We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

15

Statistical Analysis of Chemical Sensor Data

Jeffrey C. Miecznikowski[1] and Kimberly F. Sellers[2]
[1]*SUNY University at Buffalo*
[2]*Georgetown University*
USA

1. Introduction

Chemical sensors measure and quantify substances via their associated chemical or physical response, thus providing data that can be analyzed to address a scientific question of interest (Eggins; 2002). Used in a variety of applications from monitoring to medicine, chemical sensors vary vastly by construction, style, format, size and dimension, and complexity. The common, underlying feature of these sensors lies in the associated data, which are abundant with technical and structural complexities, making statistical analysis a difficult task. These data further share a common need to be measured, analyzed, and interpreted properly so that the resulting inference is accurate.

There are many image analysis algorithms available and amenable to a myriad of chemical analysis problems, thus potentially applicable to chemical sensor data problems in particular. By applying these tools to chemical sensor data, we can optimize and evaluate a chemical sensor's ability to perform its intended tasks. This chapter is designed to give an overview of the modern statistical algorithms that are commonly used when designing and analyzing chemical sensor experiments. Without focusing the discussion around a specific chemical sensor platform, our goal is to provide a general framework that will be applicable, to some degree, for all chemical sensor data.

From the beginning to the end of an experiment, various statistical methods can be employed to improve or understand the current scientific analysis. We decompose a general experiment into several facets and provide the motivation for potential statistical methods that can be applied within each component. Section 2 describes the pre-processing techniques that are available for summarizing the low-level image or signal data so that the subsequent scientific questions can be properly addressed. In this section, we particularly focus on removing background noise in order to isolate the chemical sensor signal data, quantifying this data, and normalizing it so that the resulting data are scalable across conditions. Section 3 introduces the higher-level statistical approaches that are used to analyze the pre-processed data, and Section 4 describes the statistical computational tools available for use to perform the analyses. Finally, Section 5 demonstrates and motivates the significance of these methods via a chemical sensor case study, and Section 6 concludes the chapter with discussion and summary.

2. Pre-processing

For the analysis of chemical sensor data, many of the suggested means to resolve low-level analysis problems have been posed by computer scientists or engineers, with little statistical contribution or consideration, and they remain open problems because of significant

drawbacks in the proposed approaches. Alternatively, problems have been addressed in direct association with chemical sensor data analysis without recognizing that the resulting data represent a special case from a larger context of image data whose structure has been considered by statisticians. Scientists, for example, are interested in better tools that allow for a completely automated approach to detect chemical changes. They, therefore, recognize the need for minimal inherent noise in order to gain in data reproducibility and trust the obtained summary information for subsequent statistical analysis. Statistical and/or data mining tools, and machine learning algorithms are all methods that scientists can use to remove the noise from the meaningful signal contained within an experiment.

Similar to other high throughput biological experiments, several initial steps are often necessary before analyzing the data for a scientific question. Natale et al. (2006) discuss several preprocessing steps including feature extraction, zero-centered scaling, autoscaling, and normalization. Jurs et al. (2000) outline these techniques for chemical sensor arrays, and further include discussion on background or baseline subtraction, and linearization. Meanwhile, a good introduction to the general notion of pre-processing is provided in Gentleman (2005). Although Gentleman (2005) introduce these methods motivated by a different data source, many of their techniques are general enough for chemical sensor data. These issues are all substantial problems that need to be addressed, because any subsequent chemical sensor data analyses are contingent particularly on appropriate and statistically sound low-level procedures.

2.1 Background correction

Background noise associated with chemical sensors can occur for any number of reasons, such as the nature of the chemical processes and the machines used to scan or quantify the sensor. Irrespective of the cause, the background effect must be removed in order for the data of interest to be accurate and informative.

A variety of background correction techniques exist for various data forms, depending on the dimensionality and structure of the data. A general approach for background correction is to simply subtract the blank sample response, i.e. the response which is obtained before any sample is placed within the sensor (see Jurs et al. (2000), or Sellers et al. (2007) for an analogous approach). Other general approaches subtract either the global minimum from the data, or perform some local Winsorization at a low-percentile value. Such approaches are generally accepted for sensor data that appear in spectral form (e.g. Coombes et al. (2005); Lin et al. (2005); Morris et al. (2005)). While analogous approaches can likewise be applied to two-dimensional data, other methods have also been suggested, which include filtering in the wavelet domain (Coombes et al.; 2003), and asymmetric least squares splines regression (Befekadu et al.; 2008).

2.2 Sensor detection and quantification

In the event of fluorescent- or imaging-based sensor data, there are numerous computer algorithms and summary statistics available to identify the location and size of these features in the raw image data. Peak detection and quantification, for example, are of interest as they relate to spectral data. Various methods to achieve peak detection have been proposed via simple to complex means. One can recognize these methodological developments over time as further study was devoted to this area. Coombes et al. (2003) first suggested that peaks be detected by noting the locations where a change in slope occurs, and later fine-tuned the approach by instead considering the maximum value within the kth nearest neighbors; see Coombes et al. (2005), and independent discussion by Fushiki et al. (2006). More advanced

proposals are to either apply an undecimated discrete wavelet transform (Morris et al.; 2005), or a continuous wavelet transform (Du et al.; 2006). These methods better eliminate the risk of detecting false positive peaks (i.e. data believed to represent peaks from a true signal when the data are actually, say for example, representative of residual noise).

