Drift Compensation of Gas Sensor Array Data by Common Principal Component Analysis

A. Ziyatdinov[a,b], S. Marco[e,b,f], A. Chaudry[c], K. Persaud[d], P. Caminal[a,b], A. Perera[a,b]

[a] *Centre de Recerca en Enginyeria Biomèdica. Universitat Politènica de Catalunya, Pau Gargallo 5, 08028 Barcelona, Spain*
[b] *Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN)*
[c] *Protea Ltd, 11 Mallard Court, Mallard Way, Crewe Business Park, Crewe, Cheshire, CWA 6ZQ, UK*
[d] *School of Chemical Engineering and Analytical Science, The University of Manchester, PO Box 88, Sackville St, Manchester, M60 1QD, UK*
[e] *Institute for Bioengineering of Catalonia (IBEC), Baldiri Reixac, 13, 08028 Barcelona, Spain*
[f] *Departament d'Electrònica, Universitat de Barcelona, Martí i Franqués 1, 08028-Barcelona*

Abstract

A new drift compensation method based on Common Principal Component Analysis (CPCA) is proposed. The drift variance in data is found as the principal components computed by CPCA. This method finds components that are common for all gasses in feature space. The method is compared in classification task with respect to the other approaches published where the drift direction is estimated through a Principal Component Analysis (PCA) of a reference gas. The proposed new method – employing no specific reference gas, but information from all gases –has shown the same performance as the traditional approach with the best-fitted reference gas. Results are shown with data lasting 7-months including three gases at different concentrations for an array of 17 polymeric sensors.

Key words: Gas sensor array, Drift, Common Principal Component Analysis, Component Correction, Electronic nose

1. Introduction

Chemical sensor arrays combined with read-out electronics and a properly trained pattern recognition stage are considered to be the candidate instrument to detect and recognize odors as gas mixtures and volatiles [1]. However, a strong limitation in sensor array technology, in addition to selectivity and sensitivity constraints, arise from sensor drift. This phenomenon degrades the stability of the device and makes obsolete the models built in order to recognize and quantify volatiles.

The drift phenomena, in general, are defined as gradual changes in a quantitative characteristic that is assumed to be constant over time. The drift in chemical sensor array devices (also known as e-noses) is a rather complex and inevitable effect, which is generated by different sources. Sensor aging and sensor poisoning influence the device directly through a change in the sensing layer (reorganization of sensor material and contamination). Additionally, the drift of the sensor response is also implied by experimental operation, this includes thermal and memory effects of sensors, changes in environment and odor delivery system noise.

Many efforts have been made in sensor technology and experimental design aiming to improve the stability of sensors with time. Other efforts have been focused on the data processing methods for drift counteraction that can assist these systems to enlarge their calibration lifetime.

An important assumption for drift-compensation methods in chemical

sensor signal processing is that the drift observed int the data is considered to have a preferable direction, rather than a random distribution (according to the definition of drift). This assumption reasonably conforms to the fact that the most disturbances in sensor array data are induced by the sensor side. Other sources of drift also contribute to principal directions of variance as sensors are also assumed to react similarly to the same changes in environment as temperature, humidty variations and others.

The drift evolution can be learned from calibrant samples or the reference process like wash/reference cycle, but that requires a special experimental set up. Univariate methods on baseline manipulation are simple and widely used in the industry [2]. However multivariate methods capture more complex or non-linear drift effects using the information from several sensors in order to model the drift, at the cost of increasing the number of the parameters involved in the correction. Different multivariate methods based on adaptive filters, Component Correction and System Identification theory can be found in the literature [3–5]. Most of the methods are linear, which allows capturing the most drift variance component, but more complex non-linear approaches have also been reported [6, 7].

This paper proposes a new multivariate method based on Common Principal Component Analysis (CPCA) for drift compensation. The method is compared with respect to the well-established approach of Arthursson et al. as both methods compensate the linear component drift in data by employing a Component Correction (CC) routine [5]. The performance of the algorithms is evaluated on a classification task, with a special attention given to the determination of the variance component of drift, which is a critical

point of the methods. The novelty of the CPCA approach consists in computing the drift direction explicitly as a variance common for all the odors classes, so it does not need any specific reference gas. The rest of the paper is organized as follows. The Section 2 describes the dataset and presents the details of the two methods on drift conuteraction. Section 3 reports about results and Section 4 presents the conclusions.

