b}\}$, $class_2 = \{\mathbf{b}_{\mathbb{Z}_1}, ..., \mathbf{b}_{\mathbb{Z}_z}\}$), where $\mathbf{b} = \mathbf{b}' - \sum_{i=1}^z \mathbf{b}_{\mathbb{Z}_i}$. The LDA solution (a rotation vector $\mathbf{w}$) provides the direction of maximum discrimination between the final mixture $\mathbf{b}$ and the eliminated components $\{\mathbf{b}_{\mathbb{Z}_1}, ..., \mathbf{b}_{\mathbb{Z}_z}\}$. Accordingly, wavelength selection during the exploitative stage follows a random sampling scheme with sampling probability proportional to the absolute value of the linear discriminant $|\mathbf{w}|$:

$$p(\lambda_i) \sim \frac{|w_i|}{\sum |w_i|}. \tag{9}$$

2.4. Switching between exploration and exploitation

The transition from exploration to exploitation (and vice versa) is signaled by the complexity of the solution. Namely, exploration continues for as long as the complexity of the solution continues to increase when more wavelengths are added. Denoting by $p^{(t)}$ the number of non-zero elements at step t, exploration continues for as long as $p^{(t)} > p^{(t-1)}$, and exploitation starts whenever $p^{(t)} \leq p^{(t-1)}$. The algorithm can return at any time from exploitation to exploration (i.e., if $p^{(t)} > p^{(t-1)}$), though in practice this rarely happens.

Estimation results for GPR and sNNLS are illustrated in Fig. 5(a–b) for low-resolution IR spectra from a Fabry-Perot Interferometer and a library with 8 chemicals —see section 3 for details. The example in Fig. 5(a) shows a case where GPR approximates the ground truth accurately, whereas that in Fig. 5(b) shows a case where it does not. In both cases, however, the sNLLS reconstruction is nearly indistinguishable from the ground truth, regardless of the GPR reconstruction error. The reason for this result is that GPR is only guided by the covariance matrix, whereas the sNNLS solver has access to the library matrix $\mathbf{A}$, which introduces additional constraints about which solutions are feasible. This result is further illustrated in Fig. 5(c,d) for high-resolution Fourier Transform Infrared (FTIR) spectra and a library with 500 chemicals —see section 5 for details. Despite the large discrepancy between GPR and ground truth in certain regions of the spectrum, the sNNLS solver is able to find a (sparse) solution vector that matches the ground-truth —the example in Fig. 5(d) was chosen to show that the GPR and ground-truth curves are actually different, which may not be obvious from the other three

---

[2] The utility is a scalar that indicates the importance of each wavelength at a particular time, in terms of how much uncertainty would be reduced if absorption at that wavelength was measured; see equation (5).

[3] To initialize the explorative stage of wavelength selection, the first wavelength is selected randomly from the uniform distribution $p(\lambda_i) = 1/M$.

[4] Lawson's NNLS solver generates a feasible solution that normally fits the observations within machine precision. However, for mixture identification with noisy observations, overfitting often leads to the addition of unneeded components (i.e. false positives) to the solution vector $\mathbf{x}$.

[5] Asymptotically, the BIC score reaches the lowest point when a true model is found [16].
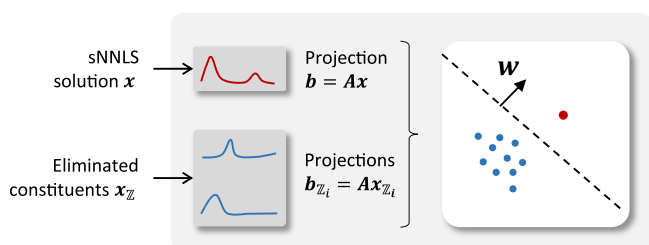
Table 1
Pseudo-code for the BIC guided shrinkage procedure.

```
Input: solution x′, observation b
Output: new sparsified solution x
 1  procedure BIC_shrinkage(x′, b)
 2    do
 3      BIC′ ← getBIC(x′,b) // equation (7)(8)
 4      x ← eliminate(x′)
 5      BIC ← getBIC(x,b) // equation (7)(8)
 6      x′ ← x
 7    while  BIC ≤ BIC'
 8    x ← x′
 9    return x
10  end
11
12  /** Find the minimal non-zero element and set it to zero **/
13  sub-procedure eliminate(x)
14    minX ← ∞
15    minI ← 0
16    for i ← 0 to length(x)-1
17      if minX > x(i) and x(i)>0
18        minxX ← x(i)
19        minI ← i
20      endif
21    endfor
22    x′ ← x
23    x′(minI) ← 0
24    return x′
  end
```


Fig. 4. Illustration of exploitative wavelength selection. Wavelengths are selected to maximize discrimination between the reconstructed spectrum for the sparse sNNLS solution and the reconstructed spectrum of those constituents that were removed from the solution during the sparsification step.

figures.