In a two-dimensional chemical sensor setting, there are several classes of algorithms that can be applied for spot detection and quantification. Determining the locations and boundaries for chemical sensors in two dimensions falls under the general research area of image segmentation. Within image segmentation there are four main approaches: threshold techniques, boundary-based techniques, region-based methods, and hybrid techniques that combine boundary and region criteria (Adams and Bischof; 1994). Threshold techniques are based on the theory that all pixels whose values lie within a certain range belong to one class. This method neglects spatial information within the image and, in general, does not work well with noisy or blurred images. Boundary-based methods are motivated by the postulate that pixel values change rapidly at the boundary between two regions. Such methods apply a gradient operator in order to determine rapid changes in intensity values. High values in a gradient image provide candidates for region boundaries which must then be modified to produce closed curves that delineate the spot boundaries. The conversion of edge pixel candidates to boundaries of the regions of interest is often a difficult task. The complement of the boundary-based approach is to work within the region of interest, e.g. the chemical sensor. Region-based methods work under the theory that neighboring pixels within the region have similar values. This leads to the class of algorithms known as "region growing", of which the "split and merge" techniques are popular. In this technique, the general procedure is to compare one pixel to its neighbor. If some criterion of homogeneity is satisfied, then that pixel is said to belong to the same class as one or more of its neighbors. As expected, the choice of the homogeneity criterion is critical for even moderate success and can be highly deceiving in the presence of noise.

Finally, the class of hybrid techniques that combine boundary and region criteria includes morphological watershed segmentation and variable-order surface fitting. The watershed method is generally applied to the gradient of the image. In this case, the gradient image can be viewed as a topography map with boundaries between the regions represented as "ridges". Segmentation is then equivalent to "flooding" the topography from local minima with region boundaries erected to keep water from different minima exclusive. Unlike the boundary-based methods above, the watershed is guaranteed to produce closed boundaries even if the transitions between regions are of variable strength or sharpness. Such hybrid techniques, like the watershed method, encounter difficulties with chemical sensor images in which regions are both noisy or have blurred or indistinct boundaries. A popular alternative is seeded region growing (SRG). This method is based on the similarity of pixels within regions but has an algorithm similar to the watershed method. SRG is controlled by choosing a small number of pixels or regions called "seeds". These seeds will control the location and formation of the regions in which the image will be segmented. The number of seeds determines what is a feature and what is irrelevant or noise-embedded. Once given the seeds, SRG divides the image into regions such that each connected region component intersects with exactly one of the seeds. The choice of the number of seeds is crucial to this algorithm's success. Fortunately, with many chemical sensor experiments, the number of chemical sensors, and thus the seed, is known beforehand; see Adams and Bischof (1994) for further details.

Feature quantification is also an important issue, as there are several options that aid in reducing data dimensionality and complexity. At the same time, one ideally wants to measure a feature in such a way that captures an optimal amount of sensor information. Pardo and

Sberveglieri (2007) compare the performance of five feature summaries in chemical sensor arrays: the relative change in resistance; the area under the curve over gas adsorption, and gas desorption; and the phase space integral over adsorption, and desorption. In their study, while they do not attain uniform results across the various datasets, they find (on average) that the phase integral over desorption performed best. Further, the integral and phase space integral over desorption performed better than the analogous computations associated with adsorption. These results are consistent with other applied fields where such feature quantification is performed by computing the associated area under the curve. Carmel et al. (2003) instead argue that focusing on such features (such as the difference between the peak and baseline, the area under the curve, the area under curve left of the peak, or the time from the beginning to the peak of a signal) does not fully capture certain sensor properties, thus limiting one's ability to perform analyses. Focusing on transient signals, the authors fit various parametric models to chemical sensors for electronic noses, namely exponential, Lorentzian, and double-sigmoid models. Their results show that the double-sigmoid models fit optimally, followed by the Lorentzian model, with the exponential model being the worst of the three but still with decent performance. The computational time needed to fit these models, however, showed that the Lorentzian and exponential models were estimated far more quickly than for the double-sigmoid model. This makes sense because the double-sigmoid model has nine parameters that require estimation, while the Lorentzian and exponential models only have two parameters. Given this tradeoff, the authors propose using the Lorentzian model to analyze such chemical sensor data.

2.3 Normalization

In a normalization step, the goal is to remove obscuring sources of variation to give accurate measurements of the desired signal. Normalization could proceed in a manner similar to that described in Sellers et al. (2007) to remove known possible sources of variation, where one can obtain associated response data based on the presence of these factors in the design.

Linearization can also be performed by considering the engineering-derived equations that drive the signal (see, e.g., Robins et al. (2005)). Some chemical sensors are ruled by a power law relationship between sensor signal and analyte concentration; this is often the case, for example, with metal-oxide semiconductor gas sensors (Natale et al.; 2006). Using least squares approaches, it is possible to estimate the parameters in a power law relationship. This is a popular approach when preprocessing chemical sensor data because many of the subsequent analyses (e.g. linear discriminant analysis, principal component analysis, principal component regression, partial least squares) assume a linear relationship between sensor response and sample class (Jurs et al.; 2000).

Relative scaling is a common practice in the normalization of chemical sensor arrays, however the approach by which it is performed may vary. Options include dividing the signal by either the maximum signal value, the Euclidean norm from the signal, or the maximum value from a reference signal. In any respect, relative scaling serves to normalize the data in order to be on the same scale.