2. Materials and Methods

2.1. Materials

The dataset was obtained the facilities of the University of Manchester. Three gases at different concentration level were measured: ammonia (0.01%, 0.02%, 0.05%), propanoic acid (0.01%, 0.02%, 0.05%), n-buthanol (0.01%, 0.1%). The experiments were repeated on a regular basis during 7 month. The sensor array was composed by 17 polymeric sensors. A total number of 3925 were acquired and labeled to aforementioned gases and concentrations. The response of the sensors has 329 seconds time-length, sampled at 1Hz frequency. The compound is induced to the sensor array at instant $t = 0$s, then the clean air enters the chamber at instant $t = 185s$.

For feature extraction, the data at instant $t = 180s$ is used from the sensor response, thus forming a 17-dimensional feature space from the array of 17 sensors. The option of using complete number of transient points (329) in the signal was discarded, because of the small improvement in class separation and it would help to exposing the method clearly. The operation on removing the outliers was performed by means of the algorithm of Filzmoser et al. with the default parameters [8]. Hence, the number of samples has been reduced

from 3925 to 3484.

2.2. Component Correction Method

The component correction method was first proposed by Arthursson at al. in 2000[5]. The drift component of variance is calculated as the principal component p (or for several components P) of a certain class, namely the reference gas. To remove the drift from the measurement matrix X, the drift direction in data is subtracted by means of a component correction (CC),

$$X_{corrected} = X - (X \cdot p)p^T \tag{1}$$

where the vector p represents the linear approximation of the drift direction. The CC operation is also linear, that in turn preserves the variance in data responsible for class separation and relationship of concentrations.

The approach strictly assumes that the drift component is highly correlated along the gas classes, such that the vector p obtained from the reference gas explains the drift variance for the rest of gases. Since drift often stands for one strong direction in data common for all classes, this strategy seems to be reasonable, but a number of assumptions are considered.

First, there is the assumption that the subspace defined by P (if more than one component is captured) captures only the variance responsible for the drift. However, the data projected onto P may correspond to the variance of the concentration induced by the odor delivery system, for example. Hence, there is the risk of that some information for gas quantification to be subtracted from the data after the CC operation. Another assumption is related with the fact that sensors can respond to the reference and other

gases differently, but this relationship is not managed by the method, as the subspace P contains only information of the reference gas.

2.3. Common Principal component Analysis

This paper proposes a generalization of the CC method through a modification on the method of computing the drift subspace of the data. The main basis is to compute the vector p so that it can express the common covariance for all classes, rather than only the variance shown in the reference gas. The Common Principal Component Analysis (CPCA) can be viewed as a generalization of PCA to k groups of classes. Under the Common Principal Component hypothesis H_C, there exists an orthogonal matrix V such that k covariance matrices Σ_i are diagonal in the data space defined by V.

$$H_c : L_i = V^T \cdot \Sigma_i \cdot V, i = 1, 2, \ldots, k \tag{2}$$

CPCA was proposed by Flury in 1984 who first derived the normal theory maximum likelihood estimates of V and L_i [9]. CPCA is also referred to as Joint Diagonalization (JD), as simultaneous diagonalization of matrices Σ_i is performed.

The exact CPCA solution exists if all the matrices Σ_i commute. When it is not the case, the approximate JD problem is stated and can be solved by optimizing different diagonality criterion [10]. In this work, the algorithm of Cardoso is used which performs the orthogonal diagonalization based on Weighted Least-Squared criteria by means of the Givens rotations[11].

Comparison between PCA and CPCA can provide insight into the application of joint diagonalization for the problem on drift compensation. On

one side, PCA finds the direction of maximum variance blindly to the class-separation information. The principal component can be useful for interpretation only in the trivial case of one class, which is the reference gas. On the other side, CPCA analyzes the between-class relationship and its principal components cover the direction of variance common for all classes.

