

Data-driven modeling of nano-nose gas sensor arrays

Tommy S. Alstrøm^a, Jan Larsen^a, Claus H. Nielsen^{b,c} and Niels B. Larsen^b

^aDTU Informatics, Technical Univ. of Denmark, DK-2800 Kongens Lyngby, Denmark;

^bDTU Nanotech, Technical Univ. of Denmark, DK-2800 Kongens Lyngby, Denmark;

^cDept. of Chemistry, Univ. of Copenhagen, DK-2100 Copenhagen Ø, Denmark.

ABSTRACT

We present a data-driven approach to classification of Quartz Crystal Microbalance (QCM) sensor data. The sensor is a nano-nose gas sensor that detects concentrations of analytes down to ppm levels using plasma polymerized coatings. Each sensor experiment takes approximately one hour hence the number of available training data is limited. We suggest a data-driven classification model which work from few examples. The paper compares a number of data-driven classification and quantification schemes able to detect the gas and the concentration level. The data-driven approaches are based on state-of-the-art machine learning methods and the Bayesian learning paradigm.

Keywords: Polymer Coated Quartz Crystal Microbalance Sensor (QCM), Gaussian Process Regression (GPR), Artificial Neural Network (ANN), Non-negative Matrix Factorization (NMF), Principal Component Analysis (PCA), Principal Component Regression (PCR), Classification, Concentration Level Estimation.

1. INTRODUCTION

The development of gas sensors is a field of great activity. Particularly the development of electronic noses for use in process control, quality control in the food and beverage industry, pollution monitoring and airport security.^{1,2} Ideally such an electronic nose should be portable, provide both a qualitative and quantitative determination of analytes as well as offer stable performance over extended periods of time. This means that the sensor element(s) must operate in a reversible manner and show an affinity towards target analytes.

Many different strategies have been pursued to develop selective gas sensors. One approach is to functionalize the physical sensor device with chemical coatings. Molecularly imprinted polymers (MIPs) seek to mimic biological antibodies. The selectivity of these “plastic antibodies” has been shown to be somewhat random and not easily controlled.³ Several macromolecules have been evaluated as potential gas sensors. One of the more popular choices are cyclodextrins, which are cage-like molecules with hydrophobic cavities and hydrophilic edges.⁴ These molecules offer both chemical and steric selectivity and serve as a platform for a vast group of derivatives that tune the specificity towards certain compounds or classes of compounds. However, the selective modification of cyclodextrins is not a simple process and requires elaborate synthesis and isolation steps.⁵ Another drawback is the fact that these substances must be dissolved in a solvent from which a film can be cast by for example spin or spray coating in order to functionalize a sensor surface.

In this work an alternative methodology known as plasma polymerization is used for sensor functionalization. This method allows for solvent free deposition of a wide range of thin polymer films on the surface of most sensors including cantilever and quartz crystal microbalance (QCM) based sensors.⁶ Plasma polymerization also offers excellent control over film thickness since it is directly proportional to polymerization time as confirmed by Kurosawa et al.⁷ A wide selection of monomers with many different functional groups can be used for plasma polymerization making it a very versatile technique. The choice of monomer naturally depends on the target analytes to be detected. As described by Grate and Abraham the choice of coatings can advantageously be based

Further author information: (Send correspondence to T.S.A.)

T.S.A.: E-mail: tsal@imm.dtu.dk, Telephone: (+45) 45 25 39 04

J.L.: E-mail: jl@imm.dtu.dk, Telephone: (+45) 45 25 39 23

C.H.N.: E-mail: chn@chn-analytical.dk

N.B.L.: E-mail: niels.b.larsen@nanotech.dtu.dk

on the main interaction parameters that are responsible for the sorption of gases into solids.⁸ According to the linear solvation energy relationship (LSER) model these parameters are: dispersion interactions, polarizability, dipolarity and hydrogen bonding. Ideally one would develop a selection of coatings where each coating is representing only one of the above interaction parameters. This is however not easily done in practice, but Grate and Abraham suggest several good candidates that can be assumed to primarily interact via one parameter only. The coatings are deposited on a QCM based sensor. The working principle of QCM sensors resemble tuning forks. The resonance frequency response of QCM sensors is close to linear with respect to mass change.⁶ Coating the quartz crystals make them selective towards various analytes making them suitable as gas sensors. In order to verify the selectivity of the coatings the sensor is tested using six different analytes. The response data is subject to data analysis methods. A number of methods have been applied to polymer coated QCM sensors. Often QCM sensors selectivity has been analyzed by using principal component analysis (PCA)^{9–11} or hierarchical cluster analysis (HCA).^{12,13} Recently the application of artificial neural networks (ANN) evolved as the dominant method.¹⁴

As described earlier the response of QCM sensors is sensitive to mass changes making them viable for detecting concentration levels of analytes. Various methods have been used to construct QCM based concentration level estimators such as partial least squares (PLS)¹¹ and multivariate linear regression.¹⁵ Artificial neural networks have been applied as concentration level estimators in a broad range of applications.^{14,16–18}

In this paper we compare classification using ANN and PCA with a Non-negative Matrix Factorization (NMF) classification scheme. NMF was introduced by Lee and Seung¹⁹ as an alternative to PCA for feature extraction applied to identifying facial features but have found use in a vast range of areas. Further we propose a QCM sensor consisting of 8 crystals with different coatings. The sensor is able to classify the test analytes with high accuracy even though the available training data is limited. In order to perform concentration level estimation the analytes are measured at various concentrations. Three different methods for concentration level estimation are evaluated; Principal component regression (PCR), ANN and Gaussian process regression (GPR). PCR is a linear method that performs well with a limited amount of training points although it handles non-linearity in data poorly. To overcome this problem neural networks has usually been applied as ANN can model non-linearities. GPR is so far an untried method for QCM data. GPR is a non-linear method as ANN but GPR should be able to perform well with a limited amount of data points.

