Research Article Vol. 33, No. 24 / 1 Dec 2025 / Optics Express 50078
Optics EXPRESS

Partial least squares regression analysis of natural gas-like mixtures absorption spectra collected using a mid-infrared supercontinuum broadband source

ANDREA ZIFARELLI,[1,*] PAOLA FORMICA,[1] ANGELO SAMPAOLO,[1,2] HONGPENG WU,[2] LEI DONG,[2] VINCENZO SPAGNOLO,[1,2] AND PIETRO PATIMISCO[1,2]

[1]*Dipartimento Interateneo di Fisica, Polytechnic of Bari & University of Bari, Via G. Amendola 173, Bari 70125, Italy*
[2]*State Key Laboratory of Quantum Optics and Quantum Optics Devices, Institute of Laser Spectroscopy, Shanxi University, Taiyuan, 030006, China*
[*]*andrea.zifarelli@uniba.it*

Abstract: We report on a broadband gas sensor based on direct absorption spectroscopy in the mid-infrared range for the simultaneous detection of methane, ethane, and propane in natural gas-like mixtures. The system employs a broadband supercontinuum light source, coupled with an absorption cell and an optical spectrum analyzer with a resolution of 0.5 cm^{-1}. This configuration enables reconstruction of the full absorption bands of the target alkanes, which exhibit significant spectral overlap in the 2.8-3.2 μm spectral region. A comparative study between multiple linear regression (MLR) and partial least-squares regression (PLSR) was conducted to determine the concentration of each individual component. The results highlight the superior performance of PLSR in the presence of unbalanced concentration ratios (1:10) among the three alkanes, achieving mean prediction accuracy of 98%, 93% and 94% for methane, ethane, and propane, respectively.

1. Introduction

Natural gas (NG) is a crucial energy resource composed primarily of light hydrocarbons, with methane (C1) as the dominant component, followed by ethane (C2) and propane (C3), and varying amounts of higher alkanes, nitrogen, carbon dioxide, and other trace gases [1]. The relative concentrations of these hydrocarbon species provide valuable insights into the NG characteristics, such as wetness and density, as well as for fluid classification and fingerprinting [2,3]. Once NG is refined for commercial use, its composition is continuously monitored to determine its calorific value, which reflects its combustibility and, consequently, its quality [4]. Gas chromatography (GC) represents nowadays the standard technique for assessing NG composition, recognized by regulatory agencies as the reference method. In combination with thermal conductivity and flame-ionization detectors, GC enables precise measurements of NG component concentrations by calibrating the system with certified reference standards [5–7]. However, due to the complex nature of NG, GC methods often require the use of several detectors, multiple columns, and backflushing techniques to ensure accurate detection of all the individual components found in NG [8]. Moreover, this approach is often limited by its bulky instrumentation, operational complexity, and response time, not suitable for on field or real-time monitoring scenarios.

In recent years, optical methods have gained prominence in the field of complex gas samples analysis due to their potential for fast, non-invasive, and highly sensitive measurements without the use of reagents and solvents [9–11]. For these reasons, optical-based approaches have been recently introduced in the field of NG analysis [12,13]. Among the others, infrared spectroscopy

#568700 https://doi.org/10.1364/OE.568700
Journal © 2025 Received 26 May 2025; revised 3 Aug 2025; accepted 10 Aug 2025; published 18 Nov 2025

Optics EXPRESS

stands out as a powerful optical technique for identifying and quantifying molecular species based on their characteristic vibrational transitions [14]. Mid-infrared (MIR) spectral range is particularly interesting for NG characterization because of the presence of fundamental C-H stretching modes within the 3-4 μm wavelength range, leading to intense absorption bands for multiple hydrocarbons [15,16]. The maturity of interband cascade lasers in this spectral region has enabled various hydrocarbon sensing techniques, such as tunable diode laser absorption spectroscopy (TDLAS) [17–19], photoacoustic spectroscopy (PAS) [20–22], and cavity-enhanced detection schemes [23–25]. However, their narrow spectral range requires isolated strong absorption lines for each species, which becomes challenging in gas mixtures like NG, where overlapping features and highly unbalanced concentrations can cause minor components to be masked by dominant ones.

The reconstruction of the full absorption band could help in better discrimination among the different constituents, even under significant concentration disparities. In this context, broadband light sources hold significant potential for reconstructing the complete absorption bands of light hydrocarbons. Near-infrared (NIR) spectrometric methods have already been proposed, employing for example a cost-effective handheld NIR spectrometer combined with a tungsten lamp and a flow cell for the quantification of methane, ethane, and propane in natural gas and biogas [26]. Moving towards the MIR region, recent studies have explored the use of frequency combs and supercontinuum (SC) sources as a new generation of broadband light sources for laser absorption spectroscopy [27]. SCs are broadband light sources generated through nonlinear optical processes in specially engineered fibers, offering continuous spectral coverage over several microns. In the mid-infrared region, the phenomenon of supercontinuum generation is typically achieved using fluoride [28] or chalcogenide fibers [29,30], enabling high-resolution and multi-component spectroscopic analysis. Their broad bandwidth makes them particularly suitable for analyzing complex gas mixtures with overlapping absorption features. Standard detection methods involve the use of a Fourier transform spectrometer or a scanning-grating-based spectrometer [31,32] Although dual-comb spectroscopy offers high spectral resolution and sensitivity, its instrumental complexity and high cost hinder field deployment. Supercontinuum sources, on the other hand, provide a simpler and more cost-effective alternative for multi-species gas detection. One example is a fully integrated and transportable sensor employing a MIR supercontinuum source spanning the 2–4 μm range. Thanks to its broad spectral coverage, the system enabled simultaneous detection of multiple VOCs using a grating-based spectrometer combined with two thermoelectrically cooled HgCdTe photodetectors [33], or with a MIR-to-NIR upconverter [34].

To benefit from the spectral information provided by the broadband sources, the spectroscopic measurements are typically coupled with machine learning-based multivariate analysis (MVA) [35,36]. Among the others, partial least squares regression (PLSR) represents a solid and reliable MVA approach for spectroscopic analysis, being based on the correlation removal and dimensionality reduction by evaluating the so-called “latent variables” underlying the collected data [37–39]. For these reasons, this MVA approach has been employed in several applications aiming to deconvolve the spectral contribution of overlapping absorption peaks [18–42].

In this work, a broadband mid-infrared supercontinuum source was employed to accurately resolve overlapping spectral features through direct absorption spectroscopy. Gas samples were passed through an absorption cell, and the spectral components of the transmitted light were analyzed using an optical spectrum analyzer employed as Fourier transform spectrometer. To mimic natural gas mixtures, laboratory samples were prepared at a 1:10 dilution using certified concentrations of the three hydrocarbons and a gas blending system, with highly unbalanced concentration ratios. This approach introduced the challenge of spectral deconvolution, where the total absorption spectrum must be mathematically separated into individual components spectra. This required multivariate calibration techniques relying on well-characterized set of

Optics EXPRESS

calibration samples with known compositions, which are used to build predictive models for individual component quantification.

2. Experimental setup

A schematic of the experimental setup is shown in Fig. 1.

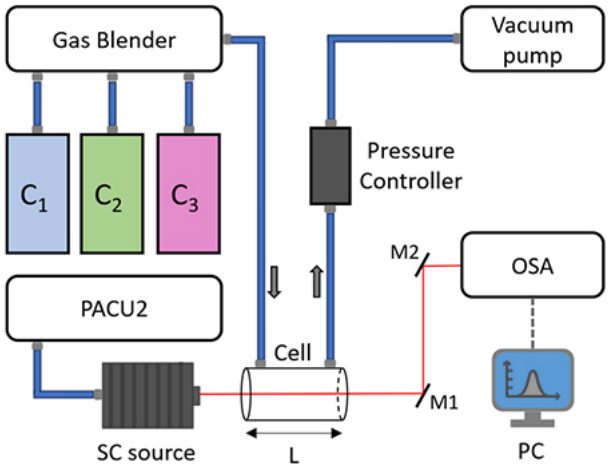

Fig. 1. Schematic of the employed experimental setup. Blue solid lines represent gas line connections, red solid line represents the optical beam, and grey dashed line represents USB connection. OSA, Optical Spectrum Analyzer; PACU2, pure air circulator unit; SC, supercontinuum source.

A broadband supercontinuum (SC) source (Thorlabs SC4500) was used as light source, providing a spectral emission covering the 1.3–4.5 μm range [43]. To ensure long-term stability and prevent degradation of the internal components, the SC source was connected to a pure air circulator unit (Thorlabs, PACU2), which continuously supplies dry air into the SC source. Two plane mirrors, M1 and M2 in Fig. 1, were used to direct the collimated output beam through a gas absorption cell with an optical path length $L = 11.6$ cm, equipped with two CaF_2 windows. The light transmitted through the cell was collected and analyzed by an optical spectrum analyzer (Thorlabs OSA207C), connected to a personal computer for data acquisition. The spectral resolution of the employed spectrometer, defined according to the Rayleigh criterion, was 0.5 cm^{-1}. This spectral resolution was selected to guarantee a high signal-to-noise ratio within the broadband spectral region targeted for the analysis. The OSA207C was operated with an averaging factor of 10 to improve the signal-to-noise ratio, and a Hann apodization function was applied to improve baseline stability. Considering the averaging operations, each sample was characterized by an acquisition time of 30 s to collect the spectrum. Background measurements were collected prior to each acquisition.