3. Materials

We evaluated the proposed active wavelength selection algorithm on a prototype chemical detection instrument based on a tunable Fabry-Perot interferometer (FPI). Illustrated in Fig. 6(a), the instrument consists of a broadband IR lamp that illuminates the sample in a gas cell, where IR radiation is absorbed at different wavelengths depending on the chemical makeup of the sample; a focusing lens that projects the transmitted radiation onto the FPI sensor; and the FPI sensor itself, which can be tuned to sample one
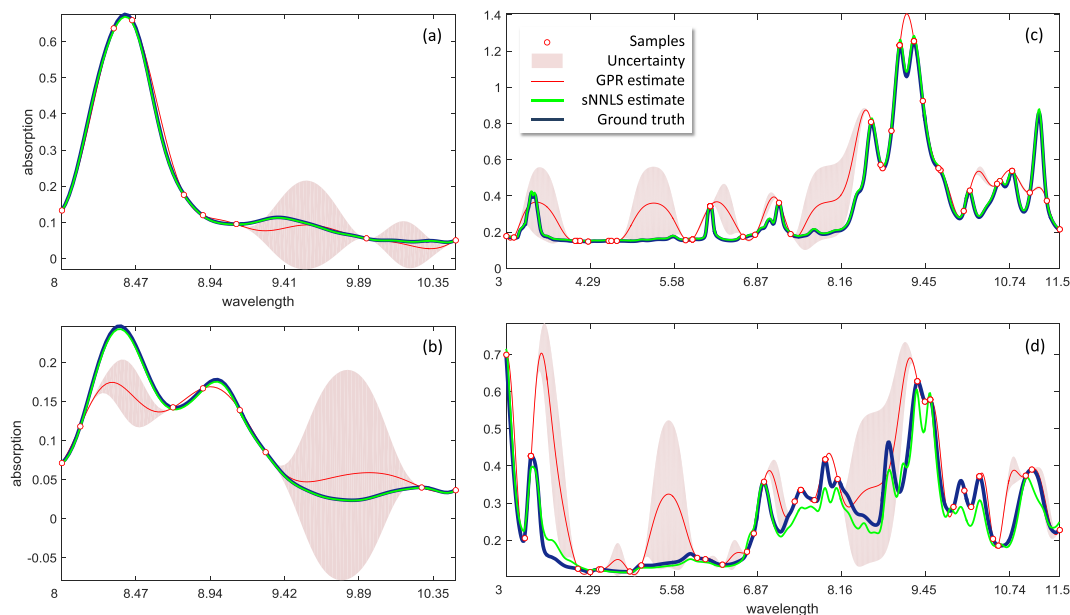
Fig. 5. Reconstruction results on (a–b) low-resolution spectra from the Fabry-Perot Interferometer described in section 3, and (c–d) high-resolution spectra from the NIST Webbook database [17].
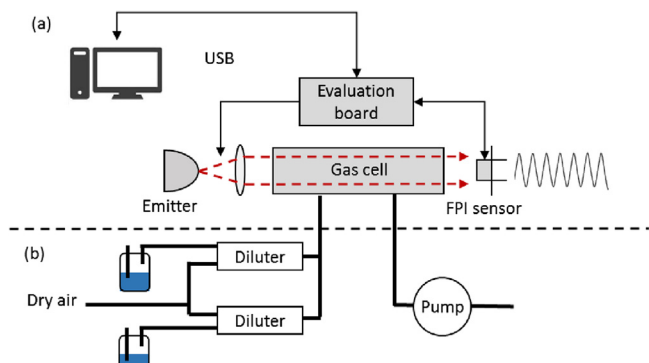
Fig. 6. Diagram of (a) the sensing system and (b) the sample delivery system.

Table 2
List of chemicals used in the experiments, and their major components.