2. SENSORS AND MEASUREMENT SYSTEM

A selection of eight monomers were chosen based on the single interaction parameter. All monomers used, dodecane (DOD), heptadecafluoro-1-decene (HDFD), maleic anhydride (MAH), methylene dioxobenzene (MDOB), methyl methacrylate (MMA), styrene (STY) and vinyl pyrrolidone (VP), were of at least $\geq 99\%$ purity except di(ethylene glycol) vinyl ether (DEGVE), which was of 98% purity, and all of them were used without further purification. The monomers were polymerized at a total pressure of 10–13 Pa and at a power of $0.5 \frac{W}{l}$ in a 50 Hz (i.e. low frequency) plasma chamber.²⁰ Argon was used as a carrier gas. Surfaces were activated for 2 min in argon plasma followed by addition of monomer to initiate polymerization. After polymerization power was turned off, but the monomer was allowed to flow for another 2 min. This “afterglow polymerization” is believed to reduce the amount of free radicals on the polymer surface by reaction with monomer species. For gas sorption measurements 10 MHz, 0.538“ diameter AT-cut quartz crystals with 10Å Ti + 1000Å Au electrodes driven by 10 MHz QCM lever oscillators (all from International Crystal Manufacturing Co.) were used. Crystals were coated on one side only by masking one side with SWT 20+ semiconductor dicing tape (from Nitto Denko). Crystals were mounted in a home built flow system made from poly(methyl methacrylate) (PMMA) which was placed on top of an aluminum block held at a 33°C(±0.1°C) throughout the duration of the experiments. The entire flow system and all oscillators were placed inside a closed box to shield the QCMs from excessive temperature variations (the temperature in the laboratory varied up to 5°C between day and night time). Figure 1 shows a schematic of the experimental setup. Analytes used were acetone ($\geq 99.8\%$), anhydrous ethanol ($\geq 99.9\%$), heptane ($\geq 99\%$), 1-pentanol ($\geq 99\%$), 1,3-Benzodioxole ($\geq 99\%$) and Milli-Q purified water (Millipore). The analyte vapors were produced by bubbling nitrogen through gas wash bottles containing analyte. Analyte concentrations were defined by adjusting the ratio of nitrogen flowing through and bypassing the wash bottle respectively. A constant flow rate of 100 sccm was maintained throughout the entire duration of the experiments using mass flow controllers

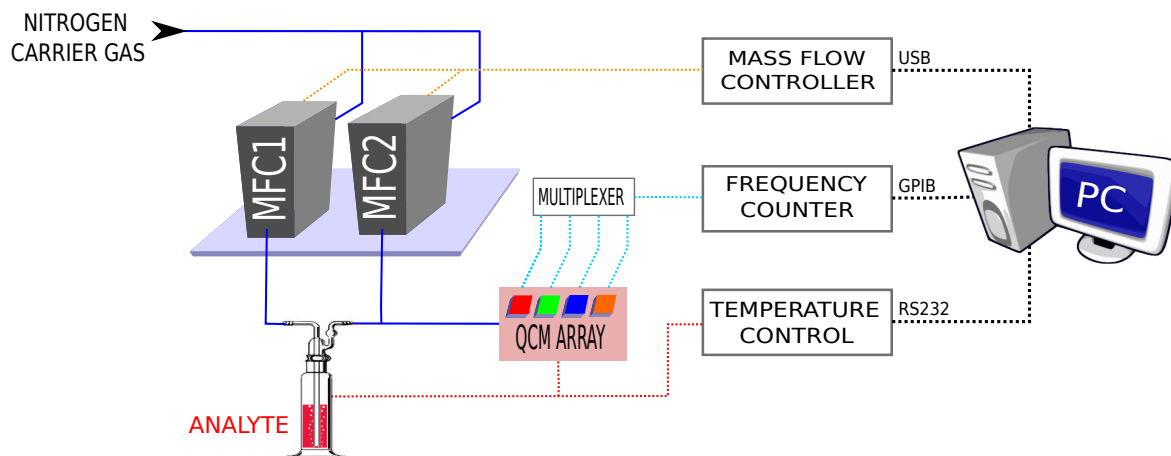


Figure 1. Experimental setup used for gas sorption measurements of QCMs coated with thin polymer films.

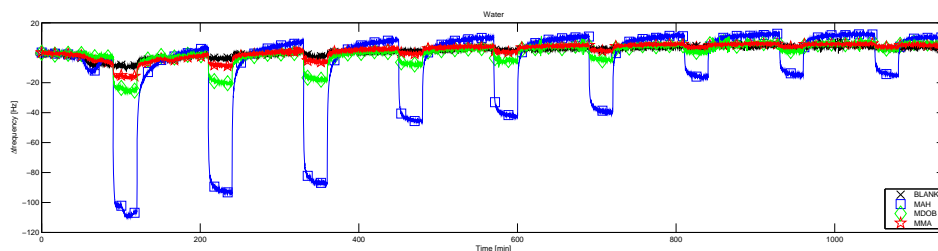


Figure 2. Δ frequency responses for MAH, MDOB, MMA and a blank quartz crystal for nine measurement cycles. The first three measurement cycles is water at 9200 parts per million (ppm), the next three cycles the concentration is at 4600 ppm and the last three cycles has a concentration level at 2300 ppm. The sign reversed peak value during a measurement cycle is recorded as representative for the cycle.

(mks Instruments, type 1179A). With an internal volume of the flow system of about 2.5 cm^3 this corresponds to about 39 exchanges per minute. For this reason the concentration of analyte in the gas phase is assumed to be constant. The setup allows for sequential measurement of four crystals. Each crystal was measured every 4 seconds and each measurement was taken as the median of 10 measurements. Of the 4 crystals being measured one was always a blank (i.e. uncoated) crystal. The gold surface of the blank crystal provided information about unspecific adsorption and deviations due to temperature changes. Resonant frequencies were measured using a Fluke PM6681 high resolution frequency counter. The sorption measurements consisted of repeated cycles of 100 min pure nitrogen flow followed by 40 min of analyte-spiked flow. The response was measured at six different concentrations, three times at each concentration, for each analyte. Figure 2 shows the response for MAH, MDOB and MMA using water as analyte. A measurement cycle consist of filling the chamber with pure nitrogen for 100 minutes allowing the coatings to get a known steady state response. After the 100 minutes have passed an analyte at a given concentration level is added to the flow maintained for 40 minutes. The peak response is recorded as representative for the measurement cycle. Figure 3 shows the measured (peak sign reversed) frequency responses for all analytes. The frequency readings show a near linear response with respect to mass for most coatings. Other coatings such as STY for heptane show a non-linear response. All responses but one are of the same sign. HDFD shows a negative response towards pentanol. This is because HDFD is a teflon-like coating onto which condensation of pentanol is less likely to occur compared to the uncoated reference quartz crystal.

