Partial least squares regression and interpretation of spectral measurements (PRISM)

Vince Paul P. Juguilon *

National Institute of Physics, University of the Philipppines Diliman, Philippines
*Corresponding author: vpjuguilon@up.edu.ph

Abstract

Partial least squares regression (PLSR) was employed to predict gas concentrations from simulated absorption spectra of mixtures containing nitrous oxide (N_2O), methane (CH_4), and carbon monoxide (CO), in single-, double-, and triple-species configurations. The absorbance spectra used for the training and test sets were generated using the HITRAN Application Programming Interface (HAPI) at arbitrary and randomized concentrations. Results showed that the model was able to predict gas concentrations with coefficients of determination $R^2 > 0.9$ and relative RMSE $< 10\%$, even under noise-levels of around 10% of the max amplitude. These metrics highlight the capability of the model to resolve overlapping spectral spectral features. Moreover, PLSR demonstrated tolerance for noisy measurements, which is an advantage compared to models that rely on discrete wavelength analysis.

Keywords: air quality monitoring, urban gas, absorbance spectroscopy, partial least squares regression

1 Introduction

Accurate measurement of atmospheric gas concentrations is essential for applications such as urban air quality monitoring, industrial emission control, and environmental research. Real-time and precise detection of urban gases like methane (CH_4), carbon monoxide (CO), and nitrous oxide (N_2O) is also critical for assessing pollution sources [1, 2].

Traditionally, gas concentrations are measured using chemometric or electrochemical methods [3]. These approaches are well-established but may require periodic calibration. Additionally, these sensors can be sensitive to environmental conditions and are sometimes limited in terms of selectivity and speed.

Spectroscopic techniques offer a non-invasive, selective, and rapid alternative based on the unique absorption features of gases in the ultraviolet (UV) to infrared (IR) spectrum [4]. However, interpreting the absorption spectra of gases requires robust statistical modeling, especially in mixtures with overlapping spectral features [5].

Partial least squares regression (PLSR) is a widely used multivariate technique that addresses this challenge by projecting both spectral data and concentration values into a shared latent space [6]. It identifies components that maximize the covariance between predictor variables (e.g., absorbance spectra) and response variables (e.g., gas concentrations). Unlike simpler regression techniques, PLSR is effective even when predictors are highly collinear or when the number of variables exceeds the number of observations [7]. This makes it especially suitable for analyzing complex or noisy spectral data, as is often the case in real-world gas sensing applications.

In this study, a partial least squares regression (PLSR) model was employed to predict gas concentrations from three types of mixtures: a single-species N_2O mixture diluted with air, a two-species mixture of N_2O and CH_4, and a three-species mixture comprising N_2O, CH_4, and CO. The selected gases exhibit strong absorption features in the mid-infrared region (2000–3200 cm^{-1} or approximately 3.1–5.0 μm) which makes them ideal candidates for spectroscopic analysis and model evaluation.

2 Methodology

2.1 Generating randomized concentration labels

A Python script (`molecular_concentrations`) was used to generate randomized gas concentrations. The concentrations were arbitrarily set to range from 1% to 30% by mol fraction for each gas in the mixture. This range is significantly higher than typical atmospheric concentrations, which are usually in the ppm to ppb range. This elevated range was chosen to ensure strong absorbance signals and to evaluate the method as a proof of concept.

The script generates blocks of concentration labels with either randomized or fixed concentrations for each gas, as shown in Figure 1. This will be used later during model training of double- and triple-species mixtures in order for the model to isolate and learn spectral features specific to each gas.

The `molecular_concentrations` script was run separately for each of the three gases, with each run producing a `labels.csv` file containing a single column of concentrations. For the double- and triple-species configuration, the corresponding `labels.csv` files were generated by concatenating the concentration columns from the CSV files of individual gases.

