Les Cahiers du GERAD ISSN: 0711–2440

Smart olfaction

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

G–2016–86

November 2016

Cette version est mise à votre disposition conformément à la politique de libre accès aux publications des organismes subventionnaires canadiens et québécois.

Avant de citer ce rapport, veuillez visiter notre site Web (https://www.gerad.ca/fr/papers/G-2016-86) afin de mettre à jour vos données de référence, s'il a été publié dans une revue scientifique.

This version is available to you under the open access policy of Canadian and Quebec funding agencies.

Before citing this report, please visit our website (https://www.gerad.ca/en/papers/G-2016-86) to update your reference data, if it has been published in a scientific journal.

Les textes publiés dans la série des rapports de recherche *Les Cahiers du GERAD* n'engagent que la responsabilité de leurs auteurs.

La publication de ces rapports de recherche est rendue possible grâce au soutien de HEC Montréal, Polytechnique Montréal, Université McGill, Université du Québec à Montréal, ainsi que du Fonds de recherche du Québec – Nature et technologies.

Dépôt légal – Bibliothèque et Archives nationales du Québec, 2016
– Bibliothèque et Archives Canada, 2016

The authors are exclusively responsible for the content of their research papers published in the series *Les Cahiers du GERAD*.

The publication of these research reports is made possible thanks to the support of HEC Montréal, Polytechnique Montréal, McGill University, Université du Québec à Montréal, as well as the Fonds de recherche du Québec – Nature et technologies.

Legal deposit – Bibliothèque et Archives nationales du Québec, 2016
– Library and Archives Canada, 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

Smart olfaction

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

[a] *GERAD & Department of Mathematics and Industrial Engineering, Polytechnique Montréal (Québec) Canada, H3C 3A7*

[b] *Department of Mathematics and Industrial Engineering, Polytechnique Montréal (Québec) Canada, H3C 3A7*

mina.mirshahi@gerad.ca
vahid.partovi.nia@gerad.ca
luc.adjengue@polymtl.ca

November 2016

Les Cahiers du GERAD
G–2016–86

Copyright © 2016 GERAD

Abstract: An electronic nose (e-nose), or artificial olfaction, is a device that analyzes the air to identify odours using an array of gas sensors. The e-nose produces multi-dimensional data for each measurement that it takes from the surrounding environment. A small sub-sample of these measurements are sent to the olfactometry, where they are analyzed for odour activities. In olfactometry, for instance, each e-nose measurement is assigned an odour concentration value which describes the odour identifiability by humans. The process of transferring the measurements to the olfactometry and analyzing their odour concentration is time consuming and costly. For this purpose, pattern recognition methods have been applied to e-nose data for automatic prediction of the odour concentration.

It is essential to assess the credibility of the measurements due to the sensitivity of the e-nose to environmental and physical changes. The imprecise measurements render the pattern recognition outcomes unreliable. Therefore, continuous monitoring of e-nose samples and taking necessary actions in case of any anomalies is vital. We devise a smart artificial olfaction that addresses these challenges. Our proposed method uses statistical methodologies for outlier detection and robust PCA. The so called *smart olfaction* is an improved variant of the existing e-noses which is capable of assessing the credibility of samples automatically in an online manner, and employing pattern recognition methods that suits the senor data.

Keywords: Artificial olfaction, electronic nose, gas sensor, graphical lasso, outlier detection, supervised learning

Acknowledgments: The project was funded by the natural sciences and engineering research council of Canada (NSERC) through the industrial partnership Engage program. We thank Masoud Asgharian for reviewing the draft manuscript and providing us with helpful comments.

1 Introduction

The ability to recognize 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: the sense of taste, the sense of smell, and other receptors distributed all over the body. All species employ their chemical senses to approach and being attracted to safe conditions and avoiding harmful ones.

In many ways, olfaction is probably the first sense and it is the core information processing system; critical for survival in wide range of living species. 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 to produce a warning.

The term "odour" specifies the action when one or more chemicals approach the receptors in the olfactory nerve and stimulate them. Odour modulates various aspects of human's life such as sexual attraction, mood, dietary preferences, and detection of danger.

Unfortunately, the human sense of smell does not respond to all harmful air pollutants. Moreover, 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 the air directly influences the environmental and human health. A crucial element in the assessment of indoor and outdoor air quality is auditing the odourants.

There are 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 in 1994 (Gardner and Bartlett, 1994). E-nose is an artificial olfactory system which consists of an array of gas sensors. The e-nose is designed to recognize complex odours of its surrounding environment. The gas sensor array receives chemical information about gaseous mixtures as the input, and converts it to measurable signals.

The inherent features of gas sensors cause unnecessary complications into the process of odour recognition. Some of these features are listed below.

- Gas sensor's performance is affected by different elements, which can make the sensor 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).
- 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.
- The behavior of a sensor is directly influenced by the surrounding chemical and physical conditions. For instance, the response of a sensor may depend on the temperature of the gas under examination. Therefore, thermal conditions around the sensing elements must be under control.

The multivariate response of gas sensor arrays undergoes different pre-processing procedures, prior to the implementation of any pattern recognition methods. Amine et al. (1999); Yan et al. (2015); Shao et al. (2015); Pardo et al. (2000); Wilson et al. (2000) have discussed various systematic feature extraction methods for gas sensor data by minimizing the redundancy in the data. They suggest the use of principal component analysis (PCA) in identifying the outliers for transformed measurements from sensors.

Our two main contributions and their importance in e-nose technology can be described as follows.

1. Sensors of the e-nose may report incorrect values or some of the sensors may 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. There is no specific outlier detection method that can be applied to all type of e-nose data, but rather, it varies depending on the sensors' measurements.
   Our first contribution is to assemble various statistical methods to be used as an algorithm for anomaly detection. This algorithm takes the statistical properties of sensors' measurements into account to assure more reliable results comparing to the existing methods in the literature.
2. Often, the sensor's output is used to quantify odour concentration. Transferring the sensor's output to olfactometry is laborious. Only small portions of data are considered for further analyses of its concentration in olfactometry. The portion of data which is tagged by their corresponding odour concentration is called *calibration set*. The pattern recognition methods employ the calibration set in order to predict the odour concentration for each set of futur sensor values. Numerous methods have been developed for modeling the gas sensor array data, including Gutierrez-Osuna (2002); Hyvarinen (1999); Kermiti and Tomic (2003); Bermak et al. (2006); Qin (1997).
   Our second contribution is employing a more flexible supervised learning model in terms of robustness and sparsity for predicting the odour concentration.

In short, the main focus of this paper is on two subjects. First, the credibility assessment for the sensors' measurements. Second, learning a supervised model on data in order to predict the odour concentration for a batch of measurements.

The paper is structured as follows. The existing structure between the sensor values is explored in Section 2. In Section 3, each sample is allocated to different zones, to quantify the credibility of the e-nose measurements. Afterwards, the odour concentration is predicted for each sample using a supervised learning method in Section 4. The validity of predictions provided by the model are authenticated through the zone's definitions. The applicability of the proposed methods is verified on simulated data in Section 5 and on real data in Section 6.