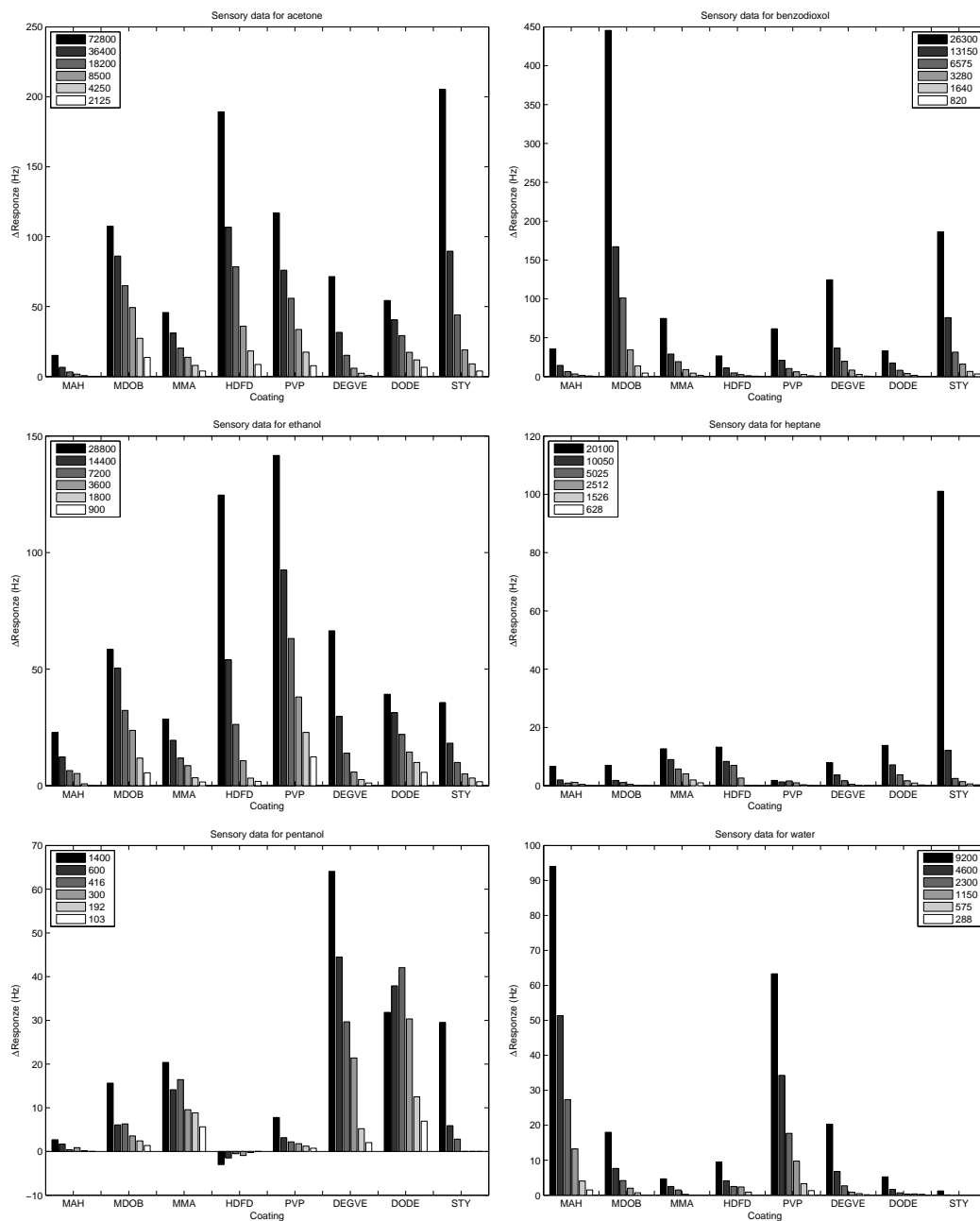


Figure 3. The bars are averages of the measured responses. Inspection of the data shows that only on one occasion does a crystal give a negative Δ frequency response, namely HDFD for pentanol. Note both the concentration levels and the y -axis scaling varies between analytes. The concentration levels is in parts per million (ppm).

3. ALGORITHM TRAINING AND EVALUATION

3.1 Data partitioning

As described in section 2 we have conducted experiments on six different analytes. Each analyte is measured at 6 different concentration levels and each experiment is repeated 3 times. Therefore the dataset has 108 data points in total, 18 data points per analyte. The data is partitioned into a training set and a test set. We want to produce learning curves for the different algorithms so the size of the training set will vary. Our minimum training set

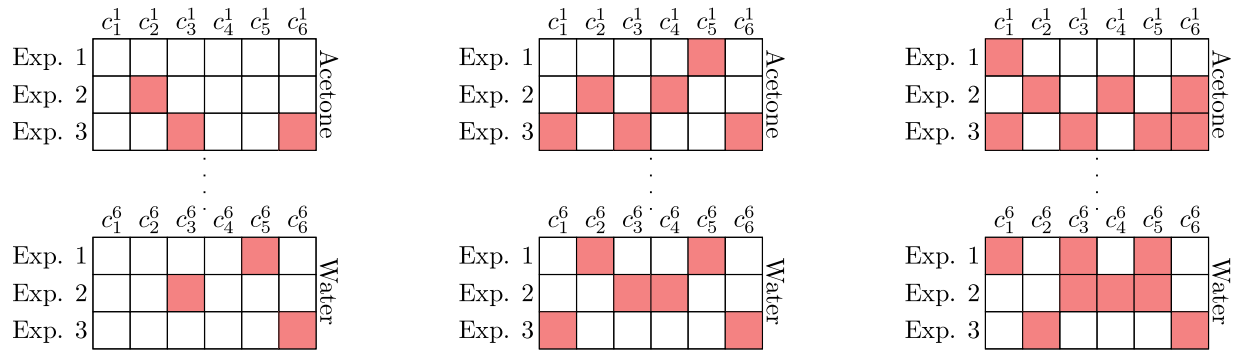


Figure 4. The figure illustrate three different data partitioning scenarios. Left has $N_{train} = 3$, center has $N_{train} = 6$ and right has $N_{train} = 8$. Each square corresponds to one data point. The test set comprise the blank squares and the training set comprise the colored squares. c_i^a denoted that concentration level i for analyte a . Here acetone corresponds to $a = 1$ and water is $a = 6$.

size will be three training points per analyte – the minimum number of points that can reveal non-linearities in the sensor response. The maximum trained set size is chosen to be twelve training points per analyte, denoted as $N_{train} = 12$. This corresponds to an experimental setup where all concentration levels per analyte is included twice. We want to increase the training set size in roughly the same manner as an experimenter would include more and more experiments. To achieve this we impose constraints on how the data is partitioned. Each analyte must be represented evenly in training set - thus the training set size must be a multiple of six. Further each concentration level within each analyte, should be represented as evenly as possible. Choosing $N_{train} = 6$ we have each concentration represented exactly once. Figure 4 shows examples of data partitioning for $N_{train} = 3$, $N_{train} = 6$ and $N_{train} = 8$. Each data partitioning and subsequent model training and evaluation is performed 100 times. This is done to ensure that we are not comparing the algorithms on a training/test set that turned out to be an exception rather than the norm.

