Les Cahiers du GERAD ISSN: 0711–2440

Self-assessed electronic nose

M. Mirshahi, V. Partovi Nia,
L. Adjengue

G–2016–28

May 2016

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

Self-assessed electronic nose

Mina Mirshahi [a]
Vahid Partovi Nia [a]
Luc Adjengue [a]

[a] Department of Mathematics and Industrial Engineering,
Polytechnique Montréal, Montréal, Québec, Canada

mina.mirshahi@polymtl.ca
vahid.partovonia@polymtl.ca
luc.adjengue@polymtl.ca

May 2016

Les Cahiers du GERAD
G–2016–28

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 identifiabilty and perceivability of an odour. It is of high importance to assess the validity of measurements during the sampling as the qualified measurements can only produce an accurate prediction for odour concentration. The physical impairment of the e-nose and/or environmental factors (including wind, humidity, temperature, etc.) can introduce significant amount of noise into sensor measurements. Inevitably, the pattern recognition results are affected. Here, we propose an online algorithm to assess the validity of sensor measurements. The algorithm enables e-nose to perform a self-assessment procedure during the sampling before utilizing the data for pattern recognition phase. The proposed algorithm is proved to be computationally cheap and easy to implement.

Key Words: Artificial olfaction, computational complexity, electronic nose, gas sensor, odour, outlier, robust covariance estimation.

1 Introduction

The ability to recognize the chemicals in the environment is a very basic and essential need for the living organisms; from a single-cell amoebae to human beings, all species are provided with a chemical awareness system. Human beings have three sensory systems to detect odours: sense of taste, sense of smell, and chemical feel with receptors all over the body. All species employ their chemical senses to approach and being attracted to possibly safe conditions, as well as avoiding and being resisted to the harmful ones. As for human beings, in every breath, the sense of smell collects a sample from its environment and forwards it to the brain for further analyses. Unlike the sense of taste, smell can be captured from a distance and assist the brain in producing a warning. Unfortunately, the human sense of smell does not respond to all harmful air pollutants. Additionally, sensitivity of humans to many air pollutants varies — one can be accustomed to a toxic smell. In the last decade, great attention has been paid to the subject of air quality because it directly influences the environmental and human health. A crucial element in assessment of indoor and outdoor air quality is auditing the odourants. There exists various odour measurement techniques such as dilution-to-threshold, olfactometers, and referencing techniques (McGinley and Inc, 2002). The performance of these approaches depend on human evaluation. Due to the high variability of individual's sensitivity, the common methods mostly lack accuracy. In 1982, the first gas multisensor array was invented as primary artificial olfaction (Persaud and Dodd, 1982). The term electronic nose (e-nose) was introduced by Gardner and Bartlett (1994). E-nose is an artificial olfactory system which consists of an array of gas sensors. The e-nose is designed for recognizing complex odours in its surrounding environment. The gas sensor array receives chemical information about gaseous mixtures as input and converts it to measurable signals. Sensors act independently and simultaneously in this device. Cross-sensitivity of gas sensors is inevitable in sensor array structure. The cross-sensitivity is the interaction among chemicals that leads to a different signal from the component in a mixture compared to the single component. Gas sensor's performance is affected by different elements which make it unstable and less sensitive to odours. One of the most serious deterioration in sensors is owing to a phenomenon called *drift*. Drift is a temporal change in sensor's response while all other external conditions are kept constant. The majority of manufactured sensor arrays are subject to drift, and several methods have been introduced to overcome this problem (Carlo and Falasconi, 2012; Artursson et al., 2000; Padilla et al., 2010; Zuppa et al., 2007). The behavior of a sensor is directly influenced by the surrounding chemical and physical conditions. For instance, the sensor response may depend on the temperature of the gas under examination. Therefore, thermal conditions around the sensing elements need to be supervised. 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. Numerous methods have been developed for analyzing the gas sensor array data, including Gutierrez-Osuna (2002); Kermiti and Tomic (2003); Bermak et al. (2006).