The gas line system consists of a gas blender (MCQ Instruments, GB100 Plus), a pressure controller (ALICAT scientific MCS3) and a vacuum pump. Three gas cylinders with certified concentration, 10% C1 in N_2, 1% C2 in N_2, and 1% C3 in N_2, respectively, were used for the generation of different gas mixtures. The gas cylinders were provided with a 3σ expanded uncertainty of 4% of the nominal concentration. Each gas cylinder was individually connected to the gas blender inlet, together with a pure nitrogen cylinder used for dilutions. In this way, different gas mixtures were dynamically prepared and introduced into the gas cell at a constant total flow rate of 100 sccm, at pressure value of 750 Torr. The pressure inside the cell was regulated using the pressure controller positioned downstream of the gas cell. Before each

Optics EXPRESS

measurement, the gas cell was evacuated using the vacuum pump to eliminate any residual gases, and background spectra were acquired. Spectroscopic measurements were performed by recording the transmitted light spectrum using the OSA.

3. Calibration procedure

Although the SC source spans a broad spectral range, the regions of interest for this study were determined by properly selecting the ranges where the highest and more relevant absorption bands of C1, C2 and C3 occur, i.e., in the spectral region around 3.3 μm [16]. For each analyte, the optimal spectral range was tailored to extract the maximum information, thus avoiding flat, near-zero absorbance regions. The absorbance spectra shown hereafter were calculated comparing the transmitted light spectrum measured by the optical spectrum analyzer with and without the target gas, I_t and I_0, respectively, according to the Lambert-Beer law equation:

$$I_t(\lambda) = I_0(\lambda) \cdot e^{-\alpha_i(\lambda) \cdot L} = \quad I_0(\lambda) \cdot e^{-A_i(\lambda)} \tag{1}$$

where $\alpha_i(\lambda)$ is the absorption coefficient of the i-th analyte at wavelength λ, L is the optical pathlength, and $A_i(\lambda)$ is the related absorbance.

Calibration measurements for the three analytes were performed individually by generating diluted gas mixtures. The cylinders with certified concentration were enabled one at a time at the gas blender inlet, with pure nitrogen N_2 used as the dilution gas. For each analyte, different concentrations were obtained by adjusting the relative flow rates at the gas blender input while maintaining a constant output flow. The absorbance spectra were retrieved at different concentrations of the analyte in the mixture, and the area under the relevant absorption features was calculated. Then, the area under the absorbance curve was computed as the sum of elementary trapezoids, each having as base the wavenumber step between two adjacent points and a pair of adjacent absorbance values as heights. The resulting values were plotted as a function of the corresponding gas concentrations to obtain the calibration curve for the analyte under investigation. The zero-point value of this dataset was determined based on the absorbance spectrum obtained when the cell was empty, at a pressure below 20 Torr. The spectrum recorded under this condition serves as the background, above which any absorption signal can be identified. In fact, given the large volume of the cell, the most effective method for eliminating all potential absorbers (including water vapor) is to reduce the pressure to the lowest achievable level. In this condition, the integrated absorbance area was measured to be 0.004 in arbitrary units (a.u.).

Figure 2(a) reports the absorbance spectrum (in absorbance units, abs. u., to be distinguished from the a.u. used for area estimation) in the spectral range of 2850 cm^{-1} to 3175 cm^{-1}, measured when a C1 concentration of 10% in N_2 was flowing through the gas cell. Figure 2(b) shows the calibration curve in terms of the integrated absorbance area, with C1 concentrations spanning from 2% to 10%.

A clear linear trend is observed between the integrated absorbance area and the C1 concentration, indicating that the Beer-Lambert law (Eq. (1)) governing optical absorption in the gas cell can be linearized under conditions of weak absorption when considering the whole absorption spectrum. A linear fit imposed to the experimental dataset yields a slope of 4.391 a.u./%, which corresponds to the sensitivity of the sensor system in detecting C1. The error bars on the data points were calculated as propagation of uncertainty considering the area of a right trapezoid with the absorbance values of two adjacent points as the bases, and the respective height.

Similarly, Fig. 3(a) reports the C2 absorption spectrum in the range of 2850 cm^{-1} to 3055 cm^{-1} referring to a gas mixture of 1% C2 in N_2, together with the C2 calibration curve in Fig. 3(b). The linear fit to the experimental data provides a sensitivity for C2 detection of 28.59 a.u./%, confirming the expected higher sensitivity relative to methane. Indeed, although Fig. 2(a) shows that methane exhibits stronger peak intensities compared to ethane in Fig. 3(a), the ethane absorption band is broader and more irregular in shape. This extended spectral profile

Optics EXPRESS

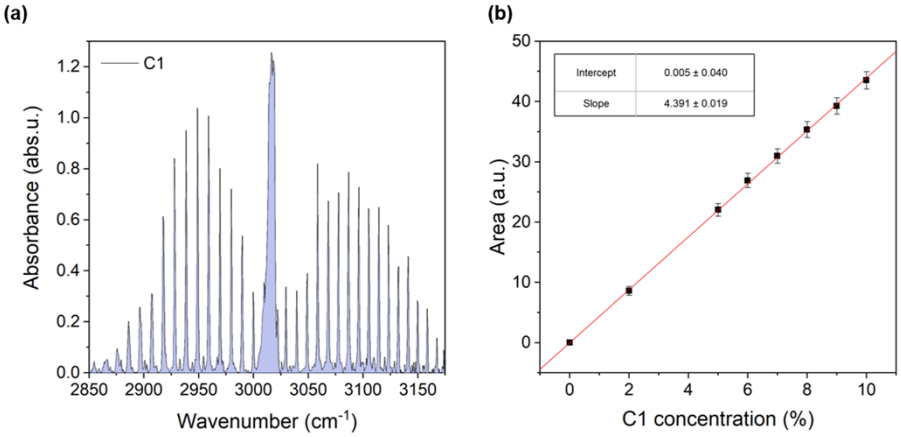

Fig. 2. (a) Absorbance spectrum in abs. u. (absorbance unit) acquired with a mixture of 10% of methane in nitrogen. (b) Area as a function of the C1 concentration (black dots) with a linear fit (red line). Linear fit parameters (intercept and slope) are shown in the insets.

significantly contributes to the total integrated absorption area, as most wavenumbers exhibit non-zero absorbance. In contrast, the regions between the narrow peaks in the methane spectrum contribute negligibly to the overall area, due to their near-zero absorbance.

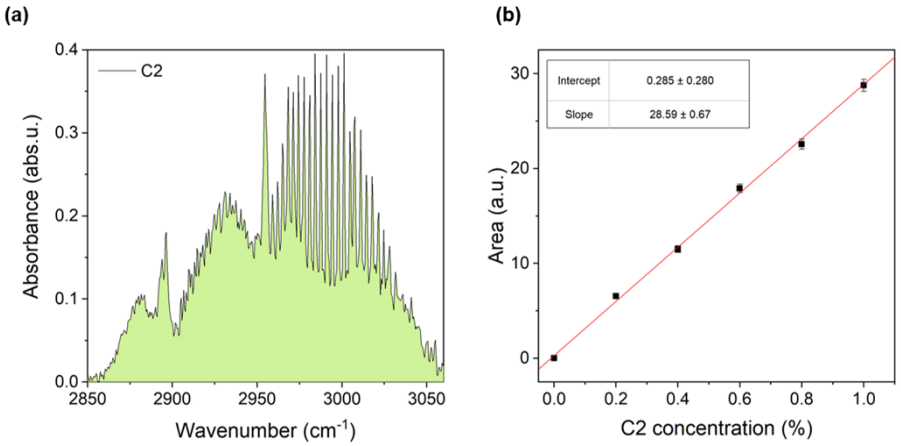

Fig. 3. (a) Absorbance spectrum of ethane. (b) Area as a function of the C2 concentration (black dots) with a linear fit (red line). Linear fit parameters (intercept and slope) are shown in the insets.

Finally, the absorbance spectrum in the spectral range of 2850 cm^{-1} to 3030 cm^{-1} was acquired with a 1% C3:N_2 mixture in the absorption cell and is shown in Fig. 4(a).

The linear fit of the integrated absorbance area at different C3 concentrations in Fig. 3(b) represents the calibration curve for C3 detection with a sensitivity of 47.60 a.u./%.