2 Data description

Depending on the application, an e-nose has varying type and number of gas sensors. The sensors detect the change in electrical resistance when they are in contact with volatile compounds. Sensors react to almost all gases in the air, but each sensor is intended to be more sensitive to a specific type of gas. Better understanding of the e-nose data is necessary for designing an effective data credibility assessment. For this reason, the existence of various common statistical assumptions should be verified.

The data under the study include 11 distinct attributes, each representing one sensor value of the e-nose. As some of the sensors measure nearly the same gases, they happen to be highly positively correlated, see Figure 1, and Figure 3 (left panel).

Suppose that $\mathbf{x}_{p\times 1}^{\top}$ is a random vector of p attributes, in which $\mathbf{a}^{\top}$ denotes the transpose of the vector $\mathbf{a}$, and its n independent realization are stored in the rows of the data matrix $\mathbf{X}_{n\times p}$. As many classical statistical methods rely on Gaussian distribution, one crucial assumption to be verified is the Gaussianity of the data. Validity of this assumption for the sensor values can be 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 Figure 2.

The aim of this research is to develop a methodology for a wide range of e-noses. For this purpose, we also discuss the inherent dependence structure of gas sensors and the sparse estimation of dependence. Sparse methods are specifically for modelling high-dimensional data. They provide better interpretability and lower the cost of modelling by selecting a subset of features. It would be of interest to explore the relationship between the sensors of e-nose for the following reasons.

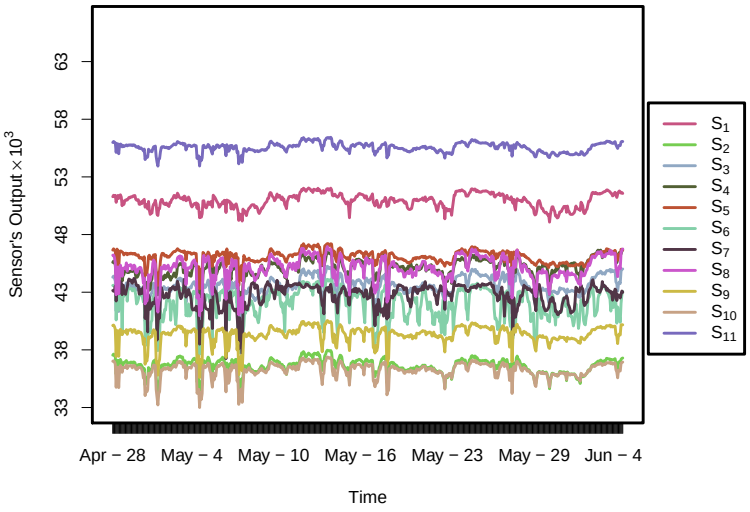

Figure 1: Senor's output during three days of sampling.

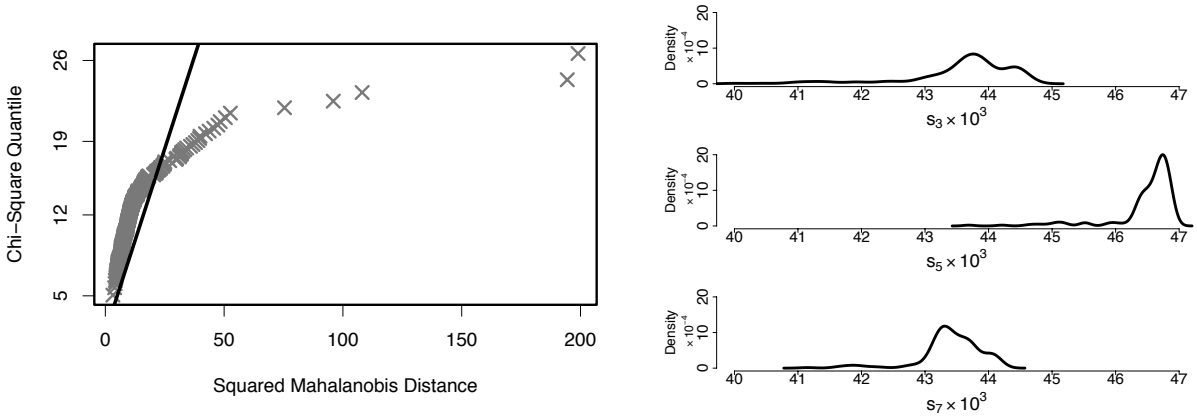

Figure 2: Left panel, the Q-Q plot of squared Mahalanobis distance supposed to follow chi-square distribution for Gaussian data. Right panel, the non-parametric marginal density estimation for some randomly chosen sensor values. Both graphs confirm the non-Gaussianity of the data.

1. To understand the sensitivity of each sensor to different types of gas. Consequently, one would be able to assign the gas sensors to distinct groups in terms of their measurements, i.e. sensors in the same group measure similar gases.
2. To replace a non-active sensor with its active counterpart. During the sampling process, it may happen that one or few sensors stop functioning for an unknown period of time. Having known the existing structure among the sensors, one could swap some of the sensors for the others within the same group with negligible effect on the analysis of the collected data from the sensors. This, in turns, means excluding the redundant sensors from the study and decreasing the dimension of data.

The covariance matrix of a random vector $\mathbf{x}_{p\times 1}$, 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 $\boldsymbol{\mu}_{p\times 1}$ denotes $\mathrm{E}(\mathbf{x}_{p\times 1})$ and E is the mathematical expectation. 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 the precision matrix, coincides with the partial correlation between the attributes. The partial correlation is the correlation between two attributes conditioning on the effect of the other attributes. Opting for a statistical model without considering the dependence between the attributes produces misleading or erroneous results due to multicolinearity.

In order to investigate the inherent dependence between the sensor values, the partial correlation must be estimated. Formally, suppose that the random vector $\mathbf{x}_{p\times 1} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and therefore $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ is the desired parameter to be estimated. To study the relationship between these p attributes, one can use the *Gaussian graphical model* (Murphy, 2012, Chapter 26) which is a graph-based representation of a non-causal structure of attributes. In this graphical model, two attributes that are conditionally dependent given all other existing attributes in graph, are connected by an *edge*. A missing edge between two attributes reveals the conditional independence of attributes. There is a one-to-one correspondence between the elements of the precision matrix, $\boldsymbol{\Theta}$, and the edges in the Gaussian graphical models. Thus non-zero elements of $\boldsymbol{\Theta}$ imply the conditional dependence and the sparse estimation of $\boldsymbol{\Theta}$ reveals the block dependent structure of attributes. The sparse estimation of $\boldsymbol{\Theta}$ set some of the off-diagonal $\boldsymbol{\Theta}$ entries exactly to zero. The *graphical lasso* (Friedman et al., 2008) sparsely estimates graphs using the Gaussian log-likelihood with a *lasso penalty* (Tibshirani, 1996). Various techniques were suggested for estimating $\boldsymbol{\Theta}$ sparsely, such as Meinshausen and Buhlmann (2006); Yuan and Lin (2007); Banerjee et al. (2008). Assuming that the attributes are centred, the log-likelihood for n realization of a random vector $\mathbf{x}_{p\times 1}$, $\mathbf{x}_{p\times 1} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$ is