In other words, the CPCA approach has the confident mathematical base to find the drift variance in accordance with the definition of long-term drift. The PCA approach with the reference class works well only if the reference gas satisfies the requirement to be physically representative. Otherwise, the principal components may capture not only the drift variance, but also the variance in data valuable for data analysis in classification, like concentration oriented components.

An important property of CPCA is that the transformation matrix V is orthogonal (a non-orthogonal formulation of CPCA also exists). Hence, this allows using only several principal components P extracted as columns of the matrix V, in the same manner as for PCA. To estimate the fitness of JD in this work, signal-to-nose ratio (SIR) is used for the components, as in signal processing on speech recognition.

3. Results

3.1. Dataset Description

The sensor data shows complexity for analysis because the influence of a strong drift effect, as shown on Figure 1, where the steady response of sensor 1 is depicted for all classes. The spikes of the signals indicate so-called short-term drift caused by some temporal changes, for example, warm-up of

sensors. The objective of study in this work is related with long-term drift that can be observed in changes in base-line of the signals similar for all the classes.

[Figure 1 about here.]

[Figure 2 about here.]

The PCA scores of the data are showed in Figure 2, where first two principal components capture about 96% of total variance. In order to generate a simple figure, PCA is computed employing a subset of data (1000 samples) rather than the complete dataset of 3484 samples. That can give some reference on how the two examined methods on drift compensation will proceed in the dataset. The depicted confidence ellipses show graphically the different covariance structure for each of the classes. The main direction of the ellipses matches the choice of the direction component for each reference in the Arthurssons approach. The advantage of the new method based on CPCA relies on the mathematical base of computing the common variance along the classes. This common variance is assumed to provide more generalization in the drift compensation process and is the main basis of this method. Two Component Correction (CC) methods on drift compensation are examined in this work. The first one is the classical approach of Arthursson where the drift subspace is computed by PCA of the reference gas. The second one is the new method proposed in this manuscript that employs CPCA to evaluate the drift component in data using the relationship along all classes. The methods are referred to as CC-PCA and CC-CPCA respectively.

To evaluate the algorithms towards drift counteraction, a simple k-NN classifier is used with the parameter of nearest neighbors $k = 3$. The choice of the classifier seems reasonable to test comparatively the performance of both approximations.

Figure 3 shows the validation scheme employed in this paper. The dataset is divided into two Training Sets (T1,T2), representing the first 1000/1200 samples of the data, respectively, and a variety of Validation Sets, which are generated with a moving sliding window as the time between training and validation set increases. The classification model and drift direction in data are computed at the beginning of the time period, in the Training Set. Then the classification ratio is measured on the data of the Validation Set, previously corrected by one of the drift counteraction methods. This follows a validation scheme specially devised for testing drift algorithms, first proposed by Gutierrez-Osuna in 2000 [12].

[Figure 3 about here.]

The number of drift components p in the CC methods can be set to any number. When one component is selected, the drift is assumed linear, whereas several drift components could hint a non-linear or more complex nature of drift. However, there is certain risk that the more components are calculated the more variance is subtracted from the data, and the subtracted variance could capture not only for drift. In this work, only one drift component is used as the final goal is a comparison of the two examined methods.

3.2. Drift Component

The two methods CC-PCA and CC-CPCA differ in the way of calculation the subspace of drift in the Training Set. On the other side, the Component Correction operation is common to both methods in order to subtract the drift (computed in the Training Set) from the data in the Validation Set. Hence, the performance on the two drift-aware methods depends only on the accuracy of calculation of the drift component.

In the CC-PCA method, the first principal component of the reference gas represents the drift direction. The fitness of this direction to the real drift in data is as good as the reference gas is representative along the others. Under CC-CPCA approach, covariance matrices of k gases (in this work $k = 3$) are jointly diagonalized, and the transformation matrix V determines the common subspace in the data, which is interpreted as the drift component. The number of components for both methods is set to one in order to ease the comparison, but this number should be optimized for each specific application. Since the number of drift component is set to one, the basis vector p with the greatest SIR ratio in the matrix V is selected. For both Training Sets T1 and T2, the SIR value is at the acceptable level and equals to 73% and 82% respectively.

