A software tool for large-scale synthetic experiments based on polymeric sensor arrays.

A. Ziyatdinov[a,b], E. Fernández Diaz[e], A. Chaudry[c], S. Marco[e,b,f], K. Persaud[d], A. Perera[a,b]

[a]*Department of ESAII. Universitat Politènica de Catalunya, Pau Gargallo 5, 08028 Barcelona, Spain*
[b]*Centro de Investigación Biomèdica en Red en Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN)*
[c]*Protea Ltd, 11 Mallard Court, Mallard Way, Crewe Business Park, Crewe, Cheshire, CW1 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 dElectrònica, Universitat de Barcelona, Martí i Franqués 1, 08028 Barcelona, Spain*

Abstract

This manuscript introduces a software tool that allows for the design of synthetic experiments in machine olfaction. The proposed software package includes both, a virtual sensor array that reproduces the diversity and response of a polymer array and tools for data generation. The synthetic array of sensors allows for the generation of chemosensor data with a variety of characteristics: unlimited number of sensors, support of multicomponent gas mixtures and full parametric control of the noise in the system. The artificial sensor array is inspired from a reference database of seventeen polymeric sensors with concentration profiles for three analytes. The main features in the sensor data, like sensitivity, diversity, drift and sensor noise, are captured by a set of models under simplified assumptions. The generator of sensor signals can be used in applications related to educational tools, neuromorphic simulations in machine olfaction, and test and benchmarking of signal processing methods. The tool is implemented in R language and can be freely accessed.

Key words: Gas Sensor Array, Conducting Polymers, Electronic Nose, Sensor Simulation, Synthetic Dataset, Benchmark, Educational Tool

1. Introduction

Instrumentation for machine olfaction traditionally comprises an array of broadly-tuned chemical sensors targeted to quantify or recognize complex odour gas mixtures [24, 26, 30]. Measurements of sensor data require an elaborate design set up, development of specific hardware and software, and weeks or months for data collection. A virtual sensor array, as a data generation tool, might help on several aspects on machine olfaction research: experimental design, prototyping of signal processing prototyping, prototyping and research of neuromorphic and neuroinspired processing, algorithm benchmarking or simply serve as an educational tool.

Unfortunately, simulations are not widely used in gas sensor data processing, although few examples can be found in the literature. Time response of virtual semiconductor gas sensors were modeled with a second-order system by Ishida et al., and simulation results suggested an optimized design of sensor array for odour source localization in a wind tunnel environment [16]. To validate methods on dynamic feature extraction, artificial signals containing exponential decays were used by Gutierrez-Osuna et al. in simulation of sensor transient responses [15]. Short-time measurements were completed by linearly simulated long-term drift data, in order to test the robustness of self-organizing maps for gas identification proposed by Marco et al. [19]. Most recent works produced more sophisticated simulated data for validation of signal processing methods. Montoliu et al. simulated eleven thermally modulated metal-oxide sensors following a dynamic Clifford-Tuma model, which allowed preliminary tests on quantification of a simulated ternary mixture, prior to the experiment with real gas mixtures [7, 8, 21]. Geng et al. validated multivariate calibration methods entirely on simulated data obtained from the model of Lei et al. for conducting polymer sensors, which permitted to explore the developed methods under linear and non-linear sensor arrays [12, 17].

Contrary to application-specific simulations with particular datasets, some machine olfaction problems such as drift compensation require standard datasets or benchmarks to evaluate and score different algorithms. This demand for benchmarking has been already mentioned by Gutierrez-Osuna [14], but a public machine olfaction dataset repository, similar to UCI Machine Learning Repository [2], is still missing to the best knowledge of the authors.

Modeling of a virtual sensor array for synthetic data generation implies a simulation of an individual sensor device and its response characteristics to some extent of detail. The transduction phenomena on polymer sensors is based on several conductive and thermodynamic properties of the polymers [6, 32, 33], and the underlying process unlikely can be explained by only one major mechanism. Hence, the sensor models published in the literature can be divided into groups according to the transduction mechanism involved. The governing phenomenon in signal transduction for the carbon-black polymer composite sensors is assumed to be polymer swelling in the presence of the analyte [9, 12, 17]. The polymer sensors targeted to detect volatile organic chemicals have been modeled based on

volatility driven partition of the analyte molecules into the polymer composite [27]. For polymeric sensors with the adsorption driven comportment, the analytes are not reactive with conducting polymers under normal conditions, instead, the analytes molecules are adsorbed on the polymer surface. Topart and Josowicz hinted analyte sorption as the driving force of analyte-polymer interaction for the polypyrrole sensors in exposure to methanol vapor [33]. The adsorption-based models supplement the Langmuir equation of sorption kinetics with the diffusion equation of mass-charge transport in the polymer [10, 18, 31].

The models for conducting polymer sensors mentioned above are not truly targeted towards simulation experiments with a chemical sensor array. Instead, the main use of these models is the characterization of the polymer material and sensor geometry in terms of their sensitivity to target analytes. Most of the models are limited to single analyte experiments without support of mixtures, which in turn, essentially restricts their use on synthetic experiments. Few works have proposed the modeling of a sensor array for simulations. Gardner et al. introduced a PSPICE model that emulates resistive gas sensors, where the strategy of modeling is based on behavioral description, rather than analytical solutions [11]. The effects of multicomponent gas mixtures and operating temperature can be emulated in these synthetic simulations. Recently, the model by Lei et al. was used to produce simulated data for validation of multivariate calibration of sensor arrays under noise-free conditions [12, 17].