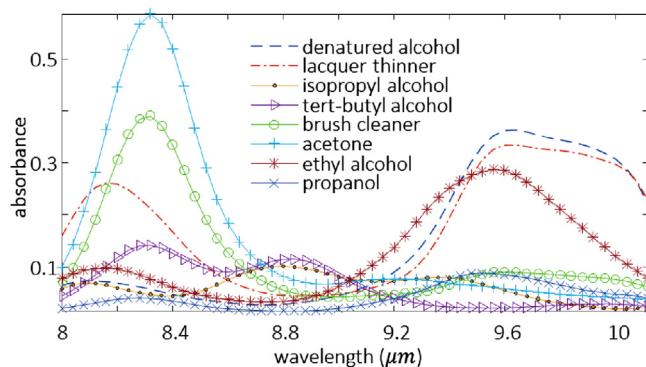| Chemical label	Components
Propanol	Propanol
Acetone	Acetone
Ethyl alcohol	Ethyl alcohol
Isopropyl alcohol	Isopropyl alcohol
Tert-Butyl alcohol	*Tert*-butyl alcohol
Denatured alcohol	Ethyl alcohol, methanol
Brush cleaner	Raffinates, acetone, methanol
Lacquer thinner	Toluene, methanol, hexane, light aliphatic naphtha

Fig. 7. The spectra of the eight chemicals collected from the FPI sensor.

individual wavelength at a time by adjusting the distance between the two mirrors in the FPI. For the experiments described here, we used a mid-infrared FPI (LFP-80105, InfraTec GmbH) with 107 tunings (absorption lines) in the range 8−10.5 μm, coupled with a collimated broadband IR source (INTX 20-1000-R; EOC, Inc.) modulated at 10 Hz and 50% duty cycle. We mounted a 10 cm gas cell (66001-10A; Specac, Inc.) with ZnSe window (602L08; Specac, Inc.) and a ZnSe focusing lens (LA7542F, Thorlabs, Inc.). The FPI, IR source and sample cell were mounted onto an opto-mechanics fixture (Thorlabs, Inc.) to ensure precise alignment. The FPI device was controlled using Matlab™ through a USB evaluation board provided by the vendor. We demodulated the FPI signals using the discrete-time Fourier transform (DTFT[6]):

$$P_f = T \sum_{n=-\infty}^{\infty} s(nT) \cdot e^{-i2\pi fnT}. \qquad (10)$$

where f corresponds to the frequency of interest (10 Hz), $s(\cdot)$ represents the time-series signal from the FPI, which contains the 10 Hz modulation from the IR source, T corresponds to the sampling interval, and n denotes the sample index.

Gas samples are delivered to the sample cell as illustrated in Fig. 6(b). Vapors from the headspace of 30 mm glass vials are drawn using negative pressure with a pump connected downstream from the sample cell. The pump is modulated at 0.125 Hz with 20% duty cycle to avoid exhausting the headspace and therefore keep the sample concentration relatively stable. Two diluters (1010 precision gas diluter, Custom Sensor Solutions, Inc.) independently mix the foreground and background sample vapors with dry air. Since water and carbon dioxide have major peaks outside of the FPI range (8 − 10.5 μm), air has a negligible contribution to the spectrum.

3.1. Chemical library and mixtures

For experimental validation purposes, we selected eight VOCs − five pure chemicals and three household paint-thinners − with absorption peaks in the range of our sensor (8−10.5 μm). Table 2 lists the eight chemicals and their effective components. Sample spectra from these chemicals, as measured by the FPI, are shown in Fig. 7. Experiments were conducted in a laboratory environment at a temperature of 22.2 °C and standard atmospheric pressure of 760 mmHg.

The large number of potential mixtures for these eight

chemicals ($2^8−1 = 255$) means that only a small portion of these testing mixtures can be tested, and even fewer of them resolved given the low spectral resolution of our FPI device. Rather than selecting mixtures randomly from this large space, we instead decided to carefully select a small number of mixtures that would allow us to characterize the active-sensing framework across a broad range of problems: from simple mixtures (single component, distinctive) to complex mixtures (multiple components, and less distinctive). Specifically, we use condition numbers[7] to measure the correlation among spectra within a set A. Given a mixture with components S, we compute their contribution to the condition number of the library A as:

$$D_S = cond(A) − cond(A_{-S}) \qquad (11)$$

where $cond(\cdot)$ denotes condition number, A contains all the reference spectra and A_{-S} contains all reference spectra *except* the S constituents in the mixture. Intuitively, the higher this number is, the more correlated these S analytes are to the rest of the constituents in the library, and therefore the harder it is to distinguish them from other constituents.

