DOI 10.5162/EUROSENSORS2025/MP44

The Dimensionality of MOX Sensors in Air Quality assessment

Luiz Miranda[1], Nathalie Redon[2], Caroline Duc[2], Marie Verriele[2] Jugurta Montalvão[3], Bernadette Dorizzi[1]
, Mossaab Hariz[1], Dijana Petrovska[1], Jerome Boudy[1]
[1] Samovar, Télécom SudParis, Institut Polytechnique de Paris, Palaiseau, France
[2] IMT Nord Europe, Institut Mines-Télécom, Univ. Lille, Center for Energy and Environment, Douai, France
[3] Universidade Federal de Sergipe, UFS, São Cristóvão, Brazil
Corresponding Author's e-mail address: luiz.miranda_cavalcante_neto@telecom-sudparis.eu

Summary: This paper investigates why air quality applications using MOX gas sensors typically rely on sensor arrays with limited number of sensors. Despite their high sensitivity, MOX sensors suffer from low selectivity, and practical constraints like power and processing capability limit the array size. By analyzing two datasets with different complexities, in terms of monitored pollutants, using three intrinsic dimensionality (ID) estimators, the study shows that the effective dimensionality of sensor data is often much lower than the total number of sensors. This suggests redundancy in sensor responses and supports the common use of 2 to 4 MOX sensors. Results also highlight that features that include some dynamic information, like relative response, yield higher information.

Keywords: Metal-oxide gas sensors, intrinsic dimensionality, air quality, sensor arrays, feature selection

Introduction

Metal oxide (MOX) gas sensors are widely used due to their low cost, high sensitivity, and long lifespan compared to technologies like electrochemical and polymer sensors [1]. However, their low selectivity makes that several sensors are often combined into arrays (electronic noses) to distinguish between gases. Applications such as breath analysis may use up to 19 sensors, but air quality monitoring typically uses fewer than 10 due to power and size constraints, especially in portable devices where each MOX sensor's micro-heater impacts battery life [2].

To address these constraints, studies often seek to optimize sensor selection, with numbers between 2 and 4 MOX sensors frequently found to be sufficient depending on the target gases [2]. In this work, we try to give a theoretical justification of these numbers by analyzing the intrinsic dimensionality (ID) of data from air quality experiments. The following sections define dimensionality in this context, introduce the ID concept, describe the datasets used, and discuss our results.

Intrinsic Dimensionality for MOX sensors

The apparent dimensionality D of multivariate signals from array of MOX sensors is determined by the number of sensors or features used to process their data. For example, a 5-sensor system is 5-dimensional with raw data, or 10-dimensional if features like relative response (R_s) and the average sensor response are used. As dimensionality increases, data visualization and computation become more challenging. However, high-dimensional data can usually be described with fewer variables, known as latent variables, the quantity of which is given by the Intrinsic Dimensionality (ID) of the underlying phenomenon, thus, the minimum number of variables needed to describe the data without loss of information [3].

ID can also be seen as the perceived dimensionality of data, indicating redundancy when ID is lower than the apparent dimensionality. Given that the sensitivity profiles of MOX sensors often overlap, systems with multiple sensors are expected to exhibit some level of redundancy.

Material and methods

Datasets

Tests were performed in two datasets, in which result from commercially available MOX sensors. Dataset 1 is an indoor activity dataset in which 10 different daily household activities were performed in a room monitored by 21 distinct MOX sensors [1]. Dataset 2 is a qualification dataset in which 4 unique MOX sensors are exposed to a fixed concentration (600 ppb) of formaldehyde and toluene injected in a 36 liters chamber. The sensors used in Dataset 1 include the sensors used in Dataset 2.

Features for both datasets were considered as the average value and the normalized relative response (NRR) during the execution of the activities, for Dataset 1, and injection of the target gases, for Dataset 2. The NRR is calculated as $R_s = (R_g - R_0)/R_0$, where R_0 and R_g are the sensors' responses at the beginning and end of activity or injection, respectively. Both datasets are available upon request.

ID estimators

ID can be estimated in several ways, here we implemented three different methods to perform this estimation. The first is based on PCA in which the ID is given by the number of eigenvectors representing more than 90 % of the data

variance. The second is an adaptation [4] of the Grassberger and Procaccia (GP) correlation dimension [5]. And the third is the Maximum Likelihood Estimator (MLE) from Levina and Bickel (2004) using the parameter $k = 20$ [6].