$$\ell(\boldsymbol{\Theta}) = -\frac{n}{2}\log(2\pi) + \frac{1}{2}\log\det(\boldsymbol{\Theta}) - \frac{1}{2}\mathrm{tr}(\mathbf{X}^\top \mathbf{X}\boldsymbol{\Theta}),$$

where det() and tr() are the determinant and the trace operators, respectively. The graphical lasso estimates the covariance matrix $\boldsymbol{\Sigma}$ under the assumption that its inverse, $\boldsymbol{\Theta}$, is sparse. The graphical lasso minimizes

$$\min_{\boldsymbol{\Theta}\succeq\mathbf{0}} f(\boldsymbol{\Theta}) = -\log\det(\boldsymbol{\Theta}) + \mathrm{tr}(\mathbf{S}\boldsymbol{\Theta}) + \lambda||\boldsymbol{\Theta}||_1, \tag{1}$$

where $\mathbf{S} = \frac{1}{n}\{\mathbf{X}^\top\mathbf{X}\}$ is the sample covariance, $||\boldsymbol{\Theta}||_1$ is the sum of the absolute entries of $\boldsymbol{\Theta}$ and λ is a regularization parameter. The larger the λ is, the more sparse the estimated precision matrix $\boldsymbol{\Theta}$ will be. Minimization problem (1) is a semi-definite programming problem— a convex optimization of a linear objective function over positive semi-definite matrices. Using the sub-gradient method one may solve the optimization problem (1)

$$-\boldsymbol{\Theta}^{-1} + \mathbf{S} + \lambda\boldsymbol{\Gamma} = \mathbf{0}, \tag{2}$$

with $\boldsymbol{\Gamma} = [\gamma_{ij}]_{i,j=1,2,\ldots,p}$ is the sign of each elements of $\boldsymbol{\Theta}$ such that $\gamma_{ij} = \mathrm{sign}(\theta_{ij})$ if $\theta_{ij} \neq 0$ or $\gamma_{ij} \in [-1, 1]$ if $\theta_{ij} = 0$. The graphical lasso employs the block-coordinate technique for solving (2). First, matrices $\boldsymbol{\Theta}$ and $\boldsymbol{\Gamma}$ are partitioned as:

$$\boldsymbol{\Theta} = \begin{bmatrix} \boldsymbol{\Theta}_{11} & \boldsymbol{\theta}_{12} \\ \boldsymbol{\theta}_{21} & \theta_{22} \end{bmatrix} \quad \boldsymbol{\Gamma} = \begin{bmatrix} \boldsymbol{\Gamma}_{11} & \boldsymbol{\gamma}_{12} \\ \boldsymbol{\gamma}_{21} & \gamma_{22} \end{bmatrix} \tag{3}$$

such that $\boldsymbol{\Theta}_{11}$ is a matrix of dimension $(p-1)\times(p-1)$, $\boldsymbol{\theta}_{12} = \boldsymbol{\theta}_{21}^\top$ is a vector of dimension $(p-1)\times 1$ and θ_{22} is a scalar. The matrix $\boldsymbol{\Gamma}$ has the same partitioning structure as $\boldsymbol{\Theta}$. Having assumed $\boldsymbol{\Sigma} = \boldsymbol{\Theta}^{-1}$, the entries of $\boldsymbol{\Sigma}$ can be calculated using the rule of inverse for a partitioned matrix. After some simplifications, the entries of $\boldsymbol{\Sigma}$ are

$$\boldsymbol{\Sigma}_{11} = (\boldsymbol{\Theta}_{11} - \frac{\boldsymbol{\theta}_{12}\boldsymbol{\theta}_{21}}{\theta_{22}})^{-1},$$

$$\boldsymbol{\sigma}_{12} = \boldsymbol{\sigma}_{21}^\top = -\frac{\boldsymbol{\Theta}_{11}^{-1}\boldsymbol{\theta}_{12}}{(\theta_{22} - \boldsymbol{\theta}_{21}\boldsymbol{\Theta}_{11}^{-1}\boldsymbol{\theta}_{12})},$$

$$\sigma_{22} = \frac{1}{(\theta_{22} - \boldsymbol{\theta}_{21}\boldsymbol{\Theta}_{11}^{-1}\boldsymbol{\theta}_{12})},$$

and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\sigma}_{12} \\ \boldsymbol{\sigma}_{21} & \sigma_{22} \end{bmatrix}.$$

Taking the first $p-1$ elements of pth column of Equation (2), we may write

$$-\boldsymbol{\sigma}_{12} + \boldsymbol{s}_{12} + \lambda\boldsymbol{\gamma}_{12} = \mathbf{0}. \tag{4}$$

Substituting $\boldsymbol{\sigma}_{12}$ in Equation (4), we have

$$\boldsymbol{\Sigma}_{11}\frac{\boldsymbol{\theta}_{12}}{\theta_{22}} + \boldsymbol{s}_{12} + \lambda\boldsymbol{\gamma}_{12} = \mathbf{0}. \tag{5}$$

The above equation is equivalent to the following ℓ_1 regularized problem,

$$\min_{\boldsymbol{\beta}}\{\frac{1}{2}\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_{11}\boldsymbol{\beta} + \boldsymbol{\beta}^\top\boldsymbol{s}_{12} + \lambda||\boldsymbol{\beta}||_1\}, \tag{6}$$

where $\boldsymbol{\beta} = \frac{\boldsymbol{\theta}_{12}}{\theta_{22}}$ and $\theta_{22} > 0$. This optimization problem corresponds to a lasso regression (Tibshirani, 1996) of pth attribute on the remaining ones where the matrix $\boldsymbol{S}_{11}$, the sub-matrix of dimensions $(p-1)\times(p-1)$ in the partitioned sample covariance matrix, is replaced by its current estimate $\boldsymbol{\Sigma}_{11}$. The solution to the above problem can be found through the element-wise coordinate descent method. Mazumder and Hastie (2012) suggested a new approach to overcome the occasional convergence issues with the graphical lasso. They proved that the graphical lasso solves the convex dual problem of Equation (1). In Figure 3 (right panel), the undirected graph depicts the estimation of $\boldsymbol{\Theta}$ with $\lambda = 0.75$ by connecting two attributes which are conditionally correlated given all the other attributes. This value of λ is chosen deliberately in order to provide a more clear and meaningful graph.