2 Problem statement

The e-nose has partially addressed the human sense of smell in diverse industrial sites. Unwanted variability may occur in sensor's output data. This happens due to environmental factors or physical impairment of the system, since e-noses are installed in outdoor fields where the conditions can dramatically fluctuate. This demands for monitoring the critical factors through adding extra sensors and temperature compensation in sensor pre-processing. The sensor's output is used to quantify odour concentration. Transferring the data to olfactometry is both time consuming and costly. Only small portions of data are appointed for further analyses of its concentration in olfactometry. Pattern recognition methods are employed in order to predict the odour concentration for each set of sensor values. To assess the accuracy of predictions, the validity of sensor values must be ensured. Sensors in the e-nose structure may report incorrect values or some 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.

3 Data description

The data under the study include 11 distinct attributes, each representing sensor values of the e-nose. Sensors react to almost all gases in the air, but they are designed so that each sensor is more sensitive to a specific type of gas. Some of the sensors are highly positively correlated with each other, see Fig. 1 and Fig. 2 (left panel).

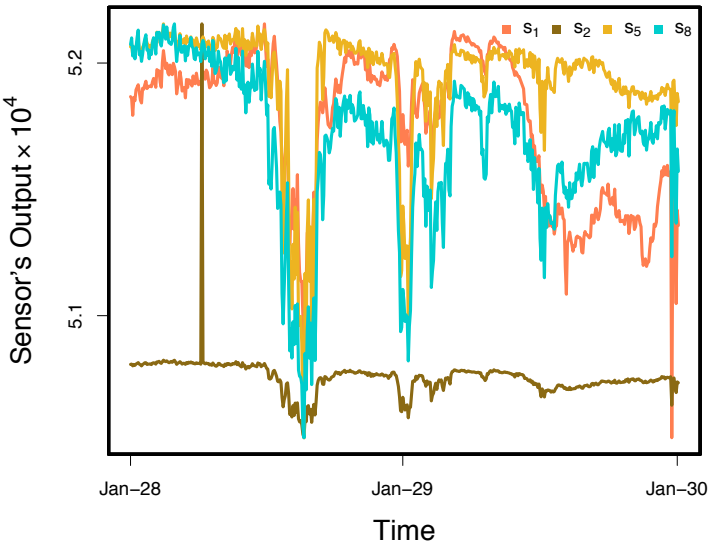

Figure 1: Sensor's output during three days of sampling for 4 randomly selected sensors.

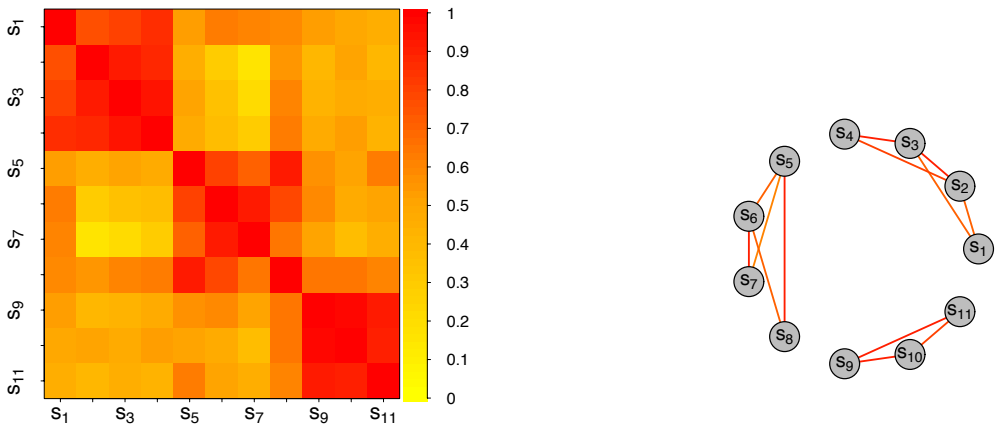

Figure 2: Left panel, heatmap of the correlation matrix of the sensor values (s_1–s_{11}). Right panel, the undirected graph of partial correlation using the graphical lasso. The undirected graph of the right panel approves the block structure of the heatmap of the left panel.