Using this measure, we ranked the 255 possible mixtures and selected a subset of them for further testing (shown in Table 3). Mixtures with difficulty scores above $D_S > 120$ were not considered since they cannot be identified reliably using the low-resolution spectra provided by our FPI device. Instead, we supplement our experimental validation (section 4) with results on synthetic data from a database of high-resolution spectra from Fourier-transform infrared spectrometry (section 5).

---

[6] We used the continuous frequency DTFT in contrast to the more common discrete frequency DTFT because the target frequency needs to coincide with the modulating frequency (10 Hz) to achieve the highest accuracy.

[7] The condition number is an indication of how stable a linear system is: how much the estimation x can change given small variations in the observation b. It is calculated as $cond(A) = \sigma_{max}(A)/\sigma_{min}(A)$, where σ_{max} and σ_{min} are the maximum and minimum singular values of A, respectively. If matrix A has collinear columns (i.e., the underlying chemicals are similar), it will have a large condition number.

Table 3
Analytes and mixtures used for experimental validation along with their Difficulty scores, computed according to equation (11).

Analyte	Abbreviation	# Components	D_S
Tert-Butyl alcohol	TBA	1	48
Ethyl alcohol	EA	1	55
Isopropyl alcohol	IA	1	57
Tert-Butyl alcohol + brush cleaner	TBA + BC	2	65
Denatured alcohol	DA	1	75
Denatured alcohol + *tert*-butyl alcohol	DA + TBA	2	96
Lacquer thinner	LT	1	98
Lacquer thinner + isopropyl alcohol	LT + ISA	2	105
Brush cleaner + acetone	BC + ACT	2	120

4. Results on experimental data

In a first experiment, we examined the global smoothness parameter ρ in equation (4). For this purpose, we collected five replicate spectra for each chemical (5×8 spectra) at 100% concentration to achieve the highest signal-to-noise ratio. For each setting of parameter ρ, we generated a smoothed spectrum for each replicate. Then, for each chemical, we removed one of the 5 replicates and computed the mean squared error (MSE) between that replicate and the average spectrum of the remaining four replicates (i.e., in a leave-one-out fashion). We repeated this process with parameters ranging from $\rho = 0.2$ μm to $\rho = 2$ μm for each chemical, and calculated the average MSE as the performance metric. As illustrated in Fig. 8, a smoothing value of $\rho = 1.2$ μm provides the optimal performance, which suggests that our sensor has an effective resolution of 1.2 μm when GPR is used for reconstruction. This value of the smoothness parameter was used for the remaining experiments.

In a second experiment, we compared our active wavelength-selection algorithm against a state-of-the-art passive algorithm, which served as a baseline. The passive algorithm was the successive projection algorithm (SPA) of Araújo et al. [7], which we implemented in Matlab™ following the description in Ref. [7]. SPA is a greedy algorithm that uses the sequential orthogonal projections of the Gram-Schmidt procedure. SPA iteratively adds one wavelength at a time: the one that is minimally correlated to the previously selected wavelengths. In this way, SPA ensures that the selected set of wavelengths is minimally redundant. To keep other variables invariant, we only replace the wavelength selection algorithm (that is, the inner loop in Fig. 2) with SPA.

To avoid making assumptions about which chemical is present, we trained SPA on reference spectra from the eight chemicals in our library, collected using our FPI device. The acquired SPA wavelength sequence then replaced the inner-loop in our framework. For evaluation purposes, we ran both algorithms until each converged
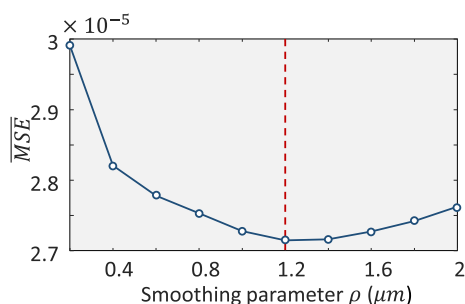
to the ground truth, which we defined as the algorithm identifying the analyte correctly for ten steps in a row, or until an upper limit of 200 steps was reached, in which case the algorithms were halted. Thus, the number of wavelengths used by each method was different, depending on how many observations were needed to reach convergence. We compared the two algorithms on the basis of three criteria:

- **Efficiency**, measured as the total number of steps needed to converge (excluding the ten steps required for confirmation). Instances where the algorithm reached the 200-step limit were not included in this measure. The lower this number, the more efficient the wavelength selection strategy is.
- **Stability**: measured as the standard deviation of the number of steps required to converge. As with efficiency, instances where the algorithm reached the 200-step limit were not considered. The lower the standard deviation is, the more stable the algorithm is.
- **Reliability**: measured as the proportion of times the algorithm converged to the correct solution within the maximum of 200 steps; i.e. converging to the wrong solution was treated as failure to converge. The higher the classification rate is, the more reliable the algorithm is.