3.2 Model training and evaluation

The proposed methods have several parameters that can have a huge impact of performance. These parameters are often set using model selection. In order to select the model that has the lowest generalization error²¹ model selection methodology must be applied. We will work with both cross validation²¹ (CV) and marginal likelihood²² (ML) depending on the method in question. Model selection for classification using SVD and NMF will use leave-one-out cross validation (LOO-CV). Due to computational limitations model selection for ANN classification will use a 6-fold cross validation scheme (6F-CV) instead of LOO-CV. The 6F-CV set will contain one data point per analyte. Model selection for PCR and ANN regression will use LOO-CV. Model selection for GPR will use ML as described by Rasmussen and Williams.²² The performance of the classifiers are evaluated based on the misclassification rate defined as

$$E_{mis} = \frac{N_{error}}{N} \quad (1)$$

where N_{error} is the amount of misclassified points and N is the total amount of points that was classified. Evaluating performance on quantifiers must be done differently than for the classifiers. The performance criterion used is the mean relative absolute error as has been used by others.^{14,17} It is defined as

$$E(RAE) = \frac{1}{N} \sum_{n=1}^N \left| \frac{y_n - \hat{y}_n}{y_n} \right| \quad (2)$$

where y_n is the true concentration, \hat{y}_n is the predicted concentration and N is the amount of data points estimated. The advantage of this performance criterion is that errors at both high and low concentrations are penalized equally.

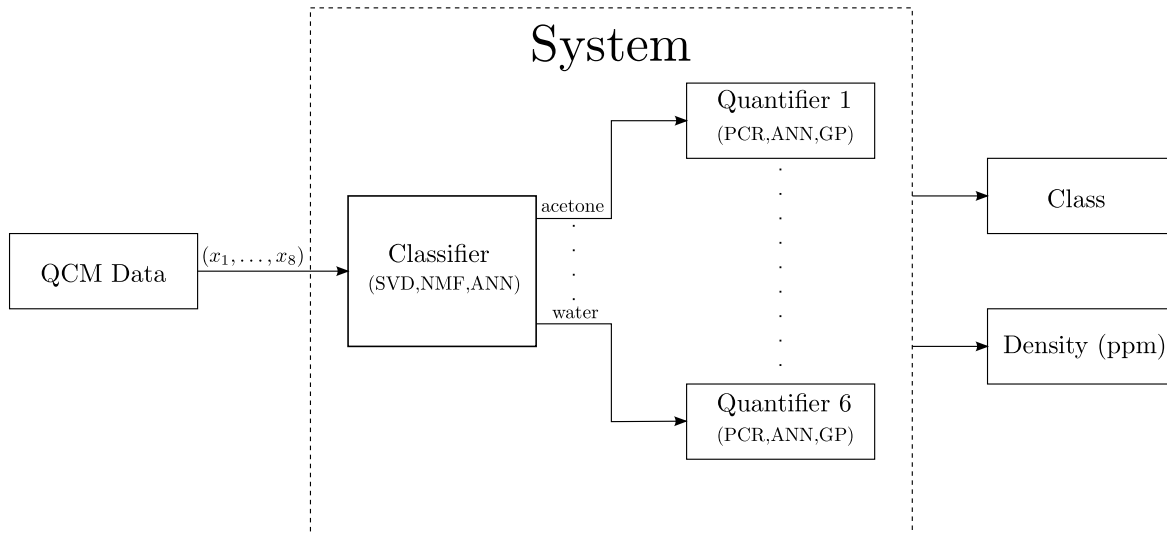


Figure 5. The schematic setup of the classification/quantification framework.

3.3 Data analysis framework

We adopt a two-tiered model where classification and concentration level estimation (quantification) is handled separately.¹⁴ Figure 5 shows how data and decisions are made throughout the data processing pipeline. Output from the QCM sensor is applied to a classifier. Based on the decision made by the classifier a given quantifier is selected. There is one quantifier trained per analyte. The output from the model is a class given as an analyte name and a concentration level given in ppm.

We will use past implementations in MATLABTM of most of the algorithms. The neural network algorithms has been developed during the past years at DTU Informatics and is freely available for download at DTU ISP toolbox.²³ The NMF algorithm is implemented by Lin²⁴ and is freely available from Lin's website but can be found in the DTU ISP Toolbox as well. The Gaussian process regression implementation used is published by Rasmussen and Williams.²⁵ The SVD and PCR algorithms are implemented by the authors.

3.4 Gaussian Process regression

Gaussian process regression is an extension of the linear Bayesian regression framework. A recent textbook on Gaussian processes in a machine learning context is written by Rasmussen and Williams.²² A Gaussian process is completely specified by its mean function and covariance function. Using Gaussian processes for regression requires two choices: a prior and a covariance function. We choose a zero mean Gaussian prior with unit variance. To make this prior plausible we subtract the empirical mean from the data and scale the data to unit variance. The empirical mean and empirical variance is calculated using the training set alone and as such the test data is not guaranteed to neither zero mean nor unit variance. We try three different covariance functions all based on the squared exponential covariance function²²

$$k(\mathbf{x}_p, \mathbf{x}_q) = \sigma_f^2 \exp\left(-\frac{1}{2}(\mathbf{x}_p - \mathbf{x}_q)^\top \mathbf{M}(\mathbf{x}_p - \mathbf{x}_q)\right) + \sigma_n^2 \delta_{pq} \quad (3)$$

where σ_f^2, σ_n^2 and \mathbf{M} are called the *hyper-parameters*. Typically σ_f^2 is called the signal variance and σ_n^2 the noise variance. Choosing $\mathbf{M}_{iso} = \ell^{-2} \mathbf{I}$ yields the isotropic squared exponential function where ℓ is called the length-scale. Choosing $\mathbf{M}_{ard} = \text{diag}(\ell)^{-2} \mathbf{I}$ yields a covariance function that implements automatic relevance determination.^{22, 26} The two approaches can be combined by adding the covariance functions together thus creating the third covariance function. Model selection consists of choosing suitable values for the hyper-parameters. This is done using the marginal likelihood approach as discussed by Rasmussen and Williams²² allowing 1000

function evaluations during the optimization. The initial values for hyper-parameters are drawn from a zero-mean Gaussian distribution with unit variance. In order to avoid bad local minima 10 restarts is tried picking the hyper-parameters that yields the best marginal likelihood.