[Figure 4 about here.]

The evolution of drift in data can be observed in Figure 4, where the complete dataset is divided into eleven consequent groups of 300 samples, and the drift direction is computed by CPCA for each group. Since the drift is modelled as a vector in the 17-dimencional feature space, the 2-dimencional

projection to the PCA plane of the first group is selected for visualization. The statistical *Wilk-Shapiro* normality test has been performed for the angle of all the eleven vectors, in order to conclude that there is a statistically significant difference along them ($p - value \leq 0.001$). Therefore, the size of the Training Set (1000-1200 samples) seems to be selected correctly to capture the as much drift variance as reasonable for the given size of the dataset (3484 samples). Additionally, one can see that the first $3 - 4$ arrows, which correspond to the region of the Training Set, are visually different.

Drift directions for the Training Sets T1 and T2 are showed in the PCA space in Figure 5. For both Sets T1 and T2, all drift vectors of three reference gases are quite different, that underline the weakness of CC-PCA approach when the choice of the reference in not clear. Comparing plots for T1 and T2, the drift direction obtained from propanoic acid 0.05% (red arrow) reference gas appeared to be sensitive to switch from 1000 to 1200 of sample size of the Training Set, while the drift vectors from n-buthanol 0.1% (grey arrow) and ammonia 0.05% (green arrow) reference mostly preserve the same direction.

The orange arrow on Figure 5 depicts the drift direction obtained via joint diagonalization (JD) of covariance matrices of three gases propanoic acid 0.05%, n-buthanol 0.1% and ammonia 0.05%, as stated in the equation (2). The JD drift vector lies very close to the drift vector obtained from the ammonia 0.05% reference gas. The mathematical explanation is illustrated on the Figure 2 , where the covariance structure (confidence ellipse) of ammonia 0.05% gas is dominant along the other two gases. This could hint a contamination from ammonia in a number of samples for the other gases.

[Figure 5 about here.]

Following the mathematical intuition of CPCA, the drift direction from joint diagonalization is believed to correspond to the actual variation of data caused by drift. Thus, the performance of the drift-aware methods based on the joint diagonalization and the ammonia 0.05% reference gas is expected to be close to each other and superior in comparison with the rest two reference gases. The numerical results will be presented in the section 3.3.

3.3. Evaluation of the Methods

The power of the examined methods on drift counteraction is performed on classification problem on the corrected data and compared in respect to the results obtained for non-corrected data. The validation process has been accomplished with help of the sliding window conforming Validation Sets with the same size as the size of the Training Set.

Figure 6 shows the main results in terms of a comparison between the two methods CC-PCA and CC-CPCA. The curve of classification ratio for non-corrected data is of black colour, the curves of the references ammonia 0.05%, propanoic acid 0.05%, n-buthanol 0.1% and joint diagonalization are marked by grey, green, red and orange respectively. The X axis represents the distance between Training and Validation Sets measured in days as state in figure 3.

[Figure 6 about here.]

For both experiments T1 and T2 the reference curve of uncorrected data (black line) indicates the low classification rate during all period of time. That means the sensor data is strongly affected by drift starting from the beginning of validation phase, 37 and 49 days respectively on the X axis.

Consequently, the drift counteraction for smaller Training Set T1 has improved the classification results significantly at the beginning and fail at the time instant of 63 days. The larger size of the Training Set T2 allows compensating the drift during the complete time period of experiments by the best drift-aware method (orange line of join diagonalization). Please note that there is some recovery on the non-corrected method after the second month for the first Validation Scheme T1/V1. This could be explained by a sudden change after month two that is recovered in month three (e.g. due to a contamination spike that is removed over time), but the true reason is unknown. This is in agreement with figure 4, were the projection of the drift direction is continuously being modified over the experiment. The direction computed at month one is no longer valid as the experiment advances in time and induce a decrease of the performance in the last segment of the experiment. In this case the performance of uncorrected data behaves better in the last part as the Component Correction could be adding more noise than correction. On the other hand, in T2/V2, as the training set includes part of the drift present in month two, the corrected models behave better than the uncorrected, as the drift component found partially includes this information. In this later case, the models show overall better figures with time, which yields to an increase in the calibration specifications over time.