We tested both algorithms 25 times for each of the analytes in Table 3, for a total of $9 \times 25 = 225$ tests. The sequence of tests was selected randomly to eliminate ordering effects, and the gas cell was purged with air before each test to avoid residual buildup. Fig. 9(a–c) summarizes the results in terms of efficiency, stability, and reliability, respectively. Our active framework outperforms SPA across analytes both in terms of efficiency and stability. However, the performance gap diminishes as the complexity of the analyte increases. This is an expected result because the more complex chemical identification problems require an increasingly broader coverage of the spectrum, which eventually becomes equivalent to using a passive strategy. Efficiency and stability allow us to compare the two strategies when they reach a solution within the allotted 200 steps. In contrast, the last measure (reliability) considers all instances – those for which the algorithm converged, and those for which it did not – to determine how often each particular strategy found the correct solution. The results in Fig. 9(c) show that the active strategy has a higher likelihood of finding the correct solution than SPA. While SPA struggles to find the correct solution, even
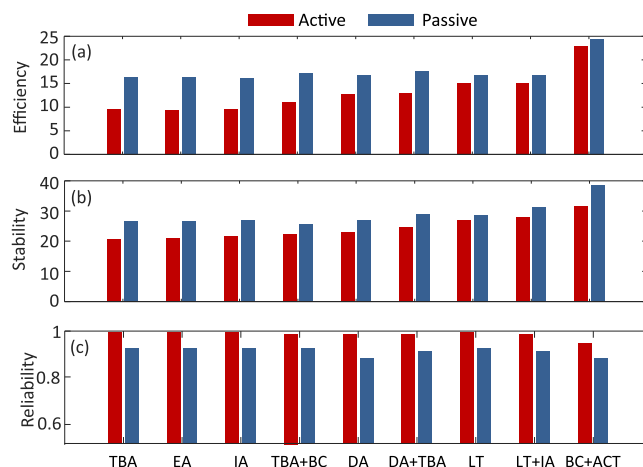

Fig. 8. Reconstruction error as a function of the smoothing parameter (ρ) in equation (4).

Fig. 9. Performance comparison between active and passive wavelength selection on experimental data in terms of (a) efficiency, (b) stability, and (c) reliability.

for simple analytes (such as TBA, EA, IA), the active strategy identifies the correct chemical the vast majority of the times.

5. Results on synthetic data

To provide a more thorough evaluation than what is possible experimentally, we also analyzed the active sensing algorithm on a large dataset of synthetic IR spectra. The dataset contained Fourier Transform Infrared (FTIR) spectra with 660 spectral lines from 500 chemicals in the NIST Webbook database [17]. To simulate the potential spectral resolution of FPIs, we convolved the FTIR spectra with a Gaussian filter of 0.1 μm spread. Each spectrum was then normalized to sum up to one. For the experiments that follow, we compared our active sensing algorithm against SPA. In all cases, we allowed the algorithms to sample each wavelength multiple times.

The difficulty of a mixture problem can vary dramatically, e.g., a badly conditioned two-component mixture can be unsolvable while a well-conditioned 20-component mixture can be easily identified. For this reason, we designed a mixture construction policy so that the chosen problems would be neither trivial nor unsolvable. Namely, we randomly selected a large number of 50-component[8] mixtures and calculated their classification rate with a set noise level.[9] We then selected five mixtures that could be correctly classified 1%–10% of the times. For each of these five 50-component mixtures, we sequentially removed one component at a time to form chemical mixtures of a lower order; this process ensures a gradual transition in problem complexity from hard to easy. For each of the resulting 250 mixtures (50×5), we evaluated the active and passive algorithms 40 times, each time with randomly added noise, for 10,000 cases. The maximum number of allowable measurements was set to 5,000.

