CANADA
EXCELLENCE
RESEARCH
CHAIR

DATA SCIENCE
FOR REAL-TIME
DECISION-MAKING

AN ONLINE DATA VALIDATION ALGORITHM FOR ELECTRONIC NOSE

Mina Mirshahi
Vahid Partovi Nia
Luc Adjengue

June 2016

DS4DM-2016-002

POLYTECHNIQUE MONTRÉAL

DÉPARTEMENT DE MATHÉMATIQUES ET GÉNIE INDUSTRIEL

Pavillon André-Aisenstadt
Succursale Centre-Ville C.P. 6079
Montréal - Québec
H3C 3A7 - Canada
Téléphone: 514-340-5121 # 3314

An Online Data Validation Algorithm for Electronic Nose

Mina Mirshahi, Vahid Partovi Nia, and Luc Adjengue

Department of Mathematics and Industrial Engineering, Polytechnique Montreal, Montreal, Quebec, Canada
{mina.mirshahi,vahid.partovinia,luc.adjengue}@polymtl.ca
http://www.polymtl.ca

Abstract. An electronic nose (e-nose) is a device that analyzes the chemical components of an odour. The e-nose consists of an array of gas sensors for chemical detection, and a mechanism for pattern recognition to return the odour concentration. Odour concentration defines the identifiability and perceivability of an odour. Given that accurate prediction of odour concentration requires valid measurements, automatic assessment of sampled measurements is of prime importance. The impairment of the e-nose, and environmental factors (including wind, humidity, temperature, etc) may introduce significant amount of noise. Inevitably, the pattern recognition results are affected. We propose an online algorithm to evaluate the validity of sensor measurements during the sampling before using the data for pattern recognition phase. The proposed algorithm is computationally efficient and straightforward to implement.

Keywords: Artificial olfaction, computational complexity, electronic nose, gas sensor, outlier detection, robust covariance estimation.

An Online Data Validation Algorithm for Electronic Nose

Mina Mirshahi, Vahid Partovi Nia, and Luc Adjengue

Polytechnique Montreal

1 Introduction

1.1 Background

The recognition of chemicals in the environment is an essential need for the living organisms. Odours are detected through millions of olfactory receptors that are located at the top of nasal cavities. The human olfactory system consists of three main components: 1) an array of olfactory receptors 2) the olfactory bulb that receives neural inputs about odours detected by the receptors and 3) the brain. The olfactory system collects a sample from its environment and transmit it to the brain, where it is recognized as a specific odour.

An olfactory system is able to detect a broad range of smells. However, the human olfacotry system fails to respond to many air pollutants; people can have different sensitivity to many air pollutants and even be accustomed to toxic smells.

The contamination of air by harmful chemicals is referred to as air pollution and is one of the biggest concerns worldwide. This is mainly because the air pollution has direct influence on the environmental and human health. Auditing odourants is a crucial element in assessment of indoor and outdoor air quality. There are various odour measurement techniques such as dilution-to-threshold, olfactometers, and referencing techniques (McGinley and Inc, 2002). The dependence of these methods on human evaluation makes them less accurate and sometimes undesirable.

The concept of an artificial olfaction was introduced by Persaud and Dodd (1982). The primary artificial olfaction rely on a gas multisensor array. The term electronic nose (e-nose) appeared for the first time in the early 1990s (Gardner and Bartlett, 1994). E-nose is designed for recognizing simple or complex odours in its environment and it comprises two main elements of hardware and software. The hardware usually include a set of gas sensors (such as metal oxide semi-conductors, conducting polymers, etc.) with partial specificity, air conditioner, flow controller, electronics, and many more components. The software consists of statistical methods for pre-processing the data and pattern recognition methods for predicting the odour concentration.

The gas sensors of e-nose should have certain features. Similar to human nose receptors, the gas sensors of e-nose need to be highly sensitive with respect to chemical compounds and less sensitive towards temperature and humidity. In addition, the sensors should be able to respond to various chemical compounds. Among the other features, one can name durability, selectivity, and easy calibration.

Gas sensor's performance is affected by various elements. One of the most serious deterioration in sensors is owing to a phenomenon called *drift*. Drift is the low frequency change in a sensor that causes offset measurements. Sensor drift, therefore, need to be detected and compensated to guarantee accurate sensor measurments. Several methods have been introduced to overcome the drift phenomenon including Carlo and Falasconi (2012); Artursson et al. (2000); Padilla et al. (2010); Zuppa et al. (2007).

The multivariate response of gas sensor arrays undergoes different pre-processing procedures before the prediction is performed using statistical tools such as regression, classification, or clustering. Gutierrez-Osuna (2002); Kermiti and Tomic (2003); Bermak et al. (2006) have discussed methods for analyzing the gas sensor array data.

1.2 Motivation

The e-nose is capable of reproducing the human sense of smell using an array of gas sensors and pattern recognition methods. Pattern recognition methods use a set of labelled data to predict the odour concentration for each set of sensor measurements. The labelled data consist of a sub-sample of sensors' outputs considered for further analyses of its concentration in olfactometry.

One of the application of e-nose is in environmental activities; e-noses provide industries with odour management plan to minimize the effect of odour in the environment. To this end, e-noses are installed in outdoor fields such as compost sites, landfill sites, waste water plants, etc., where the environmental condition can greatly fluctuate. Consequently, the occurrence of unwanted variability is very typical.

During the sampling process, sensors in the e-nose device may report incorrect values or some of the sensors stop functioning for a short period of time. These anomalies are ought to be diagnosed and reported in real time using a computationally efficient algorithm, which is the focus of this research.

We propose an online data validation algorithm which compares e-nose measurements with a set of reference samples and allocate them accordingly to different zones. The zones are distinguished from each other using distinct colors like green, yellow, red, etc., to represent the extent of the validity of the measurement. The main focus of this work is summarized in the flowchart below.