2.3.1 Quantile normalization

Quantile normalization is a very popular normalization method, because of its generality; it does not require building (non)-linear models to describe the experimental system. Let each experimental unit (e.g. subject, patient, or sample) be measured via the proposed chemical sensor(s) which produce(s) a profile for this experimental unit, and assume that our chemical sensor is, in fact, a panel of many chemical sensors. The quantile normalization thus imposes

the same empirical distribution of the chemical sensor intensity of each profile (e.g. the profile for each experimental unit will have the same quartiles, etc). The algorithm proposed in Bolstad et al. (2003) is designed so that all profiles are matched (aligned) with the empirical distribution of the averaged sample profiles.

2.4 Low-level analysis discussion

Any or all low-level analysis procedures can be performed to obtain summary information on the raw chemical sensor data. The order of operations for these algorithms, however, are inconsistent and generally unrecoverable. As a result, the resulting preprocessed chemical sensor data can vary, thus potentially causing severe repercussions in the high-level analysis (see Baggerly et al. (2004)). To this end, one should be mindful of the low-level analyses performed (along with their order of operations) and comfortable with their use in data preprocessing. Nonetheless, data preprocessing results in the $S \times I$ summary matrix, $\mathbf{X} = (x_{si})$, where x_{si} denotes the normalized measure of sensor s in sample i. This data matrix will be used for subsequent statistical analysis.

3. Data analysis

There are several approaches that can be pursued to analyze the preprocessed data, depending on the question of interest. This section introduces these high-level, downstream methods. Here, we assume that the resulting preprocessed data matrix has rows associated with the chemical sensors used for the analysis, while the columns refer to the samples or patients. Jurs et al. (2000) classifies several methods as either statistical (including linear discriminant analysis (LDA), and principal component analysis (PCA)), or using neural networks while cluster analysis tools are classified separately. Given the popularity of LDA and PCA, we focus on these statistical methods here; see Jurs et al. (2000) for added discussion regarding various alternatives.

Linear discriminant analysis (LDA) is a statistical method (credited to Fisher) for dimension reduction and potential classification in that it distinguishes between two or more groups. The discriminant functions are derived from means and covariance matrices, thus working to maximize the distance between groups (relative to the variance within respective groups). While LDA is a popular dimension reduction technique for its natural approach, it tends to overfit when the ratio of training samples to dimensionality is small; see Wang et al. (2004). Jurs et al. (2000) concur that one needs a "relatively large number of samples from each class in the training data" that is representative of the population.

Principal component analysis (PCA) is an alternative statistical approach for dimension reduction. Invented by Karl Pearson, PCA performs singular value decomposition on the data matrix, $\mathbf{X}$, where the resulting terms relate to the eigenvalue-eigenvector form of $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}\mathbf{X}'$, respectively. PCA is a popular choice in chemical sensor analysis because the first two principal components often account for at least 80% of the chemical sensor data variance (Jurs et al.; 2000), and is more robust to overfitting than LDA (Wang et al.; 2004). This method, however, may not successfully classify groups. Low-variance sensors, or nonlinear or nonadditive sensors can make classification difficult (Jurs et al.; 2000; Wang et al.; 2004).

Recognizing the limitations of these statistical methods, Wang et al. (2004) propose a "hybrid" model, termed Principal Discriminant Analysis (PDA). The hybrid matrix,

$$H = (1 - \epsilon)S_w^{-1}S_b + \epsilon S_T,$$

	Reject H_0	Fail to reject H_0
H_0 true	Type I error	Correct decision
H_0 false	Correct decision	Type II error

Table 1. Possible outcomes for a null hypothesis, H_0, and associated outcome (rejecting or failing to reject H_0).

can be interpreted as a weighted average of the within- and between-group matrices from the LDA solution, and the total data covariance associated with PCA. The optimal $\epsilon \in [0,1]$ is attained via cross-validation, where $\epsilon = 0$ attains the LDA eigenvalues, and $\epsilon = 1$ produces the PCA projection. As a result, PDA provides a compromise between the popular LDA and PCA.

Similar to Jurs et al. (2000), we explore supervised and unsupervised machine learning/ statistical techniques to understand large complex datasets. Specifically, we examine linear modeling techniques to determine significantly different chemical sensors between two or more populations (e.g. neural nets as in Hashem et al. (1995)[1]). We also explore classification (supervised) and clustering (unsupervised) techniques to explore the similarities and differences between samples or between sensors. Within classification methods, we explore methods to both build and validate (e.g. data splitting/ cross validation) the classification schemes. Ultimately, we use the concepts of sensitivity and specificity to choose among a class of classification schemes.

3.1 Multiple testing

While LDA, PCA, and PDA all work with the entire dataset, commonly researchers would like to identify a subset of chemical sensors that are associated with the outcome. In this sense, the researchers are performing a data reduction, where the goal is to choose a subset of chemical sensors that are related or associated with the outcome. In this setting, the researcher commonly uses hypothesis testing to choose the important subset of chemical sensors.

Hypothesis testing seeks to obtain statistically significant results regarding a question of interest. In this process, the null hypothesis (H_0) represents the status quo statement while the alternative hypothesis (usually denoted as H_1 or H_a) defines that which is to be potentially proven or determined. When performing a hypothesis test, one wants to make a correct decision. There are, however, four possible scenarios that can occur when performing such a test; see Table 1. Two scenarios represent correct decisions, while the other two are incorrect decisions or "errors": (1) when one rejects the null hypothesis when it is actually true, and (2) when one does not reject the null hypothesis when it is actually false. The probability associated with the first scenario is referred to as Type I error (denoted α), and the second scenario's probability is termed Type II error (denoted β). For completeness in this discussion, statistical power refers to the probability of rejecting the null hypothesis when (in fact) the null hypothesis is false. In other words, statistical power equals one minus the Type II error (i.e. $1 - \beta$). Even when performing one hypothesis test, one wants to minimize the error probabilities.