To approach the simulation of a virtual array of sensors, sensor calibration is not a sufficient procedure to gain realistic sensor signals. Sensor measurements inevitably contain noise artifacts originated from instability in particular sensor devices, variation of the ambient conditions and the sampling system and other physical and chemical processes involved. The long-term part of noise variation common to all the sensors is regarded as the drift phenomena, and represents one of the most strong noise sources in the sensor array data. A multivariate approach to estimate sensor array drift was first introduced by Artursson. Based on the observation that sensors tend to drift in similar manner Artursson and colleagues hypothesized that drift has one preferred direction in sensor space, as clearly shown in the score plot of a principal component analysis [1]. Existing multivariate methods on modeling of the drift subspace include principal component analysis of the reference gas [1], component deflation [13], component orthogonalization [22] and common principal component analysis [36]. These methods share the common technique of component correction for drift compensation introduced by Artursson, which in turn will be used inversely in this paper for drift injection in the sensor array data. Such multivariate approach differs from the published drift models, where the drift was induced individually for each sensor [19].

In this paper, we present a software tool for emulation of a virtual sensor array that features an arbitrary number of sensors, supports an arbitrary gas mixture of up to three analytes, and grants full parametric control on the noise in the sensors, including modeling of a non-trivial drift behavior. The virtual array is inspired from a real array of seventeen polymeric sensors, and substantially captures its main characteristics by a group of heuristic models developed within this paper.

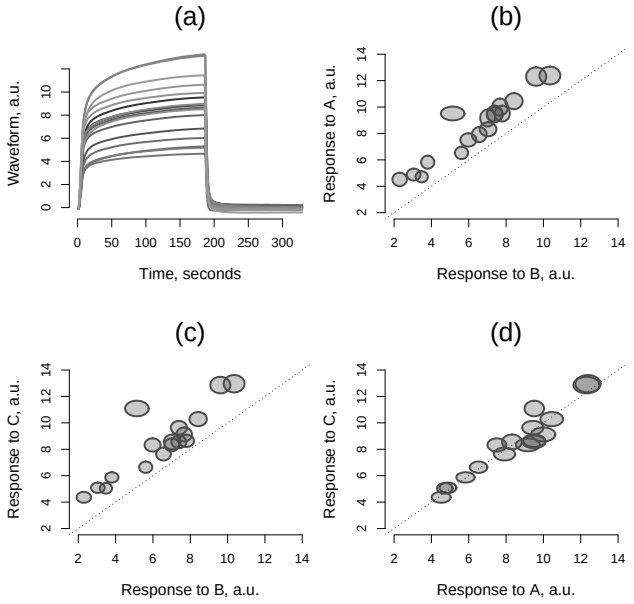

Figure 1: Example of sensor signals from UNIMAN database. Plot (a) shows waveforms of seventeen sensors from a single measurement of analyte C (n-buthanol 1 vol.%). Plots (b-d) represent seventeen sensors in the affinity space of three analytes A, B and C. The steady-state response to one analyte is plotted versus the response to another analyte. Ellipses depict sensor signals averaged over short-term UNIMAN dataset of two hundred samples in response to the three analytes at maximum concentrations.

The rest of the manuscript is organized as follows. Section 2 describes the reference sensor array database and the developed methods to simulate the virtual sensor array. Section 3 presents some examples of synthetic experiments, and Section 4 reports with concluding remarks.

2. Models and Methods

2.1. Reference Dataset

	Label	Gas	Concentration	Samples
1	A0.01	ammonia	0.01 vol.%	489
2	A0.02	ammonia	0.02 vol.%	487
3	A0.05	ammonia	0.05 vol.%	476
4	B0.01	propanoic acid	0.01 vol.%	488
5	B0.02	propanoic acid	0.02 vol.%	490
6	B0.05	propanoic acid	0.05 vol.%	487
7	C0.1	n-buthanol	0.1 vol.%	505
8	C1	n-buthanol	1 vol.%	503

Table 1: Gas classes in UNIMAN database.

The reference dataset has been measured at The University of Manchester (UNIMAN, UK). Three analytes ammonia, propanoic acid and n-buthanol, at different concentration

levels, were measured for 10 months with an array of seventeen conducting polymer sensors. The sensors were made of polypyrrole-based films, and each element in the array had different chemical selectivity characteristics. The collected records counted for 3925 samples evenly distributed over 8 analyte classes, as summarized in Table 1. The single waveform from an individual sensor has a length of 329s, acquired at 1Hz, and represents the response to a rectangular gas pulse where an analyte is introduced in the sensor chamber from 0s to 180s. Air is introduced next in a cleaning phase following 180 seconds of quasi-stabilization time. Figure 1 (XXX) (a) shows an example of the transient waveforms of seventeen sensors from a single measurement of n-buthanol at 1 vol. %. The complete data matrix from the UNIMAN sensor array has dimensionality $3925 \times 329 \times 17$.

In application to sensor array data modeling, the UNIMAN database experimental design included long-term concentration profiles of three pure analytes, which allowed to track changes of the sensitivity and selectivity performance in the presence of noise at different time-scale and analyte-dependent levels. The data had co-linearity typical for gas sensors, revealed multimodality underlying the class information, and contained a large amount of noise-related spread even for day-to-day measurements. Figure 1 (b-d)XXXX shows the performance of UNIMAN sensor array in terms of selectivity by representing the sensors in the affinity space of the three analytes [25]. The plots underline the co-linearity nature of the sensor array and indicate some difficulties in separability of analytes A and C, in comparison with the other analyte pairs. For further details, the reader can address to a number of works applied to the dataset for different machine olfaction problems, for example, fault diagnosis [23] and drift compensation [22, 36]. In this work, the UNIMAN database is used to estimate parameters of the models designed to simulate a virtual sensor array.