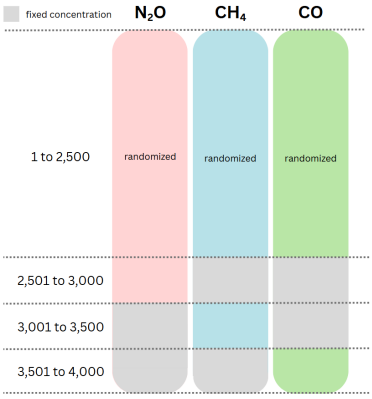


Figure 1: The script generates 4,000 concentration values each for N_2O, CH_4, and CO. Some blocks use randomized (colored) or fixed (grayed) values for the mixed gas configurations in order for the model to isolate and learn spectral features specific to each gas.

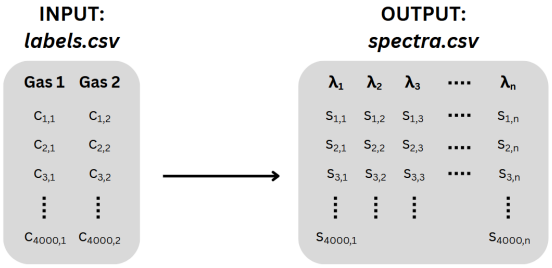


Figure 2: A `spectra.csv` file is generated, with columns representing wavelengths and rows containing the absorption spectrum for each concentration entry from `labels.csv`. The input file may correspond to single-, double-, or triple-species configurations.

2.2 Simulating the absorption spectra

The generated `labels.csv` files were used as input to the `atmospheric_spectra_simulator` script, which utilizes the HITRAN Application Programming Interface (HAPI) to simulate absorption spectra for each concentration specified in the `labels.csv` file [8]. HAPI retrieves the necessary spectral line data from the HITRAN database and reconstructs the absorption spectrum for each gas based on the specified environmental conditions: the gas concentration input, temperature of 273 K, pressure of 1 atm, and a path length of 0.013 cm. The path length was chosen to balance detectability and prevent signal saturation across the selected concentration range. The absorption spectra was calculated over the spectral range of 2000–3200 cm^{-1}, with a resolution of 1 cm^{-1}. Spectral line broadening due to the diluent (air) was also accounted for by using the Voigt profile.

The `atmospheric_spectra_simulator` script generates a `spectra.csv` file, where each column corresponds to a wavelength and each row contains the absorption spectrum associated with a concentration entry from the input `labels.csv`. Parallel processing was employed using `n_workers = 16` which enables the script to iterate through each row in the CSV file more efficiently. This transformation is illustrated in Figure 2. Afterwards, noise corresponding to a percentage of the max value is added to the spectrum.

Spectra for the double- and triple-species configurations were generated by the row-wise addition of the simulated single-species spectra. This method assumes linear and additive absorbance which is typically valid at low concentrations. Noise was added after specral mixing.

For the datasets used in the evaluation, N_2O was used as the sole absorbing gas in the single-species configuration, with air as the diluent. The double-species mixture included both N_2O and CH_4, while the triple-species mixture comprised N_2O, CH_4, and CO.

2.3 Training and prediction using PLSR model

Both the `labels.csv` containing the concentration labels, and the `spectra.csv` with the simulated spectral measurements file were loaded to the `PLS-Regression` script to train a model that will be used to predict the gas concentrations.

The data processing pipeline is as follows: spectral data was preprocessed using `StandardScaler`. The data set with 4,000 samples was split into training and test sets with a 0.9:0.1 ratio. Hyperparameter tuning for the PLSR model was also conducted on the training set to identify the optimal number of components based on the calculated coefficient of determination (R^2) and root-mean-squared error (RMSE), using 10-fold cross-validation.

The final PLSR model was then re-trained on the entire training set using the optimal number of components, and was evaluated on the withheld training set.

3 Results and Discussion

Figure 3 displays the spectral absorption lines of N_2O, CH_4, and CO, within the wavelength range of interest. The plot reveals partial overlap between the absorption features of N_2O and CO, which will be further examined in the triple-species configuration discussed in Section 3.3.