We assume that our chemical sensor is multivariate, in the sense that each experimental unit (i.e., subject, patient, sample, or animal) is measured with several chemical sensors. Thus each unit acquires a chemical sensor profile, that is a collection of signals acquired from the chemical sensor. The goal in hypothesis testing is to examine each chemical sensor in

[1] In the Appendix, we discuss neural networks as outlined in Hashem et al. (1995) as a form of regression models.

		Condition +	Condition -	
Test	+	TP	FP	PPV
	-	FN	TN	NPV
		Sensitivity	Specificity	

Table 2. Table displaying the summary measures to distinguish positives (cases) from negatives (controls). TP (FP) denotes the number of true (false) positives, while TN (FN) denotes the number of true (false) negatives. PPV and NPV denote positive and negative predictive value, respectively. See Section 3.2.2 for details.

	H_0 Retained	H_0 Rejected	Total
H_0 True	U	V	m_0
H_0 False	T	Q	$m - m_0$
	$m - R$	R	m

Table 3. A summary of results from analyzing multiple hypothesis tests, where each cell represents the number (counts) in each category with m total tests.

light of the multiple chemical sensor levels measured in each experimental unit. In this setting, researchers and statisticians usually design their tests such that rejecting H_0 will yield discoveries or chemical sensors of interest. For example, when testing case samples against control samples on a chemical sensor platform, we would like to configure our hypothesis tests such that we reject H_0 for chemical sensors that are distinct between the cases and controls; see Section 5.

Commonly these hypothesis tests are performed using (linear) regression models. In these regressions, we estimate parameters designed to measure the effects of a chemical sensor in relation to an outcome. Common outcomes might be survival times or group membership (case vs. control). Using estimates of these parameters for a given chemical sensor, we can determine its significance. The interested reader is referred to Rawlings et al. (1998) and Cohen (2003) for comprehensive discussion of linear models and associated hypothesis testing.

In light of this discussion for multiple sensors, we can define our Type I and Type II errors in terms of sensitivity and specificity. That is, sensitivity is defined as

$$\text{sensitivity} = \frac{\text{number of true positives}}{\text{number of true positives + number of false negatives}}, \tag{1}$$

while specificity is defined as

$$\text{specificity} = \frac{\text{number of true negatives}}{\text{number of true negatives + number of false positives}}. \tag{2}$$

This situation is also summarized in Table 2. Note that sensitivity and specificity are estimated values because, in any experiment, we do not know the number of true positives or true negatives. Our goal when performing multiple testing and classification is to maximize the sensitivity and specificity, thus limiting the number of errors committed.

Table 2 can be refined in light of performing multiple hypothesis tests. In Table 3, we generalize the hypothesis testing in light of performing m hypothesis tests. Note that in Table 3, U, V, T, Q denote random variables (unknowns), while we assume that m is a fixed unknown quantity of hypothesis tests.

In light of performing multiple hypothesis tests (one for each sensor), we need a method to control the Type I error across the multiple tests. A first attempt for controlling Type I error is to perform a Bonferroni correction where, given m hypothesis tests, the measure for statistical significance is now attained if the associated p-value is less than α/m. In other words, the significance level is now scaled by the number of hypothesis tests. While this approach successfully adjusts for multiple tests, the procedure is far too stringent! The following subsections detail alternative Type I error rates and the available methods to control those errors, where Table 3 provides the associated notation in terms of probabilities (Pr): the familywise error rate, FWER $= Pr(V \geq 1)$; the k-familywise error rate, k-FWER $= Pr(V \geq k)$; and the false discovery rate (FDR), which is $E(V/R)$ if $R > 0$, or otherwise 0 with $E()$ denoting the expected value function.

3.1.1 Familywise error rates

The k-FWER error rate is a generalized version of the familywise error rate (FWER). Control of FWER refers to controlling the probability of committing one or more false discoveries. If we let V denote the number of false positives from m hypothesis tests, then notationally (according to Lehmann and Romano (2005)), α control of FWER can be expressed as

$$Pr(V \geq 1) \leq \alpha, \tag{3}$$

or equivalently,

$$Pr(V = 0) \geq 1 - \alpha. \tag{4}$$

Note that α is usually chosen to be small, e.g., 0.05. Often Equation (3) is abbreviated as FWER $\leq \alpha$. In k-FWER, the equation becomes

$$Pr(V \geq k) \leq \alpha, \tag{5}$$

where $k \geq 1$ and α are usually determined prior to the analysis. Similar to FWER, control of k-FWER can be expressed as k-FWER $\leq \alpha$. Note that there is the potential for ambiguity in control of k-FWER since, occasionally (as in Gentleman (2005)), k-FWER may be expressed as $Pr(V > k) \leq \alpha$ for $k \geq 0$.

The adjusted Bonferroni method to control k-FWER is a generalized version of the Bonferroni correction designed to control FWER (Lehmann and Romano; 2005). The Bonferroni correction is designed to control the FWER at level α by doing each individual test at level α/m, where m is the number of tests. The adjustment given in Lehmann and Romano (2005) to control k-FWER at α is done by performing each test at level $k\alpha/m$. By performing each test at this level, the probability against k or more false positives is no larger than α; that is, k-FWER $\leq \alpha$. The proof is supplied in Lehmann and Romano (2005) and is a generalization of the proof for the original Bonferroni method designed to control FWER. For a description of other methods to control k-FWER and a power comparison of k-FWER methods, see Miecznikowski et al. (2011).