In process of modeling of the array we make the distinction between short-term (STD) and long-term reference data (LTD). Two hundred samples from the first 6 days are used to characterize the array assuming the absence of drift. The long-term reference data counts for the complete number of samples covering a 10 month time-span. A pre-processing procedure on outliers removal was applied to the long-term reference data. The standard method based on the squared Mahalanobis distance was used with quantile equal to 0.975%.

2.2. Simulation models

Figure 1 XXXX presents a block scheme of the virtual sensor array. The input concentration matrix C_0 of three columns (three analytes, corresponding to analyte A, B and C) encodes the profile of gas mixture exposed to the array over time. Two basic blocks, Sorption Model and Calibration Model, emulate the response of sensors in the array over time. The output raw data matrix X has a number of columns equal to the number of sensors and a number of rows equal to the length of the simulation. Three blocks at the bottom line of the scheme introduce the noise affecting the array at different levels of its workflow.

The Sorption and Calibration Models have different roles in the simulation flow. The Sorption Model controls the amount

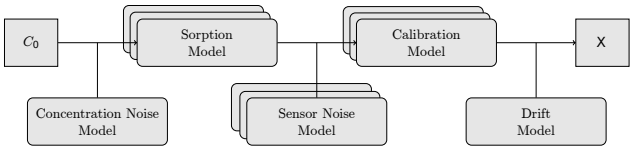

Figure 2: The block scheme shows the basic simulation models and underlines the workflow on data synthesis, from the concentration matrix C_0 to the sensor array data X.

of gas absorbed by the sensor following the Langmuir relation, which also emulates the intrinsic non-linearity nature of the polymeric sensors. The Calibration Model regulates the relationship between this amount of the absorbed gas and the output sensor signal. Both Sorption and Calibration Models utilize STD reference data.

2.2.1. Sorption Model

The Sorption Model is based on the extended Langmuir isotherm for a multi-component gas mixture [20] given in Equation 1.

$$c_i = \frac{q_i \, k_i \, c_{0i}}{1 + \sum_{j=1}^{N} k_j \, c_{0j}}, \; i, j = 1, 2, ..., N \tag{1}$$

where c_{0i} is the concentration in air of analyte i in the mixture, and c_i is its concentration in the adsorbed form. The isotherm has two parameters per analyte, q_i denotes the sorption capacity and k_i stands for the sorption affinity of the analyte i. Indeed, the isotherm extends the Langmuir isotherm for a single gas under a simplified assumption that molecules of the analytes in mixture do not interact with each other. Such property allows us to describe the adsorption process in the gas mixture explicitly by computing a single-adsorption Langmuir isotherm per analyte.

Hence, three pairs of parameters q_i and k_i for the seventeen UNIMAN sensors are estimated independently for three analytes A, B and C by rewriting the equation 1 for one component ($N = 1$). Given that x is the sensor signal and c_0 is the analyte concentration, we estimate the parameters of the Langmuir isotherm by fitting a linear model $1/x \sim 1/c_0$. The resulting coefficients of determination R^2 of the models are not below than 0.973 for analyte C, and slightly worse for analytes A and B giving a minimum value 0.779. Comparison of the obtained affinity terms for the seventeen UNIMAN sensors gives an insight into the sensor responses to a mixture of analytes, in terms of the Langmuir isotherm abstraction. The affinity numbers indicate that both analytes A and C dominates over the analyte B in mixture for all the sensors, and most of the sensors show substantially greater affinity to the analyte C than A. This observation along the UNIMAN sensors will be also observed for simulated arrays described in the next Section 3.

In practice, the assumption posed by the extended Langmuir isotherm about the component independence in mixture is valid at low concentrations. However, the deviation from the Langmuir proportion is critical for a detailed physical sensor model, which can be accepted in our modeling aimed to emulate the

complexity and non-linearity in the sensor behavior. One of the benefits of using the Langmuir relation is its flexibility to extension, widely studied in the gas separation discipline [3–5, 35].

2.2.2. Calibration Model

The Calibration Model is composed of two consecutive parts. First, a static model derives the steady-state signal from the concentration of the adsorbed analyte. Second, a dynamic model proceeds to the transient signal from the value of the steady-state. Equations 2 and 3 define the static model.

$$x_{ss} = \sum_{i=1}^{N} x_{ss,i} = \sum_{i=1}^{N} f_i(c_i) \tag{2}$$

where x_{ss} is the steady-state sensor signal in response to a N-component mixture with analyte concentrations c_i. The model explicitly assumes that the response x_{ss} to a mixture of analytes is a sum of responses to the individual analyte components $x_{ss,i}$.

A function $f(\cdot)$ specifies the law of sensor sensitivity to the analytes.

$$f_i(c_i) = \sum_{k=1}^{M} \beta_{i,k}\ B_k(c_i),\ \ i = 1, 2, ..., N \tag{3}$$

where the concentration range can be divided into M regions, and coefficients $\beta_{i,k}$ stands for the sensitivity coefficients to analyte i at the concentration level k. Such implementation represents a broken stick regression at M concentration intervals with base functions $B(\cdot)$.