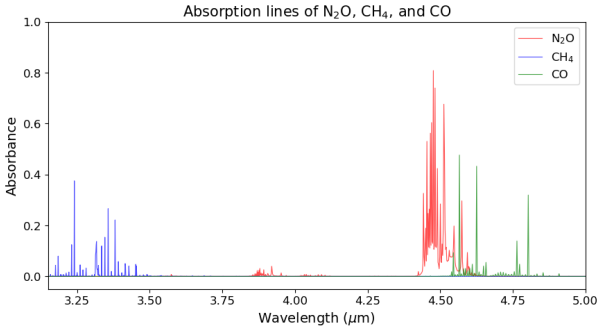
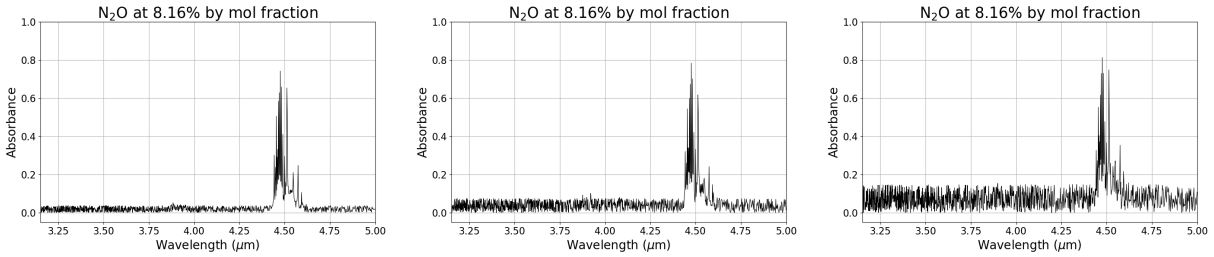


Figure 3: Absorption lines of nitrous oxide (N_2O), methane (CH_4), and carbon monoxide (CO) in the mid-IR

3.1 Analysis of PLS model for single-species (N_2O) gas mixture

The simulated absorption spectra for N_2O at a representative concentration of 8.16% mol fraction is shown in Figure 4. Air was used as the diluent in this single-species mixture, which also contributed to the broadening of spectral lines. The plots illustrate the spectra at noise levels of 5%, 10%, and 20%, respectively.



(a) 5% noise (b) 10% noise (c) 20% noise

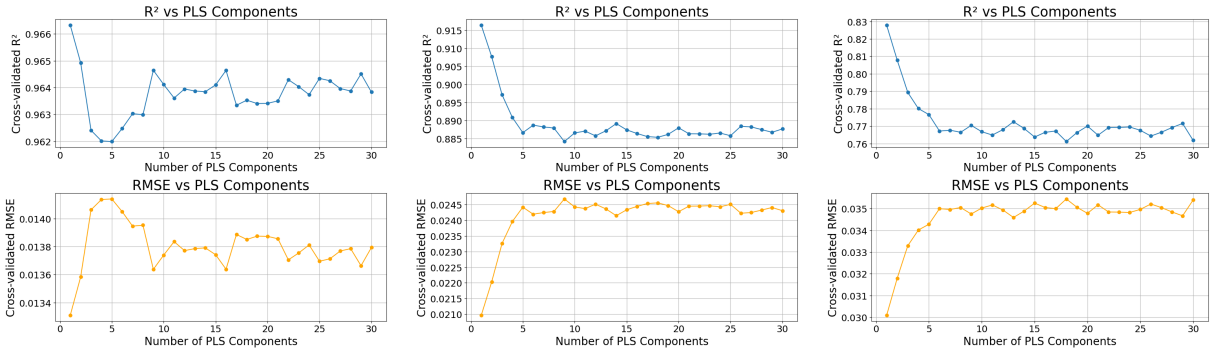
Figure 4: Simulated absorption spectra of 8.16% N_2O with air as diluent at various noise levels

The results of hyperparameter tuning across different noise levels are shown in Figure 5. In all cases, selecting `n_components` = 1 for yielded the highest average R^2 and lowest RMSE based on cross-validation of the training set.