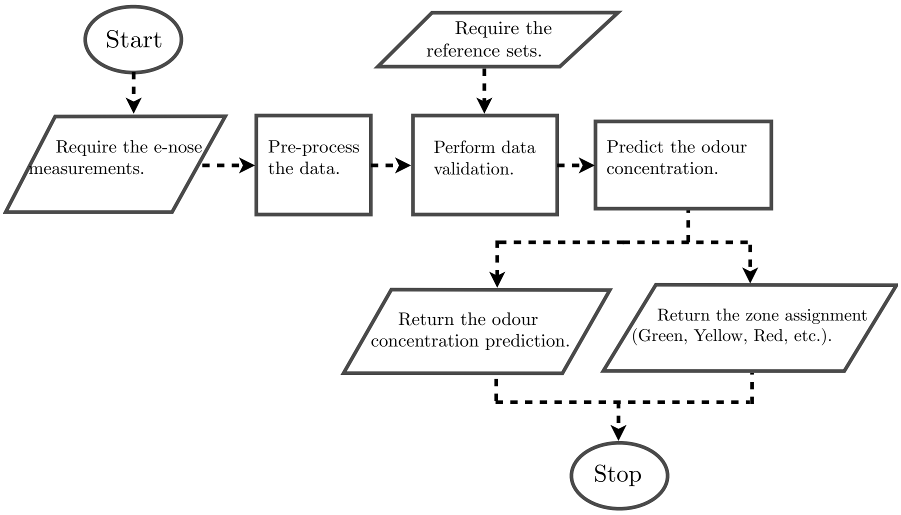
Start
Require the reference sets.
Require the e-nose measurements.
Pre-process the data.
Perform data validation.
Predict the odour concentration.
Return the odour concentration prediction.
Return the zone assignment (Green, Yellow, Red, etc.).
Stop

Fig. 1: A schematic flowchart of the proposed online task for an e-nose.

1.3 Data Preparation

The e-nose relies on a sensor array consists of several gas sensors. The number of the sensors depends on the purpose of analysis. Each sensor represents an attribute; the more the sensors are, the better the e-nose discriminate among analytes. Nonetheless the inclusion of too many sensors can lead to unnecessary data and a complex system.

The e-nose under the study includes 11 senors each designed to be responsive to a specific chemical compound in the air. However, senors react to almost all gases as they may not be highly selective. As a result, some of the sensors are highly positively correlated with each other, see Fig. 2 and Fig. 3 (left panel). Consider the data matrix $\mathbf{X}_{n\times p}$ with its rows being n independent realization of 11 sensor values, $\mathbf{x}^{\top}_{p\times 1}$ in which $\mathbf{a}^{\top}$ indicates the transpose of the vector $\mathbf{a}$.

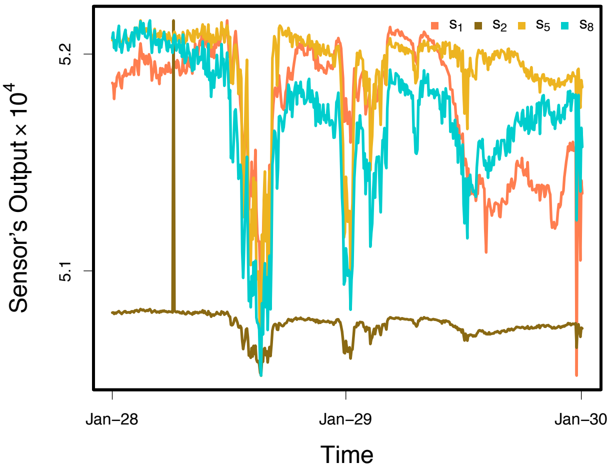

Fig. 2: Senor's output during three days of sampling for 4 randomly selected sensors.

The covariance matrix of $\mathbf{x}_{p\times 1}$, say $\boldsymbol{\Sigma} = [\sigma_{ij}]_{i,j=1,2,\ldots,p}$, is defined as

$$\boldsymbol{\Sigma}_{p\times p} = \mathrm{Cov}(\mathbf{x}) = \mathrm{E}\{(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\top}\},$$

where $\boldsymbol{\mu}$ represents the mean of $\mathbf{x}$, and E is the mathematical expectation. The covariance, σ_{ij}, measures the degree to which two attributes are linearly associated. However, in order to have a better idea about the relationship between two attributes, one needs to eliminate the effect of other attributes. The partial correlation is the correlation between two attributes, while controlling for the effects of other attributes. The inverse of covariance matrix is commonly known as precision or concentration matrix. The entries of $\boldsymbol{\Sigma}^{-1}$ have an interpretation in terms of partial correlation. Non-zero elements of $\boldsymbol{\Sigma}^{-1}$ implies conditional dependence. Therefore, the sparse estimation of $\boldsymbol{\Sigma}^{-1}$ pinpoints the block structure of attributes. Sparse estimation of $\boldsymbol{\Sigma}^{-1}$ set some of the $\boldsymbol{\Sigma}^{-1}$

entries to zero. Investigation of the inherent dependence between the sensor values is then performed by means of the partial correlation.

Here, the *graphical lasso* (Friedman et al., 2008) is considered for a better understanding of the existing relationship between the sensor values. Friedman et al. (2008) proposed estimating the covariance matrix such that its inverse, $\boldsymbol{\Sigma}^{-1}$, is sparse by applying a *lasso penalty* (Tibshirani, 1996). In Fig. 3 (right panel), the undirected graph connects two attributes which are conditionally correlated given all other attributes. The sensors 5 to 8 are correlated with each other conditioning on the effect of the others. This is also reflected in the heatmap of the correlation matrix Fig. 3 (left panel). This dependence must be taken into account while modelling the data. Gaussianity of the data is another crucial assumption that should be verified. The validity of this assumption for the sensor values is tested using various methods such as analyzing the distribution of individual sensor values, scatter plot of the linear projection of data using principal components, estimating the multivariate kurtosis and skewness, and also multivariate Mardia test, see Fig. 4.

2 Data Analysis

We aim to verify the validity of e-nose measurements by considering some reference samples for the purpose of comparison. These reference samples are collected when the e-nose functions normally, and the conditions are fully under control. The e-nose measurements are compared with reference samples and are allocated to various zones accordingly. These zones are distinguished by various colors, like green, yellow, red, etc., to indicate the status of e-nose measurements (Mirshahi et al., 2016).

Two distinct reference sets, if applicable, are recommended for data validation. *Reference* 1 consists of data in a period of sampling defined by an expert after installation of the e-nose. The data in this period of sampling is called as *proposed set*. *Reference* 2, upon its availability, is manually gathered samples from the field that are brought to the laboratory for quantification of their odour concentration. The data in this period of sampling is called *calibration set*, to emphasize that it can be incorporated for data modelling using a supervised learning algorithm.