The simplest case of linear relationship between sensor signal and analyte concentration is achieved if M is equal to 1 and $B(\cdot)$ is the identity function. The non-linear implementation of the function $f(\cdot)$ is based on spline base functions $B(\cdot)$. The use of the non-linear calibration model is reasonable when the sorption model is disabled. In this case, the function $f(\cdot)$ is an abstraction that can fit an arbitrary non-linear law of sensor sensitivity. Such approach of pure modeling includes a number of benefits. Monotonicity of the sensitivity curve can be guaranteed by use of the integrated version of B-splines (I-splines), due to the monotonic nature of these base functions. To simulate the saturation behavior of the sensor at high concentrations of analyte i, it is sufficient to set up the last coefficient $\beta_{i,M}$ to zero. On the other side, fixing the first coefficient $\beta_{i,1}$ to zero makes the sensor not responding to analyte i at very low concentration levels. Both implementations of the calibration model, ordinary linear regression and spline-based regression, are available in the released software tool.

The monotone spline regression [29] is formulated as a least-squares optimization problem as pointed in Equation 4.

$$\min_{f} \{\ [x_{ss,i} - f_i(c_i)]^2 + \lambda \int [f''(c_i)]^2 dc_i]\ \} \tag{4}$$

where the second term in the sum is a smoothing penalty expression for the second derivative $f''(c_i)$, and parameter λ controls the level of smoothing. The quadratic programming approach is used to control the non-negativity of the coefficients $\beta_{i,k}$ to assure the curve monotonicity.

Finally, the dynamic part of the Calibration Model to emulate the transient behavior of sensor is given in Equation 5.

$$x(t) = AR(P\ ,\ \tau_{i,p} |\ x_{ss}), p = 1, 2, ..., P \tag{5}$$

where AR stands for a auto-regressive filter of the order P. This simulates the transient sensor response $x(t)$ based on the steady-state value x_{ss}, previously derived from the equations 1 and 2. Such approach allows to produce an arbitrary signal waveform, not only a response to the rectangular gas pulse, parametrized with P time constants $\tau_{i,p}$ per analyte.

Overall, the group of the Equations 1, 2 and 5 define the design of noise-free sensor array data, where the set of parameters controls the creation of a synthetic sensor instance. We especially treat the parameters k_i to code an abstract sensor type expressing such sensor properties, as selectivity to analytes and sorption affinity in mixtures, that are mostly related to characteristics of the polymer composite. The other parameters, sensitivity coefficients $\beta_{i,k}$ and time constants $\tau_{i,p}$, are assumed to contain more variability along sensor instances, caused by many factors in the sensor design, for example, the geometry of the device. The sorption capacity q_i is set to 1 for all the sensor instances, because the coefficients $\beta_{i,k}$ play the same role.

When the user creates an array of arbitrary number of sensors, a new sensor instance is derived by varying the model parameters pre-computed for the seventeen UNIMAN sensors. Parameters $\beta_{i,k}$ and $\tau_{i,p}$ are generated from an univariate uniform distribution with control for non-negative values and the level of spread, while parameters k_i are copied from the seventeen UNIMAN sorption profiles, this preserving the number of sensor types.

2.3. Noise models

Three noise models introduced in the Figure 1 XXX represent three types of noise injected into the sensor array data. We referred these types of noise as additive, multiplicative and common, which correspond to Concentration Noise Model, Sensor Noise Model and Drift Model, respectively. The implementation details will be given later on, but, jointly for all the models these noise parameters are generated with multivariate normal distribution of independent variables with diagonal covariance Σ-matrices and zero mean.

2.3.1. Concentration Noise Model

The Concentration Noise Models emulate perturbations in the analyte delivery system. Equation 6 defines the noise term ΔC_0 induced into the input matrix C_0 of analyte concentrations in air.

$$\Delta C_0 = \mathcal{N}_{\Sigma_c}\ log(1 + C_0) \tag{6}$$

where $\mathcal{N}_{\Sigma_c}$ is the normally distributed noise with zero-mean and diagonal covariance matrix Σ_c for three analytes. The diagonal structure of the covariance matrix implies that the concentration noises of analytes do not affect each other. The logarithm term expresses an additional scaling of the amplitude applied to simulate more noise on high levels of concentration.

2.3.2. Sensor Noise Model

The Sensor Noise Model simulates a degradation in the performance of an individual sensor by injecting noise in the sensitivity coefficients $\beta_{i,k}$ from the Equation 4. Equation 7 defines a noise term $\Delta\beta_{i,k}$ induced into the coefficient $\beta_{i,k}$ for analyte i at the concentration interval k.

$$\Delta\beta_{i,k} = R_{n,\,\sigma_{i,k}} \tag{7}$$

where $R_{n,\sigma_{i,k}}$ is an one-dimensional random walk of n steps based on the normal distribution with zero-mean and standard deviation $\sigma_{i,k}$. The length of the output noise vector is equal to n, and its root mean square is proportional to $\sqrt{n}$. The other parameter of random walk $\sigma_{i,k}$ controls the noise amplitude individually for each coefficient $\beta_{i,k}$. The LTD set is used to compute the standard deviation statistics for the sensitivity coefficients derived from the Equation 4, which gives an estimation of the parameters $\sigma_{i,k}$.

2.3.3. Drift Model

The Drift Model generates the drift noise in two steps. A preliminary step involves quantification of drift-related data presented in the LTD set. In the next step the noise is injected into the sensor array signals by means of the component correction technique. The primary question arising in drift modeling is related to the way of one defines the drift phenomena for gas sensor arrays. In this work we rely on our previous study of a drift compensation method, where the multivariate representation of the UNIMAN dataset suggests that the data contain a drift-related subspace P [36]. The subspace was evaluated via common principal component analysis.