Optics EXPRESS

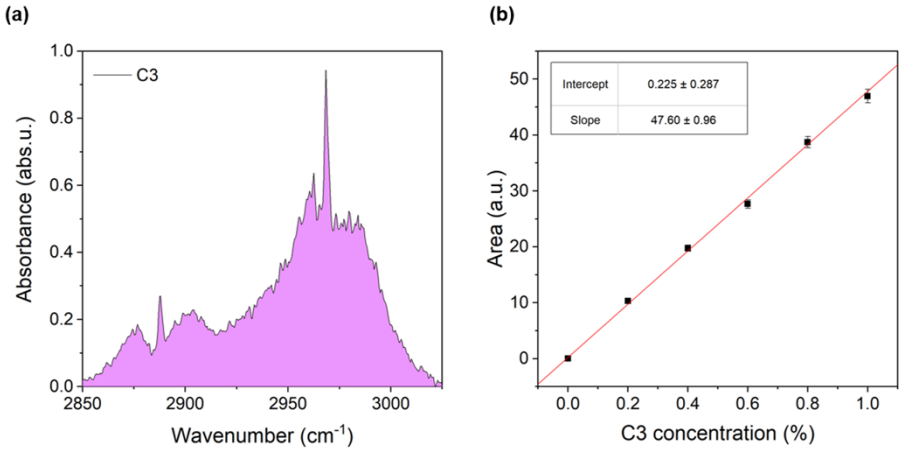

Fig. 4. (a) Absorbance spectrum of propane. (b) Area as a function of the C3 concentration (black dots) with a linear fit (red line). Linear fit parameters (intercept and slope) are shown in the insets.

4. Selection of multivariate regression approach

The calibration curves obtained in the previous sections constitute the basis for the quantitative analysis of gas mixtures containing two or three alkanes simultaneously. To simulate a typical natural gas matrix, the gas samples will be deliberately prepared with a highly unbalanced composition, featuring C1 at significantly higher concentrations than C2 and C3, mimicking the natural gas composition in a 1:10 dilution. The primary reason for using dilution is that excessively high concentrations can contaminate the absorption cell by promoting the adhesion of substances to its internal walls. In such cases, a simple cleaning with a vacuum pump may not be sufficient to remove all residues, which can compromise the accuracy of concentration measurements for each gas mixture analyzed. This imbalance poses considerable challenges in the quantitative interpretation of overlapping absorption spectra, particularly in accurately resolving the spectral contributions of the minor components. To evaluate the most suitable computational approach for this task, two multivariate regression techniques were implemented: Multiple Linear Regression (MLR) and Partial Least Squares Regression (PLSR).

4.1. Multiple linear regression

MLR models the relationship between the spectral response of a mixture and the concentrations of its analytes by expressing the response as a linear combination of independent variables, under the assumption that the absorbance at each wavenumber is linearly dependent on the concentration of the analytes [44]. To build the regression model, individual reference spectra for each of the three gases at known nominal concentrations were used: 10% C1:N_2, 1% C2:N_2 and 1% C3:N_2, respectively, as reported in Fig. 2(a), 3(a) and 4(a). Then, the absorbance spectra of the mixtures were then recorded within the spectral range from 2850 cm^{-1} to 3175 cm^{-1}, to include the main absorption bands of all target gases. The underlying idea in MLR is that the absorbance spectrum A_{TOT} of a test gas mixture containing the three analytes C1, C2 and C3 can be expressed as the weighted sum of the reference spectra:

$$A_{TOT}(\lambda) = A_{C1}(\lambda) \cdot \frac{c_{C1}}{c^r_{C1}} + A_{C2}(\lambda) \cdot \frac{c_{C2}}{c^r_{C}} + A_{C3}(\lambda) \cdot \frac{c_{C3}}{c^r_{C3}} \tag{2}$$

Optics EXPRESS

In this expression, each c_i in the numerator represents the unknown concentration to be estimated, while c_i^r with superscript "r" refers to the reference concentration used to obtain the corresponding $A_i(\lambda)$.

The c_i values can be determined by solving the linear regression problem:

$$\boldsymbol{Y} = \boldsymbol{X} \cdot \boldsymbol{B} + \boldsymbol{E} \tag{3}$$

where $\boldsymbol{Y}$ is the matrix of the dependent variables, i.e., the concentration of each analyte, $\boldsymbol{X}$ is the matrix of the independent variables, i.e., the spectral data points, $\boldsymbol{B}$ is the regression coefficient matrix, and $\boldsymbol{E}$ is the residuals matrix. In MLR, the $\boldsymbol{B}$ matrix is estimated by means of the ordinary least squares methods [45]. Finally, these parameters are used to estimate the c_i concentrations. However, this approach is characterized by several limitations, providing unstable regression weights and poor repeatability mainly due to correlation among the collected data [46].

4.2. Partial least squares regression

The PLSR model offers an alternative to traditional MLR, with improved capability to handle highly correlated spectral data, making it particularly suitable for complex mixtures [37]. This approach requires a training phase using the full set of calibration spectra collected for each analyte at different concentrations. This method relies on two matrices: the predictor matrix $\mathbf{X}$, built from the absorbance data, and a response matrix $\mathbf{Y}$, containing the corresponding gas concentration values. The algorithm reduces the dimensionality of the data by projecting it into a latent space, decomposing both matrices according to the following relationships:

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^{\boldsymbol{T}} + \boldsymbol{E} \tag{4}$$

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{Q}^{\boldsymbol{T}} + \boldsymbol{F} \tag{5}$$

where $\boldsymbol{T}$ and $\boldsymbol{U}$ are the score matrix for $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively containing the latent variables (LVs); $\boldsymbol{P}$ and $\boldsymbol{Q}$ are the loading matrices for $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively; and $\boldsymbol{E}$ and $\boldsymbol{F}$ are the residual matrices. For consistency, the spectral interval of 2850–3175 cm^{-1} was also selected for this analysis. Different spectral range were also considered, e.g., the interval 2850–3005 cm^{-1} to reduce the influence of dominant C1 features, but no increase in prediction accuracy was observed (see Supplement 1 A).

To identify the optimal number of LVs, which are a linear combination of the predictors, a 10-fold cross-validation approach was applied [47]. The performance of the model was assessed by evaluating both the cumulative explained variance and the root mean square error (RMSE) as a function of the number of PLS components, as shown in Figs. 5(a) and 5(b), respectively.

The results indicate that over 99% of the variance is captured by the first four components, while the RMSE significantly decreases up to the fourth latent variable (LV4) and then stabilizes. Therefore, four PLS components were identified as a good compromise, adequately capturing the relevant variance while minimizing the prediction error and avoiding overfitting. The corresponding loading plots, shown in Fig. 5(c), clearly prove that the first three LVs capture meaningful spectral contributions across the wavenumber range coming from the three analytes.

Although LV4 appears relatively flat, it still captures spectral characteristics related to C2 that would not be captured using only three LVs.

Once the latent space is defined, predictions can be evaluated using a regression model that relates $\boldsymbol{Y}$ to $\boldsymbol{X}$ through the score matrix $\boldsymbol{T}$. The resulting model is expressed as:

$$\boldsymbol{Y} = \boldsymbol{\beta} \cdot \boldsymbol{X} \tag{6}$$

where $\boldsymbol{\beta}$ is the matrix of regression coefficients derived from the LVs.

Optics EXPRESS

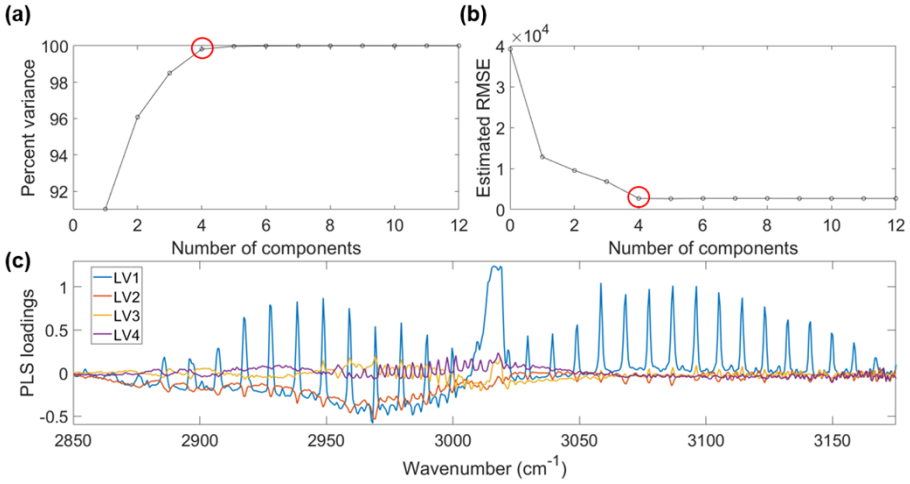

Fig. 5. Percent variance (a) and RMSE (b) as functions of the number of components obtained using the 10-fold CV approach. (c) Visualization of the contribution given from each latent variable (LV).