If a new datum does not follow the overall pattern of data previously observed, then it is marked as an outlier and is assigned to Red zone. This zone represents a dramatic change in the pattern of samples and is referred to as “risky” observations. If the new datum is not an outlier and it is also located within the data polytope of the Reference 1 or the Reference 2, it is allocated to Green or Blue zone respectively. These zones represent the “safe” observations. If the new datum is not an outlier, but outside of the area of Green and Blue zones, they are assigned to Yellow zone. This zone displays potentially “critical” observations.

If large proportion of samples belong to the Yellow and Red zones, the reliability of the system should be suspected. Undesirable measurements can be the outcome of physical complications, such as sensor loss in the e-nose, or sudden changes in the chemical pattern of the environment. Zone assignment, therefore, require some outlier detection algorithms. For the Green and the Blue zones, the new samples are projected onto a subspace with lower dimension. Dimension reduction methods such as principal

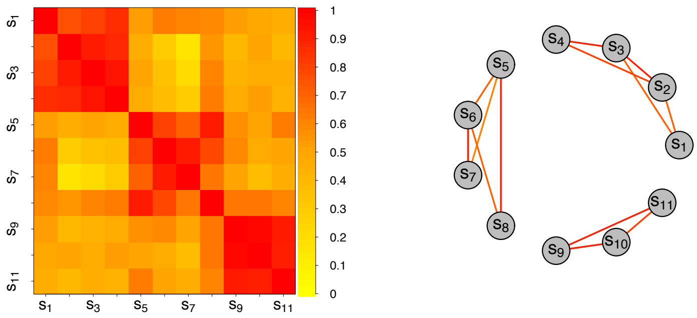

Fig. 3: Left panel, heatmap of the correlation matrix of the sensor values (s_1–s_{11}). Right panel, the undirected graph of partial correlation using the graphical lasso. The undirected graph of the right panel approves the block structure of the heatmap of the left panel.

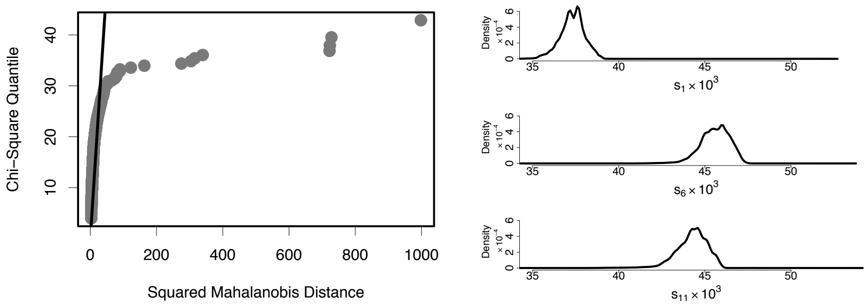

Fig. 4: Left panel, the Q-Q plot of squared Mahalanobis distance supposed to follow the chi-squared distribution for Gaussian data. Right panel, the marginal density for some randomly chosen sensor values. Both graphs confirm the non-Gaussianity of data.

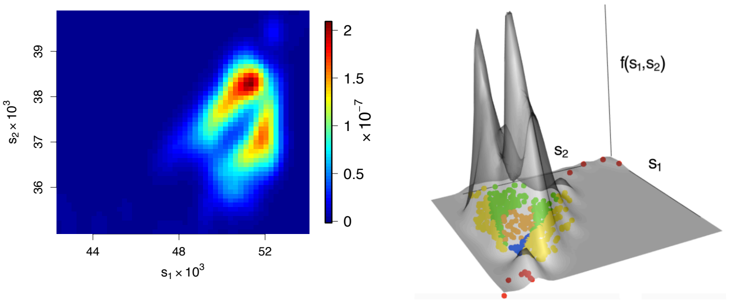

Fig. 5: Validity assessment for about 700 samples based on 2 sensor values. Left panel, the plot illustrates the contour map of estimated density function for the 2 sensors. Right panel, the density function of the samples demonstrated in $3D$ with zones identified for each of the samples in the sensor 1 (s_1) versus sensor 2 (s_2) plane. Higher density is assigned to the Green, Blue, and Orange zones compared to the Yellow and Red zones.

component analysis (PCA) can serve for this purpose (Jolliffe, 2002). PCA attempts to explain the data covariance matrix, $\hat{\boldsymbol{\Sigma}}$, by a set of components; these components are the linear combination of the primary attributes. PCA, basically, converts a set of possibly correlated attributes into a set of linearly uncorrelated axes through orthogonal linear transformations. The first k ($k < p$) principal components are the eigenvectors of the covariance matrix $\boldsymbol{\Sigma}$ associated with the k largest eigenvalues. The classical estimation of covariance matrix, $\hat{\boldsymbol{\Sigma}}$, is strongly influenced by outliers (Prendergast, 2008). As producing outlier is typical of sensor data, robust covariance estimation must be applied to avoid misleading results.

Robust principal component analysis (Hubert et al., 2005) is employed for dimension reduction purpose throughout this article. This robust PCA computes the covariance matrix through projection pursuit (Li and Chen, 1985) and minimum covariance determinant (Croux and Haesbroeck, 2000) methods. The robust PCA procedure can be summarized as follows:

1. The matrix of data is pre-processed such that the data spread in the subspace of at most $\min(n-1, p)$.
2. In the spanned subspace, the most obvious outliers are diagnosed and removed from data. The covariance matrix is calculated for the remaining data, $\hat{\boldsymbol{\Sigma}}_0$.
3. $\hat{\boldsymbol{\Sigma}}_0$ is used to decide about the number of principal components to be retained in the analysis, say k_0 ($k_0 < p$).
4. The data are projected onto the subspace spanned by the first k_0 eigenvectors of $\hat{\boldsymbol{\Sigma}}_0$.
5. The covariance matrix of the projected points is estimated robustly using minimum covariance determinant method and its k leading eigenvalues are computed. The corresponding eigenvectors are the robust principal components.