In terms of performance of the methods, the CC-PCA approach with the reference gas ammonia 0.05% performs as well as CC-CPCA method, as expected from their almost coincident drift directions, for the Training Set T1. In the experiments with the Training Set T2 the CC-CPCA method shows superior efficiency along the others, and the time-stability of the classification

is increased being at the level not below than 80%.

4. Conclusion

In this manuscript we have proposed a new method based on common-class variance CC-CPCA that has proven to perform at same level of confidence as CC-PCA approach with the best reference gas (ammonia 0.05%). Moreover, the drift direction captured by Joint Diagonalization is able to show better stability to artifacts in the training set for a particular class, as it employs information from all classes, as seen from Training Set T2 experiment. The direction of future work will be related with application of CPCA to the drift compensation problem optimizing the dimension of the subspace obtained by CPCA, using higher order components rather than the first one.

Acknowledgment

This work was partially funded from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 216916: Biologically inspired computation for chemical sensing (NEUROCHEM) and the Ramon y Cajal program from the Spanish Ministerio de Educación y Ciencia. CIBER-BBN is an initiative of the Spanish ISCIII.

References

[1] K. Persaud, G. Dodd, Analysis of discrimination mechanisms in the mammalian olfactory system using a model nose., Nature 299 (5881) (1982) 352–355.

[2] J. Haugen, O. Tomic, K. Kvaal, A calibration method for handling the temporal drift of solid state gas-sensors, Analytica Chimica Acta 407 (1-2) (2000) 23–39.

[3] M. Holmberg, F. Winquist, I. Lundström, F. Davide, C. DiNatale, A. D'Amico, Drift counteraction for an electronic nose, Sensors & Actuators: B. Chemical 36 (1-3) (1996) 528–535.

[4] M. Holmberg, F. Davide, C. Di Natale, A. D'Amico, F. Winquist, I. Lundström, Drift counteraction in odour recognition applications: lifelong calibration method, Sensors & Actuators: B. Chemical 42 (3) (1997) 185–194.

[5] T. Artursson, T. Eklov, I. Lundstrom, P. Martensson, M. Sjostrom, M. Holmberg, Drift correction for gas sensors using multivariate methods, Journal of chemometrics 14 (5-6) (2000) 711–723.

[6] C. Natale, F. Davide, A. D'Amico, A self-organizing system for pattern classification: time varying statistics and sensor drift effects, Sensors & Actuators: B. Chemical 27 (1-3) (1995) 237–241.

[7] S. Marco, A. Ortega, A. Pardo, J. Samitier, Gas identification with tin oxide sensor array and self-organizing maps: adaptive correction of sensor drifts, IEEE Transactions on Instrumentation and Measurement 47 (1) (1998) 316–321. doi:10.1109/19.728841.

[8] P. Filzmoser, R. Maronna, M. Werner, Outlier identification in high dimensions, Computational Statistics and Data Analysis 52 (3) (2008) 1694–1711.

[9] B. Flury, Common principal components in k groups, Journal of the American Statistical Association 79 (388) (1984) 892–898.

[10] X.-L. Li, X.-D. Zhang, Nonorthogonal joint diagonalization free of degenerate solution, IEEE Transactions on Signal Processing 55 (5) (2007) 1803–1814. doi:10.1109/TSP.2006.889983.

[11] J. Cardoso, A. Souloumiac, Jacobi angles for simultaneous diagonalization, SIAM Journal on Matrix Analysis and Applications 17 (1) (1996) 161–164.