The hypothesis of common principal component analysis H_c states that exists an orthogonal matrix V such that the covariance matrices of K groups have the diagonal form simultaneously, as formulated in Equation 9.

$$H_c : L_i = V^T\,\Sigma_i\,V,\;\; i = 1, 2, ..., K \tag{8}$$

where Σ_i is covariance matrix of group i, and L_i is its diagonalized form when the linear transformation with matrix V is applied. The resulting eigenvectors (columns of the matrix V) define the subspace common for the groups, and the eigenvalues (different for each class) assist to evaluate the statistical significance of the estimator. In practice, the common hypothesis H_c is not feasible, because the exact solution exists only in the case the covariance matrices commute. To find an approximate solution, we employ a new algorithm recently published by Trendafilov [34]. The algorithm imitates standard principal component analysis iteratively via the well-known power method. The advantage of the new method is its computational efficiency and the availability of score-values for the common eigenvectors in terms of the captured variance. As a result of common principal component analysis, the drift subspace P contains the few columns of the matrix V.

The Drift Model generates the noise just in the multivariate subspace defined by P, as pointed in Equation 9.

$$\Delta X_P = R_{n,\Sigma_d} \tag{9}$$

where R_{n,Σ_d} is a multi-dimensional random walk of n steps based on the multivariate normal distribution with zero-mean and diagonal covariance matrix Σ_d. The relative proportion along the diagonal elements in Σ_d is specified by the importance of drift components in terms of projected variance. For the LTD set first three components captures the data variance in percentage as 86.23% 7.25% and 3.26%.

Finally, equation 10 shows the component correction operation that allows to induce the generated noise ΔX_P back into the complete multivariate space of the sensor array data.

$$\Delta X = (\Delta X_P\; P)\; P^T \tag{10}$$

where ΔX represents the final drift noise matrix that will be added to the output data matrix X. The control on the number of drift components (the number of columns in P) indicates the level of non-linearity for the drift noise and can be parametrized in the simulation.

2.4. Summary

The user creates a virtual sensor array given two groups of parameters. The first group of parameters k_i, $\beta_{i,k}$ and $\tau_{i,p}$ encodes the type of sensor instances the array will be composed of. The number of sensor types is equivalent to number of the UNIMAN sensors instantiated. In practice, it is sufficient to pass a number from 1 to 17 to encode the sensor instance. The second group of parameters controls the volume of noise that will be induced into the sensor array data. The dimensionless parameters σ_c, σ_β and σ_d scale the magnitude of the concentration noise, sensor noise and drift noise, respectively. The three parameters lie in the range from 0 to 1, where zero values allow the noise-free mode, and one corresponds to the maximum level of noise, equal to the noise presented over the LTD set. In addition, the structure of the drift noise can be parametrized with the number of drift components (modeling the dimension of the drift subspace) used in the multivariate data space.

All software is coded in open source R language for statistical computing [28] and packed in the library *chemosensors* soon to be available on the R public repository CRAN. The released package makes use of an object-oriented programming paradigm, supports parallel computing and contains the datasets depicted in the manuscript. Figure 3 outlines a typical example of R code to emulate a virtual sensor array and then generate the raw data from a predefined concentration matrix.

```
sa <- new("SensorArray", nsensors=10,
  num=1:5, csd=0.1, ssd=0.1, dsd=0.1, ndcomp=1)
sdata <- sdataModel(sa, conc, nclusters=2)
```

Figure 3: An example of R code to generate parametrized synthetic sensor array data from a predefined concentration matrix *conc*.

The first command creates a virtual array as an object *sa* of class *SensorArray*. The first two parameters *nsensors* and *num* indicate 10 sensor elements in the array derived from 5 sensor types. The next three parameters *csd*, *ssd* and *dsd* (parameters σ_c, σ_β and σ_d mentioned above) set up the level of noise at 10% for all noise sources. The last parameter *ndcomp* fixes the only

one drift component for the drift noise. The second command calls the class method *sdataModel*, which produces the sensor array data from predefined concentration matrix *conc*. The parameter *nclusters* points out to run calculations in parallel with 2 CPU cores, if available.

Figure 3 XXX visualizes the results of execution of the R code given above on a particular concentration matrix. This concentration matrix of three columns encodes the concentration of analytes in the form of rectangular pulses of 60s. Three lines on the plot (a) encode the change in concentration over time for three analytes A (solid), B (dashed) and C (dotted), that results in a total of seven different pulses of the three pure analytes and their binary and ternary mixtures. The lines on the plot (b) show the response of the array of ten sensors created according to the example code. Visually, the sensor signals have diverse time dynamics in response to gas pulses, reach the steady-state at different levels along the analytes and tolerate a certain drift effect according to the change in the base line.

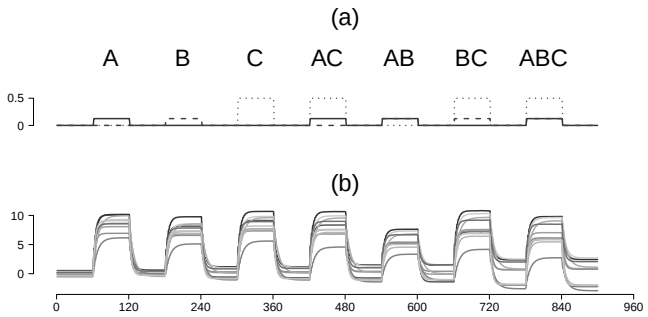

Figure 4: The input and output of the simulation models. The concentration matrix of three analytes A (solid line), B (dashed) and C (dotted) encodes the change in concentration over time on plot (a). The raw data show the response of the sensor array on the plot (b).