After retraining the model using the optimal number of components, its performance was then evaluated on the withheld test set. The predicted concentrations were plotted against the true values and compared to the ideal (red) line which represents perfect prediction.

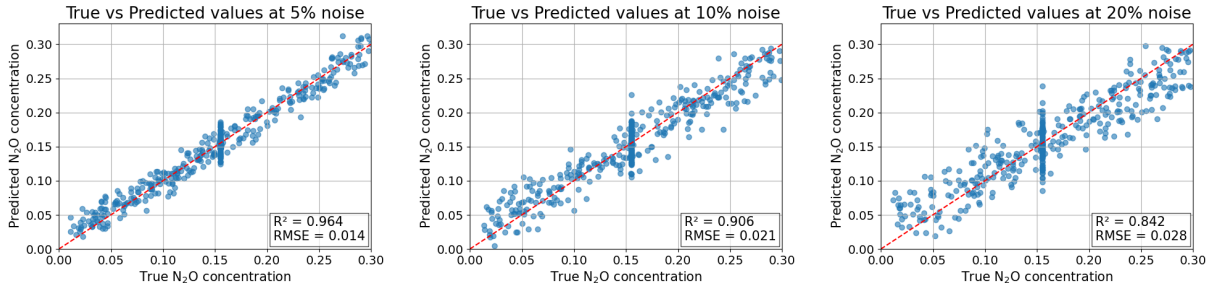
The results demonstrate strong model performance even in the presence of noise. The model achieved an $R^2 = 0.964$ and RMSE = 0.014 (relative RMSE with respect to full range, rRMSE = 4.7%) at a 5% noise level. Meanwhile, the model still performed reasonably well with $R^2 = 0.842$ and RMSE = 0.028 (rRMSE = 7.0%) at a higher noise level of 20%. These results highlight the robustness of the PLS model. By analyzing information across a range of wavelengths, the model is able to effectively suppress the influence of noise and maintain high prediction accuracy.

Moreover, limiting the number of components during training ensures that the model captures only the most relevant variance in the data which helps filter out noise. This also reduces the risk of overfitting and improves generalization to unseen samples.



(a) 5% noise (b) 10% noise (c) 20% noise

Figure 5: Hyperparameter tuning reveals the optimal components for the PLS model based on R^2 and RMSE.



(a) 5% noise, `n_components` = 1 (b) 10% noise, `n_components` = 1 (c) 20% noise, `n_components` = 1

Figure 6: Comparison of true and predicted concentrations of N_2O using the trained PLS model, with corresponding R^2 and RMSE values. The red line represents the line of perfect prediction ($R^2 = 1$).

3.2 Analysis of PLS model for double-species (N_2O and CH_4) gas mixture

The double-species gas mixture consisting of N_2O and CH_4 were also analyzed using the same pipeline for the single-species configuration.

Table 1 summarizes the performance metrics of the model after evaluation on the training set at 5%, 10%, and 20% noise levels. Relevant plots for the representative spectra, parameter tuning, and visualization of model prediction are compiled in the Appendix.

Table 1: Summary of the models' performance for double-gas configuration based on calculated R^2 and RMSE. The number of components used in PLS training are specified for each noise level.

	5% noise level (`n_components` = 15)		**10% noise level** (`n_components` = 10)		**20% noise level** (`n_components` = 10)	
	R^2	RMSE	R^2	RMSE	R^2	RMSE
N_2O	0.963	0.014	0.911	0.023	0.712	0.037
CH_4	0.986	0.008	0.983	0.009	0.988	0.008

The results indicate that the trained model predicted CH_4 concentrations more accurately than N_2O, despite N_2O exhibiting stronger absorption lines as shown in Figure 3. At a noise level of 20%, the model achieved an R^2 of 0.988 and RMSE of 0.008 for CH_4. Meanwhile, performance dropped for N_2O, with an R^2 of 0.712 and RMSE of 0.037.