Following procedures described in section 4, we used the average number of steps to converge, variance, and classification rate as measures of efficiency, stability, and reliability, respectively. Fig. 10(a–b) shows the average and standard deviation of the number of measurements required to converge, averaged over 200 tests (40×5) per mixture. These results are consistent with those in the experimental section, and show that active sensing outperforms its passive counterpart across all analytes. As before, the "active-sensing advantage" diminishes as the number of constituents in the mixture grows; for a mixture of 50 chemicals, active and passive algorithms have similar efficiency. In terms of stability, however, the active algorithm remains significantly more stable than the passive approach across all mixtures. This is an added advantage of using an approach that adapts to the characteristics of each analyte. Lastly, Fig. 10(c) shows the reliability of the two algorithms, measured as the classification rate across all tests —regardless of whether or not they converged. Active sensing maintains a 100% classification rate up until 50-component mixtures, whereas its passive counterpart fluctuates and rarely reaches a 100% classification rate —see Fig. 10(d) for a close-up view. Both algorithms collapse for mixtures containing more than 50 components,[10] which suggests that the measurements have reached the intrinsic dimensionality of the underlying linear system.

In a final experiment, we examined the exploration and exploitation stages independently. If we consider the concentration
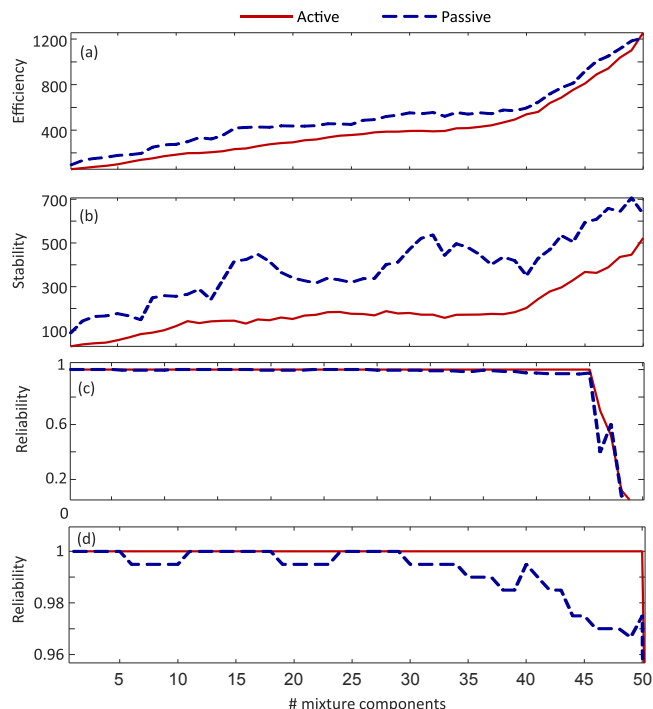

Fig. 10. Performance comparison between active and passive algorithms as a function of the number of components in the mixture: (a) efficiency, (b) stability, and (c) reliability. (d) is a close-up view of (c).
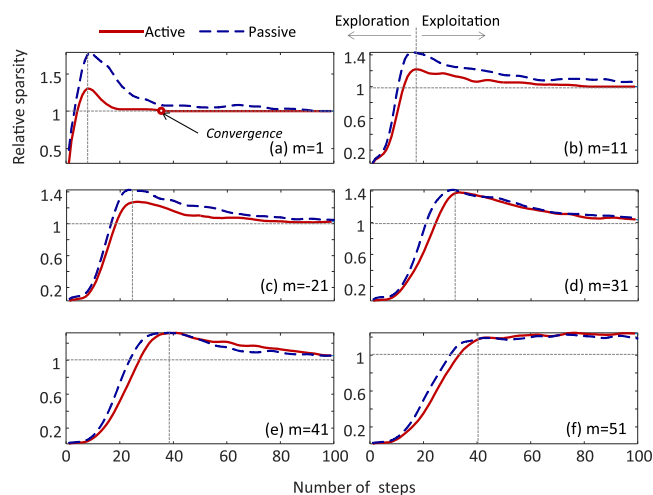

Fig. 11. Relative sparsity of the solution for the active and passive algorithms during the first 100 steps for mixtures with increasing number of components $m = \{1, 11, 21, 31, 41, 51\}$.

space $\boldsymbol{x}$, the exploration stage tends to add more constituents to reduce the reconstruction error of the spectrum, whereas the exploitative stage tends to remove constituents to accelerate convergence. Thus, the sparsity of the solution vector $\boldsymbol{x}$ is a good indicator of how well each stage performs. Let the relative sparsity of solution $\boldsymbol{x}$ be measured as $p = \|\boldsymbol{x}\|_0 / \|\boldsymbol{x}_{true}\|_0$ where $\boldsymbol{x}_{true}$ denotes the ground truth and $\|\cdot\|_0$ denotes the l_0 norm of a vector (i.e., number of non-zero entries). Fig. 11(a) shows the average relative sparsity for 1-component analytes when we increase the number of steps from 1 to 100 for both methods. During exploration, the complexity of the solution increases monotonically until

---

[8] We also evaluated mixtures containing more than 50 constituents. However, classification rates for both algorithms collapse to nearly zero as the number of constituents goes beyond 50. This suggests that the solver has reached the maximum effective resolvability of the spectral library.