3.5 Artificial Neural Networks

The neural network classifier is a two-layered feed forward network (one hidden layer). Neural networks with one hidden layer are general approximators²¹ so it is sufficient to use just one hidden layer. The network has eight inputs – one per polymer, and six outputs – one per analyte. Prior to training the entire data set whitened (made zero mean and scaled to unit variance). The hidden units use tangent hyperbolic sigmoidal function as transfer function. The network is trained using the BFGS optimization algorithm as described by Nielsen.²⁷ The cost function (that is used by the optimizer) is a quadratic cost function augmented with outliers detection and weight decay. The cost function $S(\mathbf{w})$ has two hyper-parameters α and β .

$$S(\mathbf{w}) = E_D(\mathbf{w}, \beta) + \alpha E_W(\mathbf{w}) \quad (4)$$

where $E_D(\mathbf{w}, \beta)$ is the cross-entropy error function and $E_W(\mathbf{w})$ is a regularization term.²⁸ Both hyper-parameters are updated after each iteration using MacKays Bayesian maximum likelihood II (MLII) scheme.^{29,30} The hyper-parameters are initialized to $\alpha = 8$ and $\beta = 1$ and the network weights are initialized using a zero mean Gaussian with variance equal to α . The stopping criterion for the network training is either 100 iterations or if the hyper-parameters are updated with a margin lower than 10^{-5} . Each network training cycle is repeated 10 times and the network with the lowest training error $E_D(\mathbf{w}, \beta)$ is selected for concentration level estimation. The network training and cost function is described in details by Sigurdsson et. al.²⁸ The neural network regressor is almost identical to the neural network classifier. The cost function does not have outliers detection and the regressor uses just one output neuron which hold the estimated concentration level.

The tunable parameters subject to cross validation is the number of hidden units. For the model selection part two to six hidden units were tested. The most often number of hidden units chosen was three which was chosen in 52.3% of the runs. The number of hidden units chosen for the regressor turned out to depend heavily on the analyte and the amount of training points available. Making a histogram on all runs put together showed an almost uniform histogram. However if the histogram is calculated for each value of N_{train} the most common choice for hidden units was 2 (for N_{train} in range 6–9)

3.6 Non-negative Matrix Factorization

Classification using NMF is a linear vector-space method. The data is represented in a 8-dimensional vector space and training consist of finding a suitable subspace as representative for each analyte. When classifying an unknown data point, the data point is projected into each subspace. The subspace that represents the data point best (calculated as residuals) is chosen as the correct analyte. This is done as follows; the training points for each analyte is stored as columns in a data matrix $\mathbf{X}_{analyte} \in \mathbb{R}^{8 \times N_{train}}$. NMF factorizes \mathbf{X} into two new matrices

$$\mathbf{W}\mathbf{H} \approx \mathbf{X}, \quad \mathbf{W} \in \mathbb{R}_{0+}^{8 \times b}, \quad \mathbf{H} \in \mathbb{R}_{0+}^{b \times N_{train}}, \quad \mathbf{X} \in \mathbb{R}_{0+}^{8 \times N_{train}} \quad (5)$$

NMF assumes that all values in \mathbf{X} is non-negative so all negative values in the data set are set to zero. Effectively it means pentanol will be classified without using the contribution from HDFD, see figure 3. The matrix \mathbf{W} is the basis and the columns of matrix \mathbf{H} is the data points represented in the basis \mathbf{W} . After training we have one \mathbf{W} matrix per analyte (it is enforced that all subspaces are of the same size). When classifying data points the NMF factorization is run with \mathbf{W} held fixed. This procedure is carried out for all six \mathbf{W} matrices resulting in six different \mathbf{H} matrices. The \mathbf{H} matrix that represent the data point best is chosen as the correct analyte.

The factorization is calculated using the approach by Lin.²⁴ Each element of matrices \mathbf{W} and \mathbf{H} are initialized to a value drawn from a uniform distribution in the interval $]0; 1[$. The stopping criterion for the algorithm is either 30 iterations or 10^{-4} as relative error. The amount of basis vectors to use per subspace b is a model parameter that is allocated by LOO-CV. Allowable subspace size is in the interval $b \in \{2, \dots, \min(N_{train} - 1, 10)\}$. The most often value chosen was $b = 2$. Note that b must be less than the amount of training points – so if $N_{train} = 3$ the subspace size is automatically chosen to $b = 2$. If we look at the case where $N_{train} = 12$ the subspace size was mostly chosen to $b = 2$ but only for 20% of the runs.

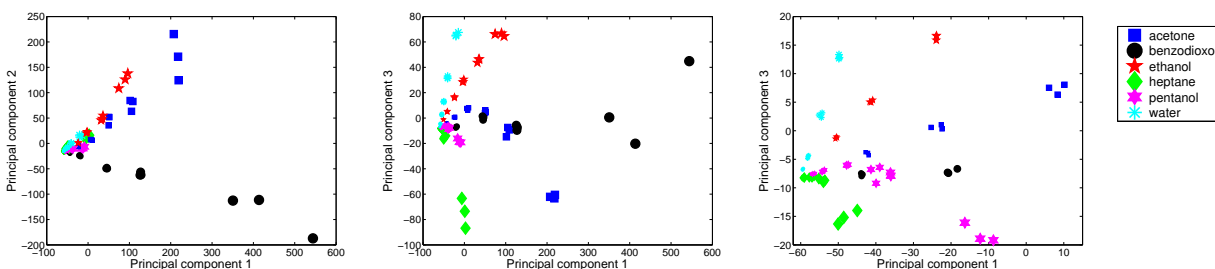


Figure 6. QCM data visualized using PCA. The size of the squares represent the concentration. Panel 3 is a zoomed in version of panel 2.

3.7 Principal component regression

Principal component regression is a linear regression method that is related to regularized least squares regression and partial least squared regression (PLS).^{21,31} The parameters to be chosen when using principal component regression is which principal components to include in the regression. We include principal components in the model ordered by their variance choosing the highest variance component. To determine the amount of components k to include in the model we use LOO-CV. The number of allowable components is $k \in \{1, \dots, \min(N_{train}, 8)\}$. For $N_{train} = 12$ the most commonly chosen value was $k = 6$, but only in 22.3% of the runs.

3.8 Singular Value Decomposition

Classification using SVD is identical to the NMF classification method except for the matrix factorization. The matrix factorization is carried out using SVD instead and basis vectors are included based on their eigenvalues, always choosing the basis vector with the highest eigenvalue. The SVD classification method used is described in details by Eldén.³¹ As with NMF the model selection parameter is the subspace size. For model selection the allowable subspace size of is $b \in \{1, \dots, \min(N_{train} - 1, 7)\}$. For $N_{train} = 12$ the most commonly chosen subspace size was $b = 2$ which was chosen in 90% of the runs.