3. Simulation Examples

3.1. UNIMAN Replicas

One of the main features of the proposed simulation models lies in the fact that the virtual arrays are designed to replicate the reference sensor array. The given simulation creates a virtual array of the same number of sensors (seventeen) and then generates the signals in response to the concentration matrix of the gas classes from the Table IXXX. Another goal of this simulation is to test different noise parametrization in the sensor array data.

Figure 4 XXX presents a principal component analysis scoreplot of two hundred samples and eight gas classes of the UNIMAN dataset on plot (a) in comparison with synthetic data from two virtual arrays on plots (b) and (c). Both virtual arrays are parametrized at the noise level of 50%, but the array on plot (b) is drift-free and contains only concentration noise and sensor noise, as the array on plot(c) is influenced by all three noise sources. The figure denotes the matching multivariate structure of class-dependent information within the similar numbers of captured variance on principal components.

The distribution of noise in data on plot (b) is consistent with the nature of the simulation models. The concentration noise is oriented towards concentration direction of the analytes, and the sensor noise adds more distortion to the data. The effect of the sensor noise to high-concentration classes is stronger, as the noise model affects the sensitivity coefficients directly. The major drift direction in the multivariate space presented on plot (c) coincides with the drift direction given in the UNIMAN data on plot (a). Hence, the synthetic data generated with all three noise sources is capable to reproduce well the reference dataset.

3.2. Case Study

A main advantage of the proposed virtual array under real arrays is the flexibility for extending simulations to an arbitrary number of sensors and support of multi-component mixtures. We conducted a synthetic experiment with an array combined of one hundred sensors of seventeen UNIMAN types. Three analyte concentrations were selected from the middle of the range given in the Table IXXX, and the gas classes included three pure analytes A at 0.025 vol.%, B at 0.025 vol.% and C at 0.5 vol.%, their three binary mixtures AB, BC and AC, and one ternary mixture ABC. The typical transient response of ten sensors to the gases is shown on the Figure 3XXX in the previous section. An instance of a long-term synthetic dataset of 3500 samples is re-generated in the presence of concentration noise, sensor noise and drift noise parametrized at the maximum 100% level.

We formulated a classification problem for seven gas classes under conditions of strong drift, following the experimental scheme presented in our previous work for UNIMAN dataset [36]. Particularly, the training set was selected of the same size of 1000 samples, and a kNN classifier ($k = 3$) was used to train a classification model. The validation procedure selected a validation set of 100 samples by means of the sliding window operation, so that the classification metric is measuring the validity of the training model as the validation is further in the future. The data in both training and validation sets is preprocessed by component correction to counteract to the drift. The examined correction methods are the method of the reference gas by Artursson [1] and the method based on common component analysis [36] with one and two components. Figure 5 XXX presents the experimental results. Plot (a) depicts the principal component analysis scores of the training data after component correction, and plot (b) shows the performance of the classifier in combination with the different drift counteraction methods.

The sensor array data presented on Figure 5 (a)XXX shows the distribution for gas classes expected from the parametrization of the simulation models. The simulated sensors adopt the affinity profiles found for the seventeen UNIMAN sensors, that particularly leads to the prevalence of analyte *C* and suppression of analyte *B* in all mixtures. However, the relation of analyte forces in mixtures can be changed by selection of proper sensors (for example, sensors with more affinity to *A* than *C*) or

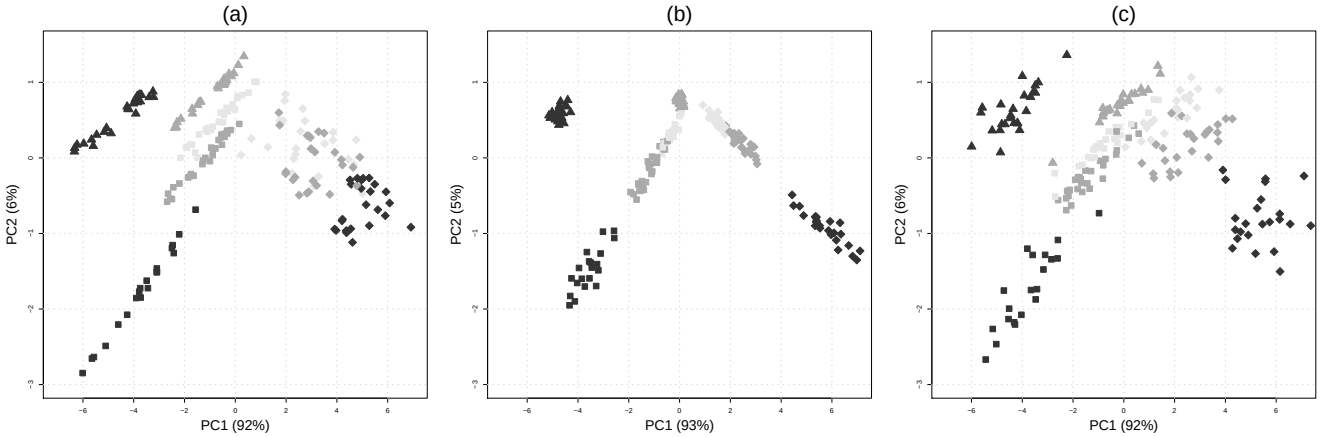

Figure 5: The UNIMAN dataset of two hundred samples (a) is compared with two synthetic data (b) and (c) from virtual arrays of seventeen sensors with different noise parametrization. The array (b) is drift-free and contains only concentration noise and sensor noise at the 50% level. The array (c) is influenced by all three noise sources at the 50% level. The synthetic data (c) replicates the reference data (a) matching both the class-dependent and noise-related multivariate structure.