Suppose that $\mathbf{x}_{p\times 1}^{\top}$ is a random vector of $p = 11$ attributes, in which $\mathbf{a}^{\top}$ illustrates the transpose of vector $\mathbf{a}$, and its n independent realization are stored in the rows of data matrix $\mathbf{X}_{n\times p}$. The covariance matrix of $\mathbf{x}_{p\times 1}$, say $\mathbf{\Sigma} = [\sigma_{ij}]_{i,j=1,2,\ldots,p}$, is defined as

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

where $\boldsymbol{\mu}$ represents the mean of $\mathbf{x}$, E is the expectation operator. The covariance, σ_{ij}, measures the degree to which two attributes are linearly associated. It is well-known that the inverse of covariance matrix, commonly

known as precision matrix, yields the partial correlation between the attributes. The partial correlation is the correlation between two attributes conditioning on the effect of other attributes. Non-zero elements of $\mathbf{\Sigma}^{-1}$ implies the conditional dependence. Therefore, the sparse estimation of $\mathbf{\Sigma}^{-1}$ pinpoints the block dependent structure of attributes. The sparse estimation of $\mathbf{\Sigma}^{-1}$ set some of the $\mathbf{\Sigma}^{-1}$ entries exactly to zero. Investigation of the inherent dependence between the sensor values is then performed by means of the partial correlation. In order to obtain a clear image of sensors which are potentially grouped together, the *graphical lasso* (Friedman et al., 2008) is used. Friedman et al. (2008) considered estimating the inverse of covariance matrix, $\mathbf{\Sigma}^{-1}$, sparsely by applying a *lasso penalty* (Tibshirani, 1996). In Figure 2 (right panel), the undirected graph connects two variables which are conditionally correlated given all other attributes. For instance, the sensors 9, 10, and 11 are conditionally correlated with each other. This also agrees with the heatmap of the correlation matrix Figure 2 (left panel). Thus, this dependence must be taken into account while modeling data. Another vital assumption that should be verified is the Gaussianity of the data. The non-Gaussianity of the sensor values is established 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 Figure 3.

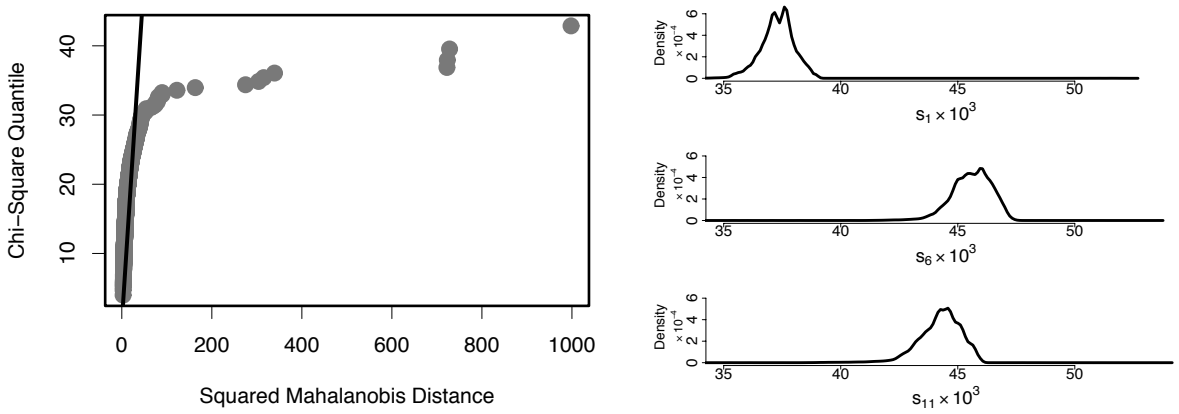

Figure 3: Left panel, the Q-Q plot of squared Mahalanobis distance supposed to follow chi-square distribution for Gaussian data. Right panel, the marginal density for some randomly chosen sensor values. Both graphs confirm the non-Gaussianity of data.