4. RESULTS

Principal component analysis is used to visualize the entire data set more effectively, see figure 6. The plots illustrate that the analytes lying on their own almost linear manifold thus motivating the use of linear methods. The plots imply (but do not verify) that clustering methods such as K -nearest neighbor²¹ would not perform well given a small amount of training points. If all concentrations are not represented by a cluster a substantial amount of data points would likely be misclassified. Table 1 lists the classification accuracy for the tested classification algorithms. The performance of SVD classification is superior to both ANN and NMF. Using only 3 training

Table 1. Classification error on test sets calculated as the mean over a 100 runs \pm the standard deviation of the mean. Note N_{train} is the amount of training points per analyte – the total amount of training points is N_{train} multiplied with six.

Algorithm	N_{train}				
	3	4	5	6	12
ANN	0.611 ± 0.027	0.200 ± 0.024	0.076 ± 0.003	0.023 ± 0.002	0.003 ± 0.001
NMF	0.039 ± 0.004	0.028 ± 0.004	0.025 ± 0.003	0.017 ± 0.003	0.009 ± 0.002
SVD	0.032 ± 0.004	0.012 ± 0.002	0.006 ± 0.001	0.002 ± 0.001	0.001 ± 0.000

points per analyte we gain remarkable accuracy – only misclassifying 3% of the test points. ANN is quite poor initially but as the amount of training point increases ANN classification catches up. NMF performance is similar to SVD although not quite as good. Figure 7 shows learning curves for the three algorithms. The SVD classifier is the preferred method regardless of the amount of training points. Concentration level estimation results are

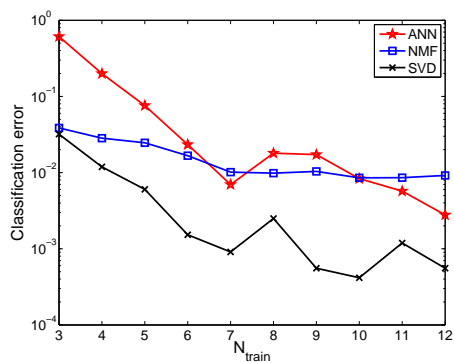


Figure 7. Learning curves for the classification algorithms on a test set. Note N_{train} refers to the amount of training points per analyte. The total amount of training points is $6N_{train}$.

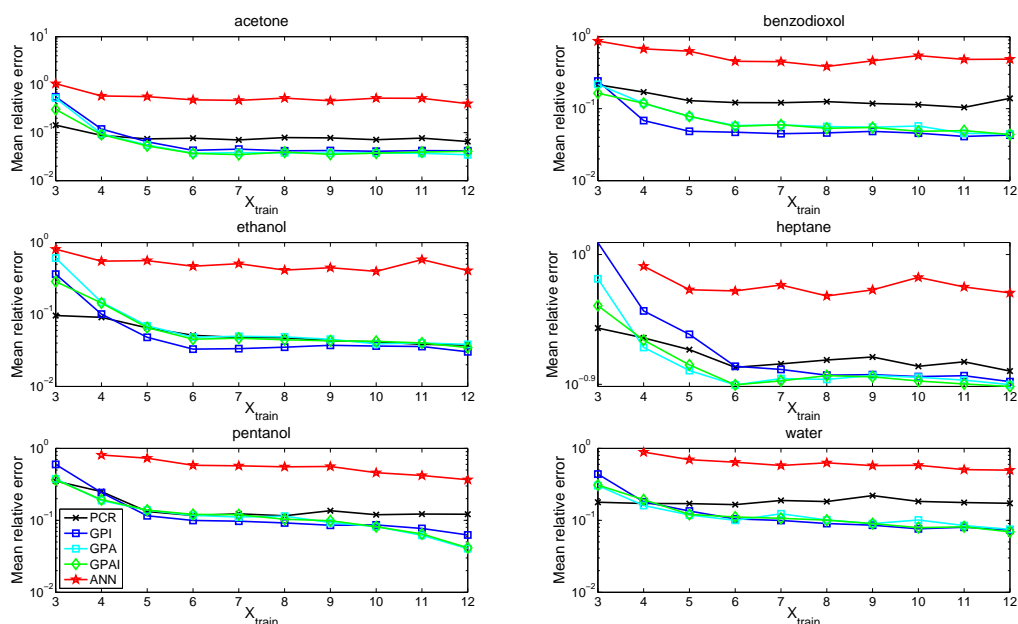


Figure 8. Learning curves for the concentration level estimation.

listed in table 2. There is no clear indication of which algorithm should be the preferred choice although it shows that ANN is consistently the poorest performer. As with classification there are too few training points to train an adequate neural network solution. The best regressor for $N_{train} = 3$ is PCR regardless of analyte except for Benzodioxol where the GPAI approach is the best. At $N_{train} = 4$ the GPR method performs similar to PCR and for $N_{train} \geq 5$ GPR is either the superior method or on par for all choices of analyte. At $N_{train} = 12$ the GPR method offers superior accuracy. There is no significant indication of which of the three covariance functions is the best although the combined covariance function (GPAI) seems slightly better. Figure 8 shows learning curves for the concentration level estimators. The learning curves clearly illustrates the troubles of ANN using a limited amount of training points. Another interesting observation is that PCR does not seem to improve significantly once N_{train} reaches 6.

Table 2. Estimation error on test sets calculated as the mean over a 100 runs \pm the standard deviation of the mean. The three instances where ANN is listed as *NaN* means that there was occurrences where the network failed to train due to an unlucky choice of training set. GP* refers to Gaussian process regression using the different covariance functions. GPI is the isotropic covariance function, GPA is the ARD covariance function and GPAI is the combined covariance function.