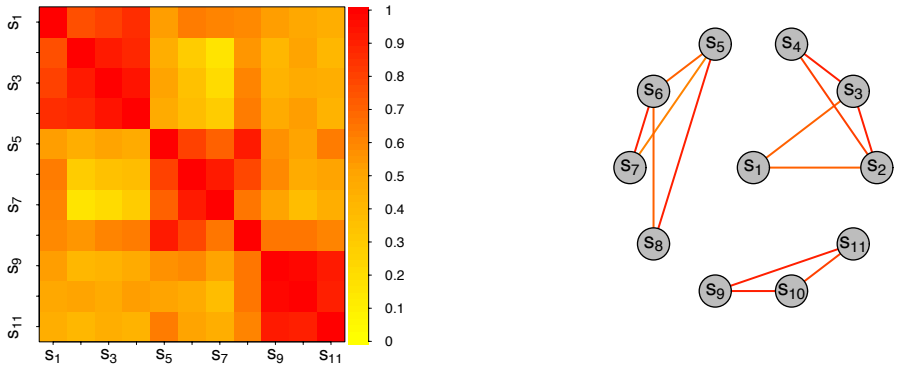

Figure 3: Left panel, heatmap of the correlation matrix of the sensor values (s_1–s_{11}). Right panel, the undirected graph of partial correlation using the graphical lasso. The undirected graph of the right panel approves the block diagonal structure of the heatmap of the left panel.

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 3 (left panel). The conditional correlation among some of the sensors implies that these sensors are measuring similar gases. Thus this dependence must be taken into account while modelling the e-nose data. In other words, we may consider the exclusion of one or few of the senors between the conditionally correlated group of sensors to avoid statistically ill-conditioned models.

3 Credibility assessment

To be able to verify the credibility of the measurements automatically, it is necessary to have some reference samples for the purpose of comparison. Our first task is to allocate each sample to a meaningful measurement zone, say Green, Yellow or Red, etc. The 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 consists of a subset 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 the *proposed set*. The Reference 1 contains the pre-processed data from the 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 set, the *calibration set* to emphasize that it can be used for data modelling using supervised learning.

If a new datum diverges 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 is referred to as "risky" samples. If the new datum is not an outlier and is located within the data convex polytope of the Reference 1, it is assigned to Green zone. This zone represents the "safe" samples. Two more zones, say Blue and Orange, are defined based on the location of convex polytope of Reference 2 which they both belong to the "safe" samples, see Mirshahi et al. (2016) for more details . If the new datum is not an outliers, but outside of the area of "safe" samples, 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. A schematic flowchart for the credibility assessment is provided in Figure 6. The idea behind using the convex polytope as a criterion for outlier detection is based on the assumption of log-concavity of density functions for sensors' measurements (Walther, 2002; Bagnoli and Begstrom, 2005). Log-concave distribution is a wide class that contains many commonly used parametric distributions, like Gaussian, Gamma, Beta, and many others. Suppose f is a density function on $\mathbb{R}^d$, $d \geq 1$, such that

$$f(\mathbf{x}) \propto \exp\{-\varphi(\mathbf{x})\}, \tag{7}$$

where φ is a strictly convex function. The class of all densities f on $\mathbb{R}^d$ of the form (7) is called log-concave densities. Log-concave densities are unimodal. The level sets for a log-concave density function, $\{\mathbf{x}|f(\mathbf{x}) = c\}$ for a constant c, is always a closed convex set (Marshahll and Olkin, 1979). Therefore, the measurements that fall outside the convex polytope of reference sets fail to follow the same log-concave distribution as the reference sets.

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. The zone assignment in smart olfaction, therefore, requires some robust outlier detection algorithms. Figure 4 shows the credibility assessment during the sampling process for 700 sensors' measurements.

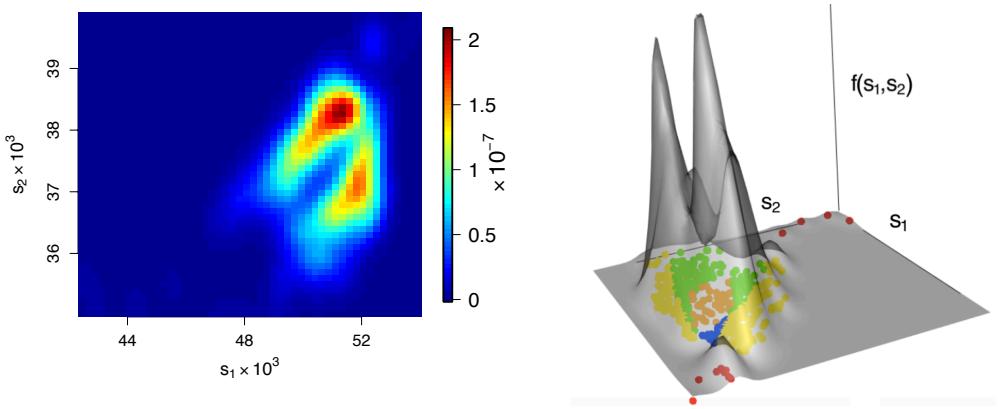

Figure 4: Credibility assessment for about 700 samples based on two sensor measurements. Left panel, the plot illustrates the contour map of estimated density function for two sensors. Right panel, the density function of the samples demonstrated in $3D$ with zones identified for each of the samples in the sensor 1 (s_1) versus sensor 2 (s_2) plane. Higher densities are assigned to "safe" zones compared to "critical" and "risky" zones.

4 Supervised data learning

The ultimate goal of this section is to suggest a suitable model for predicting odour concentration. The credibility assessment serves as a method for analyzing the quality of obtained predictions.

During odour testing, the most common variable of interest is the odour concentration which is evaluated by the olfactometer. The odour concentration of a gaseous sample of odourants is determined by presenting the sample to a panel of selected and screened humans. In order to determine the dilution factor at the 50% detection threshold, the concentration of sample is varied by diluting with neutral gas. At that dilution factor the odour concentration is 1 ou_E/m^3 (European odour unit per cubic meter). The odour concentration of the examined sample is then expressed as a multiple of 1 ou_E/m^3 at standard conditions for olfactometry. Only small proportion of the samples are selected for the examination of their concentrations (calibration set). Consequently, small proportion of data are available for the modelling stage. Here, *sparse partial robust M-regression* (SPRM) (Hoffman et al., 2015) is used for modelling the data. SPRM is a new method of modelling which combines sparseness and robustness with the classical partial least square regression. This regression is claimed to be robust with respect to both response and leverage outliers. Although sparse methods are mostly designed for high-dimensional data, they can be advantageous if applied to low dimensional data as well (Filzmoser et al., 2012).

In a linear regression setting, the relationship between the attributes and the response variable is formulated as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is the vector of measurement errors. The estimate of regression coefficients, $\boldsymbol{\beta}$, is computed through the ordinary least squares $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$. However, there are often situations where the matrix $\mathbf{X}^\top\mathbf{X}$ is not invertible: 1) the attributes are highly correlated and 2) the number of attributes, p, is larger than the number of samples, n. Partial least squares (PLS) regression (Wold, 1996) is used as an alternative to the ordinary least squares regression while $\mathbf{X}^\top\mathbf{X}$ is ill-conditioned. The PLS method projects the data onto a number of latent components and then models the components by one dimensional linear regression, see Manne (1987); Hoskuldsson (2005). Chun and Keles (2010) combined the feature selection with dimension reduction techniques which led to *sparse partial least squares regression*. This sparse PLS regression produces sparse linear combination of original attributes based on the *least angle regression* of Efron et al. (2004).