4 Methodology

In order to demonstrate the validity of the e-nose measurements, we aim to allocate each sample to different zones. To be able to verify the validity of the measurements, it is necessary to have some reference samples for the purpose of comparison. These reference samples are collected while the e-nose is at its best performance, and the conditions are fully under control. For the data set under the study, there are two distinct reference sets. *Reference* 1 is constituted of data in a period of sampling defined by an expert after installation of the e-nose. We call the data in this period of sampling as *proposed set*. *Reference* 2, upon its availability, is manually gathered samples from the field and brought to the laboratory to quantify the odour concentration. We call the latter data, *calibration set* to emphasize that it can be used for data modelling using supervised learning. If new data diverge greatly from the overall pattern of data previously seen, then it is marked as an outlier and is allocated to the red zone. This zone represents a dramatic change in the pattern of samples and refer to "risky" samples. If new data is non-outlier and it is also located within the data polytope of the Reference 1 or the Reference 2, it is assigned to green or blue zone respectively. These zones represent the "safe" samples. If new data is non-outlier, but outside of the area of green and blue zones, it is assigned to yellow zone. This zone displays potentially "critical" samples.

Producing many samples belonging to the yellow and the red zones is an indication of a major flaw in the system. Physical complications, such as sensor loss in the e-nose, or sudden changes in the chemical pattern of the environment, account for all undesirable measurements. Zone assignment, therefore, require some outlier detection algorithms. To define the green and the blue zones, the new samples are projected onto a lower dimension subspace. Dimension reduction methods such as principal component analysis (PCA) can serve this purpose (Jolliffe, 2002). PCA transforms a collection 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 $\mathbf{\Sigma}$ associated with the k largest eigenvalues. PCA exploits empirical covariance matrix, $\hat{\mathbf{\Sigma}}$, which is extremely sensitive to outliers (Prendergast, 2008). Since the data contain many outliers, 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 paper. 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, $\hat{\mathbf{\Sigma}}_0$.
3. $\hat{\mathbf{\Sigma}}_0$ 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 $\hat{\mathbf{\Sigma}}_0$.
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.

To define the red zone, it is required to find the outliers of data as it is being measured by the e-nose through time. As the data fail to follow a Gaussian distribution, outlier detection methods that rely on the assumption of elliptical contoured distribution should be avoided. Here, outliers are flagged by means of *adjusted outlyingness* (AO) criterion (Brys et al., 2006). If a sample is detected as an outlier by AO measure, it 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. A short description of each zone is provided in Table 1. Before determining the color tag for each new data, the samples are checked for missing values and are imputed in case needed by *multivariate imputation* methods such as Josse et al. (2011). The idea behind the validity assessment is visualized in Fig. 4. For simplicity, only 2 sensors are used for all computations in Fig. 4 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

Table 1: Description of each zones in validity assessment procedure.

Zone	Description
Red	Observations that are outliers in terms of AO measure.
Green	Observations that are non-outliers in terms of AO measure. Moreover, they fall into the polytope of the Reference 1.
Blue	Observations that are non-outliers in terms of AO measure. Moreover, they fall into the polytope of the Reference 2.
Orange	Observations that are non-outliers in terms of AO measure. Moreover, they fall into the polytopes of both the Reference 1 and the Reference 2.
Yellow	Observations that are non-outliers in terms of AO measure. Moreover, they do not fall into the polytope of neither the Reference 1 nor the Reference 2.