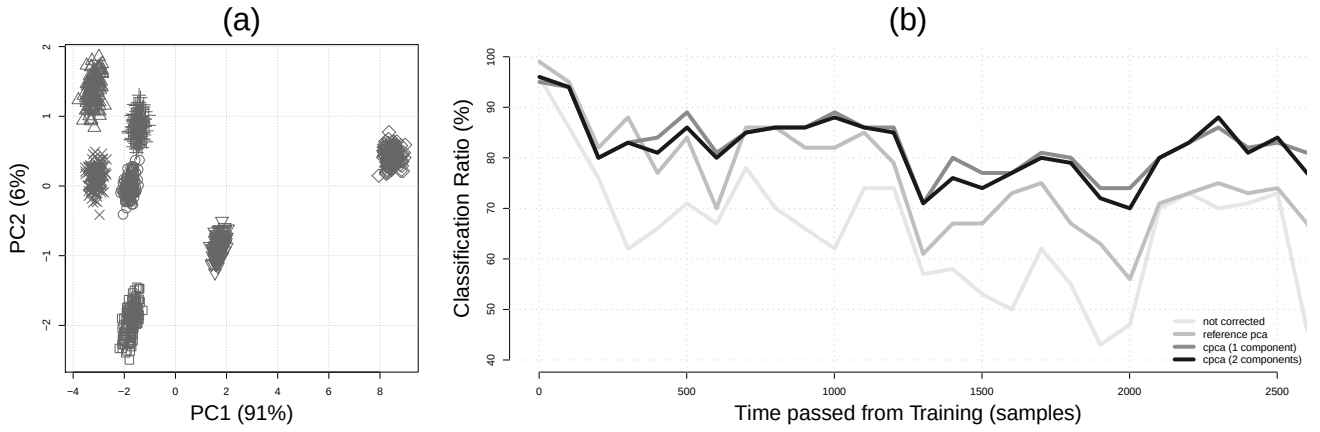

Figure 6: Evaluation of the case study experiment for drift compensation. Plot (a) show the principal component analysis (PCA) score plot of the training set after drift correction. Eight gases are coded with symbols □ (A), ◇ (B), △ (C), ∇ (AC), + (BC), × (AC) and *o* (ABC). Plot (b) shows changes in the performance of a kNN classifier (k = 3) on validation sets, where the time on *X* axis is the distance between the training and validation sets in samples.

tuning of concentration volume along the components in mixture.

Figure 5 (b)XXX shows the change in classification ratio as the time difference between training and validation sets increases. The comparison of performance numbers over time for non-corrected and drift-corrected experiments clearly shows that drift in data plays a crucial role in the degradation of the classifier performance. The application of the reference gas method by Artursson gas slightly improves the classification results, although the method likely captures the portion of noise related to the sensor noise of the reference gas, that leads to confusion with drift. The best classification results are achieved by the method based on the common principal component analysis that marks out the drift part of noise more accurately. The method is not capable to reach the 100% performance, although the method of drift injection for the virtual array is the same. That is explained by the strong and irreversible changes induced by the noise models emulating the effect of sensor noise for individual sensors degrading the overall performance of the array.

4. Conclusions

The main contribution of this work is a software framework for large-scale synthetic experiments in machine olfaction, that features sensor arrays with arbitrary number of elements, concentration profiles of arbitrary mixtures of three analytes and parametric control of the level of noise in data. The framework was implemented in the software library *chemosensors* in R, under the Neurochem project, funded by the European Commission. The developed software package contains models to emulate virtual sensor arrays inspired from a reference array of seventeen conducting polymer sensors, including the extension model based on the Langmuir isotherm to simulate mixtures of analytes and realistic models of long-term drift.

The authors believe that public synthetic dataset generators are especially interesting for applications regarding optimization problems, like benchmarking of machine olfaction algorithms or searching for an optimal design of sensor arrays. The package is also able to generate large (over 1k sensor) sensor arrays suitable for the prototyping of neuromorphic signal processing.

Acknowledgment

This work was funded from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 216916: Biologically inspired computation for chemical sensing (NEUROChem), the Ramon y Cajal program from the Spanish Ministerio de Educación y Ciencia and TEC2010-20886-C02-02. CIBER-BBN is an initiative of the Spanish ISCIII.

References

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

[2] Asuncion, A. and Newman, D. (2007). UCI Machine Learning Repository.

[3] Avila, M. A. S. and Breiter, R. (2008). Competitive sorption of cis-DCE and TCE in silica gel as a model porous mineral solid. *Chemosphere*, 72(11):1807–15.

[4] Bai, R. and Yang, R. T. (2001). A Thermodynamically Consistent Langmuir Model for Mixed Gas Adsorption. *Journal of colloid and interface science*, 239(2):296–302.

[5] Bai, R. and Yang, R. T. (2003). Improved Multisite Langmuir Model for Mixture Adsorption Using Multiregion Adsorption Theory. *Langmuir*, (12):2776–2781.

[6] Charlesworth, J. M., Partridge, A. C., and Garrard, N. (1993). Mechanistic studies on the interactions between poly(pyrrole) and organic vapors. *The Journal of Physical Chemistry*, 97(20):5418–5423.

[7] Clifford, P. and Tuma, D. (1983a). Characteristics of semiconductor gas sensors I. Steady state gas response. *Sensors & Actuators*, pages 233–254.