The classical least squares method suits Gaussian errors. In the case of heavy-tailed errors, Cauchy distribution or ε-contaminated normal distributions, the *M-estimators* tend to provide more promising results (Huber, 1981). Serneels et al. (2005) introduced *partial robust M-regression* by embedding the M-estimators in the PLS.

The sparse partial robust M-regression (SPRM) has the characteristics of both partial robust M-regression and sparse PLS in its inner nature. Here, we briefly explain the SPRM regression procedure. The latent linear components, say $\mathbf{T}$, in PLS are defined as linear combinations of the original attributes, $\mathbf{T} = \mathbf{XA}$. The columns of $\mathbf{A}$, the direction vectors $\mathbf{a}_h$, maximizes

$$\begin{aligned} \mathbf{a}_h &= \underset{\mathbf{a}}{\operatorname{argmax}}\ \operatorname{Cov}^2(\mathbf{Xa}, \mathbf{y}) \text{ for } h = 1, \ldots, h_{max} \\ &s.t.\ ||\mathbf{a}_h||_2 = 1 \text{ and } \mathbf{a}_h^\top\mathbf{X}^\top\mathbf{Xa}_i = 0, \end{aligned} \tag{8}$$

for $1 \leq i < h$ where $||.||_2$ is the ℓ_2-norm ($||\mathbf{x}||_2 = \sqrt{\sum_{i=1}^{p} x_i^2}$). The h_{max} refers to the maximum number of components we prefer to keep in the study. It is assumed that all the attributes and the corresponding response variable $\mathbf{y}$ are centred such that $\hat{\operatorname{Cov}}^2(\mathbf{Xa}, \mathbf{y}) = \frac{1}{(n-1)^2}\mathbf{a}^\top\mathbf{X}^\top\mathbf{y}\mathbf{y}^\top\mathbf{Xa}$. On the other hand, $\mathbf{y}$ can be decomposed as $\mathbf{y} = \mathbf{Tv} + \boldsymbol{\varepsilon}$. This equation can be rewritten as $\mathbf{y} = \mathbf{XAv} + \boldsymbol{\varepsilon}$, where $\mathbf{Av}$ is the vector of coefficients, $\boldsymbol{\beta}$, that relates $\mathbf{y}$ to the original attributes in $\mathbf{X}$. Once the matrix $\mathbf{A}$ is found and $\mathbf{v}$ is estimated through the ordinary least squares method, the estimate of the coefficients are obtained by $\hat{\boldsymbol{\beta}} = \mathbf{A}\hat{\mathbf{v}}$. To make the PLS results robust in the presence of outliers, some weights, $\omega_i \in [0, 1]$; $i = 1, 2, \ldots, n$, are assigned to each row of $\mathbf{X}$ and $\mathbf{y}$. Outliers are given a weight smaller than one. Suppose that $\mathbf{t}_i$ is the ith column of the matrix $\mathbf{T}$ and $r_i = \mathbf{y}_i - \mathbf{t}_i^\top\hat{\mathbf{v}}$ is the residual of the latent variable regression model. The weights, ω_i, is;

$$\omega_i^2 = \omega_R(\frac{r_i}{\hat{\sigma}})\omega_T\Big(\frac{||\mathbf{t}_i - \operatorname{median}_j(\mathbf{t}_j)||_2}{\operatorname{median}_i||\mathbf{t}_i - \operatorname{median}_j(\mathbf{t}_j)||_2}\Big),$$

where $\hat{\sigma}$ is the median absolute deviation of the residuals, ω_R and ω_T are the *Hampel weighting function* with quantiles of standard normal and chi-square distribution (Hampel et al., 1986). In order to obtain a robust PLS, Equation (8) should be rewritten in terms of $\tilde{\mathbf{X}} = \boldsymbol{\Omega}\mathbf{X}$ and $\tilde{\mathbf{y}} = \boldsymbol{\Omega}\mathbf{y}$ where $\boldsymbol{\Omega}$ is a diagonal matrix with

diagonal elements of ω_i, $i = 1, 2, \ldots, n$. A fully robust version of PLS requires estimating $\mathbf{v}$ robustly using *M-estimators*. Moreover, if an ℓ_1 penalty is imposed while computing direction vectors, $\mathbf{a}_h$, the product is a sparse version of the PLS. Zou et al. (2006) suggest penalization on a surrogate direction vector, say $\mathbf{c}$, yields sufficiently sparse estimates. Therefore, using Zou et al. (2006) suggestion, (8) can be transformed to

$$\begin{aligned} \min_{\mathbf{a},\mathbf{c}} & -\kappa \mathbf{a}^\top \tilde{\mathbf{X}}^\top \tilde{\mathbf{y}} \tilde{\mathbf{y}}^\top \tilde{\mathbf{X}} + (1-\kappa)(\mathbf{c}-\mathbf{a})^\top \tilde{\mathbf{X}}^\top \tilde{\mathbf{y}} \tilde{\mathbf{y}}^\top \tilde{\mathbf{X}}(\mathbf{c}-\mathbf{a}) + \lambda_1 ||\mathbf{c}||_1 \\ & s.t.\ ||\mathbf{a}_h||_2 = 1 \text{ and } \mathbf{a}_h^\top \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} \mathbf{a}_i = 0. \end{aligned} \tag{9}$$

The desired direction vector is given by $\mathbf{a}_h = \frac{\hat{\mathbf{c}}}{||\hat{\mathbf{c}}||_2}$, where $\hat{\mathbf{c}}$ is the estimate of the surrogate vector acquired from (9). For more details on SPRM see Chun and Keles (2010) and Hoffman et al. (2015).

Figure 5 presents the flowchart of the algorithm we propose for smart olfaction.