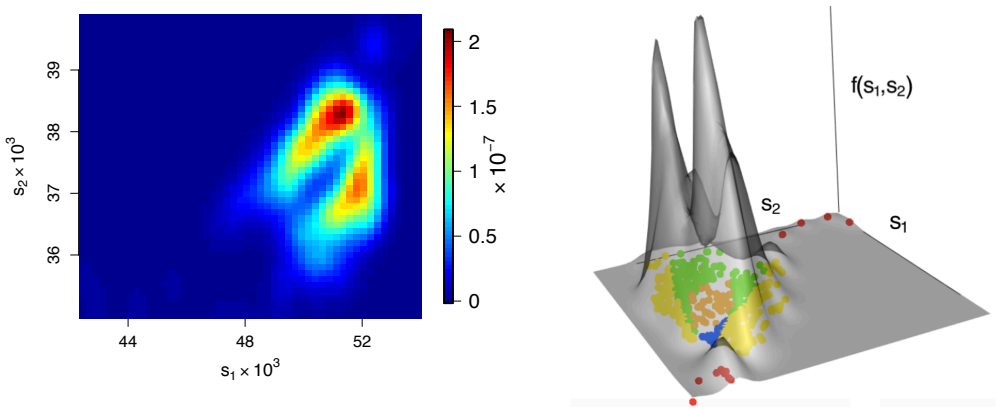

Figure 4: Validity assessment for about 700 samples based on 2 sensor values. Left panel, the plot illustrates the contour map of estimated density function for the 2 sensors. Right panel, the density function of the samples demonstrated in $3D$ with zones identified for each of the samples in the sensor 1 (s_1) versus sensor 2 (s_2) plane. Higher density is assigned to green, blue, and orange zones compared to yellow and red zones.

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 necessary 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 details. 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 the 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, $\mathbf{X}_{n_1\times 11}$.
$\text{ConvexHull}^{(1)} \leftarrow$ 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, $\mathbf{y}_{n_2}$.
$\text{ConvexHull}^{(2)} \leftarrow$ 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.

The above steps are implemented over 8 months of data collected by the e-nose in Section 6. In order to justify our choice of statistical techniques, the proposed methodology is run over a set of simulated data in a following section.

5 Simulation

To emphasize on the importance of the assumptions such as non-eliptical contoured distribution and robust estimation considered in our methodology, we examine the methodology on a set of simulated data. Assume the matrix of data $\mathbf{X}_{N\times 2}$, where $\mathbf{x}_l^\top = (x_{l1}, x_{l2})$; $l = 1, 2, \ldots, N$, are generated according to the mixture of Gaussian and the Student's t-distributions, Fig. 5 (top left panel). Ignoring the distribution of data and seeking for any classical approach toward outlier detection, renders some observations as outliers mistakenly, Fig. 5 (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. 5 (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 which deal with contaminated data appropriately. Adjusted outlyingness (AO) measure identifies the outliers of the data correctly. In the Main Algorithm, suppose we take the Gaussian sub-sample as the Reference 1. Fig. 5 (bottom right panel) shows the result of our algorithm on the simulated data.

6 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 run-time 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 computational complexity of order $O(n)$. In other words, the performance of the sum and mean grow linearly and in 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 4. 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).
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 $O(k_0 p^2)$ (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 sub-linear in n, for fixed p. This is $O(n)$ (Rousseeuw and Driessen, 1999). The calculation of the spectral decomposition of the final covariance matrix has at maximum $O(nk_0)$ time complexity.

Considering the worst case complexity in the above steps, one concludes that the computational complexity of robust PCA is $O(\max\{pn \log n, np^2\})$, or $O(p^2 n \log n)$.