This observation could be attributed to the slightly broader distribution of CH_4 absorption lines throughout the spectrum, which provides more distinct features (wavelengths). Furthermore, the PLS components likely captured more variance related to CH_4 wavelengths.

3.3 Analysis of PLS model for triple-species (N_2O, CH_4, and CO) gas mixture

The triple-species mixture of N_2O, CH_4, and CO were also trained and evaluated using a PLS model. Results of the model performance are summarized in Table 2. The relevant plots are included in the Appendix.

Table 2: Summary of the models' performance for triple-gas configuration based on calculated R^2 and RMSE. The number of components used in PLS training are specified for each noise level.

	5% noise level (`n_components` = 15)		**10% noise level** (`n_components` = 10)		**20% noise level** (`n_components` = 7)	
	R^2	RMSE	R^2	RMSE	R^2	RMSE
N_2O	0.965	0.014	0.910	0.022	0.829	0.030
CH_4	0.986	0.009	0.969	0.012	0.949	0.016
CO	0.854	0.027	0.566	0.048	0.200	0.065

The results indicate that the model maintained strong performance in predicting CH_4 concentrations, even in the presence of background noise as with the double-species configuration. At 5% noise, the model still achieved reasonable accuracy in predicting CO concentrations. However, increasing the noise level to 10% and 20% significantly affected CO prediction performance, with the R^2 values dropping to 0.566 and 0.200, respectively.

Nitrous oxide (N_2O) and carbon monoxide (CO) exhibit overlapping absorption features within the spectral range of interest as shown in Figure 3. This overlap may have led to ambiguity in signal attribution, which made it more difficult for the model to predict CO concentrations.

In such cases, additional techniques may be required to better isolate the spectral signatures of CO. For example, feature engineering such as identifying non-overlapping wavelength regions could enhance model performance. Additionally, applying pre-processing methods such as noise smoothing can also improve prediction accuracy.

4 Conclusions

This study showed that partial least squares regression (PLSR) can accurately predict gas concentrations from simulated broadband absorption spectra across single-, double-, and triple-species gas mixtures. The model demonstrated good accuracy in predicting CH_4 concentrations across all configurations. This is likely due to methane's broader and more distinct spectral features, which enabled the model to better capture variance during training. In contrast, prediction accuracy for CO declined significantly under higher noise levels, primarily due to spectral overlap with N_2O.

These results highlight the importance of spectral feature distribution in multivariate regression models. To improve performance in such cases, additional techniques like noise smoothing or selective wavelength filtering may be utilized. Overall, the findings support the use of PLSR as a reliable and noise-tolerant method for quantitative gas analysis using spectral measurements.

References

[1] W. Zhang, H. Li, Q. Xiao, and X. Li, Urban rivers are hotspots of riverine greenhouse gas (N2O, CH4, CO2) emissions in the mixed-landscape chaohu lake basin, *Water Res.* **189**, 116624 (2021).

[2] S. V. Williams, R. Close, F. B. Piel, B. Barratt, and H. Crabbe, Characterising carbon monoxide household exposure and health impacts in high- and middle-income countries-a rapid literature review, 2010-2024, *Int. J. Environ. Res. Public Health* **22**, 110 (2025).

[3] D. E. Williams, Electrochemical sensors for environmental gas analysis, *Curr. Opin. Electrochem.* **22**, 145 (2020).

[4] M. Y. Bacaoco, V. P. Juguilon, A. I. Cafe, C. A. Tugado, M. A. B. Faustino, G. Bagtasa, and E. Estacio, Design of a low-cost differential optical absorption spectroscopy set-up for simultaneous monitoring of atmospheric NO2 concentration and aerosol optical thickness (2020).

[5] M. S. I. Sagar, N. R. Allison, H. M. Jalajamony, R. E. Fernandez, and P. K. Sekhar, Review–Modern data analysis in gas sensors, *J. Electrochem. Soc.* **169**, 127512 (2022).