The Red zone represents the outliers of the samples as being measured by the e-nose through time. One common approach for detecting outliers in multivariate data is to use the Mahalonobis ditstance

$$D_m(\mathbf{x}_i) = \sqrt{(\mathbf{x}_i - \hat{\mu})^\top \hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}_i - \hat{\mu})}. \tag{1}$$

The large value of $D_m(\mathbf{x}_i)$ for $i = 1, 2, \ldots, n$, indicates that the observation $\mathbf{x}_i$ locates away from the centre of the data $\hat{\mu}$. As the estimation of μ and $\boldsymbol{\Sigma}$ is itself affected by outliers, the use of equation (1) is inadvisable for outlier detection. Even the robust plugin estimation of μ and $\boldsymbol{\Sigma}$ do not lead to any improvement as long as the associated outlier detection cut-offs are based on elliptical distributions. Hubert and Van der Veeken (2008); Brys et al. (2006) suggested an outlier detection method which does not assume any elliptical distribution for data. Their method is formed on a modified version of Stahel-Donoho outlyingness measure (Stahel, 1981; Donoho, 1982) and is called *adjusted outlyingness* (AO) criterion. For the observation $\mathbf{x}_i$, the Stahel-Donoho measure is

$$\mathrm{SD}(\mathbf{x}_i) = \sup_{\mathbf{a} \in \mathbf{R}^p} \frac{|\mathbf{a}^\top \mathbf{x}_i - \mathrm{median}(\mathbf{X}_n \mathbf{a})|}{\mathrm{mad}(\mathbf{X}_n \mathbf{a})}, \tag{2}$$

where $\mathrm{mad}(\mathbf{X}_n\mathbf{a}) = 1.483\ \mathrm{median}_i|\mathbf{a}^\top \mathbf{x}_i - \mathrm{median}(\mathbf{X}_n\mathbf{a})|$ is the median absolute deviation. The SD measure essentially looks for outliers by projecting each observation on many univariate directions. As it is not applicable to look for all possible directions, it is suggested that considering $250p$ directions, where p is the number of attributes, suffices and produces efficient results. Taking into account the effect of skewness in the SD measure results in the following AO

$$\mathrm{AO}_i = \sup_{\mathbf{a} \in \mathbf{R}^p} \begin{cases} \frac{\mathbf{a}^\top \mathbf{x}_i - \mathrm{median}(\mathbf{X}_n\mathbf{a})}{w_2 - \mathrm{median}(\mathbf{X}_n\mathbf{a})} & \text{if } \mathbf{a}^\top \mathbf{x}_i > \mathrm{median}(\mathbf{X}_n\mathbf{a}), \\ \frac{\mathrm{median}(\mathbf{X}_n\mathbf{a}) - \mathbf{a}^\top \mathbf{x}_i}{\mathrm{median}(\mathbf{X}_n\mathbf{a}) - w_2} & \text{if } \mathbf{a}^\top \mathbf{x}_i < \mathrm{median}(\mathbf{X}_n\mathbf{a}), \end{cases} \tag{3}$$

where w_1 and w_2 are the lower and upper whiskers of the adjusted boxplot (Hubert and Vandervieren, 2008). If the AO_i exceeds the upper whisker of the adjusted boxplot, it is then detected as an outlier.

The sample that is rendered as an outlier by AO measure, belongs to the Red zone. For the specification of the remaining zones, we need to define the polytopes of the samples in Reference 1 and Reference 2. These polytopes are built using the convex hull of the robust principal component *scores*. More specifically, the boundary of the Green zone is defined by computing the convex hull of the robust principal component scores of the Reference 1.

Before determining the color tag for each new data, the samples are checked for missing values and are imputed if needed by *multivariate imputation* methods such as Josse et al. (2011). The idea behind the validity assessment is visualized in Fig. 5. For simplicity, only 2 sensors are used for all computations in Fig. 5 and a $2D$ presentation of zones is plotted using the sensors' coordinates. Suppose that $\mathbf{X}_{n\times 11}$ represents the matrix of sensor values for n samples, $\mathbf{y}_n$ the vector of corresponding odour concentration values and $\mathbf{x}_l^\top$ is the lth row of $\mathbf{X}_{n\times 11}$, $l = 1, 2, \ldots, n$. Furthermore, suppose that n_1 refers to the number of samples in the proposed set of the sampling and n_2 refers to the number of samples in the calibration set. The samples of the proposed set are always available, but not necessarily the calibration set. Two different scenarios occur based on the availability of the calibration set.

If the calibration set is accessible, then Scenario 1 happens. Otherwise, we only deal with Scenario 2. Scenario 1 is a general case which is explained more in detail. The data undergo a pre-processing stage, including imputation and outlier detection, before any further analyses. Having done the pre-processing stage, data are stored as Reference 1, $\mathbf{X}_{n_1\times 11}$, and Reference 2, $\mathbf{X}_{n_2\times 11}$. The first k, e.g. $k = 2,3$, robust principal components of $\mathbf{X}_{n_1\times 11}$ are calculated and the corresponding *loading* matrix is denoted by $\mathbf{L}_1$. The pseudo code of two algorithms for Scenario 1 is provided below. Scenario 2 is a special case of Scenario 1 in which Sub-Algorithm (Scenario 1) is used with $\text{ConvexHull}^{(2)} = \varnothing$ that eliminates the Blue and Orange zones. Consequently, there is no model for odour concentration prediction in the Main Algorithm.

Sub-Algorithm (Scenario 1)

```
if the point $\mathbf{x}_l^\top, l = 1,2,\ldots,N$ is identified as an outlier by $AO$ measure then
    $\mathbf{x}_l^\top$ is in Red zone,
else if $\mathbf{x}_l^\top\mathbf{L}_1 \in \text{ConvexHull}^{(1)}$ AND $\mathbf{x}_l^\top\mathbf{L}_1 \notin \text{ConvexHull}^{(2)}$ then
    $\mathbf{x}_l^\top$ is in Green zone,
else if $\mathbf{x}_l^\top\mathbf{L}_1 \notin \text{ConvexHull}^{(1)}$ AND $\mathbf{x}_l^\top\mathbf{L}_1 \in \text{ConvexHull}^{(2)}$ then
    $\mathbf{x}_l^\top$ is in Blue zone,
else if $\mathbf{x}_l^\top\mathbf{L}_1 \in \text{ConvexHull}^{(1)}$ AND $\mathbf{x}_l^\top\mathbf{L}_1 \in \text{ConvexHull}^{(2)}$ then
    $\mathbf{x}_l^\top$ is in Orange zone,
else
    $\mathbf{x}_l^\top$ is in Yellow zone.
end if
```