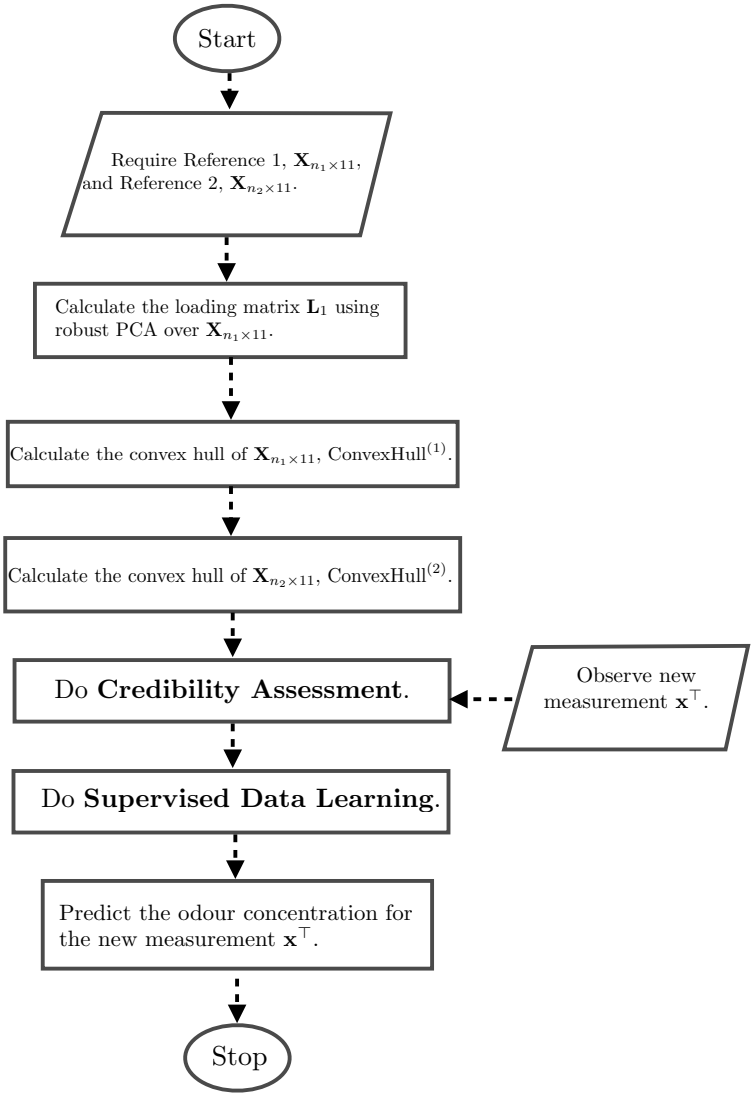

Figure 5: The main algorithm performed by smart olfaction at each sampling iteration.

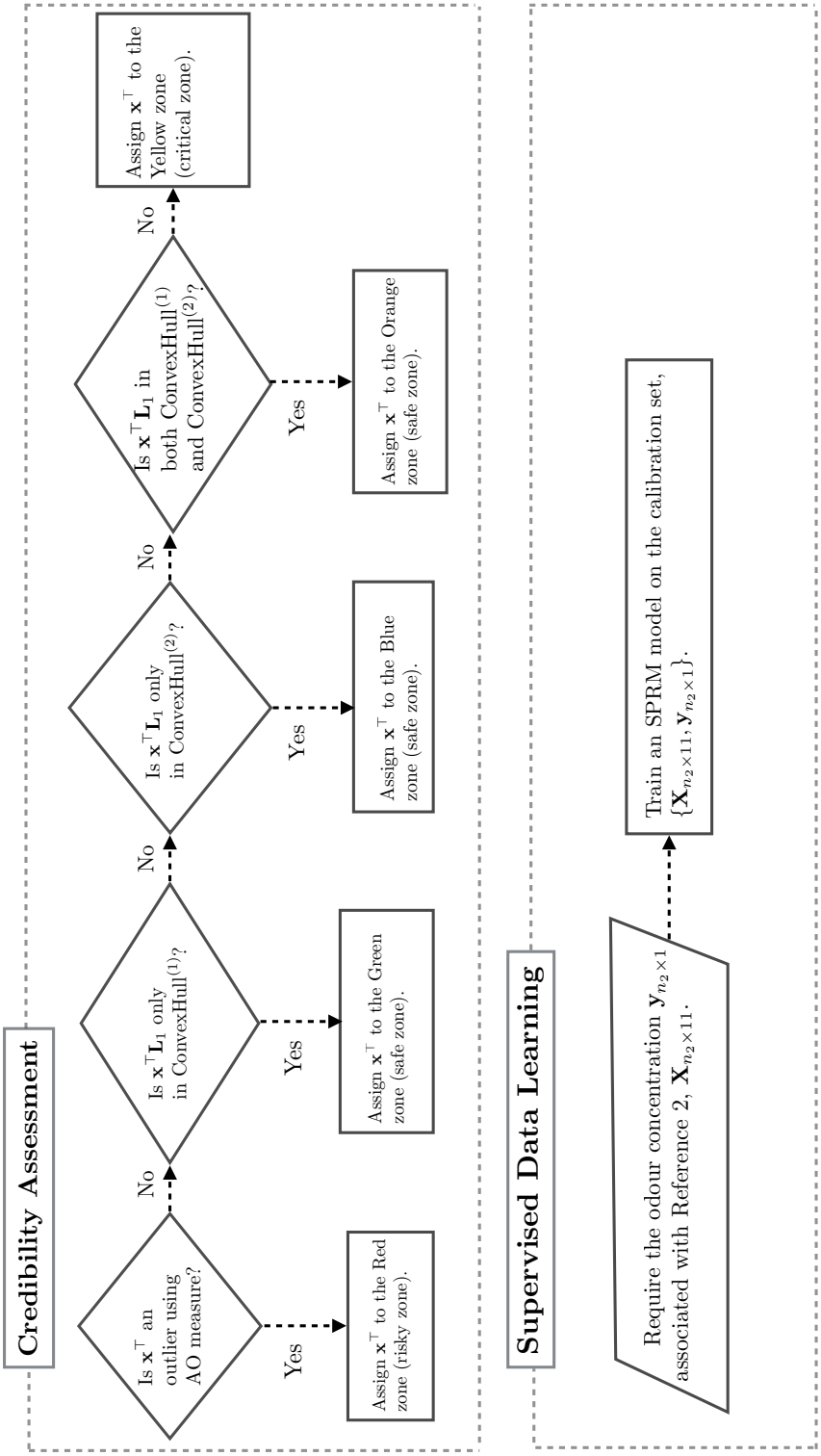

Figure 6: The two sub-algorithms involved in the main algorithm of smart olfaction, Figure 5.

5 Simulation

We examine the applicability of our algorithm on a simulated data set. The data are simulated using the same setup appeared in Hoffman et al. (2015) such that the data resemble e-nose measurements. Consider the linear model $\mathbf{y} = \mathbf{XAv} + \boldsymbol{\varepsilon}$, the details of which are explained in Section 4. Let $\mathbf{X}_{500\times 11}$ be a data matrix generated according to the multivariate Gaussian distribution with 30% contamination and a random covariance matrix, such the final data are highly correlated over some of the attributes. The matrix of direction vectors, $\mathbf{A}$, is generated such that only the first 4 attributes are predictive of the response variable $\mathbf{y}$, that is:

$$\mathbf{A}_{11\times h_{max}} = \begin{bmatrix} \mathbf{A}_{4\times 4} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

The non-zero part of $\mathbf{A}$ is the eigenvectors of $\mathbf{X}_{n\times 4}^{\top}\mathbf{X}_{n\times 4}$ which is the design matrix of the first 4 attributes, each measured over $n = 500$ samples. The components of $\mathbf{v}$ are generated from the uniform distribution on the interval $(0.5, 1.5)$. The error terms, ε_i's $i = 1, 2, \ldots, 500$, are simulated from standard Gaussian distribution. In order to add some additional outlier effects in our study, 10% of the error terms are generated from $\mathrm{N}(3, 0.3)$ instead of $\mathrm{N}(0, 1)$, giving a contaminated Gaussian mixture overall, see Figure 7.