Analyte	Algorithm	N_{train}				
		3	4	5	6	12
Acetone	PCR	0.14 ± 0.009	0.09 ± 0.004	0.07 ± 0.003	0.08 ± 0.003	0.07 ± 0.003
	GPI	0.56 ± 0.094	0.12 ± 0.010	0.06 ± 0.004	0.04 ± 0.001	0.04 ± 0.001
	GPA	0.52 ± 0.088	0.09 ± 0.020	0.06 ± 0.006	0.04 ± 0.001	0.03 ± 0.002
	GPAI	0.30 ± 0.047	0.09 ± 0.008	0.05 ± 0.003	0.04 ± 0.001	0.04 ± 0.002
	ANN	1.05 ± 0.060	0.58 ± 0.021	0.56 ± 0.026	0.48 ± 0.015	0.40 ± 0.024
Benzodioxol	PCR	0.22 ± 0.008	0.17 ± 0.008	0.13 ± 0.005	0.12 ± 0.003	0.14 ± 0.011
	GPI	0.24 ± 0.024	0.07 ± 0.004	0.05 ± 0.001	0.05 ± 0.001	0.04 ± 0.001
	GPA	0.22 ± 0.027	0.12 ± 0.005	0.08 ± 0.003	0.06 ± 0.002	0.04 ± 0.002
	GPAI	0.16 ± 0.011	0.12 ± 0.006	0.08 ± 0.004	0.06 ± 0.003	0.04 ± 0.001
	ANN	0.87 ± 0.057	0.68 ± 0.043	0.63 ± 0.036	0.46 ± 0.018	0.49 ± 0.034
Ethanol	PCR	0.10 ± 0.005	0.09 ± 0.003	0.07 ± 0.002	0.05 ± 0.002	0.04 ± 0.001
	GPI	0.36 ± 0.053	0.10 ± 0.006	0.05 ± 0.002	0.03 ± 0.000	0.03 ± 0.001
	GPA	0.61 ± 0.108	0.15 ± 0.021	0.07 ± 0.004	0.05 ± 0.001	0.04 ± 0.001
	GPAI	0.29 ± 0.049	0.14 ± 0.019	0.07 ± 0.003	0.05 ± 0.001	0.03 ± 0.001
	ANN	0.81 ± 0.047	0.55 ± 0.015	0.56 ± 0.041	0.47 ± 0.021	0.41 ± 0.023
Heptane	PCR	0.31 ± 0.014	0.26 ± 0.009	0.22 ± 0.007	0.17 ± 0.004	0.16 ± 0.008
	GPI	1.21 ± 0.161	0.41 ± 0.034	0.28 ± 0.015	0.17 ± 0.004	0.13 ± 0.002
	GPA	0.68 ± 0.098	0.23 ± 0.012	0.16 ± 0.006	0.12 ± 0.002	0.13 ± 0.003
	GPAI	0.44 ± 0.059	0.26 ± 0.019	0.17 ± 0.008	0.12 ± 0.003	0.12 ± 0.003
	ANN	$NaN \pm NaN$	0.83 ± 0.043	0.57 ± 0.016	0.56 ± 0.022	0.54 ± 0.032
Pentanol	PCR	0.35 ± 0.007	0.25 ± 0.010	0.13 ± 0.004	0.12 ± 0.002	0.12 ± 0.003
	GPI	0.60 ± 0.058	0.24 ± 0.019	0.12 ± 0.004	0.10 ± 0.002	0.06 ± 0.002
	GPA	0.38 ± 0.026	0.19 ± 0.008	0.14 ± 0.005	0.12 ± 0.003	0.04 ± 0.005
	GPAI	0.37 ± 0.024	0.19 ± 0.008	0.14 ± 0.004	0.12 ± 0.003	0.04 ± 0.004
	ANN	$NaN \pm NaN$	0.81 ± 0.024	0.73 ± 0.026	0.58 ± 0.002	0.37 ± 0.003
Water	PCR	0.18 ± 0.002	0.17 ± 0.002	0.17 ± 0.003	0.17 ± 0.003	0.17 ± 0.007
	GPI	0.44 ± 0.038	0.18 ± 0.012	0.13 ± 0.007	0.10 ± 0.003	0.07 ± 0.003
	GPA	0.30 ± 0.037	0.16 ± 0.013	0.12 ± 0.009	0.10 ± 0.005	0.07 ± 0.003
	GPAI	0.31 ± 0.037	0.19 ± 0.018	0.12 ± 0.011	0.11 ± 0.010	0.07 ± 0.003
	ANN	$NaN \pm NaN$	0.89 ± 0.032	0.70 ± 0.021	0.64 ± 0.012	0.50 ± 0.026

5. CONCLUSION

Eight different polymer coatings have been evaluated for their suitability as gas sensors by exposing polymer coated QCMs to volatile organic compounds. The high classification accuracy implies that the coatings do indeed represent the range of interactions described by the LSER model very well. Each polymer film seems to contribute with unique data. According to our analysis the coatings are useful as gas sensing elements and can thus advantageously be used to functionalize many other types of sensors. Deposition of multiple types of sensor coatings by plasma polymerization is compatible with practically any sensing device, and circumvents the use of solvents and subsequent drying processes during deposition which can often damage fragile device components.

We presented a two-tiered data analysis framework that successfully classified responses from an eight-dimensional polymer coated QCM sensor. By using linear methods such as SVD and NMF we get remarkable high accuracy ($> 96\%$) with only 3 training points per analyte. Having one training point per analyte per concentration level SVD have an classification accuracy of 99.8%. NMF does not perform to the level of SVD, but is still better than ANN for a limited amount of training points. The NMF classification algorithm is fairly

simple and more sophisticated approaches could be developed. One could perform NMF on the entire data set instead of per analyte and then map the basis vectors to classes using a probabilistic approach.³² The model selection for NMF and SVD can be improved further as well. If there is more than one choice of b tied for the best model, b is assigned randomly to a value among the best choices. One could construct a cost function that takes the residual values into account and then select the value of b based on the cost function.

For concentration level estimation the Gaussian process regression is better than the linear method PCR (when $N_{train} \geq 5$). Both of these methods are markedly better than the ANN. The issue with ANN regression being that the system is underspecified. For the smallest possible neural network with two hidden units, the number of parameters in the model is 18 (16 weights in the first layer, 2 weights in the second layer), but the network only have up to 12 training points. One could possibly improve ANN performance by reducing the dimension of the data before applying the data to the neural network (for example using PCA) thus reducing the number of input neurons.

The best case scenario for GPR is a mean relative error of 3%. However the GPR method does have difficulties when having just three training points per analyte. This can be explained by the zero mean Gaussian prior and the way data partitioning is performed. For training sets where the lowest and/or highest concentrations levels are omitted the GPR will perform poorly. In general the GPR will return to zero when estimating points outside the training interval. The performance outside the training interval could probably be improved by adding a growing term to the prior or using a thin-plate spline kernel as covariance function.³³

6. ACKNOWLEDGMENTS

We acknowledge the support from the Danish Agency for Science and Technology's, Program Commission on Nanoscience Biotechnology and IT (NABIIT), grant: 2106-07-0031 and 2106-05-0039.