3.1.2 False discovery rate

Multiple statistical testing procedures began to be reexamined in the early 1990s with the advent of high-throughput genomic technologies. The Benjamini and Hochberg (BH) method was proposed to control the false discovery rate (FDR), or the expected rate of false test positives (see Benjamini and Hochberg (1995)). In the BH multiple testing procedure, the FDR is controlled by the following scheme:

1. Let $p_{(1)} < \cdots < p_{(m)}$ denote the m ordered p-values (smallest to largest).
2. Denote $\hat{t} = p_{(k)}$ for the largest k such that $p_{(k)} \leq \frac{k\alpha}{m}$.
3. Reject all null hypotheses, H_{0i}, for which $p_i \leq \hat{t}$.

Note that we define FDR such that

$$FDR \equiv E[V/R], \tag{6}$$

where E denotes the expected value function. Benjamini and Hochberg (1995) proves that, if the above procedure is applied, $FDR \leq \alpha$. Storey (2002) further show that, for p-value threshold t,

$$\text{FDR}(t) = \frac{(1-\pi)t}{(1-\pi)t + \pi F(t)}, \tag{7}$$

where π is the probability that an alternative hypothesis is true, and $F(t)$ is the distribution of p-values given the alternative. FDR performance has been evaluated for sensor detection given a variety of scenarios, for example, in the presence of correlation (Benjamini and Yekutieli; 2001; Shao and Tseng; 2007). Importantly, note that FDR analysis does not control what Genovese and Wasserman (2004) call "the realized FDR" (rFDR), i.e. the number of false rejections V divided by the number of rejections R (assuming at least one rejection) which, in fact, can be quite variable (as shown in Gold et al. (2009)).

3.2 Classification

This section provides an overview of classification models. Throughout this section, we assume that the outcome variable of interest is binary. This is commonly the situation with case/control experiments where, for example, the goal may be to predict the presence or absence of a disease.

The simplest and most direct approach to classification with a binary outcome variable is to estimate the regression function, $r(x) = E(Y|X = x)$, and use the classification rule,

$$\hat{h}(x) = \begin{cases} 1 & \text{if } \hat{r}(x) > 0.5 \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

Here, the simplest regression model is the linear regression model,

$$Y = r(x) + \epsilon = \beta_0 + \sum_{j=1}^{d} \beta_j X_j + \epsilon, \tag{9}$$

where the errors, ϵ, have mean 0. A simple example of this classifier is provided in Section 5. Other examples of classifiers include linear discriminant analysis, support vector machines, and ensemble classifiers using bootstrapping and bagging techniques; see Hastie et al. (2005) and Wasserman (2004) for a more complete treatment of classification models. Similar to hypothesis testing, we want an accurate classifier that commits relatively few errors; i.e. we would like a classifier with a high sensitivity and specificity (see Equations (1) and (2)). To estimate sensitivity and specificity for a given classification model, we commonly use cross validation methods, as described in the next section.

3.2.1 Cross validation

Cross validation can be classified under the general realm of sample splitting. Its objective is to obtain an estimate of the prediction qualities of the model when using that same data to build the prediction model. The simplest version of cross-validation involves randomly splitting the data into two pieces: the training set, and the validation set. The classifier is constructed from the training set, and the associated error is estimated using the validation set; the error is defined as

$$\text{Error} = \text{number of misclassifications} / \text{number of predictions}. \tag{10}$$

Two extensions of this method are g-fold cross validation, and leave-one-out cross validation (LOOCV). Note that LOOCV is a special case of g-fold cross validation, where g is equal to the number of objects in the dataset. As described in Wasserman (2004) in a g-fold cross validation, we do the following:

1. Randomly divide the data into G groups of approximately equal size.
2. For $g = 1$ to G,
   (a) delete group g from the data.
   (b) fit or compute the classifier from the remaining data.
   (c) use the classifier to predict the data in group g, and let L denote that observed error rate.
3. Let the overall error rate be estimated from averaging over the error rates from the previous step.

See Section 3.2.2 for a discussion of other summary measures (e.g. sensitivity and specificity) that are commonly estimated with cross validation.

3.2.2 Summary measures in a population

Naturally, we want our classifiers to make accurate predictions. We have seen that specificity and sensitivity as estimated via cross validation are reasonable measures to summarize our classification models. In this section, however, we highlight some of the measures used to evaluate our classification models in a population. The measures introduced in this section are often crucial in deciding the utility of a chemical sensor.

When determining a classifier's effectiveness, analysts usually calculate the sensitivity and specificity using cross-validation. To understand the potential utility of the chemical sensor-derived classifier, however, it is important to calculate the positive predictive value (PPV) and negative predictive value (NPV). The PPV is the proportion of subjects with positive test results who are correctly diagnosed. It reflects the probability that a positive test reflects the truly positive underlying condition. The PPV depends heavily on the prevalence of the outcome of interest, which is usually unknown. Using Bayes Theorem (see Wasserman (2004)) and Table 2, we can derive the positive predictive value as

$$PPV = \frac{(\text{sensitivity})(\text{prevalence})}{(\text{sensitivity})(\text{prevalence}) + (1 - \text{specificity})(1 - \text{prevalence})}. \tag{11}$$

Note that we define the prevalence in terms of epidemiologic factors, i.e. prevalence. Prevalence (of disease) is the total number of (disease) cases in the population divided by the number of individuals in the population. Prevalence is (essentially) an estimate of how common the underlying condition is within a population over a certain period of time.