[12] R. Gutierrez-Osuna, Drift reduction for metal-oxide sensor arrays using canonical correlation regression and partial least squares, in: Electronic Noses and Olfaction 2000: Proceedings of the Seventh International Symposium on Olfaction and Electronic Noses, Held in Brighton, UK, July 2000, Institute of Physics Publishing, 2000, p. 147.

List of Figures

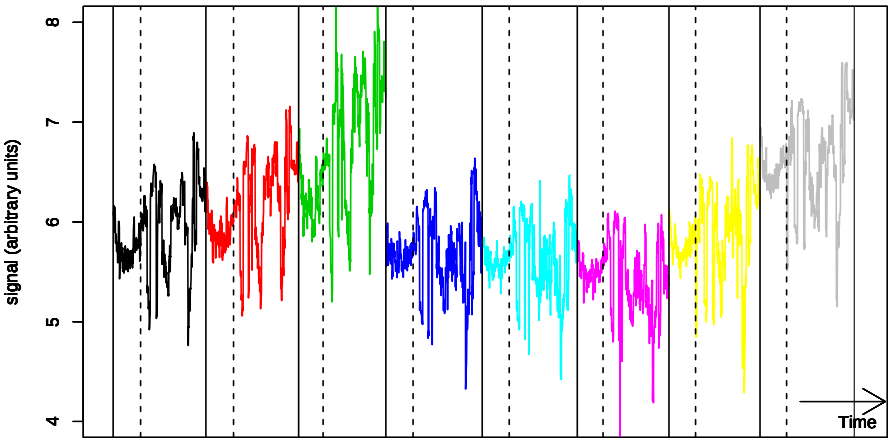

Figure 1: Trajectories of steady-state point recorded from the sensor 1 during 7 months of experiments for all three gasses at their concentration levels. The plots of trajectories are listed from left to right: ammonia (0.01% in black, 0.02% in red, 0.05% in green), propanoic acid (0.01% in blue, 0.02% in cyan, 0.05% in magenta), n-buthanol (0.01% in yellow, 0.1% in grey). For each of eight plots the dashed lines separates the samples to the training set on the left and the validation set on the right.