Results and discussion

Table 1 summarizes the results of the tests performed. The GP method sometimes provides two values, as it estimates the ID in several scales of signal intensity (multiscale analysis). Thus, when two stable estimates are observed, for two different signaling scales, the method reports two numbers. For both datasets, most resulting intrinsic dimensionalities (d) are smaller than the original dimensionality (D), indicating redundancy, which is typical in multidimensional experimental data. The experiment for Dataset 1, which involves multiple volatile compounds, results in a higher ID compared to the experiment that generated Dataset 2, which only tests two gases. This shows that more complex experiments tend to generate data with a higher degree of freedom, reflected in the higher ID values.

Although providing different results, the different ID estimators provide mostly agreeing values, with the exception of PCA. We believe that it comes from the linear nature of the PCA method, which is not able to estimate the local dimension of curved structures suggested by observations, thus, causing an overestimation [3].

The different features tested also give clues of which one can provide most information about the phenomenon behind observations. While the average measurement from the sensors can be useful for some application, the relative response introduces an idea of time progression which is often more important to characterize the phenomenon. This is reflected in the resulting ID of these features. Therefore, as higher ID represents a higher complexity (degrees of freedom), more information can be extracted from the features.

Finally, to try to understand why the number of sensors in air quality applications are often limited to 4, as previously mentioned, we can go back to the richness of the experiment. As air quality applications using these sensors often try to identify and quantify concentration of a handful of target gases, the complexity of these applications is closer to the experiment for Dataset 2 than for Dataset 1. As ID is the lower bound of the number of "ideal" sensors, we see that this low number of sensors (and ID) is reflected in these types of experiments.

Tab. 1: Comparison of dimensionality estimation methods across two datasets.

Estimator	**Dataset 1** ($D = 21$)		**Dataset 2** ($D = 4$)	
	Mean	R_s	Mean	R_s
PCA	6	5	3	4
MLE	4.4	6.5	1.9	2.8
GP	3.2	3.5 or 5.8	1.4 or 2.4	2.5

Conclusion

This study investigated the intrinsic dimensionality (ID) of data from MOX gas sensor arrays using three estimation methods across two datasets of differing richness. Results showed that the effective dimensionality is significantly lower than the total number of sensors, confirming that only a few ideal sensors (or very diverse actual ones) would be enough to capture the essential information in air quality monitoring applications.

These findings give a new perspective for most practical systems that use 2 to 4 MOX sensors, especially under power and processing constraints. Additionally, the use of relative response as a feature yielded higher IDs than the average response of the sensors, suggesting that some idea of time progression can provide more useful information for measuring relevant gases in air quality applications.

References

[1] L. Miranda, C. Duc, N. Redon, J. Pinheiro, B. Dorizzi, J. Montalvão, M. Verriele, and J. Boudy, "Automatic detection of indoor air pollution-related activities using metal-oxide gas sensors and the temporal intrinsic dimensionality estimation of data," *Indoor Environments*, vol. 1, no. 3, 2024.

[2] Z. Yuan, X. Luo, and F. Meng, "Machine Learning-Assisted Research and Development of Chemiresistive Gas Sensors," *Advanced Engineering Materials*, vol. 26, no. 20, p. 2400782, 2024.

[3] F. Camastra and A. Staiano, "Intrinsic dimension estimation: Advances and open problems," *Information Sciences*, vol. 328, pp. 26–41, 2016.

[4] J. Montalvão, J. Canuto, and L. Miranda, "Bias-Compensated Estimator for Intrinsic Dimension and Differential Entropy," *Journal of Communication and Information Systems*, vol. 35, pp. 300–310, Dec. 2020.

[5] P. Grassberger and I. Procaccia, "Characterization of Strange Attractors," *Physical Review Letters*, vol. 50, pp. 346–349, Jan. 1983.

[6] E. Levina and P. Bickel, "Maximum likelihood estimation of intrinsic dimension," in *Advances in Neural Information Processing Systems*, vol. 17, MIT Press, 2004.

Acknowledgments

The authors gratefully acknowledge the support of ADEME and the Hauts-de-France Region for funding the extraction and processing of the datasets as part of the doctoral work of the first author. This work is currently supported by the European Union through the HORIZON-CL4-2023-RESILIENCE-01-33 project "Smart sensors for the Electronic Appliances market (RIA)," grant number 101130159 – AMUSENS.