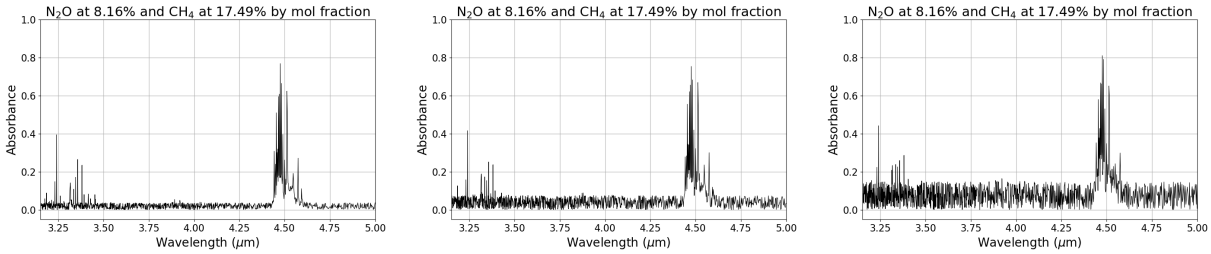
[6] H. Abdi and L. J. Williams, Partial least squares methods: partial least squares correlation and partial least square regression, *Methods Mol. Biol.* **930**, 549 (2013).

[7] P. L. de Micheaux, B. Liquet, and M. Sutton, A unified parallel algorithm for regularized group PLS scalable to big data, *arXiv [stat.ML]* (2017).

[8] R. V. Kochanov, I. E. Gordon, L. S. Rothman, P. Wcisło, C. Hill, and J. S. Wilzewski, HITRAN application programming interface (HAPI): A comprehensive approach to working with spectroscopic data, *J. Quant. Spectrosc. Radiat. Transf.* **177**, 15 (2016).