Main Algorithm (Scenario 1)

```
Require: $\mathbf{X}_{n_1\times 11}$, $\mathbf{X}_{n_2\times 11}$, and the loading matrix $\mathbf{L}_1$ using robust PCA over Reference 1,
    $\mathbf{X}_{n_1\times 11}$.
ConvexHull$^{(1)}$ $\leftarrow$ the convex hull of the projected values of the Reference 1, $\mathbf{X}_{n_1\times 11}\mathbf{L}_1$.
Train a supervised learning model on Reference 2, $\mathbf{X}_{n_2\times 11}$, and its odour concentration vector, $\mathbf{y}_{n_2}$.
ConvexHull$^{(2)}$ $\leftarrow$ the convex hull of the projected values of the Reference 2, $\mathbf{X}_{n_2\times 11}\mathbf{L}_1$.
Do Sub-Algorithm for new data $\mathbf{x}^*$.
Predict the odour concentration for new data $\mathbf{x}^*$ using the trained supervised learning model.
```

In Section 3, a set of simulated data is used to verify the relevancy of our proposed algorithm and the choice of statistical methods. The applicability of our algorithm is also tested based on 8 months sampling from the e-nose in Section 4.

3 Simulation

We examine the methodology on two sets of simulated data to highlight the importance of the assumptions such as non-elliptical contoured distribution and robust estimation considered in our methodology. In each example, we stored the simulated data in the matrix $\mathbf{X}_{n\times 2}$, where $\mathbf{x}_l^\top = (x_{l1}, x_{l2});\ l = 1,2,\dots,n$.

In the first example, the data is simulated from a mixture distribution with 10% contamination. The elements of mixture distribution are chosen arbitrarily from Gaussian and the Student's t-distribution.

We simulated data from the bivariate skew t-distribution (Gupta, 2003) in the second example in order to test the effect of skewness on our algorithm, .

Using classical approaches for outlier detection without considering the actual data distribution, mistakenly renders many observations as outliers, Fig. 6 and Fig. 7 (top right panel). The parameters of interest, the mean vector and the covariance matrix, need to be estimated robustly, otherwise the confidence region misrepresents the underlying distribution. In Fig. 6 and Fig. 7 (bottom left panel), the classical confidence region is pulled toward the outlier observations. On the contrary, the robust confidence region perfectly unveil the distribution of the majority of observations because of the robust and efficient estimation of the mean and the covariance matrix. Consequently, the classical principal components are affected by the inefficient estimation of the covariance matrix. We proposed using methods that deal with asymmetric data appropriately. Adjusted outlyingness (AO) measure identifies the outliers of the data correctly. Considering a sub-sample of data as Reference 1 in each of the examples, the result of the Main Algorithm can be observed in the right bottom panel of Fig. 6 and Fig. 7.

4 Experiment

In order to evaluate the performance of our data validation method, we implement the Main Algorithm on a collection of e-nose measurements. We decide to keep the the first 3 robust principle components of the data $PC1$, $PC2$, $PC3$ for simplification and the easy visualization. The 3 principal components correspond to the 3 largest eigenvalues of the robust covariance matrix. Prior to the implementation of the Main Algorithm, the data undergoes a pre-processing stage including the imputation of the missing values.

The validity of the e-nose measurements are identified using the Main Algorithm for the 8 months of sampling. In favor of more readable graphs, only a subset of 500 samples out of 200 thousands of observations are plotted. In Fig. 8, the sample points are drawn in gray and each zone is highlighted using its corresponding color. The circles in Fig. 8 are also illustrated on $PC1$ and $PC2$ plane for a better demonstration of the zones.

The interpretation of a zones is heavily depends on its definition. For instance, the Green, Blue, and Orange zones, represent samples that are very close the samples that have already been observed in either Reference 1 or Reference 2. As the observations in reference sets were entirely under control, the Green, Blue, and Orange zones affirm the validity of the samples. In addition, the accuracy of the gas concentration predicted for these zones is certified. On the other hand, the gas concentration prediction for samples

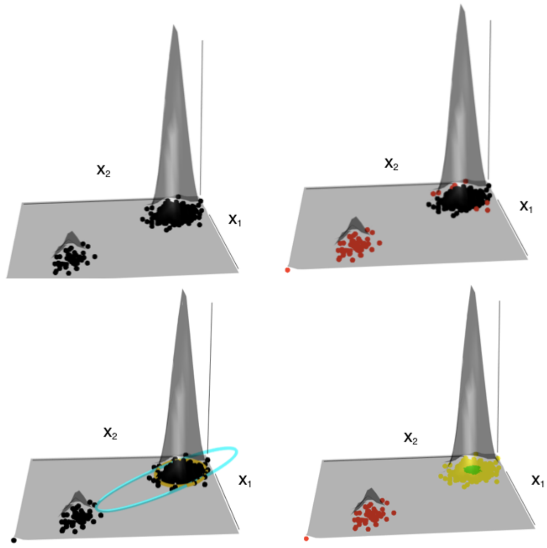

Fig. 6: Top left panel, the simulated data from the mixture distribution $f(x) = (1-\varepsilon)f_1(x) + \varepsilon f_2(x)$ with contamination proportion of $\varepsilon = \frac{1}{10}$, and f_1 and f_2 being the Gaussian and Student's t-distribution respectively. Top right panel, the outliers of data are identified and highlighted with red using the classical Mahalonobis distance and 95th percentile of the chi-squared distribution with two degrees of freedom. Bottom left panel, the 95% confidence region for the data is computed using the classical estimates of parameters (cyan) and the robust estimates (gold). Bottom right panel, the Main Algorithm is implemented and the zones are plotted using their associated color tag.

in the Red zone is less accurate compared with that of the Green, Blue, and Orange zones.

The data that are significantly dissimilar to the already observed data deserve further attention. These data are outliers and are reported in the Red zone. Similarly, the gas concentration predictions associated with samples in the Red zone can be very misleading. Generating a remarkable percentage of samples belonging to the Yellow and the Red zones refers to the possible failure of the e-nose equipment.