[8] Clifford, P. and Tuma, D. (1983b). Characteristics of semiconductor gas sensors II. Transient response to temperature change. *Sensors & Actuators*, pages 255–281.

[9] Cole, M. and Ulivieri, N. (2003). Parametric model of a polymeric chemoresistor for use in smart sensor design and simulation. *Architecture*, 34:865–875.

[10] Gardner, J., Bartlett, P., and Pratt, K. (1995). Modelling of gas-sensitive conducting polymer devices. *IEE Proceedings - Circuits, Devices and Systems*, 142(5):321.

[11] Gardner, J., Llobet, E., and Hines, E. (1999). SPICE model for resistive gas and odour sensors. *IEE Proceedings - Circuits, Devices and Systems*, 146(3):101.

[12] Geng, Z., Yang, F., and Wu, N. (2011). Optimum design of sensor arrays via simulation-based multivariate calibration. *Sensors and Actuators B: Chemical*, 156(2):854–862.

[13] Gutierrez-Osuna, R. (2000). Drift Reduction For Metal-Oxide Sensor Arrays Using Canonical Correlation Regression And Partial Least Squares. *Analysis*, pages 1–7.

[14] Gutierrez-Osuna, R. (2002). Pattern analysis for machine olfaction: a review. *IEEE Sensors Journal*, 2(3):189–202.

[15] Gutierrez-Osuna, R. (2003). Transient response analysis for temperature-modulated chemoresistors. *Sensors and Actuators B: Chemical*, 93(1-3):57–66.

[16] Ishida, H., Yamanaka, T., Kushida, N., Nakamoto, T., and Moriizumi, T. (2000). Study of real-time visualization of gas/odor flow image using gas sensor array. *Sensors and Actuators B: Chemical*, 65(1-3):14–16.

[17] Lei, H., Pitt, W. G., Mcgrath, L. K., and Ho, C. K. (2007). Modeling carbon black / polymer composite sensors. *Sensors And Actuators*, 125:396–407.

[18] Lin, C. W., Hwang, B. J., and Lee, C. R. (1999). Characteristics and sensing behavior of electrochemically codeposited polypyrrole-poly(vinyl alcohol) thin film exposed to ethanol vapors. *Journal of Applied Polymer Science*, 73(11):2079–2087.

[19] Marco, S., Ortega, A., Pardo, A., and Samitier, J. (1998). Gas identification with tin oxide sensor array and self-organizing maps: adaptive correction of sensor drifts. *IEEE Transactions on Instrumentation and Measurement*, 47:316.

[20] Markham, E. C. and Benton, A. F. (1931). The adsorption of gas mixtures by silica. *Journal of the American Chemical Society*, 53.

[21] Montoliu, I., Tauler, R., Padilla, M., Pardo, a., and Marco, S. (2010). Multivariate curve resolution applied to temperature-modulated metal oxide gas sensors. *Sensors and Actuators B: Chemical*, 145(1):464–473.

[22] 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. *Chemometrics and Intelligent Laboratory Systems*, 100(1):28–35.

[23] Padilla, M., Perera, A., Montoliu, I., Chaudry, A., Persaud, K. C., and Marco, S. (2007). Poisoning fault diagnosis in chemical gas sensor arrays using multivariate statistical signal processing and structured residuals generation LiIi # D. *2007 Ieee International Symposium on Intelligent Signal Processing*.

[24] Pearce, T., Schiffman, S., Nagle, H., and Gardner, J. (2003). *Handbook of Machine Olfaction - Electronic Nose Technology*. John Wiley & Sons.

[25] Perera, A., Yamanaka, T., Gutierrez Galvez, A., Raman, B., and Gutierrez-Osuna, R. (2006). A dimensionality-reduction technique inspired by receptor convergence in the olfactory system. *Sensors and Actuators B:*

Chemical, 116(1-2):17–22.

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

[27] Persaud, K. C., Bissell, R. A., and Travers, P. (2002). The influence of non-specific molecular partitioning of analytes on the electrical responses of conducting organic polymer gas sensors. *Physical chemistry chemical physics : PCCP*.

[28] R Development Core Team (2010). R: A Language and Environment for Statistical Computing.

[29] Ramsay, J. O. (1988). Monotone Regression Splines in Action. *Statistical Science*, 3(4):425–441.

[30] Röck, F., Barsan, N., and Weimar, U. (2008). Electronic nose: current status and future trends. *Chemical reviews*, 108(2):705–25.

[31] Stussi, E., Stella, R., and Rossi, D. D. (1997). Chemoresistive conducting polymer-based odour sensors : influence of thickness changes on their sensing properties. *Sensors And Actuators*, 43:180 – 185.

[32] Topart, P. and Josowicz, M. (1992a). Characterization of the interaction between poly(pyrrole) films and methanol vapor. *The Journal of Physical Chemistry*, 96(19):7824–7830.

[33] Topart, P. and Josowicz, M. (1992b). Transient Effects In the Interaction between Polypyrrole and Methanol Vapor. *The Journal of Physical Chemistry*, 96(21):8662–8666.

[34] Trendafilov, N. T. (2010). Stepwise estimation of common principal components. *Computational Statistics & Data Analysis*, 54(12):3446–3457.

[35] Yang, R. T. (2003). *Adsorbents: fundamentals and applications*. John Wiley and Sons, Inc., Hoboken, New Jersey.

[36] Ziyatdinov, A., Marco, S., Chaudry, A., Persaud, K., Caminal, P., and Perera, A. (2010). Drift compensation of gas sensor array data by common principal component analysis. *Sensors and Actuators B: Chemical*, 146(2):460–465.