5. Results

This section presents the results of the quantitative analysis conducted on laboratory-prepared binary and ternary mixtures with concentrations representative of the typical natural gas composition in dilution 1:10. Both MLR and PLSR approaches were employed to interpret complex and overlapped the broadband absorbance spectra and estimate the concentrations of the individual analytes.

The two regression methods were applied to the same experimental datasets, and the predicted concentrations were compared to the nominal reference values to evaluate the accuracy and reliability of each model.

5.1. Binary mixtures

Two types of binary mixtures were prepared in laboratory using the gas blender with certified cylinders (Fig. 1): one combining C1 and C2, and the other with combinations of C1 and C3. Table 1 reports the concentrations predicted by MLR and PLSR, along with the corresponding nominal (expected) values and relative accuracies for two representative test sets.

The uncertainties in the predicted concentrations were calculated from the standard deviation of the regression parameters derived from residual variance. For the PLSR model, the Root Mean Squared Error of Calibration (RMSEC) was used to quantify the average deviation between the predicted and actual concentrations within the training set. Uncertainties in the expected values were calculated by accounting for the nominal concentration tolerances of each gas cylinder and an additional 1% uncertainty associated with the gas mixer setpoint.

Figure 6(a) reports the measured spectrum of the 7% C1: 0.3% C2:N_2 mixture together with the simulated spectrum reconstructed by summing the individual analyte contributions estimated using PLSR, which are separately shown in Fig. 6(b).

For comparison, the same plots are shown in Fig. 7 using concentrations extracted through the MLR method for the same C1 and C2 mixture. In this case, the methane concentration appears to be overestimated, resulting in a better reconstruction of the peak at 3018 cm^{-1}, but at the cost of a worse fit of the P and R branch peaks.

Optics EXPRESS

Table 1. Comparison between the MLR and PLSR predicted and the expected concentrations for the different test set of the two-species natural gas-type mixture (C1 and C2) between 2850 cm^{-1} and 3175 cm^{-1}. Expected concentrations are reported with the calculated uncertainty, MLR predicted concentrations are reported with calculated RMSE of regression parameters, and PLSR predicted concentrations are reported with calculated RMSEC. Arrow symbols indicate over- and under-estimation of the retrieved concentrations.

			MLR			PLSR		
		Expected (%)	Predicted (%)		Accuracy ($\%_{REL}$)	Predicted (%)		Accuracy ($\%_{REL}$)
Test set #1	C1	8.0 ± 0.2	8.8 ± 0.1	↑	87	8.3 ± 0.5	↑	96
	C2	0.20 ± 0.02	0.16 ± 0.02	↓	79	0.20 ± 0.06	↔	100
	C3	0	0.00 ± 0.01		-	0.01 ± 0.06		-
Test set #2	C1	7.0 ± 0.1	7.8 ± 0.1	↑	88	6.9 ± 0.2	↓	99
	C2	0.30 ± 0.02	0.21 ± 0.02	↓	72	0.29 ± 0.02	↓	95
	C3	0	0.00 ± 0.01		-	0.01 ± 0.02		-

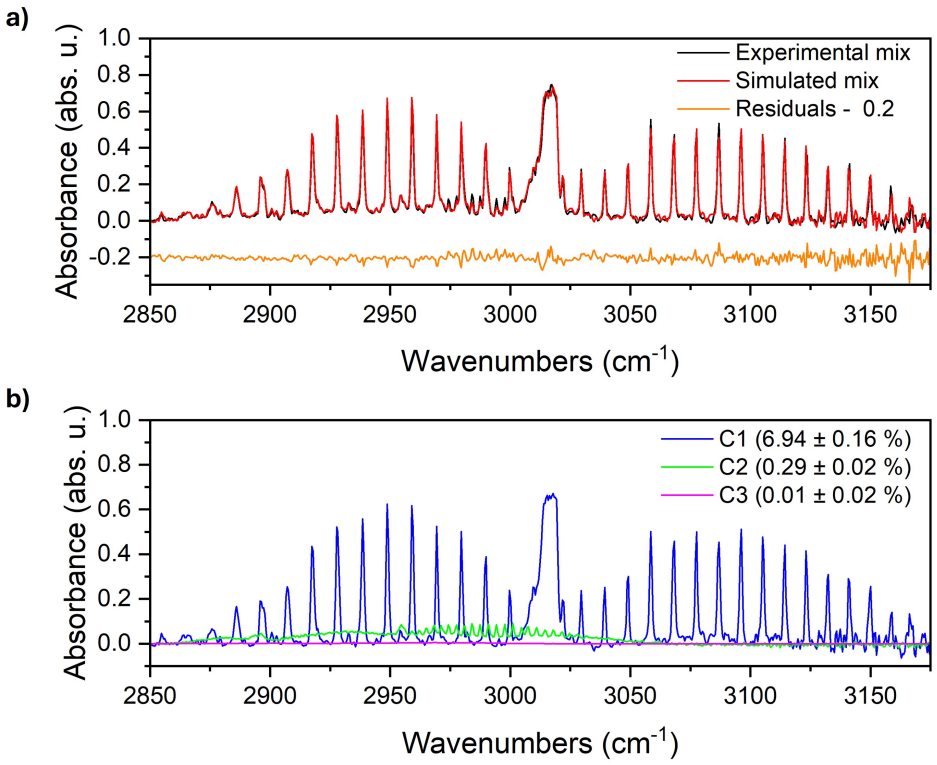

Fig. 6. (a) Comparison between experimental absorbance spectrum of a 7% C1 and 0.3% C2 mixture (black) and simulated spectrum reconstructed from a sum of PLSR estimated concentrations (red). Residuals are plotted with a visual shift of -0.2 (orange) (b) Individual spectral contributions of C1 (blue), C2 (green), and C3 (magenta) based on PLSR estimates.

A similar analysis was conducted with binary mixtures of C1 and C3 in N_2. Table 2 summarizes the predicted concentrations for two representative test sets. As in the previous case, the PLSR method demonstrates high accuracy, exceeding 97%.

Optics EXPRESS

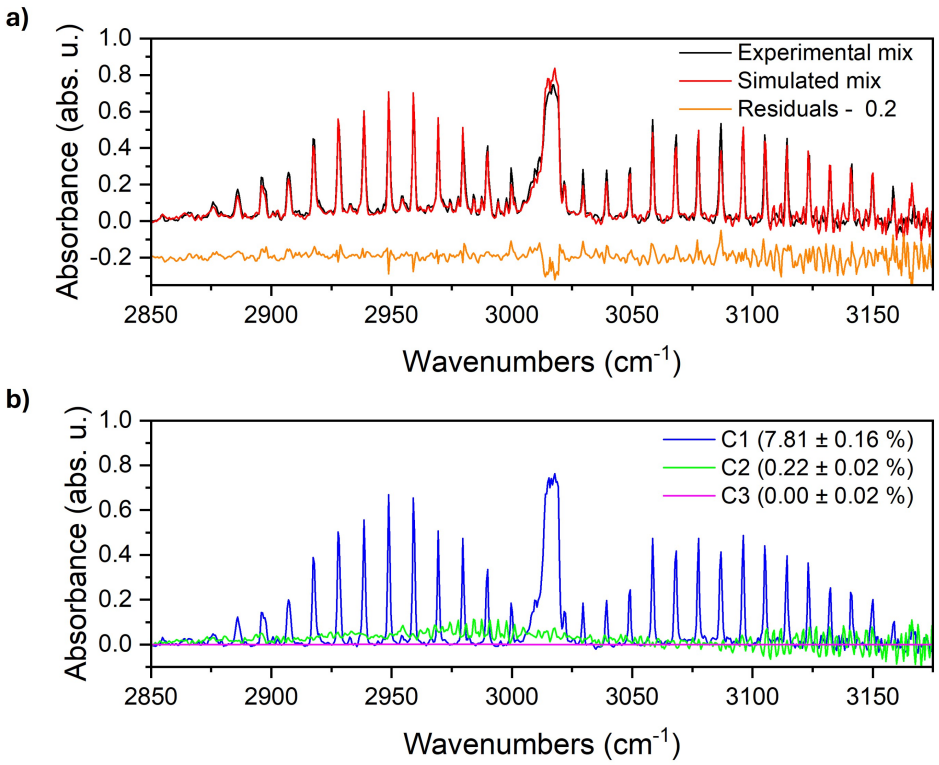

Fig. 7. (a) Comparison between experimental absorbance spectrum of a 7% C1 and 0.3% C2 mixture (black) and simulated spectrum reconstructed from a sum of MLR estimated concentrations (red). Residuals are plotted with a visual shift of -0.2 (orange) (b) Individual spectral contributions of C1 (blue), C2 (green), and C3 (magenta) based on MLR estimates.