5 Computational Complexity

Here, we discuss the computational complexity of our proposed algorithm (Main Algorithm). First, a brief introduction to computational complexity is given to facilitate the understanding.

The computational complexity of an algorithm is studied asymptotically by the big O-notation (Arora and Barak, 2009). The big O-notation explains how quickly the runtime of an algorithm grows relative to its input. For instance, sum of n values require $(n-1)$ operations. Consequently, the mean requires n operations reserving one for the division of the sum by n. As they are both bounded by a linear function, they have com-

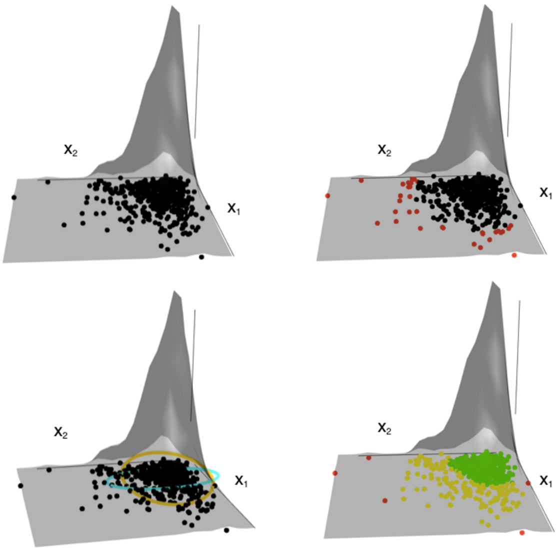

Fig. 7: Top left panel, the simulated data from the bivariate skew t-distribution. Top right panel, the outliers of data are identified and highlighted with red using the classical Mahalonobis distance and 95th percentile of the chi-squared distribution with two degrees of freedom. Bottom left panel, the 95% confidence region for the data is computed using the classical estimates of parameters (cyan) and the robust estimates (gold). Bottom right panel, the Main Algorithm is implemented and the zones are plotted using their associated color tag.

putational complexity of order $O(n)$. In other words, the performance of the sum and mean grow linearly and in a direct proportion to the size of the input. Note that not all algorithms are computationally linear. Computational complexity of covariance matrix, for instance, is $O(np^2)$ where n is the sample size and p is the number of attributes. Since each covariance calls for sum of the pairwise cross-products each of complexity $O(n)$. In total, there are $\frac{p(p-1)}{2}$ off-diagonal cross products and p square sums for the diagonal entries of the covariance matrix. This yields $n\{p(p-1)+p\}$ operations. For a fixed number of attributes p, the computation is of order $O(n)$. Likewise, for a fixed number of observations the computation is of order $O(p^2)$. Another nontrivial example for non-linear algorithm is PCA or the robust PCA. Computation of robust principal components involves various operations that has been briefly discussed in Section 2. Computational complexity of robust PCA is discussed below. Computation of robust PCA comprises the following steps:

1. Reducing the data space to an affine subspace spanned by the n observations using singular value decomposition of $(\mathbf{X}-\mathbf{1}_n\hat{\boldsymbol{\mu}})^\top(\mathbf{X}-\mathbf{1}_n\hat{\boldsymbol{\mu}})$, where $\mathbf{1}_n$ is the column vector of n dimension with all entries equal to 1. This step is of order $O(p^3)$, see Golub and Loan (1996) and Holmes et al. (2007).

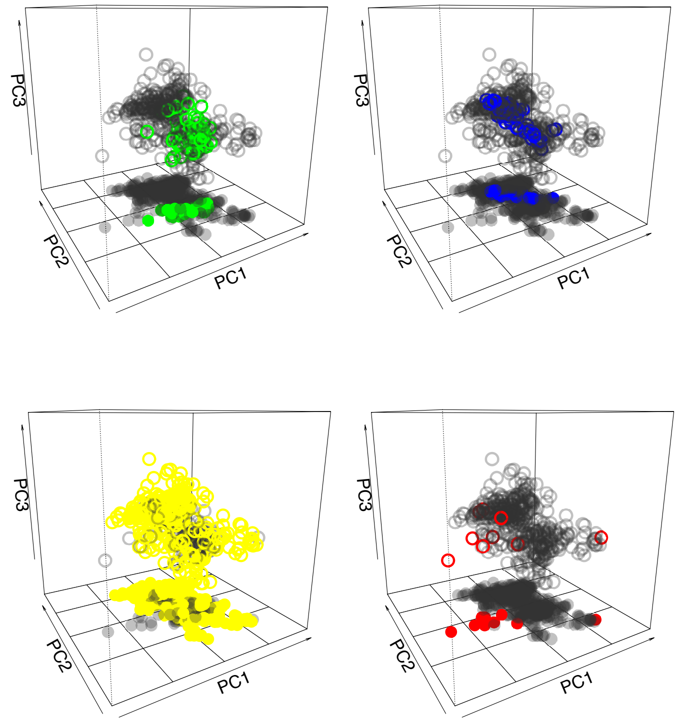

Fig. 8: A random sample of size $n = 500$ is plotted over the first three robust principal components coordinates. From top left panel to bottom right panel, the colored blobs represent green, blue, yellow, and red zones respectively.

2. Finding the least outlying points using the Stahel-Donoho affine-invariant outlyingness (Stahel, 1981; Donoho, 1982). Adjusting this outlyingness measure by the minimum covariance determinant location and scale estimators is of order $O(pn\log n)$, see Hubert and Van der Veeken (2008) and Hubert et al. (2005). Then the covariance matrix of the non-outliers data, $\hat{\boldsymbol{\Sigma}}_0$, is calculated which is computationally less expensive.
3. Performing the principal component analysis on $\hat{\boldsymbol{\Sigma}}_0$ and choosing the number of projection components (say $k_0 < p$) to be retained. Computing the $\hat{\boldsymbol{\Sigma}}_0$ needs np^2 operations. Thus its complexity is $O(np^2)$. The spectral decomposition of the covariance matrix is achieved by applying matrix-diagonalization method, such as singular value decomposition or Cholesky decomposition. This results in $O(p^3)$ computational complexity. Determining the k_0 largest eigenvalues and their corresponding eigenvectors has time complexity of $O(k_0p^2)$ (Du and Fowler, 2008). As a result, the time complexity of this step is $O(np^2)$.