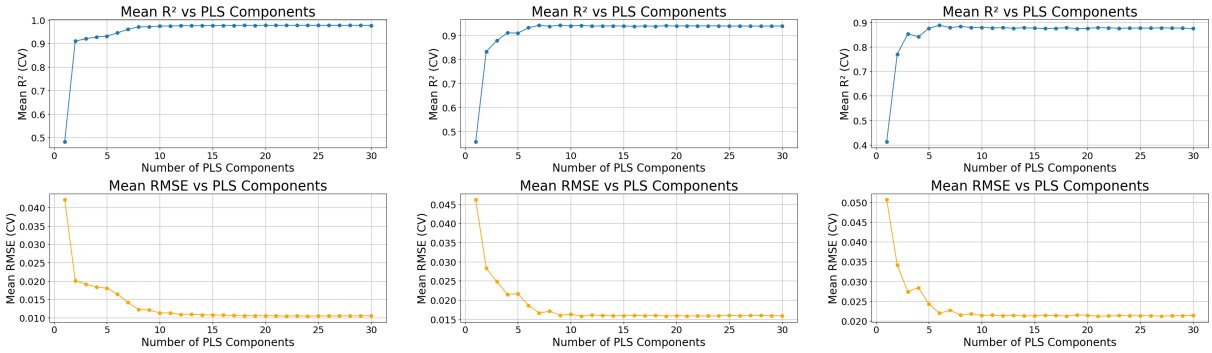
Appendix

Plots for double-species mixture



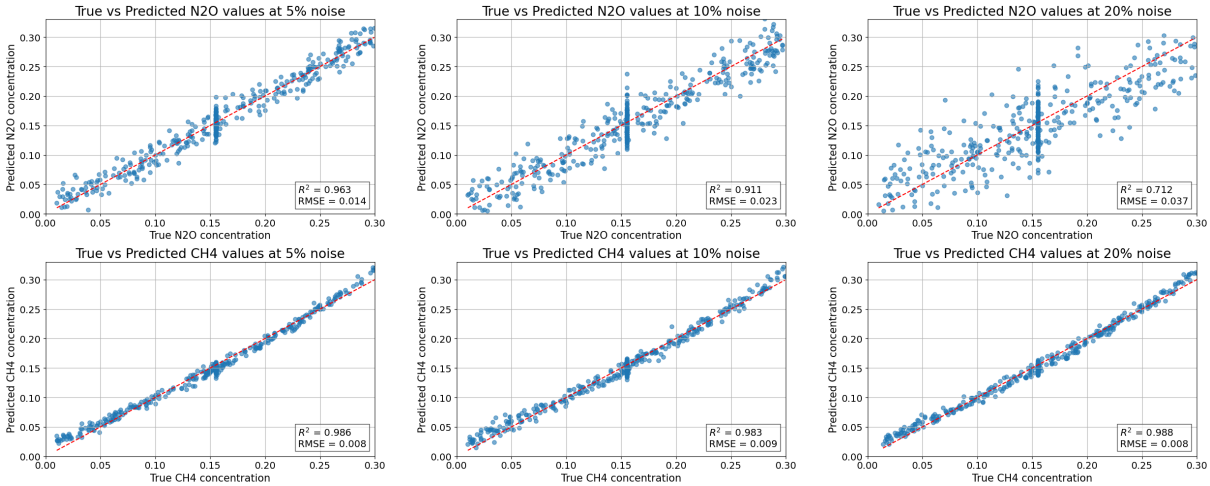
(a) 5% noise (b) 10% noise (c) 20% noise

Figure 7: Simulated absorption spectra of 8.16% N_2O and 17.49% CH_4 with air as diluent at various noise levels



(a) 5% noise (b) 10% noise (c) 20% noise

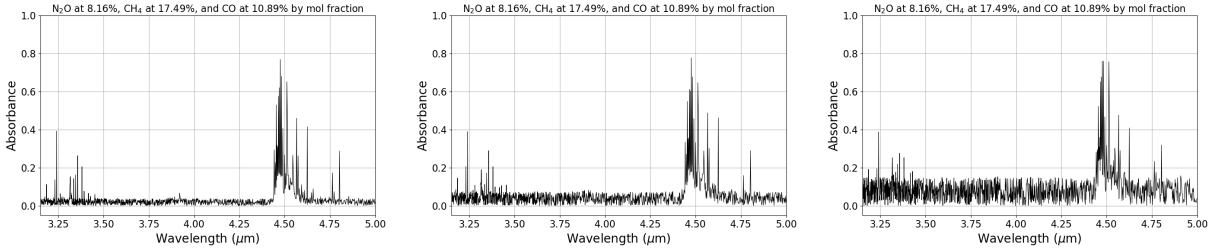
Figure 8: The optimal number of components is determined based on the mean R^2 and RMSE for N_2O and CH_4 during 10-fold cross validation.



(a) 5% noise (b) 10% noise (c) 20% noise

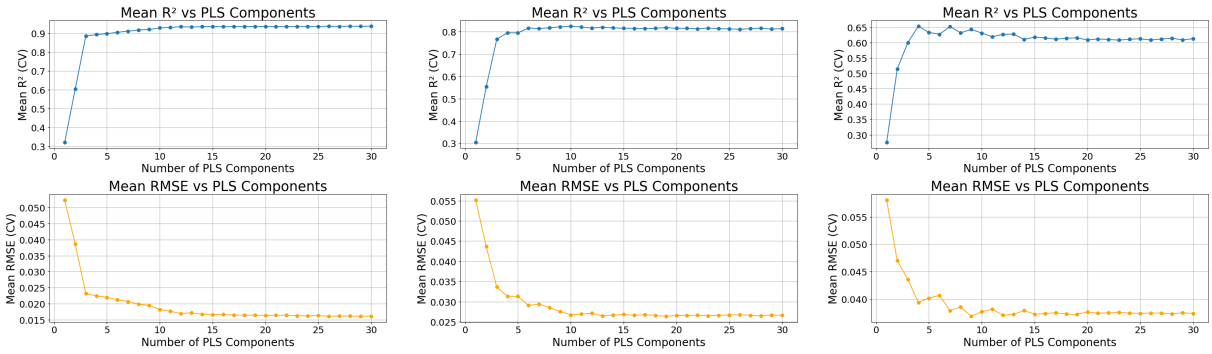
Figure 9: Comparison of true and predicted concentrations of N_2O and CH_4 using the trained PLS model