Table 2. Comparison between the MLR and PLSR predicted and the expected concentrations for the different test set of the two-species natural gas-type mixture (C1 and C3) between 2850 cm^{-1} and 3175 cm^{-1}. Expected concentrations are reported with the calculated uncertainty, MLR predicted concentrations are reported with calculated RMSE of regression parameters, and PLSR predicted concentrations are reported with calculated RMSEC. Arrow symbols indicate over- and under-estimation of the retrieved concentrations.

			MLR			PLSR		
		Expected (%)	Predicted (%)		Accuracy ($\%_{REL}$)	Predicted (%)		Accuracy ($\%_{REL}$)
Test set #1	C1	9.0 ± 0.2	9.3 ± 0.1	↑	97	8.7 ± 0.2	↓	97
	C2	0	0.00 ± 0.01		-	-0.01 ± 0.02		-
	C3	0.10 ± 0.01	0.07 ± 0.01	↓	68	0.10 ± 0.02	↔	100
Test set #2	C1	7.0 ± 0.1	7.8 ± 0.1	↑	88	6.8 ± 0.4	↓	97
	C2	0	0.00 ± 0.02		-	0.02 ± 0.06		-
	C3	0.30 ± 0.02	0.23 ± 0.01	↓	77	0.29 ± 0.06	↓	97

In contrast, while the MLR method accurately predicts C1 concentrations, as expected due to its much higher levels (9% and 7%), it performs poorly for C3, with accuracies dropping below 70%.

Figure 8 compares the sum of the individual spectral contribution extracted by the PLSR model with the experimental spectrum, both referred to a mixture of 9% C1:0.1% C3:N_2.

Optics EXPRESS

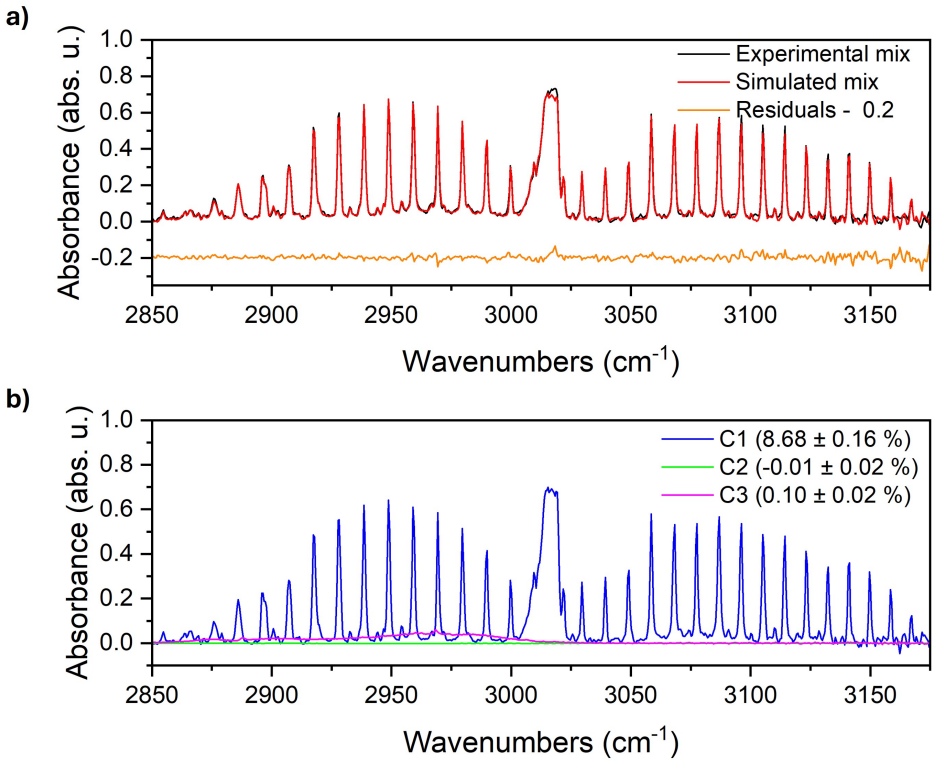

Fig. 8. (a) Comparison between experimental absorbance spectrum of a 9% C1 and 0.1% C3 mixture (black) and simulated spectrum reconstructed from a sum of PLSR estimated concentrations (red). Residuals are plotted with a visual shift of -0.2 (orange) (b) Individual spectral contributions of C1 (blue), C2 (green), and C3 (magenta) based on PLSR estimates.

5.2. Ternary mixtures

A more intricate scenario was explored by preparing ternary mixtures of C1, C2 and C3 at varying concentrations in N_2. Using the same approach as before, Table 3 compares the expected and predicted concentrations of the three analytes for three test sets, along with their respective accuracy values.

Even in mixtures containing the three analytes, PLSR provides higher accuracy in estimating the concentrations of individual components within a natural gas-type mixture. It is worth noting that MLR results are characterized by higher precision compared to PLSR ones, as also reported in Table 1 and 2. However, this degree of precision does not correspond to a high degree of accuracy. This behavior is indicative of overfitting, a well-known limitation of MLR, particularly when applied to large datasets.

An important observation emerges to guide the upcoming discussion. The test set #1, consisting of 8% C1, 0.1% C2, and 0.1% C3, yields only 82% accuracy in predicting the concentration of C2. In contrast, test set #3, with the same concentration of C2 but reduced C1 (7%) and slightly increased C3, achieves nearly perfect prediction accuracy (≈100%) for C2. This suggests that the primary challenge in the prediction capability lies in distinguishing C2 from C1, due to their highly overlapping spectral signatures. Reducing the dominance of C1 in the mixture improves the accuracy of detectability of C2, even when the C3 contribution increases.

Optics EXPRESS

Table 3. Comparison between the MLR and PLSR predicted and the expected concentrations for the different test set of the three-species natural gas-type mixture (C1, C2 and C3) between 2850 cm^{-1} and 3175 cm^{-1}. Expected concentrations are reported with the calculated uncertainty, MLR predicted concentrations are reported with calculated RMSE of regression parameters, and PLSR predicted concentrations are reported with calculated RMSEC. Arrow symbols indicate over- and under-estimation of the retrieved concentrations.

			MLR			PLSR		
		Expected (%)	Predicted (%)		Accuracy ($\%_{REL}$)	Predicted (%)		Accuracy ($\%_{REL}$)
Test set #1	C1	8.0 ± 0.2	8.4 ± 0.1	↑	94	8.1 ± 0.5	↑	99
	C2	0.10 ± 0.01	0.07 ± 0.02	↓	69	0.08 ± 0.06	↓	82
	C3	0.10 ± 0.01	0.09 ± 0.01	↓	90	0.11 ± 0.06	↑	89
Test set #2	C1	7.5 ± 0.2	8.0 ± 0.1	↑	93	7.4 ± 0.5	↓	99
	C2	0.15 ± 0.01	0.11 ± 0.02	↓	72	0.13 ± 0.06	↓	88
	C3	0.10 ± 0.01	0.06 ± 0.01	↓	64	0.09 ± 0.06	↓	90
Test set #3	C1	7.0 ± 0.1	7.6 ± 0.1	↑	91	6.7 ± 0.2	↓	96
	C2	0.10 ± 0.01	0.05 ± 0.02	↓	50	0.10 ± 0.02	↔	100
	C3	0.20 ± 0.02	0.15 ± 0.01	↓	77	0.19 ± 0.02	↓	94

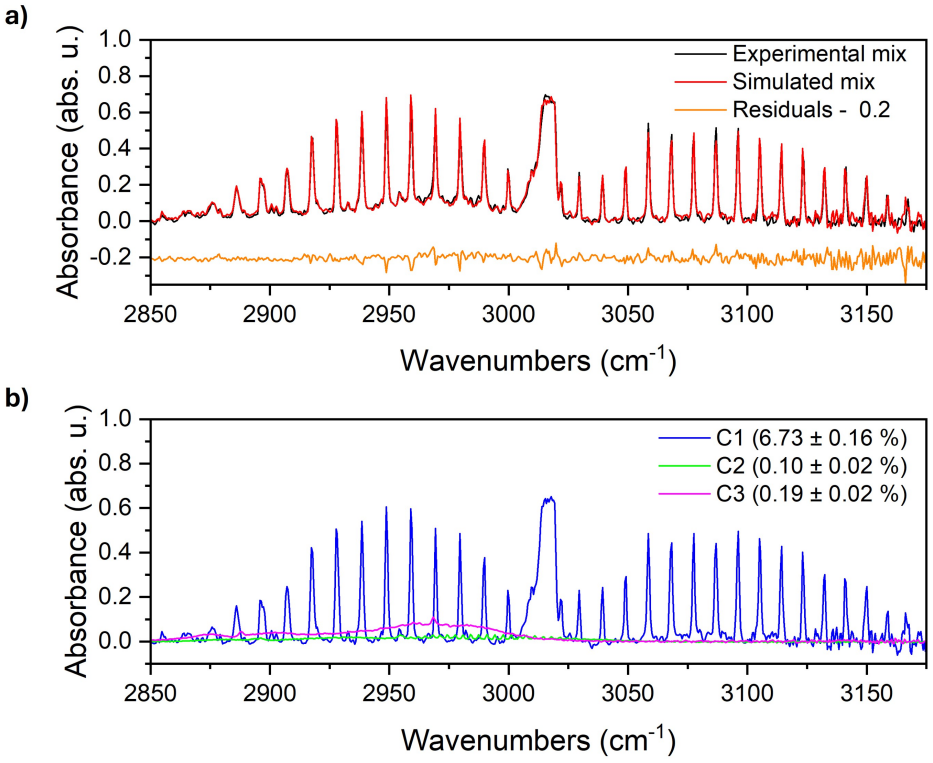

Fig. 9. (a) Comparison between experimental absorbance spectrum of a 7% C1, 0.1% C2 and 0.2% C3 mixture (black) and simulated spectrum reconstructed from a sum of PLSR estimated concentrations (red). Residuals are plotted with a visual shift of -0.2 (orange) (b) Individual spectral contributions of C1 (blue), C2 (green), and C3 (magenta) based on PLSR estimates.