Defining a as the number of individuals in a given population with the disease at a given time, and b as the number of individuals in the same population at risk of developing the disease at this given time (not including those already with the disease), the prevalence is specified by

$$\text{prevalence} = \frac{a}{a+b}. \tag{12}$$

Similarly, we can define the NPV as the proportion of subjects with a negative test result who are correctly diagnosed. A high NPV means that, when the test yields a negative result, it is uncommon that the result should have been positive. Mathematically, the NPV is computed as

$$NPV = \frac{(\text{specificity})(1\text{-prevalence})}{(\text{specificity})(1\text{-prevalence}) + (1-\text{sensitivity})(\text{prevalence})}. \tag{13}$$

While sensitivity and specificity play a role in assessing a chemical sensor, we stress that NPV and PPV are often the measures used when deciding the clinical and medical utility of a potential chemical sensor panel. A thorough handling of the topic of estimation with regard to sensitivity, specificity, PPV, and NPV is provided in Pepe (2004), Cai et al. (2006), and Pepe et al. (2008).

4. Software development

`Bioconductor` (Gentleman et al.; 2004) and *R* (R Development Core Team; 2008) are two statistical computing tools that can be used for statistical programming and data analysis. Both are freeware tools that are downloadable from the internet. Researchers developing novel chemical sensors should consider writing a computational package for analyzing their chemical sensor data in *R*. This will enable the statistical methods and algorithms to be used by the scientific community. For various examples of *R* packages, see Gaile et al. (2010) and Gentleman (2005).

For example, Chandrasekhar et al. (2009) authored a software package specific to Xerogel chemical sensor images such as those shown in Figure 1 (A). These images and the Xerogel technology are further described in Chandrasekhar et al. (2009). The *Xerogel R* package consists of routines that import the tagged image file format (TIFF) image, read the image into a matrix, binarize the image matrix, identify the position and structure of the spots, and return statistics such as the mean, median, and total intensity for these spots. The summary statistics represent the results after preprocessing and, ultimately, provide information on how the intensity of light varies with varying amounts of the volatile organic compounds (VOCs). A summary set of images demonstrating the pre-processing of a representative Xerogel image is shown in Figure 1 (A)-(D).

5. Case study

In this case study, we examine a subset of the data from Schröder et al. (2010) which represents a chemical sensor dataset designed to classify pancreatic cancer patients from normal patients. This dataset is publicly available and can be downloaded from `ArrayExpress` (see Parkinson et al. (2009)) with ID accession number E-MEXP-3006; see `http://www.ebi.ac.uk/arrayexpress/`. This dataset is representative of a two dimensional chemical sensor dataset. The experiment employs protein antibody microarrays with two color channels on an array consisting of 1800 features (proteins).

The preprocessing for this data is consistent with the methods described in Section 2. The data in this study were preprocessed using the following scheme:

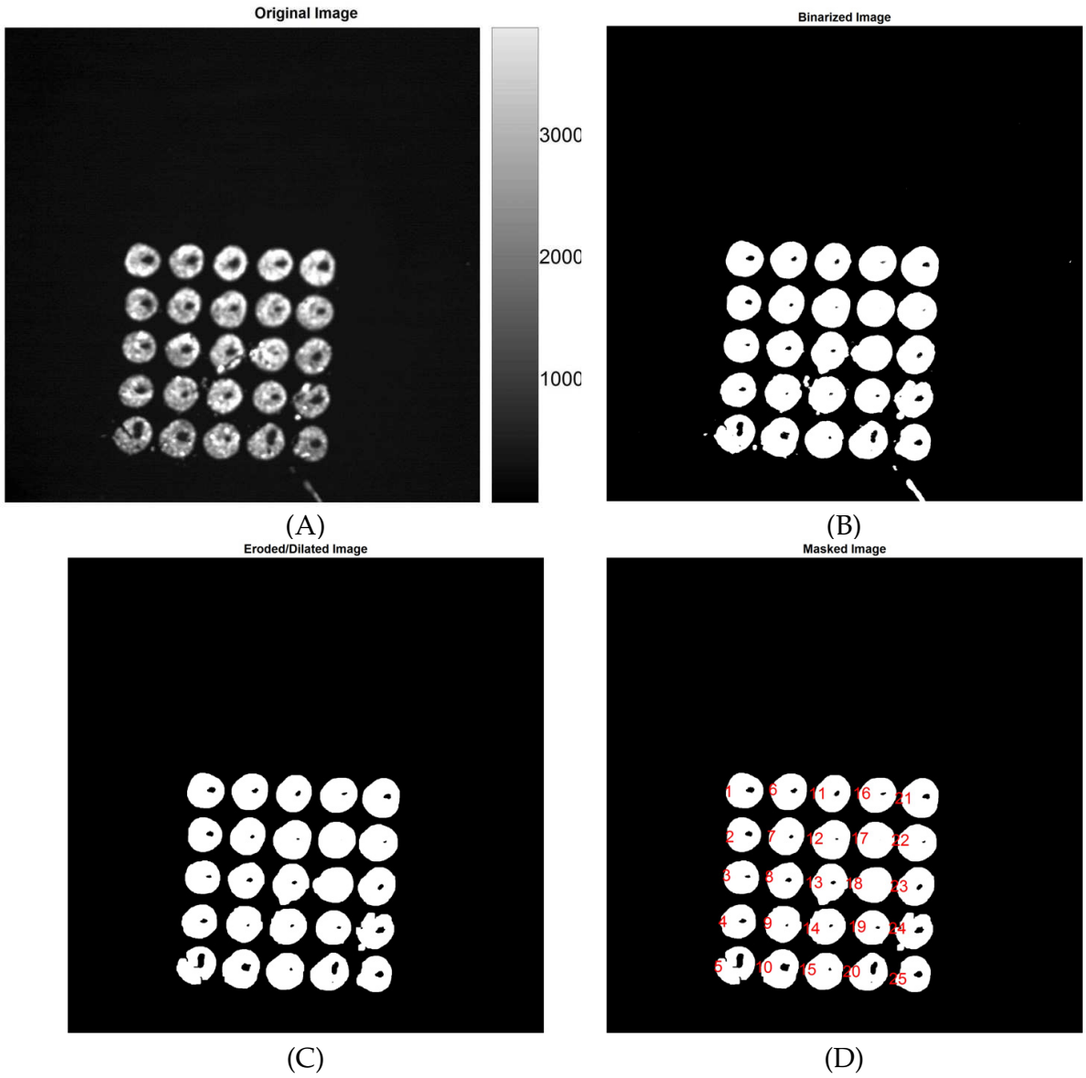


Fig. 1. **Xerogel Preprocessing:** Xerogel Images showing the preprocessing steps. (A) Original Image (B) Binarized Image (C) Eroded/Dilated Image (D) Masked Image. For more details on Xerogel chemical sensors see Chandrasekhar et al. (2009).

- Background Correction - as recommended in Ritchie et al. (2007), a convolution of normal and exponential distributions is fitted to the foreground intensities using the background intensities as a covariate, and the expected signal given the observed foreground becomes the corrected intensity.
- Normalization - `lowess` is applied as proposed in Yang et al. (2002) and Smyth and Speed (2003). Here, the signal in each array is adjusted to account for the intensity bias with a nonlinear curve fitting method, lowess.

All preprocessing steps were performed using the *limma* package in *R* (see Smyth (2004)). The experimental design for this experiment is fully described in Schröder et al. (2010). For our subset of data, we have a total of 12 subjects, specifically three patients in each of four groups (male controls, female controls, males with pancreatic cancer, and females with pancreatic

cancer). The patients were named in terms of their disease status (healthy = h, cancer = c) and gender (male = m, female = f); e.g. hf_1 refers to the sample for the first healthy female, while cm_1 refers to the sample for the first pancreatic cancer male. Midstream urine samples were collected from each patient and pH was adjusted to 7. After sample preparation, the samples were dye-labeled and incubated to antibody microarrays containing 1,800 features. Figure 2 shows a representative fluorescence array from this study where a urine sample and reference consisting of a pool of samples from diseased and healthy subjects were labeled with different dyes.

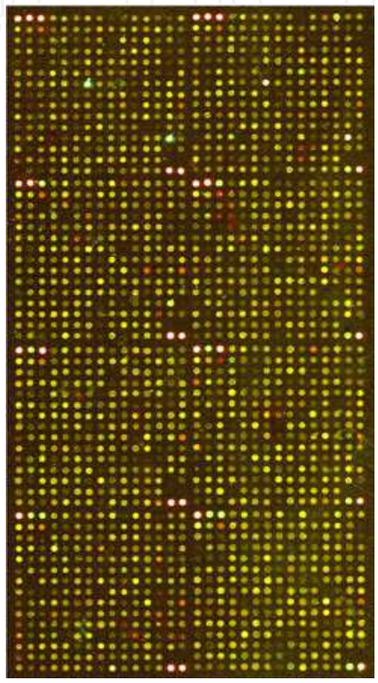


Fig. 2. Protein antibody image from our case study (figure taken from Schröder et al. (2010)).

After preprocessing, we arrive at a matrix containing 1678 rows (proteins) and 12 columns (samples), where each entry represents the logarithmic intensity (amount) of the protein present in the sample (relative to the reference channel). The heatmap representing this data is shown in Figure 3.

In performing an exploratory data analysis, we looked at clustering the samples as well as the distance between the samples. The clustering results in terms of a hierarchical clustering are shown in Figure 4. From this figure, we see that sample hf_2 may represent an outlier in this study; this is further confirmed in Figure 5. From this figure, we see that hf_2 is widely separated from the other samples in the study (white band). Note that the Euclidean distance metric was used to calculate the distance between each sample. For an overview of distance metrics, see Chapter 12 of Gentleman (2005). This potential outlier, hf_2, is also confirmed by studying the protein profile in Figure 3, where we see a pattern that is inconsistent with the other samples.