We go through the zone assignment step to describe the procedure more in details. For this purpose, we need to define the reference sets initially. For an easy understanding and a better visualization, the only first two attributes are used for the computation of the zone assignments. The proposed set is a random sub-sample of size $n_1 = 200$ from $\mathbf{X}_{500\times 2}$ data matrix. The calibration set corresponds to another random sub-sample of size $n_2 = 50$ from $\mathbf{X}_{500\times 2}$ data matrix and it contains only $\frac{2}{3}$ of contaminated data. Both sets are pre-processed to form the Reference 1 and the Reference 2 respectively. The zones are defined for the simulated samples imitating the procedure explained in Mirshahi et al. (2016), here visualized in Figure 7.

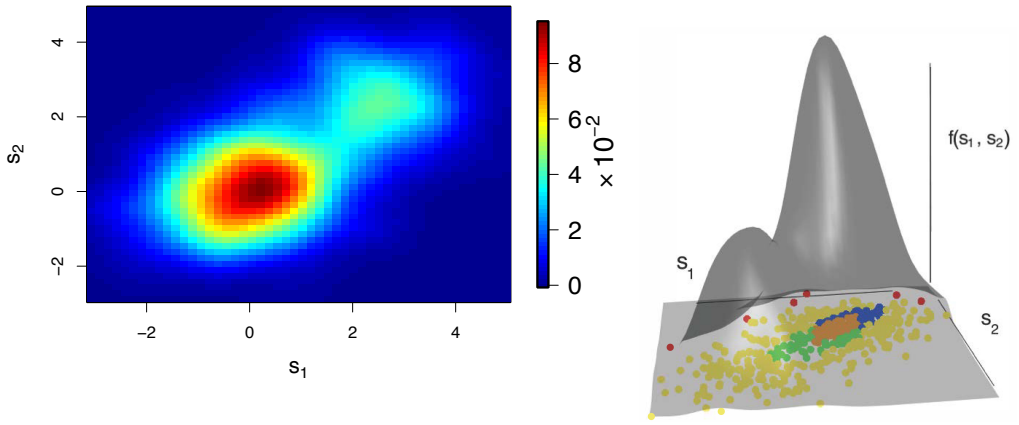

Figure 7: Credibility assessment for about 500 samples based on two attributes generated from bivariate Gaussian distribution with 30% contamination. Left panel, the plot illustrates the contour map of the estimated density function for two attributes. Right panel, the density function of the samples demonstrated in $3D$ with zones identified for each of the samples in the attribute s_1 versus attribute s_2 plane.

For the supervised modelling stage, the two models of PLS and SPRM are tried. Primarily, the optimum values of the parameters for each model is computed using 5-fold cross-validation proposed in the literature (Hastie et al., 2008, Chapter 7). As an example, for SPRM model, computation is run over a grid of different values of components (h_{max}) and the shrinkage parameter (λ_1). The 15% trimmed means squared error of prediction (MSE_{pred}) is used for the final selection of the parameters in the 5-fold cross-validation procedure. Once the parameters are determined, models are compared in terms of their prediction error in 200 rounds of computations. The obtained results are summarized in Table 1. It shows that the optimum number of components are set to 4 for both models, while SPRM suggests a shrinkage parameter $\lambda_1 = 0.71$. In terms of prediction power, two models compete closely with each other.

The main advantage of SPRM is its feature selection ability while modelling, and this counts as a great asset in high-dimensional data— perhaps for e-nose equipment with more gas sensors. The SPRM model produces a more parsimonious and easy to interpret direction vectors compared with the ordinary PLS. In addition, SPRM models estimate the coefficients which are robust with respect to various types of outliers.

Model	h_{max}	λ_1	MSE_{pred} (s.d.)	Average number of zero β's (s.d.)
PLS	4	.	2.0,1 (0.18)	0 (0)
SPRM	4	0.71	2.38 (0.45)	2.6 (1.45)

Table 1: Specification of the parameters for the PLS and SPRM models in 200 **repetitions.**

6 Application

The algorithms of Section 3 and 4 are implemented using over 8 months of data collected by the e-nose equipment. The first 3 robust principle components of the data are used for the zone assignment stage. These components correspond to the 3 largest eigenvalues of the covariance matrix, see Mirshahi et al. (2016). The odour concentration of the sample, $\hat{y}$, evaluated using PLS and SPRM models. The zone color associated with each set of sensors' measurements and their corresponding odour concentration are plotted in Figure 8.

The data contain no measurements on the odour concentration of samples, but rather there is some prior information on its habitual behaviour at the specific field that the e-nose was installed. Given the prior knowledge, it is expected that odour maintains high levels of concentration for the month of April until the end of July. The odour concentration anticipated to drop to small values for the month of August and then increase steadily over the next months. It is also known that odour concentration should not be over 1000 ou_E/m^3 for the industrial site where the data were collected. Using 10-fold cross-validation (another common choice in the literature (Hastie et al., 2008, Chapter 7)), the optimum number of components for two models is $h_{max} = 2$. For the SPRM model, $\lambda_1 = 0.1$ and the first hidden component is only a function of 10 sensor values; the s_2 was eliminated by this choice of λ_1. The predictions based on two SPRM and PLS models closely follow each other, see Figure 8. However, the predictions for the SPRM model do not have as high peaks as the PLS model. The predicted values of both models failed to increase to higher levels for the months of September to November.

The zones' definition is helpful in interpreting the results. As an example, the Green, the Blue and the Orange zones reveal 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 Green, the Blue and the Orange zones justify the credibility 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. The potential outliers and are reported in the Red zone. Our described methodology reveals that the predicted values of such data can be misleading; producing a noticeable percentage of samples belonging to the Yellow and the Red zones. Such findings indicate a possible failure of the e-nose equipment, and hence the need for an on-site visit by a technician.