Optics EXPRESS

As in the case of binary mixtures, Fig. 9 reports the comparison between the experimental and PLSR-simulated spectrum for test set #3.

6. Discussion

In the previous analysis, the physical parameter selected to represent the absorbance was the area under the absorbance spectrum. This approach accounts for the cumulative effect of all spectral features, both weak and intense, associated with the analyte, with each one contributing with its weight to the total integrated absorbance. To isolate and assess the individual contributions to the integrated absorbance, the peak intensity of each absorption feature composing the overall spectral signature can be calibrated against the analyte concentration, following the typical approach used in absorption spectroscopy when a narrow-band diode laser targets a specific feature within an absorption band. This is reasonable, as the concentration range is not sufficiently wide to observe a noticeable spectral broadening of individual features with increasing concentration, as experimentally verified, due also to the poor spectral resolution of the OSA207C. Following the observations reported in the previous section, a calibration of absorbance values was performed for the most prominent C1 peaks at 3018 cm^{-1} (Q-branch) and 2959 cm^{-1} (P-branch), as shown in Figs. 10(a) and 10(b), respectively. For comparison, Figs. 10(c) and 10(d) report the corresponding calibration curves for these C1 peaks within a lower concentration range, obtained starting from a certified cylinder with a mixture of 0.5%:C1:N_2 at the gas blender input (Fig. 1).

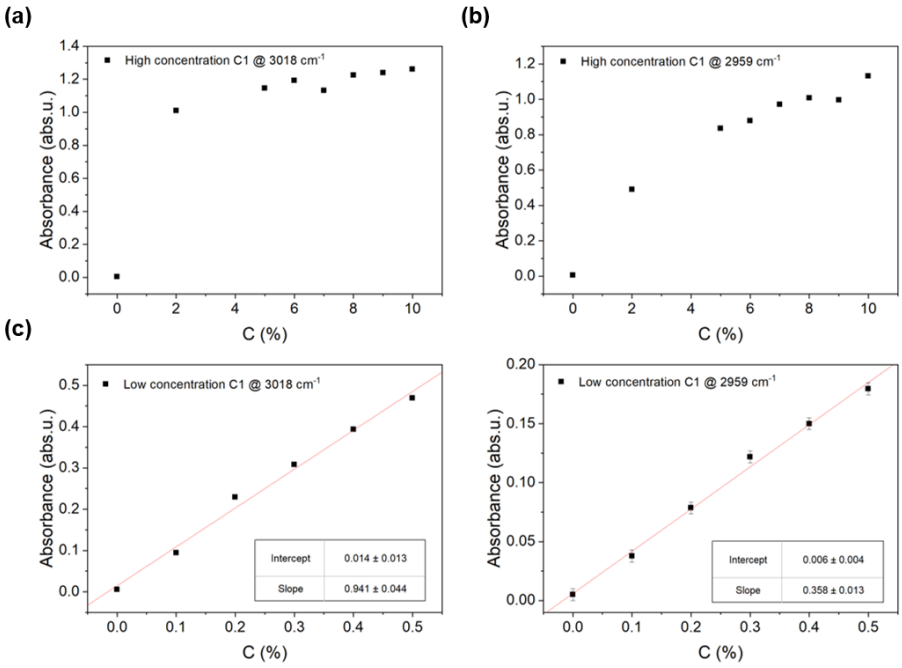

Fig. 10. Calibration curves of absorbance as a function of C1 concentration (black dots) at two characteristic wavenumbers. Panels (a) and (b) show the absorbance response at 3018 cm^{-1} (Q branch) and 2959 cm^{-1} (P branch), respectively, for high C1 concentrations (0–10%). Panels (c) and (d) report the corresponding linear calibration curves at low concentration (0–0.5%) for the same peaks.

Optics EXPRESS

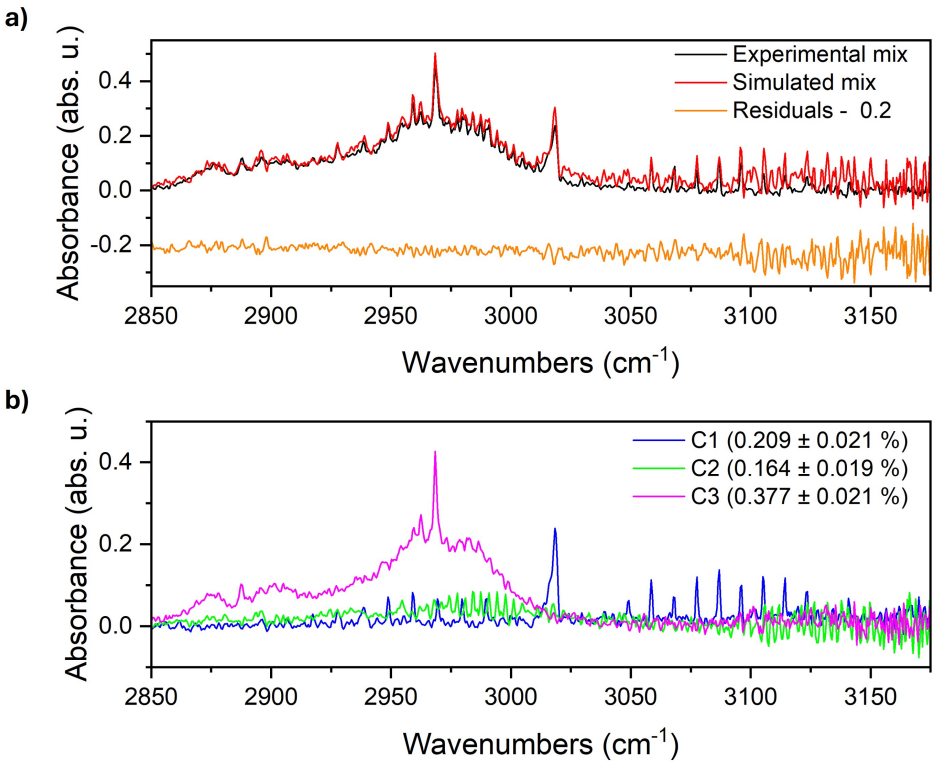

Fig. 11. (a) Comparison between experimental absorbance spectrum of a 0.2% C1, 0.2% C2 and 0.4% C3 mixture (black) and simulated spectrum reconstructed from a sum of MLR estimated concentrations (red). Residuals are plotted with a visual shift of -0.2 (orange) (b) Individual spectral contributions of C1 (blue), C2 (green), and C3 (magenta) based on MLR estimates.

At high C1 concentrations (Figs. 10(a) and 10(b)), the two most intense C1 peaks exhibit a clear saturation in intensity rather than scaling linearly with concentration. This occurs already at concentrations even higher than 2% for C1 peaks at 3018 cm^{-1}. When the concentration range is reduced (Figs. 10(c) and 10(d)), both peaks maintain linearity, indicating that the saturation effect arises solely at higher concentrations. By comparing Fig. 10(a) and Fig. 10(c) related to the most intense peak of the C1 Q-branch, it can be seen that linearity is maintained up to 0.5%, after which the absorbance levels off at a value of 1.2. Since the C1 band contains numerous additional spectral features at lower intensity (Fig. 2(a)), it is reasonable to assume that their peak values scale nearly linearly with concentration. As a result, their cumulative contribution to the total integrated absorbance area is substantial enough to mask the non-linear behavior of the more intense peaks. This explains why the integrated absorbance area for C1 shows a linear relationship with concentration (Fig. 2(b)), and why the prediction accuracy for C1 concentration has consistently remained high (> 96%) across all cases analyzed in this study. Therefore, the reason behind the reduced predictivity of C2 is straightforward and can be directly attributed to its wide spectral interference with C1. This is more critical in ternary mixtures because almost all spectral contribution of C3 fall within that spectral range. A substantial portion of the C2 absorption band overlaps with the intense spectral region associated with C1 (Fig. 6(b)). As a result, the strong and saturating response from the dominant C1 component masks the linear behavior of the C2 signal, effectively overwhelming its contribution and in turn compromising the