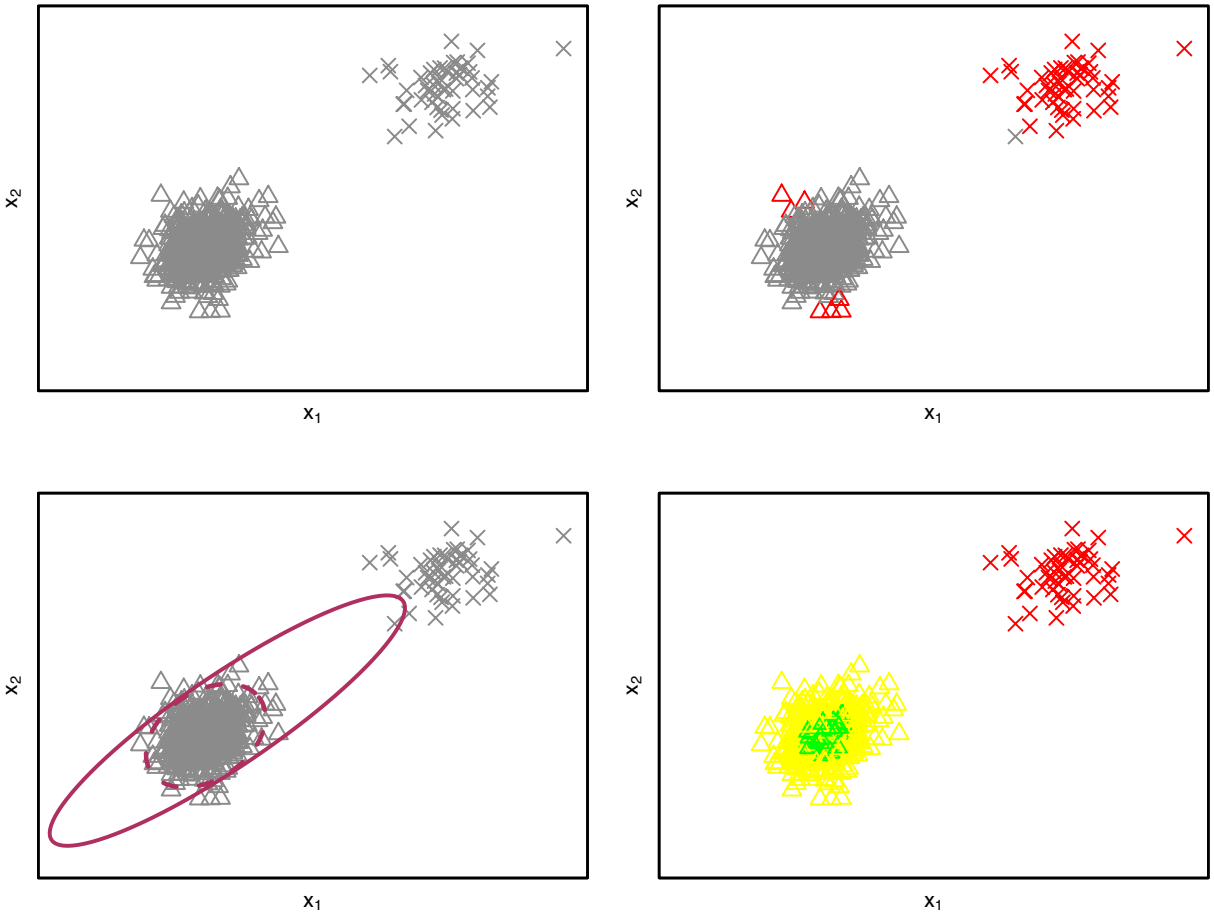

Figure 5: Top left panel, the simulated data from the mixture distribution $f(x) = (1-\varepsilon)f_1(x) + \varepsilon f_2(x)$ with contamination proportion of $\varepsilon = \frac{1}{10}$, and f_1 and f_2 being the Gaussian and Student's t-distribution respectively. The data from f_1, and f_2 are plotted in triangles and crosses correspondingly. Top right panel, the outliers of data are identified and highlighted with red using the classical Mahalonobis distance and 95th percentile of the Chi-square distribution with two degrees of freedom. Bottom left panel, the 95% confidence region for the data is computed using the classical estimates of parameters (solid line) and the robust estimates (dashed line). Bottom right panel, the Main Algorithm is implemented and the zones are graphed by colors described in Table 1.

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 pre-processing 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, $\mathbf{X}_{1:t\times p}^{\top} = [\mathbf{X}_{1:t-1\times p}^{\top} : \mathbf{x}_{t\times p}]$ 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.