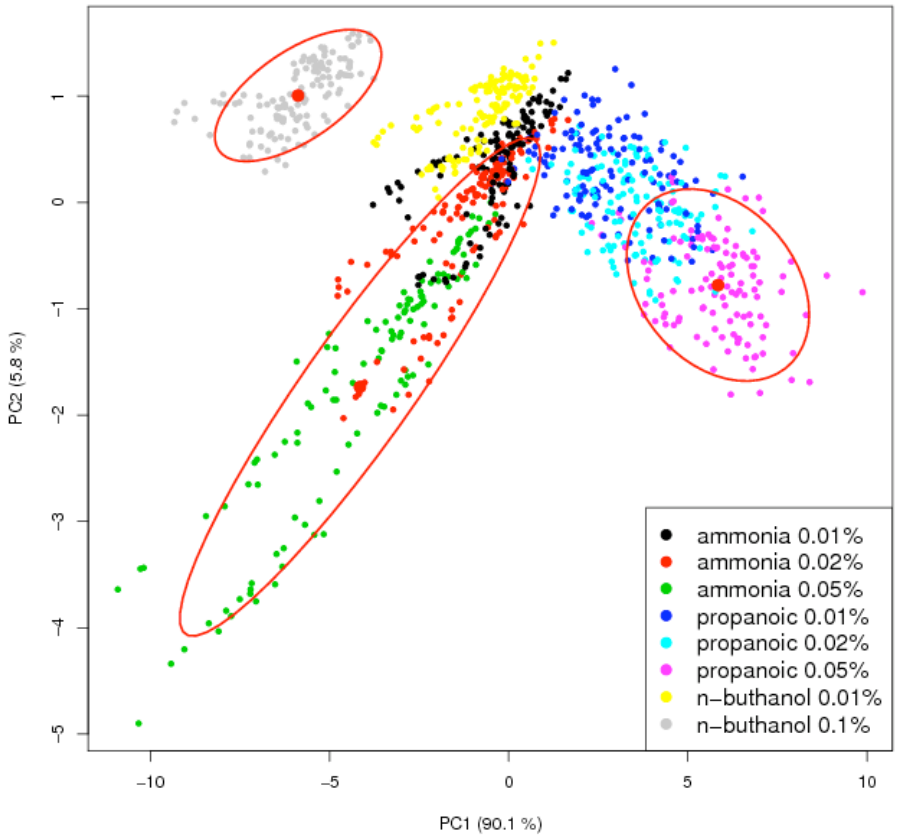

Figure 2: PCA scores of the first 1000 samples with the depicted confidence regions depicted for three classes: ammonia 0.05%, propanoic acid 0.05%, n-buthanol 0.1%.

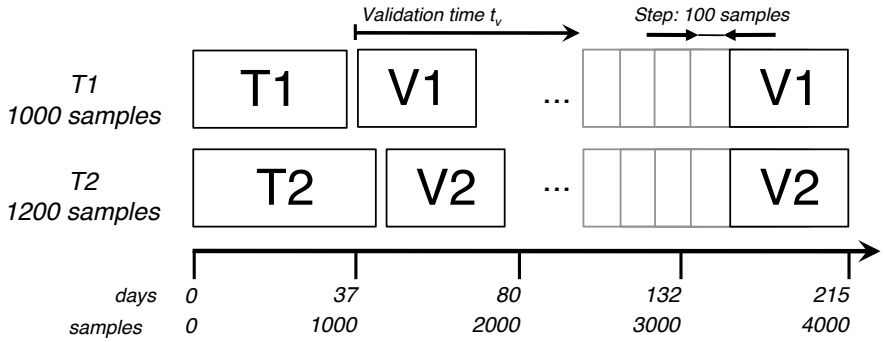

Figure 3: The data is split into Training and Validation Sets, where the last is moving as a sliding window with step of 100 samples. The performance of each algorithm is tested as the distance to the training set increases, following a validation scheme proposed by Gutierrez-Osuna in 2000 [12]

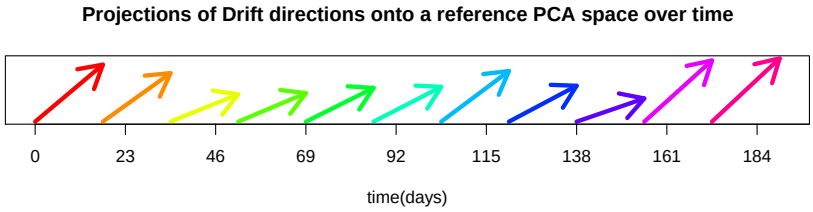

Figure 4: Given the complete dataset divided into eleven consequent groups, the drift direction computed by CPCA is plotted as a projection onto the first two principal components (PCA) of the first group.

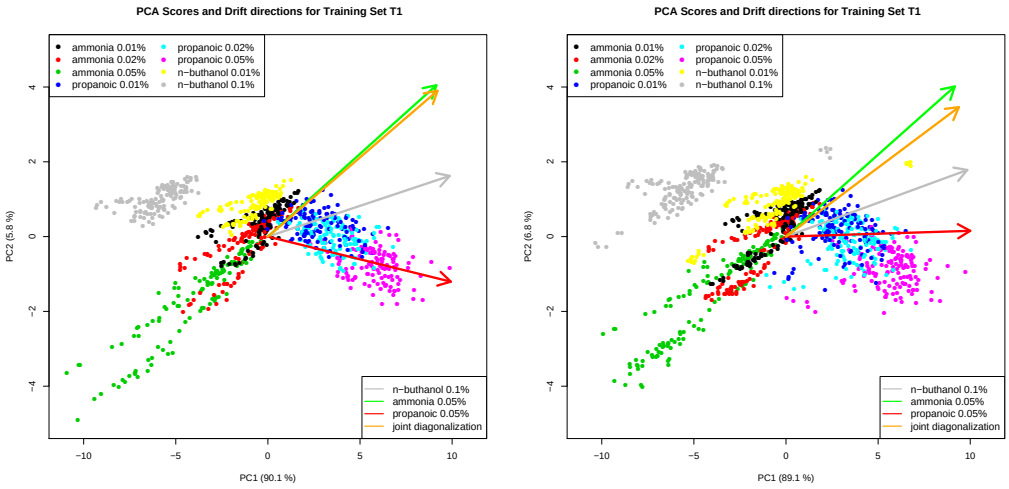

Figure 5: PCA scores and principal drift directions for the Training Set T1/T2 (left/right).