Optics EXPRESS

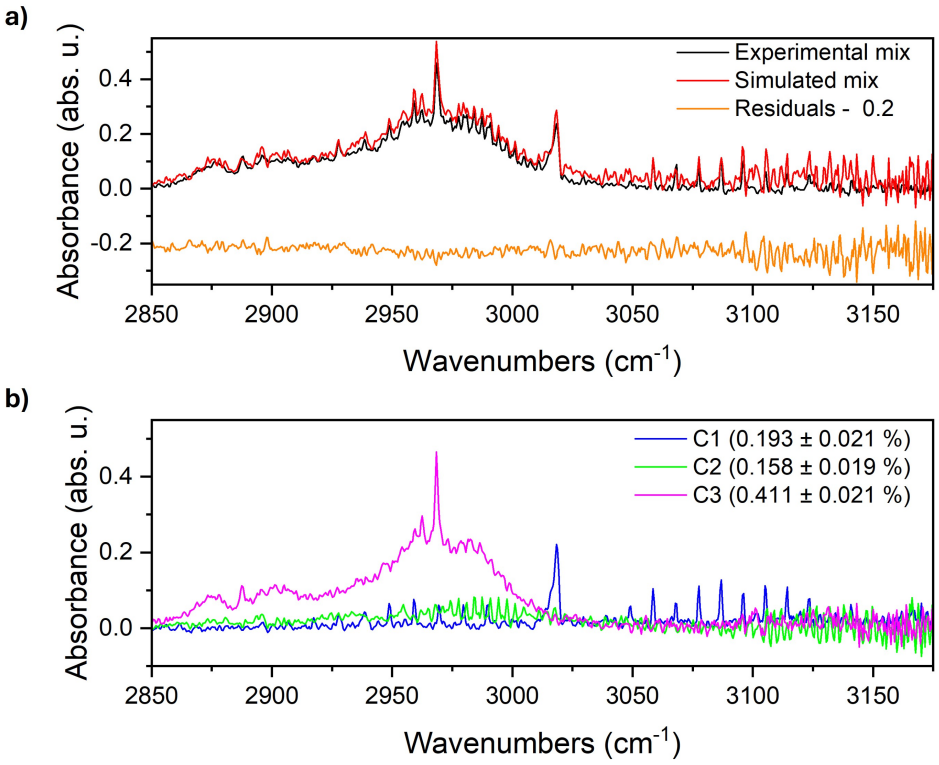

Fig. 12. (a) Comparison between experimental absorbance spectrum of a 0.2% C1, 0.2% C2 and 0.4% C3 mixture (black) and simulated spectrum reconstructed from a sum of PLSR estimated concentrations (red). Residuals are plotted with a visual shift of -0.2 (orange) (b) Individual spectral contributions of C1 (blue), C2 (green), and C3 (magenta) based on PLSR estimates.

accuracy of its identification. These effects can thus explain the reduced prediction accuracy of MLR algorithm, which is highly sensitive to spectral correlation. In the case of PLSR, this results in an optimal number of latent variables (LVs) identified by cross-validation (4) that exceeds the number of analytes (3), indicating that the fourth LV captures both co-existing linear and nonlinear spectral features present in the data. The effect of saturation in the C1 Q-branch is particularly evident in the single-peak sensor calibration. Because MLR linearly scales the C1 reference spectrum, which already includes saturation, this effect contributes to the overestimation of C1 concentration. However, excluding the zero-C1 data point, the calibration approaches a linear behavior with a non-zero offset in the explored concentration range. Such a trend can be still modelled by MLR, as it is demonstrated by comparing the results achieved analyzing a reduced range (2850–3005 cm^{-1}, without the C1 Q-branch) and the full range (2850–3175 cm^{-1}), as reported in Table S1-S3. For C1 retrieval accuracy, the reduced range returned to a value of 89% value, compared to 91% obtained with the full range, as reported in Table S4. Conversely, PLSR is less affected by this issue, as it identifies and prioritizes spectral features with strong covariance, effectively down-weighting the influence of saturated components. It is worth noting that, in more complex scenarios where saturation affects broader spectral regions, a hybrid multivariate strategy has been proposed to address saturation effects by identifying and excluding saturated regions prior to calibration [48]. Nonetheless, since our study already achieves good performance using PLSR combined with spectral window selection, the added benefit of implementing the

Optics EXPRESS

full hybrid method may be limited for the gas mixtures considered here. Specifically, this study focuses on natural gas analysis at a 1:10 dilution ratio, where only methane reaches concentrations sufficient to cause saturation, and even then, only in the Q-branch.

To further support this, Fig. 11(a) shows a comparison between the experimental spectrum and the MLR-simulated spectrum for a "low concentration" mixture containing 0.2% C1, 0.2% C2, and 0.4% C3 in N_2. The individual spectral contributions of each component are also presented in Fig. 11(b).

For comparison, the PLSR-simulated spectrum for the same mixture is presented in Fig. 12.

In this case, the PLSR-CV analysis individuates 3 as the optimal number of LVs for the regression problem, supporting the previous discussion.

7. Conclusions

In this study, a gas sensing system was developed for the broadband detection and quantification of the three main alkanes composing natural gas, using MIR direct absorption spectroscopy. The system combines a broadband supercontinuum light source with an absorption cell with an optical path length $L = 11.6$ cm and an optical spectrum analyzer (OSA). This configuration allows for the reconstruction of the full absorption bands of light hydrocarbons in the 2850-3200 cm^{-1} range, enabling the identification of strongly overlapping spectral features. For quantitative analysis, the performance of two regression techniques, such as MLR and PLSR, was evaluated on binary and ternary mixtures with unbalanced component concentrations (ratio 1:10), representative of natural gas compositions. The results demonstrated that PLSR outperformed MLR in accuracy. Specifically, MLR yielded mean prediction accuracies of 91% for C1, 68% for C2, and 75% for C3, whereas PLSR achieved 98%, 93%, and 94%, respectively. The PLSR-reconstructed spectrum perfectly aligns with experimental data, highlighting the superior ability of PLSR to decompose individual contributions and accurately identify components that are most predictive of concentration levels. This is particularly evident in cases with significant spectral overlap and unbalanced proportionality among the components, conditions that an MLR approach struggles to manage. However, when non-linearities are introduced, the predictivity of the model is slightly affected even with the PLSR approach.

In summary, the main advantages of MLR are its simplicity, ease of interpretation, and computational efficiency. It provides clear coefficients that directly reflect the influence of each predictor on the outcome, making the results straightforward to understand. However, MLR relies on the assumption that predictors are independent, and tends to perform poorly when multicollinearity is present or when the number of predictors is close to or exceeds the number of observations. On the other hand, PLSR excels in handling datasets with highly collinear predictors or when there are more predictors than samples. Its strength lies in reducing dimensionality while maximizing the covariance between predictors and responses, which enhances model robustness and predictive accuracy in complex, multivariate settings. Nonetheless, PLSR is generally less interpretable than MLR because its latent variables are combinations of the original predictors, complicating the attribution of effects to individual variables. Additionally, selecting the appropriate number of latent components requires careful cross-validation to prevent overfitting or underfitting.

Funding. Ministero dell'Università e della Ricerca (PE0000023-NQSTI).

Acknowledgments. Authors from Dipartimento Interateneo di Fisica acknowledge funding from PNRR MUR project PE0000023-NQSTI, MUR-Dipartimenti di Eccellenza 2023–2027 project "Quantum Sensing and Modeling for One-Health" (QuaSiModO) and THORLABS GmbH, within PolySense, a joint-research laboratory.

Disclosures. The authors declare no conflicts of interest.

Data availability. Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

Optics EXPRESS

Supplemental document. See Supplement 1 for supporting content.