REFERENCES

- [1] Saraoglu, H. M. and Edin, B., "E-nose system for anesthetic dose level detection using artificial neural network," *JOURNAL OF MEDICAL SYSTEMS* **31**, 475–482 (DEC 2007).
- [2] Yinon, J., "Detection of explosives by electronic noses," *ANALYTICAL CHEMISTRY* **75**, 99A–105A (MAR 1 2003).
- [3] Mahony, J. O., Nolana, K., Smytha, M. R., and Mizaikoff, B., "Molecularly imprinted polymers-potential and challenges in analytical chemistry," *Analytica Chimica Acta* **534**(1), 31–39 (2005).
- [4] Szejtli, J., "Introduction and general overview of cyclodextrin chemistry," *Chem. Rev.* **98**(5), 1743–1753 (1998).
- [5] Khan, A. R., Forgo, P., Stine, K. J., and D'Souza, V. T., "Methods for selective modifications of cyclodextrins," *Chem. Rev.* **98**(5), 1977–1996 (1998).
- [6] SAUERBREY, G., "Verwendung von schwingquarzen zur wagung dunner schichten und zur mikrowagung," *ZEITSCHRIFT FUR PHYSIK* **155**(2), 206–222 (1959).
- [7] Kurosawa, S., Hirokawa, T., Kashimaa, K., Aizawaa, H., Hana, D.-S., Yoshimic, Y., Okadaa, Y., Yasea, K., Miyakeb, J., Yoshimotoe, M., and Hilbornf, J., "Detection of deposition rate of plasma-polymerized films by quartz crystal microbalance," *Thin Solid Films* **374**(2), 262–267 (2000).
- [8] Grate, J. and Abraham, M., "Solubility interactions and the design of chemically selective sorbent coatings for chemical sensors and arrays," *SENSORS AND ACTUATORS B-CHEMICAL* **3**(2), 85–111 (1991).
- [9] Lu, H.-H., Rao, Y. K., Wu, T.-Z., and Tzeng, Y.-M., "Direct characterization and quantification of volatile organic compounds by piezoelectric module chips sensor," *SENSORS AND ACTUATORS B-CHEMICAL* **137**, 741–746 (APR 2 2009).
- [10] Rosengren, A. M., Golker, K., Karlsson, J. G., and Nicholls, I. A., "Dielectric constants are not enough: Principal component analysis of the influence of solvent properties on molecularly imprinted polymer-ligand rebinding," *BIOSENSORS & BIOELECTRONICS* **25**, 553–557 (NOV 15 2009).
- [11] Si, P., Mortensen, J., Kornolov, A., Denborg, J., and Moller, P. J., "Polymer coated quartz crystal microbalance sensors for detection of volatile organic compounds in gas mixtures," *ANALYTICA CHIMICA ACTA* **597**, 223–230 (AUG 10 2007).

- [12] Sepcic, K., Josowicz, M., Janata, J., and Selby, T., "Diagnosis of used engine oil based on gas phase analysis," *ANALYST* **129**(11), 1070–1075 (2004).
- [13] Ying, Z., Jiang, Y., Du, X., Xie, G., Yu, J., and Tai, H., "Polymer coated sensor array based on quartz crystal microbalance for chemical agent analysis," *EUROPEAN POLYMER JOURNAL* **44**, 1157–1164 (APR 2008).
- [14] Gulbag, A., Temurtas, F., and Yusubov, I., "Quantitative discrimination of the binary gas mixtures using a combinational structure of the probabilistic and multilayer neural networks," *SENSORS AND ACTUATORS B-CHEMICAL* **131**, 196–204 (APR 14 2008).
- [15] Wang, Y.-L. and Shih, J.-S., "Multi-channel piezoelectric crystal gas sensor with principal component analysis for organic solvent pollutants from polymer plants," *JOURNAL OF THE CHINESE CHEMICAL SOCIETY* **53**, 1427–1437 (DEC 2006).
- [16] Saraoglu, H. M. and Kocan, M., "Determination of blood glucose level-based breath analysis by a quartz crystal microbalance sensor array," *IEEE SENSORS JOURNAL* **10**, 104–109 (JAN 2010).
- [17] Mumykmaz, B., Oezmen, A., Ebeoglu, M. A., and Tasaltin, C., "Predicting gas concentrations of ternary gas mixtures for a predefined 3d sample space," *SENSORS AND ACTUATORS B-CHEMICAL* **128**, 594–602 (JAN 15 2008).
- [18] Ozmen, A., Tekce, F., Ebeoglu, M., Tasaltin, C., and Ozturk, Z., "Finding the composition of gas mixtures by a phthalocyanine-coated qcm sensor array and an artificial neural network," *SENSORS AND ACTUATORS B-CHEMICAL* **115**, 450–454 (MAY 23 2006).
- [19] Lee, D. and Seung, H., "Learning the parts of objects by non-negative matrix factorization," *NATURE* **401**, 788–791 (OCT 21 1999).
- [20] Winther-Jensen, B. and Glejbl, K., "Method and apparatus for the excitation of a plasma." US Patent 6,628,084 B1 (September 2003).
- [21] Bishop, C. M., [*Pattern Recognition and Machine Learning*], Springer-Verlag New York, Inc., Secaucus, NJ, USA (2006).
- [22] Rasmussen, C. E. and Williams, C. K. I., [*Gaussian Processes for Machine Learning*], MIT Press (2006).
- [23] DTUInformatics, "Dtu:toolbox – <http://isp.imm.dtu.dk/toolbox>."
- [24] Lin, C.-J., "Projected gradient methods for non-negative matrix factorization," *Neural Computation* **19**, 2756–2779 (2007).
- [25] Rasmussen, C. E. and Williams, C. K. I., "Gpml code – <http://www.gaussianprocess.org/gpml>," (2006).
- [26] Neal, R. M., *BAYESIAN LEARNING FOR NEURAL NETWORKS*, PhD thesis (1995).
- [27] Nielsen, H., "Ucminf - an algorithm for unconstrained, nonlinear optimization," Tech. Rep. IMM-TEC-0019, IMM, Technical University of Denmark (2001).
- [28] Sigurdsson, S., Larsen, J., Hansen, L., Philipsen, P., and Wulf, H., "Outlier estimation and detection application to skin lesion classification," *Acoustics, Speech, and Signal Processing, 2002. Proceedings. (ICASSP '02). IEEE International Conference on* **1**, I-1049 – I-1052 vol.1 (2002).
- [29] Mackay, D. J. C., "A practical bayesian framework for backpropagation networks," *NEURAL COMPUTATION* **4**, 448–472 (MAY 1992).
- [30] Mackay, D. J. C., "The evidence framework applied to classification networks," *NEURAL COMPUTATION* **4**, 720–736 (SEP 1992).
- [31] Eldén, L., [*Matrix Methods in Data Mining and Pattern Recognition*], Society for Industrial and Applied Mathematics, Philadelphia, PA, USA (2007).
- [32] Kolenda, T., Hansen, L. K., Larsen, J., and Winther, O., "Independent component analysis for understanding multimedia content," in [*Proceedings of IEEE Workshop on Neural Networks for Signal Processing XII*], Boursard, H., Adali, T., Bengio, S., Larsen, J., and Douglas, S., eds., 757–766, IEEE Press, Piscataway, New Jersey (2002). Martigny, Valais, Switzerland, Sept. 4-6, 2002.
- [33] Wahba, G., [*Spline Models for Observational Data*], Series in Applied Math., Vol. 59, SIAM, Philadelphia (1990).