Plots for triple-species mixture



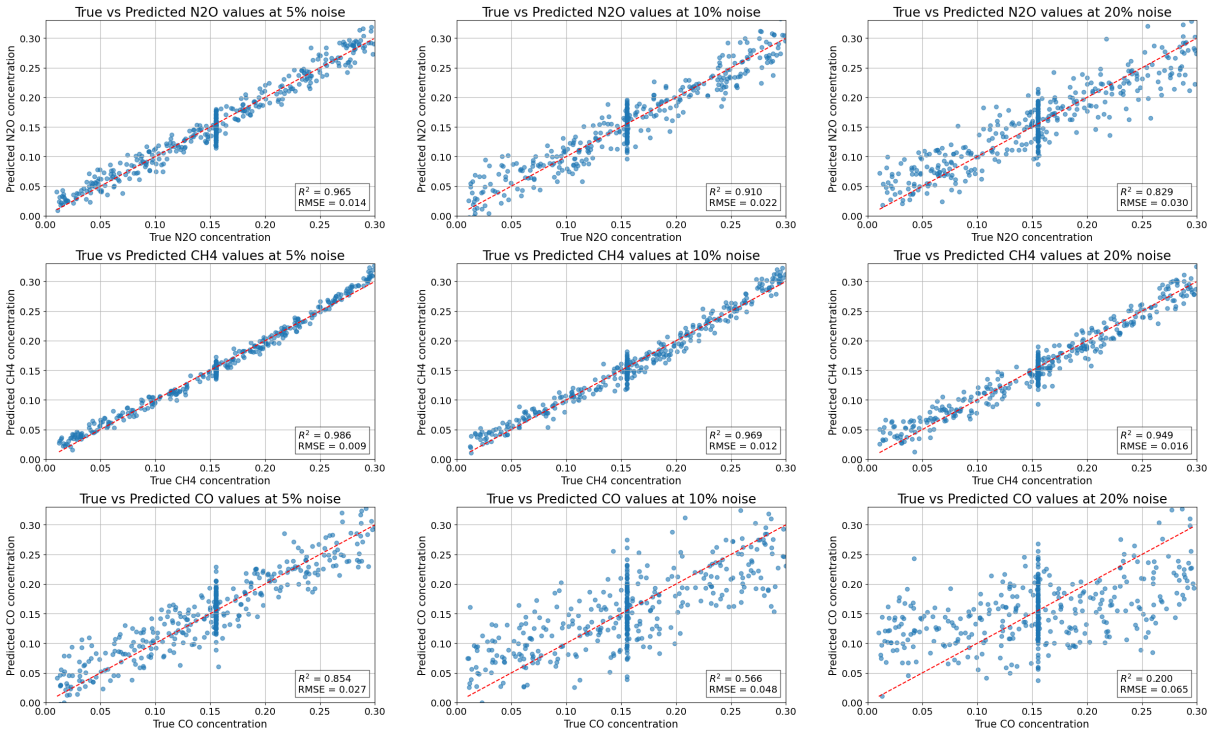
(a) 5% noise (b) 10% noise (c) 20% noise

Figure 10: Simulated absorption spectra of 8.16% N_2O, 17.49% CH_4, and 10.89% CO at various noise levels



(a) 5% noise (b) 10% noise (c) 20% noise

Figure 11: The optimal number of components is determined based on the mean R^2 and RMSE for N_2O, CH_4, and CO during 10-fold cross validation.



(a) 5% noise (b) 10% noise (c) 20% noise

Figure 12: Comparison of true and predicted concentrations of N_2O, CH_4, and CO using the trained PLS model