[9] We added white noise with standard deviation at 1% of the median value of all absorption spectra in the library.

[10] In reality, because of nonlinear interactions between constituents, the linear model we are using here will not be reliable enough for higher-order mixtures.
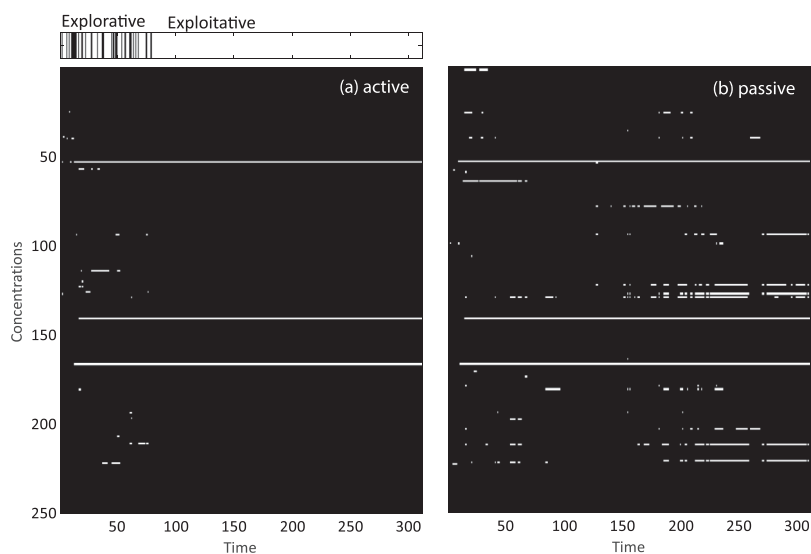
Fig. 12. Evolution of the solution space for (a) active and (a) passive wavelength selection. The inset on top of (a) denotes when the active sensing algorithm switches between explorative and exploitative search.

approximately eight measurements; this result suggests that as few as eight steps (out of 660 wavelengths) suffice to capture the overall structure of the spectrum using GPR. Compared to the passive algorithm, active sensing grows the solution more slowly during exploration, which reduces the overshoot and consequently accelerates convergence during the exploitation stage. During the subsequent exploitation stage, the active algorithm arrives at the ground truth relatively early (less than 40 steps) whereas the passive method takes an extra 60 steps. Fig. 11(b—f) shows the average sparsity over time for mixtures containing 11, 21, 31, 41 and 51 components. The "active-sensing advantage" is most prominent at lower-order mixtures; as the complexity of the mixture grows, both methods take more steps to convergence. In the extreme case of a 51-component mixture, the exploitative stage becomes ineffective. This result suggests that the system has reached the maximum resolvability of the spectrum for the given spectral library and noise level. Interestingly, even in this extreme case, the explorative stage maintains a more controlled complexity growth rate than the passive algorithm.

Finally, we illustrate how the solution space evolves for the two algorithms: passive and active. For this purpose we choose a 15 chemical mixture problem, and only show the first 250 chemicals in the solution space. Results are shown in Fig. 12; white horizontal traces represent non-zero elements in the solution. As shown, the active sensing algorithm converges faster to the solution and its intermediate solutions contain fewer false positives than those of the passive sensing algorithm. As active sensing aggressively leverages any intermediate results to sparsify the false-positives, additional measurements are beneficial. In contrast, since the passive algorithm always assumes each constituent is equally likely, it is more likely to introduce irrelevant wavelengths as the number of sensing steps increases.

6. Conclusions

Off-line wavelength selection strategies work well for multi-component calibration, where the identity of the target analytes is known and one seeks to estimate their relative concentration in the mixture. Because the optimal set of wavelengths is analyte

dependent, knowledge of the mixture components ensures that an optimal set does exist. Wavelength selection becomes more challenging when the mixture components are unknown, as is the case in mixture identification problems. In mixture identification, the goal is to identify the constituents in a mixture from among a long list of chemicals in a large spectral library. The combinatorial nature of this problem makes it a poor match for off-line wavelength selection strategies. Instead, mixture identification problems require an adaptive strategy to wavelength selection.