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. 6.

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

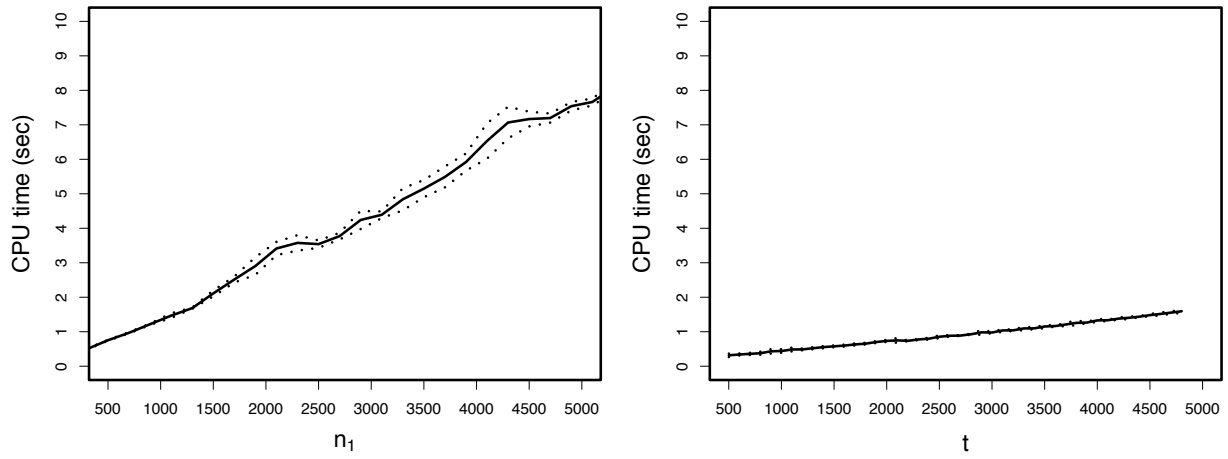

Figure 6: The solid line shows the mean CPU time in seconds as a function of input being run on 1.3 GHz i5 processor. The dashed lines depict the lower and the upper bound of the 95% confidence interval for the mean CPU time. Left panel, the run-time corresponding to Step Require, Step 1, and Step 3 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. 6 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.

7 Application

For the easy visualization, the first 3 robust principle components of the data are used, $PC1$, $PC2$, $PC3$. These components correspond to the 3 largest eigenvalues of the covariance matrix. In case of sensor failures, the data contain missing values that need to be imputed. First, data are imputed to replace all the missing values, and then the validity of the measurements are identified over the 8 months sampling. Only a subset of 500 samples out of 200 thousands of observations are plotted to make the graphs more readable. In Fig. 7, the sample points are drawn in gray and each zone is highlighted using its corresponding color of Table 1. The circles in Fig. 7 are also illustrated on $PC1$ and $PC2$ plane for a better demonstration of the zones. The zones' definition is helpful in interpreting the results. As an example, the green or the blue zone reveals the fact that the sampling points are very close to the samples that have already been observed in either Reference 1 or Reference 2. The observations in reference sets were entirely under control, therefore, the blue and green zones justify the validity of samples. Consequently, the prediction obtained over these samples is expected to be more accurate. On the contrary, the prediction values for the points in the yellow zone are less accurate compared with the green and the blue zones. In other words, the data that are dissimilar to the already observed data deserve further attention. These points are the potential outliers and are reported in the red zone. Additionally, this also reveals that the predictions values associated with such data can be misleading. Producing a noticeable percentage of samples belonging to the yellow and the red zones referring to the possible failure of the e-nose equipment.