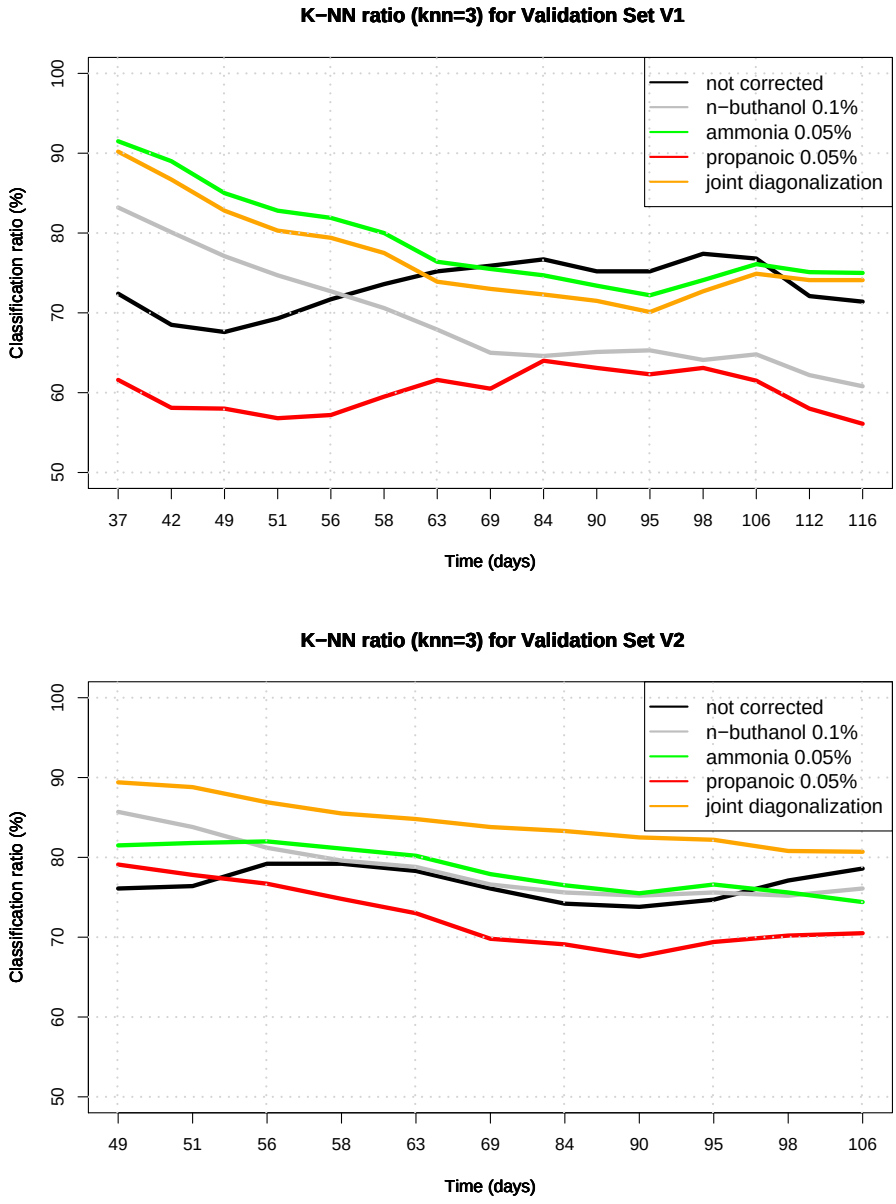

Figure 6: k-NN performance as a function of the distance of the Validation Set from the Training Set T1/T2 (top/bottom).