4. Projecting the data onto the subspace spanned by the first k_0 eigenvectors, i.e $(\mathbf{X} - \mathbf{1}_n\hat{\boldsymbol{\mu}})\mathbf{P}_{p\times k_0}$ where $\mathbf{P}_{p\times k_0}$ is the matrix of eigenvectors corresponding to the first k_0 eigenvalues. This step has $O(npk_0)$ time complexity.
5. Computing the covariance matrix of the projected points using the method of fast minimum covariance determinant has the computational complexity which is sublinear in n, for fixed p. This is $O(n)$ (Rousseeuw and Driessen, 1999). The calculation of the spectral decomposition of the final covariance matrix is bounded by $O(nk_0)$ time complexity.

Remark 1 *The computational complexity of robust PCA is $O(\max\{pn\log n, np^2\})$, or $O(p^2 n\log n)$ considering the worst case complexity.*

To ascertain the complexity of the Main Algorithm, one needs to analyze each step separately. The measurement validation in e-nose broadly necessitates the calculation of certain steps of the Main Algorithm including Step Require, Step 1, Step 3, and Step 4. All these tasks excluding Step 4 of the Main Algorithm (Sub-Algorithm) must be run only once. Step 4 duplicates upon the arrival of the new observations.

First, we start by evaluating the complexity of Step Require, Step 1, and Step 3 that should be run once. Afterwards Step 4 is analyzed in a similar fashion. Note that for the e-nose data, the number of samples is generally much greater than the number of sensors p. In addition, as the number of sensors p is fixed in an e-nose equipment, the computational complexity is reported as the function of number of samples only.

The Main Algorithm starts with the robust PCA over the Reference 1. As a result, Step Require has $O(\{n_1 \log n_1\})$ complexity assuming p to be fixed. Step 1 requires $O(n_1 k_0)$ computing time for computing $\mathbf{X}_{n_1\times 11}\mathbf{L}_1$ where k_0 stands for the the number of eigenvectors retained in the loading matrix $\mathbf{L}_1$. Computing the convex hull of these projected values for $k_0 \leq 3$ is of order $O(n_1 \log n_1)$. For $k_0 > 3$, the computational complexity of hull increases exponentially with k_0, see Ottmann et al. (1995) and Chan (1996). Similarly, the same complexity is valid for Step 3. Performing some preprocessing steps on the Reference sets including outlier detection using AO measure has $O(n_1 \log n_1)$ complexity (Hubert and Van der Veeken, 2008) assuming that $n_1 > n_2$, which is common in practice. As a result, Step Require, Step 1, and Step 3 which is performed only once take $O(n_1 \log n_1)$ run-time.

Now, we analyze Step 4 in terms of its computational complexity. Step 4 mainly does the following three tasks.

i) Accumulating the new observations with the past history, $\mathbf{X}^{\top}_{1:t\times p} = [\mathbf{X}^{\top}_{1:t-1\times p} : \mathbf{x}_{t\times p}]$ where $n_1 < t \leq n$, and identifying outliers using AO measure. This has computational complexity of $O(t\log t)$.
ii) Projecting the observations onto the space of Reference 1, $\mathbf{x}_l^{\top}\mathbf{L}_1$. This is a simple matrix product and has the computational complexity of $O(k_0 p)$.
iii) Verifying whether the projection of data, $\mathbf{x}_l^{\top}\mathbf{L}_1$, locates within the convex hull of either Reference 1 or Reference 2 which is equivalent to solving a linear optimization with linear constraints (Kan and Telgen, 1981; Dobkin and Reiss, 1980). The algorithm used for this purpose has computational complexity which varies quadratically with respect to the number of vertices of the convex hull, and has $O(n_1^2 k_0)$

complexity in the worst case. The R code used for solving this linear program resembles the MATLAB code [1] and is available upon the request.

Thus, the computational complexity of Step 4 is $O(t\log t)$ as in practice the convex hull of Reference 1 is computed, in Step 1, and kept fixed prior to this step.

Remark 2 *The computational complexity of Main Algorithm is $O(t\log t)$.*

The mean CPU time in seconds for Step Require, Step 1, and Step 3 that need to be run once and Step 4 which duplicates for each new sample, are reported in Fig. 9.

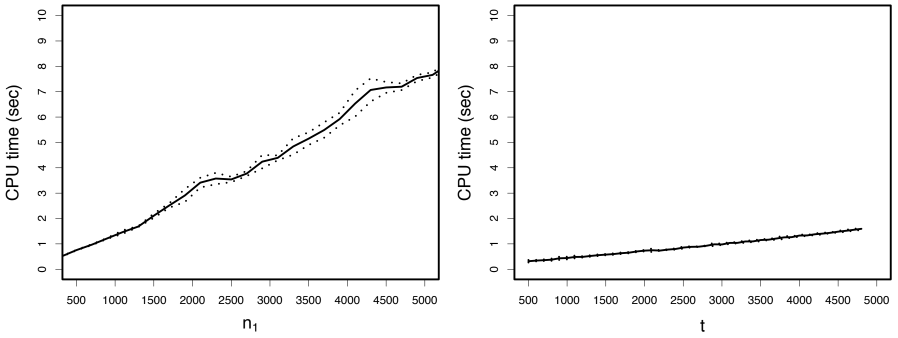

Fig. 9: The solid line shows the mean CPU time in seconds as a function of input being run on 1.3 GHz i5 processor. The dashed lines depict the lower and the upper bound of the 95% confidence interval for the mean CPU time. Left panel, the run-time corresponding to Step Require, Step 1, and Step 3 as the function of the number of samples in Reference 1, n_1. Right panel, the run-time associated with Step 4 as a function of the total number of samples upto the moment, t. In each iteration, 100 new observations are sampled.

Fig. 9 confirms that the run-times for the ensemble of the steps Require, 1, and 3 and the step 4 agree with the computational complexity evaluated theoretically earlier. This implies that measurement validation can be achieved with $O(t\log t)$ time complexity employing our proposed method.