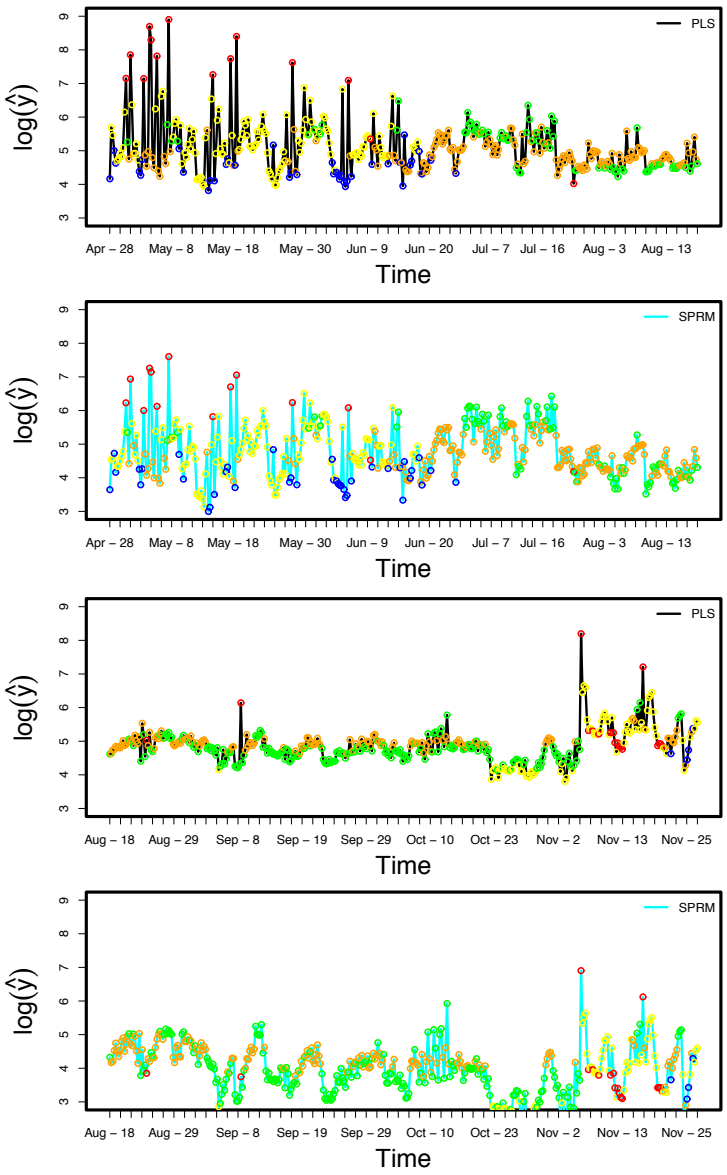

Figure 8: A random sample of size $n = 800$ over time and their predicted odour concentrations according to the SPRM and the PLS models. The coloured circles show the associated zone color to each of the samples. The black lines and the cyan lines illustrate the predictions based on the PLS and SPRM models, respectively. The number of hidden components used in the study is $h_{max} = 2$ and $\lambda_1 = 0.1$.

7 Conclusion

Electronic nose devices have received continuous attention in the field of sensor technology recently. The applications of e-nose are in industrial production, processing, and manufacturing including quality control, grading, processing controls, gas leak detection, and monitoring odours. The measurement quality of the e-nose depends on its sensor's performance. Due to the high variability of the gases in the air and the sensitivity of the sensor values, e-nose measurements can fluctuate very often and fail to maintain a certain level of precision. An automatic procedure that detect the samples credibility in an online fashion has been a technical shortage for a long time and was addressed in this work. The smart olfaction provides an automated process for assessing the credibility of samples and predicting the odour concentration accurately during the sampling procedure and eliminates the need for unnecessary personnel.

References

Amine, H., Bazzo, S., Labreche, S., 1999. Intensity and quality discrimination using the fox4000 gas sensor array system in: Electronic noses and sensor array bases system. In: Design and Applications. W.J. Hurst (Ed.).

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

Bagnoli, M., Begstrom, T., 2005. Log-concave probability and its applications. Economic Theory 26, 445–469.

Banerjee, O., Ghaoui, L. E., d'Aspremont, A., 2008. Model selection through sparse maximum likelihood estimation. Journal of Machine Learning Research 9, 458–516.

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

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

Chun, H., Keles, S., 2010. Sparse partial least squares regression for simulataneous dimension reduction and variable selection. Journal of Royal Statistics Society, Series B 72, 3–25.

Efron, B., Hastie, T. J., Johnstone, I. M., Tibshirani, R. J., 2004. Least angle regression. Annals of Statistics 32, 407–499. http://www.jstor.org/stable/3448465

Filzmoser, P., Gschwandtner, M., Todrov, V., 2012. Review of sparse methods in regression and classification with application in chemometrics. Journal of Chemometrics 26, 42–51.

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

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

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

Hampel, F., Ronchetti, E., Rousseeuw, P., Stahel, W., 1986. Robust Statistics: the approach based on influence functions. Wiley.

Hastie, T., Tibshirani, R., Friedman, J., 2008. The Elements of Stistical Learning. Springer, New York.

Hoffman, I., Serneels, S., Filzmoser, P., Croux, C., 2015. Sparse partial robust m-regression. Chemometrics and Intelligent Laboratory Systems 149 (A), 50–59. Available at SSRN: http://ssrn.com/abstract=2619056

Hoskuldsson, A., 2005. Pls regression methods. Journal of Chemometrics 2 (3), 211–288.

Huber, P., 1981. Robust Statistics. Wiley, New York.

Hyvarinen, A., 1999. Fast and robust fixed-point algorithms for independent component analysis. IEEE Transactions on Neural Networks 10, 626–634.

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

Manne, R., 1987. Analysis of two partial least squares algorithms for multivariate calibration. Chemometrics and Intelligent Laboratory Systems 2, 187–197.

Marshahll, A. W., Olkin, I., 1979. Inequalities: Theory of majorization and its applications. Academic Press.

Mazumder, R., Hastie, T., 2012. The graphical lasso: new insights and alternatives. Electronic Journal of Statistics 6, 2125–2149.

McGinley, P. C., 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.

Meinshausen, N., Buhlmann, P., 2006. High-dimensional graphs and variable selection with the lasso. Annals of Statistics 34, 1436–1462.

Mirshahi, M., Partovi Nia, V., 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. pp. 407–414.

Murphy, K. P., 2012. Machine Learning: A probabilistic perspective. The MIT press.

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

Pardo, M., Niederjaufner, G., Benussi, G., Comini, G., Faglia, E., 2000. Data preprocessing enhances the classification of different brands of Espresso coffee with an electronic nose. Sensors and Actuators B 69, 359–365.

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

Qin, J. S., 1997. Neural networks for intelligent sensors and control - practical issues and some solutions. CiteSeer.

Serneels, S., Croux, C., Filzmoser, P., Espen, P. J. V., 2005. Partial robust m-regression. Chemometrics and Intelligent Laboratory Systems 79, 55–64.

Shao, X., Li, H., Wang, N., Zhang, Q., 2015. Comparison of different classification methods for analyzing electronic nose data to characterize sesame oils and blends. Journal of Sensors 15, 26726–26742.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B 58, 267–288. http://www.jstor.org/stable/2346178

Walther, G., 2002. Detecting the presence of mixing with multiscale maximum likelihood. Journal of American Statistical Association 97, 508–513.

Wilson, D. M., Dunman, K., Roppel, T., Kalim, R., 2000. Rank extraction in tin-oxide sensor arrays. Sensors and Actuators B: Chemical 62 (3), 199–210.

Wold, H., 1996. Estimation of principal components and related models by iterative least sqaure. New York: Academic Press.

Yan, J., Guo, X., Duan, S., Jia, P., Wang, L., Peng, C., Zhang, S., 2015. Electronic nose feature feature extraction methods: A review. Journal of Sensors 15, 27804–27831.

Yuan, M., Lin, Y., 2007. Model selection and estimation in the guassian graphical model. Biometrika 94 (1), 19–35.

Zou, H., Hastie, T., Tibshirani, R., 2006. Sparse principal component analysis. Journal of Computational and Graphical Statistics 15, 265–286.

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