References

1. S. Faramawy, T. Zaki, and A. A. E. Sakr, "Natural gas origin, composition, and processing: A review," J. Nat. Gas. Sci. Eng. **34**, 34–54 (2016).
2. B. O. Pixler, "Formation Evaluation by Analysis of Hydrocarbon Ratios," Journal of Petroleum Technology **21**(06), 665–670 (1969).
3. H. Devold, Oil and Gas Production Handbook An Introduction to Oil and Gas Production, Transport, Refining and Petrochemical Industry (n.d.).
4. M. Jaeschke, P. Schley, and R. Janssen van Rosmalen, "Thermodynamic Research Improves Energy Measurement in Natural Gas," Int. J. Thermophys. **23**(4), 1013–1031 (2002).
5. S. A. Baylis, K. Hall, and E. J. Jumeau, "The analysis of the C1–C5 components of natural gas samples using gas chromatography-combustion-isotope ratio mass spectrometry," Org. Geochem. **21**(6–7), 777–785 (1994).
6. Y. Yürüm and M. Levy, "Quantitative determination of shale oil compounds by gas chromatography-mass spectrometry-selected ion monitoring," Fuel Processing Technology **11**(1), 59–69 (1985).
7. A. L. Sessions, "Isotope-ratio detection for gas chromatography," J. Sep. Sci. **29**(12), 1946–1961 (2006).
8. L. M. L. Nollet and D. A. Lambropoulou, eds., *Chromatographic Analysis of the Environment* (CRC Press, 2017).
9. R. K. Jha, "Non-Dispersive Infrared Gas Sensing Technology: A Review," IEEE Sens. J. **22**(1), 6–15 (2022).
10. G. Ma, Y. Wang, W. Qin, *et al.*, "Optical sensors for power transformer monitoring: A review," High Voltage **6**(3), 367–386 (2021).
11. J. Hodgkinson and R. P. Tatam, "Optical gas sensing: a review," Meas. Sci. Technol. **24**(1), 012004 (2013).
12. B. Wang, A. Y. Bukhamsin, B. S. Ooi, *et al.*, "A Review of Distributed Fiber–Optic Sensing in the Oil and Gas Industry," J. Lightwave Technol. **40**(5), 1407–1431 (2022).
13. B. Chemisky, F. Menna, E. Nocerino, *et al.*, "Underwater Survey for Oil and Gas Industry: A Review of Close Range Optical Methods," Remote Sens. **13**(14), 2789 (2021).
14. J. M. Thompson, *Infrared Spectroscopy* (Jenny Stanford Publishing, 2018).
15. B. Smith, *Infrared Spectral Interpretation: A Systematic Approach* (CRC Press, 2018).
16. I. E. Gordon, L. S. Rothman, R. J. Hargreaves, *et al.*, "The HITRAN2020 molecular spectroscopic database," J. Quant. Spectrosc. Radiat. Transf. **277**, 107949 (2022).
17. J. Scheuermann, P. Kluczynski, K. Siembab, *et al.*, "Interband Cascade Laser Arrays for Simultaneous and Selective Analysis of 336–342 Hydrocarbons in Petrochemical Industry," Appl. Spectrosc. **75**(3), 336–342 (2021).
18. Y. Wang, Y. Wei, T. Liu, *et al.*, "TDLAS Detection of Propane/Butane Gas Mixture by Using Reference Gas Absorption Cells and Partial Least Square Approach," IEEE Sens. J. **18**(20), 8587–8596 (2018).
19. A. Zifarelli, A. Sampaolo, P. Patimisco, *et al.*, "Methane and ethane detection from natural gas level down to trace concentrations using a compact mid-IR LITES sensor based on univariate calibration," Photoacoustics **29**, 100448 (2023).
20. A. F. P. Cantatore, G. Menduni, A. Zifarelli, *et al.*, "Methane, Ethane, and Propane Detection Using a Quartz-Enhanced Photoacoustic Sensor for Natural Gas Composition Analysis," Energy Fuels **39**(1), 638–646 (2025).
21. H. Mei, G. Wang, Y. Xu, *et al.*, "Simultaneous measurement of methane, propane and isobutane using a compact mid-infrared photoacoustic spectrophone," Photoacoustics **39**, 100635 (2024).
22. A. Sampaolo, G. Menduni, P. Patimisco, *et al.*, "Quartz-enhanced photoacoustic spectroscopy for hydrocarbon trace gas detection and petroleum exploration," Fuel **277**, 118118 (2020).
23. N. L. Miles, D. K. Martins, S. J. Richardson, *et al.*, "Calibration and field testing of cavity ring-down laser spectrometers measuring CH4, CO2, and δ13CH4 deployed on towers in the Marcellus Shale region," Atmos Meas. Tech. **11**(3), 1273–1295 (2018).
24. K. M. Manfred, G. A. D. Ritchie, N. Lang, *et al.*, "Optical feedback cavity-enhanced absorption spectroscopy with a 3.24 μm interband cascade laser," Appl. Phys. Lett. **106**(22), 14 (2015).
25. S. M. Defratyka, J. D. Paris, C. Yver-Kwok, *et al.*, "Ethane measurement by Picarro CRDS G2201-i in laboratory and field conditions: Potential and limitations," Atmos. Meas. Tech. **14**(7), 5049–5069 (2021).
26. M. F. Barbosa, J. R. B. Santos, A. N. Silva, *et al.*, "A cheap handheld NIR spectrometric system for automatic determination of methane, ethane, and propane in natural gas and biogas," Microchem. J. **170**, 106752 (2021).
27. A. Schliesser, N. Picqué, and T. W. Hänsch, "Mid-infrared frequency combs," Nat. Photonics **6**(7), 440–449 (2012).
28. I. Zorin, P. Gattinger, A. Ebner, *et al.*, "Advances in mid-infrared spectroscopy enabled by supercontinuum laser sources," Opt. Express **30**(4), 5222 (2022).
29. C. R. Petersen, U. Møller, I. Kubat, *et al.*, "Mid-infrared supercontinuum covering the 1.4–13.3 μm molecular fingerprint region using ultra-high NA chalcogenide step-index fibre," Nat. Photonics **8**(11), 830–834 (2014).
30. G. Woyessa, K. Kwarkye, M. K. Dasa, *et al.*, "Power stable 1.5–10.5 μm cascaded mid-infrared supercontinuum laser without thulium amplifier," Opt. Lett. **46**(5), 1129–1132 (2021).
31. P. Masłowski, F. Adler, K. C. Cossel, *et al.*, "Mid-infrared Fourier transform spectroscopy with a broadband frequency comb," Opt. Express **18**(21), 21861–21872 (2010).
32. M. Halloran, N. Traina, J. Choi, *et al.*, "Simultaneous Measurements of Light Hydrocarbons Using Supercontinuum Laser Absorption Spectroscopy," Energy Fuels **34**(3), 3671–3678 (2020).

Optics EXPRESS

33. K. E. Jahromi, Q. Pan, A. Khodabakhsh, *et al.*, "A Broadband Mid-Infrared Trace Gas Sensor Using Supercontinuum Light Source: Applications for Real-Time Quality Control for Fruit Storage," Sensors **19**(10), 2334 (2019).
34. K. E. Jahromi, Q. Pan, L. Høgstedt, *et al.*, "Mid-infrared supercontinuum-based upconversion detection for trace gas sensing," Opt. Express **27**(17), 24469 (2019).
35. W. Zhang, L. C. Kasun, Q. J. Wang, *et al.*, "A Review of Machine Learning for Near-Infrared Spectroscopy," Sensors **22**(24), 9764 (2022).
36. C. A. Meza Ramirez, M. Greenop, L. Ashton, *et al.*, "Applications of machine learning in spectroscopy," Appl. Spectrosc. Rev. **56**(8-10), 733–763 (2021).
37. H. Wold, "Partial Least Squares," in *Encyclopedia of Statistical Sciences* (John Wiley & Sons, Inc., 2004), **6**, pp. 581–591.
38. P. Geladi and B. R. Kowalski, "Partial least-squares regression: a tutorial," Anal. Chim. Acta. **185**(C), 1–17 (1986).
39. H. Abdi, "Partial least squares regression and projection on latent structure regression (PLS Regression)," WIREs Computational Statistics **2**(1), 97–106 (2010).
40. D. Cassanelli, N. Lenzini, L. Ferrari, *et al.*, "Partial Least Squares Estimation of Crop Moisture and Density by Near-Infrared Spectroscopy," IEEE Trans. Instrum. Meas. **70**, 1–10 (2021).
41. A. Zifarelli, M. Giglio, G. Menduni, *et al.*, "Partial Least-Squares Regression as a Tool to Retrieve Gas Concentrations in Mixtures Detected Using Quartz-Enhanced Photoacoustic Spectroscopy," Anal. Chem. 92(16), 11035–11043 (n.d.).
42. G. Menduni, A. Zifarelli, A. Sampaolo, *et al.*, "High-concentration methane and ethane QEPAS detection employing partial least squares regression to filter out energy relaxation dependence on gas matrix composition," Photoacoustics **26**, 100349 (2022).
43. "Mid-Infrared Supercontinuum Laser," https://www.thorlabs.com/newgrouppage9.cfm?objectgroup_id=10819.
44. A. C. Olivieri, *Introduction to Multivariate Calibration: A Practical Approach* (Springer Netherlands, 2018).
45. G. K. Uyanık and N. Güler, "A Study on Multiple Linear Regression Analysis," Procedia Soc. Behav. Sci. **106**, 234–240 (2013).
46. A. E. Maxwell, "Limitations on the use of the multiple linear regression model," Brit. J. Math. Statis. **28**(1), 51–62 (1975).
47. J. Shao, "Linear Model Selection by Cross-Validation," J. Am. Stat. Assoc. **88**(422), 486–494 (1993).
48. R. Krebbers, L. A. AE. Sluijterman, J. Meurs, *et al.*, "Optimizing data analysis for broadband mid-infrared absorption spectroscopy: A hybrid dataset approach," Anal. Chim. Acta. **1367**, 344303 (2025).