8 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 with reference samples in order to predict odour concentration. Consequently, the accuracy of

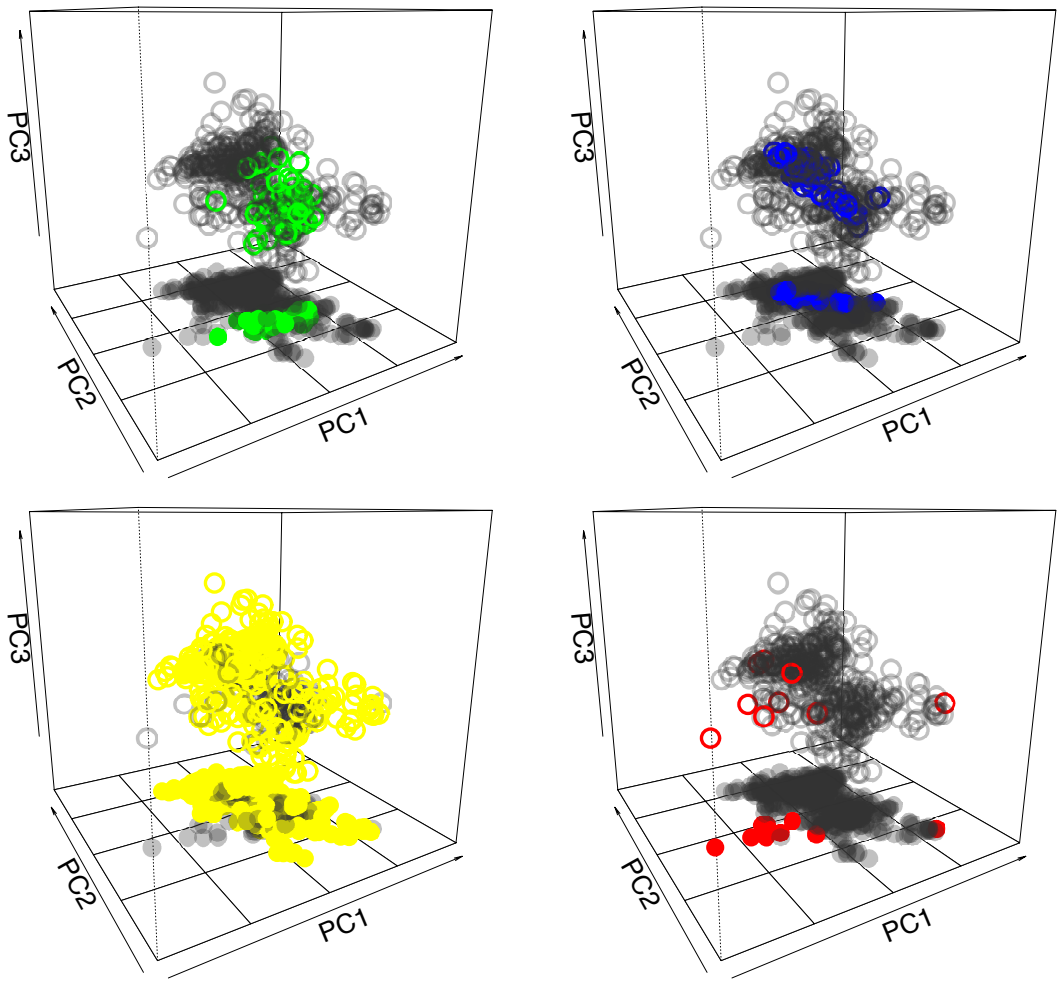

Figure 7: A random sample of size $N = 500$ is plotted over the first three robust principal components coordinates. From top left panel to bottom right panel, the colored blobs represent green, blue, yellow, and red zones respectively.

predicted odour concentration depends heavily on the validity of sensor's output. An automatic procedure that detects the samples' validity in an online manner has been a technical shortage that is addressed in this work. A measurement validation process provides the administrator with the possibility of attaching a margin of error to the predicted odour concentrations. Furthermore, it allows them to take the subsequent actions such as re-sampling to re-calibrate the models or checking the e-nose structure for the possible sensor failures. The proposed measurement validation algorithm in this work hopefully initiate a new era to automatic odour detection by minimizing the manpower involvement.

References

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.

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.

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, 13–15.

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.