6 Conclusion

An electronic nose device, which mainly consists of a multi-sensor array, attempts to mimic the human olfactory system. The sensor array is composed of various sensors selected to react to a wide range of chemicals to distinguish between mixtures of analytes. Employing the pattern recognition methods, the sensor's output are compared

[1] http://www.mathworks.com/matlabcentral/fileexchange/10226-inhull

with reference samples to predict odour concentration. Consequently, the accuracy of predicted odour concentration depends heavily on the validity of sensor's output. An automatic procedure that detects the samples' validity in an online fashion has been a technical shortage and is addressed in this work. A measurement validation process provides the possibility of attaching a margin of error to the predicted odour concentrations. Furthermore, it allows taking the subsequent actions such as re-sampling to re-calibrate the models or checking the e-nose device for possible sensor failures. The proposed measurement validation algorithm initiates a new development in automatic odour detection by minimizing the manpower intervention.

Acknowledgement

The project was funded by the natural sciences and engineering research council of Canada (NSERC) through the industrial partnership Engage program. Vahid Partovi Nia is partially supported by the Canada excellence research chair in data science for real-time decision making.

Bibliography

Arora, S. and Barak, B. (2009). *Computational complexity: A Moden approach*. Cambridge University Press.

Artursson, T., Eklov, T., Lundstrom, I., Martensson, P., Sjostrom, M., and Holmberg, M. (2000). Drift correction methods for gas sensors using multivariate methods. *Journal of Chemometrics*, 14:711–723.

Bermak, A., Belhouari, S. B., Shi, M., and Martinez, D. (2006). Pattern recognition techniques for odor discrimination in gas sensor array. *Encyclopedia of Sensors*, X:1–17.

Brys, G., Hubert, M., and Rousseeuw, P. J. (2006). A robustification of independent component analysis. *Chemometrics*, 19:364–375.

Carlo, S. D. and Falasconi, M. (2012). Drift correction methods for gas chemical sensors in artificial olfaction systems: techniques and challenges. *Advances in Chemical Sensors*, 14:305–326.

Chan, T. M. (1996). Output-sensitive results on convex hulls, extreme points, and related problems. *Dicrete and Computational Geometry*, 16(4):369–387.

Croux, C. and Haesbroeck, G. (2000). Principal components analysis based on robust estimators of the covariance or correlation matrix: Infulence functions and efficiencies. *Biometrika*, 87:603–618.

Dobkin, D. P. and Reiss, S. P. (1980). The complexity of linear programing. *Theoritical Computer Science*, 11:1–18.

Donoho, D. L. (1982). Breakdown properties of multivariate location estimators. *Ph.D. qualifying paper Harvard University*.

Du, Q. and Fowler, J. E. (2008). Low-complexity principal component analysis for hyperspectral image compression. *International Journal of High Performance Computing Applications*, 22:438–448.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9:432–441.

Gardner, J. and Bartlett, P. (1994). A brief history of electronic noses. *Sensors and Actuators B: Chemical*, 18:211–220.

Golub, G. H. and Loan, C. F. V. (1996). *Matrix Computations*. The John Hopkins University Press, 3rd edition.

Gupta, A. (2003). Multivariate skew t-distribution. *Statistics*, 37:359–363.

Gutierrez-Osuna, R. (2002). Pattern analysis for machine olfaction: A review. *IEEE Sensors Journal*, 2:189–202.

Holmes, M. P., Gray, A. G., and Isbell, C. L. (2007). Fast SVD for large-scale matrices. *Workshop on Efficient Machine Learning at NIPS*, 58.

Hubert, M., Rousseeuw, P. J., and Branden, K. V. (2005). ROBPCA: A new approach to robust principal component analysis. *Thechnometrics*, 47:64–79.

Hubert, M. and Van der Veeken, S. (2008). Outlier detection for skewed data. *Journal of Chemometrics*, 22:235–246.

Hubert, M. and Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Computational Statistics and Data Analysis*, 52:5186–5201.

Jolliffe, I. (2002). *Principal Component Analysis*. Springer.

Josse, J., Pagès, J., and Husson, F. (2011). Multiple imputation in principal component analysis. *Advances in Data Analysis and Classifications*, 5:231–246.

Kan, A. R. and Telgen, J. (1981). The complexity of linear programming. *Statistica Neerlandica*, 2.

Kermiti, M. and Tomic, O. (2003). Independent component analysis applied on gas sensor array measurement data. *IEEE Sensors Journal*, 3:218–228.

Li, G. and Chen, Z. (1985). Projection-pursuit approach to robust dispersion matrices and principal components: primary theory and monte carlo. *Journal of the American Statistical Association*, 80:759–766.

McGinley, P. C. and Inc, S. (2002). Standardized odor measurement practices for air quality testing. *Air and Waste Management Association Symposium on Air Quality Measurement Methods and Technology, San Francisco, CA*.

Mirshahi, M., Partovi Nia, V., and Adjengue, L. (2016). Statistical measurement validation with application to electronic nose technology. In *Proceedings of the 5th International Conference on Pattern Recognition Applications and Methods*, pages 407–414.

Ottmann, T., Schuierer, S., and Soundaralakshmi, S. (1995). Enumerating extreme points in higher dimensions. *STACS 95: 12th Annual Symposium on Theoretical Aspects of Computer Science, Lecture Notes in Computer Science*, 900:562–570.

Padilla, M., Perera, A., Montoliu, I., Chaudry, A., Persaud, K., and Marco, S. (2010). Drift compensation of gas sensor array data by orthogonal signal correction. *Journal of Chemometrics and Intelligent Labrotory System*, 100:28–35.

Persaud, K. and Dodd, G. (1982). Analysis of discrimination mechanisms in the mammalian olfactory system using a model nose. *Nature*, 299:352–355.

Prendergast, L. (2008). A note on sensitivity of principal component subspaces and the efficient detection of influential observations in high dimensions. *Electronic Journal of Statistics*, 2:454–467.

Rousseeuw, P. J. and Driessen, K. V. (1999). A fast algorithm for the minumum covariance determinant estimator. *Technometrics*, 41:212–223.

Stahel, W. A. (1981). Robust estimation: Infinitesimal optimality and covariance matrix estimators. *Ph.D. thesis, ETH, Zurich*.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.

Zuppa, M., Distante, C., Persaud, K. C., and Siciliano, P. (2007). Recovery of drifting sensor responses by means of DWT analysis. *Journal of Sensors and Actuators*, 120:411–416.