To address this need, we have proposed an active sensing strategy that interleaves wavelength selection with the sensing process. The algorithm operates in two distinctive regimes: an exploration stage (based on Gaussian processes) that selects wavelengths to minimize the reconstruction error of the spectrum; and an exploitation stage (based on Linear Discriminant Analysis) that selects wavelengths to sparsify the intermediate least squares solutions. Both stages are unsupervised, and do not require the conventional training-validation process. The method is also computationally efficient, making it suitable for portable platforms with limited computation resources.

We evaluated our approach on experimental data from a Fabry-Perot interferometer and synthetic data from high-resolution FTIR spectra, and compared it against a state-of-the-art passive strategy (the successive projection algorithm). For the experimental comparison, we use single components and binary mixtures from a library of eight chemicals. For a more comprehensive evaluation, we used synthetic data containing up to 50-component mixtures from a library of five hundred chemicals. Both experimental and simulation results suggest that the active approach outperforms passive approach in terms of efficiency, stability, and reliability. Our results — both experimental and in simulation — show an "active-sensing advantage" for mixtures containing a few[11] components.

---

[11] In our experience, this number is in the order of 10% of the number of chemicals in the library, but the number depends on the complexity of the library. On a library containing many similar spectra the number would be lower than in a library in which spectra are orthogonal.

Acknowledgments

References

[1] C.H. Spiegelman, M.J. McShane, M.J. Goetz, M. Motamedi, Q.L. Yue, G.L. Gote, Theoretical justification of wavelength selection in PLS Calibration: development of a new algorithm, Anal. Chem. 70 (1998) 35—44.

[2] R.M. Balabin, S.V. Smirnov, Variable selection in near-infrared spectroscopy: benchmarking of feature selection methods on biodiesel data, Anal. Chim. Acta 692 (2011) 63—72.

[3] P.M. Narendra, K. Fukunaga, A branch and bound algorithm for feature subset selection, Comput. IEEE Trans. 100 (1977) 917—922.

[4] R. Leardi, Genetic algorithms in chemometrics and chemistry: a review, J. Chemom. 15 (2001) 559—569.

[5] H. Swierengaa, F. Wülfertb, O.E. de Noordc, A.P. de Weijerd, A.K. Smildeb, L.M.C. Buydens, Development of robust calibration models in near infra-red spectrometric applications, Anal. Chim. Acta 311 (2000) 121—135.

[6] M. Shamsipur, V. Zare-Shahabadi, B. Hemmateenejad, M. Akhond, Ant colony optimisation: a powerful tool for wavelength selection, J. Chemom. 20 (2006) 146—157.

[7] M.C.U. Araújo, T.C.B. Saldanha, R.K.H. Galvão, T. Yoneyama, H.C. Chame, V. Visani, The successive projections algorithm for variable selection in spectroscopic multicomponent analysis, Chemom. Intell. Lab. Syst. 57 (2001) 65—73.

[8] W. Cai, Y. Li, X. Shao, A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra, Chemom. Intell. Lab. Syst. 90 (2008) 188—194.

[9] J. Huang, R. Gutierrez-Osuna, Active wavelength selection for mixture analysis with tunable infrared detectors, Sens. Actuat. B Chem. 208 (2015) 245—257.

[10] D.R. Jones, M. Schonlau, W.J. Welch, Efficient global optimization of expensive black-box functions, J. Glob. Optim. 13 (1998) 455—492.

[11] J. Sacks, W.J. Welch, T.J. Mitchell, H.P. Wynn, Design and analysis of computer experiments, Stat. Sci. (1989) 409—423.

[12] C.E. Rasmussen, C.K.I. Williams, Squared exponential covariance function, in: Gaussian Processes for Machine Learning, The MIT Press, 2006, pp. 83—84.

[13] C.E. Rasmussen, C.K.I. Williams, Dot product covariance functions, in: Gaussian Processes for Machine Learning, The MIT Press, 2006, pp. 89—90.

[14] C.L. Lawson, R.J. Hanson, Problem NNLS, in: Solving Least Squares Problems, Prentice-hall, Inc, 1974, pp. 160—165.

[15] H. Akaike, A new look at the statistical model identification, Automat. Control, IEEE Trans. 19 (1974) 716—723.

[16] G. Schwarz, Estimating the dimension of a model, Ann. Stat. 6 (1978) 461—464.

[17] P. M. Chu, F. R. Guenther, G. C. Rhoderick, and W. J. Lafferty, "Quantitative infrared database," in NIST Chemistry WebBook, NIST Standard Reference Database Number 69, N. Eds. P.J. Linstrom and W.G. Mallard, Ed., ed.