To further the analysis in this case study, we build linear model(s) to discover potential differentially expressed proteins between cancer patients and normal patients; see Section 3.1. We explored univariate protein models and multivariate protein models that adjusted

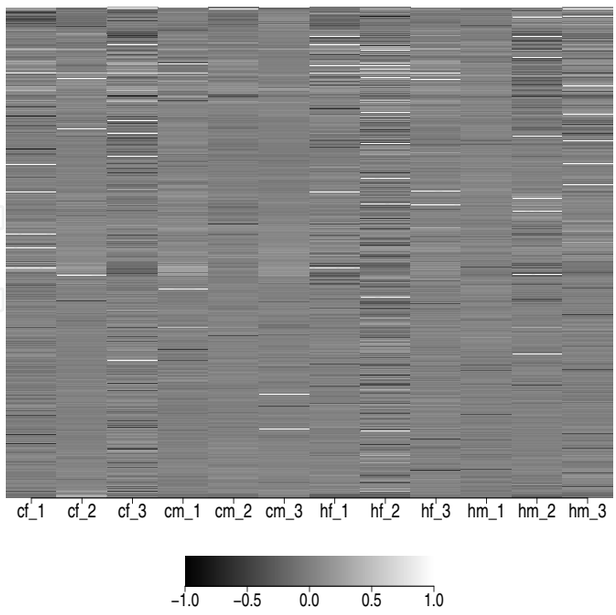


Fig. 3. Heatmap showing the proteins levels for each sample after preprocessing. Note sample hf_2 appears to be an outlier in this subset of data.

for gender. Building these models allows us to test each protein for association with disease status. In particular, we let

$$y_{ji} = \mu_j + \beta_{dz} x_{dz} + \epsilon_{ji} \tag{14}$$

denote the observed protein level for protein j in sample i ($i = 1, \ldots, 12$), with x_{dz} equaling 1 if sample j is a cancer sample and 0 otherwise and ϵ_{ji} is assumed to be normally distributed with mean 0 and variance σ_j^2. In this model, we are interested in estimating the parameters for each protein. For protein j, the pancreatic cancer samples have a mean of $\mu_j + \beta_{dz}$ while the healthy samples have a mean of μ_j. We are especially interested in proteins where β_{dz} is significantly different from 0, as this indicates proteins that are significantly different between diseased patients and healthy patients. Accordingly, for each protein, we wish to test the hypothesis,

$$H_0 : \beta_{dz} = 0. \tag{15}$$

Using an empirical Bayes test described in Smyth (2004), we obtain a test statistic and p-value for each protein corresponding to the test in Equation (15). Using a Šidàk control method described in Miecznikowski et al. (2011), we control k-FWER such that the probability of committing no more than five false positives is no larger than 0.05. Under this scheme, we discover three significant proteins; see Table 4.

In Equation (16), we introduce a more complex model. We include the explanatory variable x_{gen}, which is 1 if sample j is a female and 0 otherwise. By incorporating a variable for the patient's gender, this model is more complex than the model described in Equation (14). This model is described as

$$y_{ji} = \mu_j + \beta_{gen} x_{gen} + \beta_{dz} x_{dz} + \epsilon_{ji}. \tag{16}$$

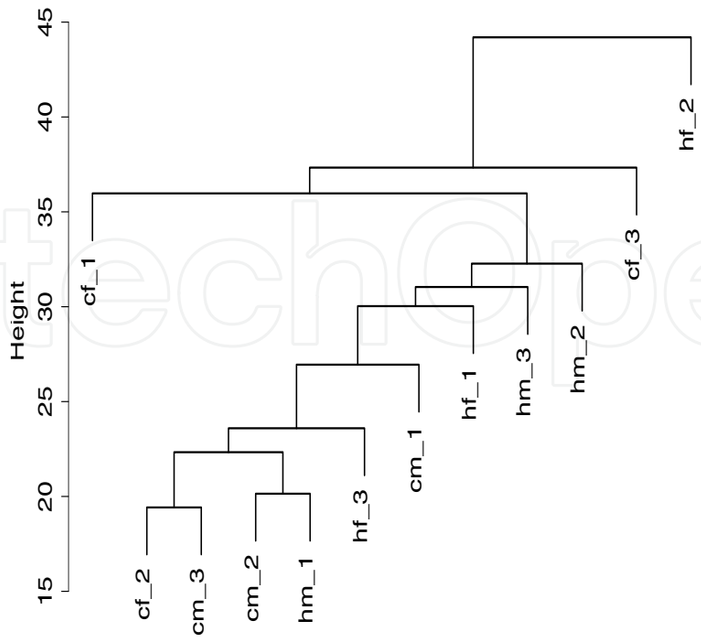


Fig. 4. Hierarchical clustering showing the similarity of the samples. From this figure it appears that sample hf_2 is an outlier.

	k-FWER Control via Šidàk Method		
	k	α	# Sig Proteins
Univariate Model	5	0.05	2
	10	0.05	3
Multivariate Model	5	0.05	1
	10	0.05	4

Table 4. Table displaying the significant proteins from an analysis with univariate and multivariate models as described in Equations (14) and (16), respectively.

After estimating the parameters in Equation (16), we are also interested in testing the null hypothesis in Equation (15). With this model, however, the parameter β_{dz} represents proteins that are significantly different in disease states (case/control) after adjusting for potential gender biases. For protein j, the female pancreatic cancer patients have estimates, $\mu_j + \beta_{gen} + \beta_{dz}$; while the female healthy patients have estimates, $\mu_j + \beta_{gen}$. Similarly, the male pancreatic cancer patients have estimates $\mu_j + \beta_{dz}$, while the male healthy patients have estimates, μ_j.

Using the model described in Equation (16) for each protein, we obtain the test statistic and p-value for β_{dz} from an empirical Bayes test (Smyth; 2004). As with the model in Equation (14), we use a Šidàk method to control k-FWER such that the probability of committing 10 or more false positives to be no larger than 0.05. Under this scheme, we obtain four significant proteins. Figure 6 displays the heatmap of the significant proteins under this setting and Table 4 displays the number of significant proteins under different configurations for controlling k-FWER.

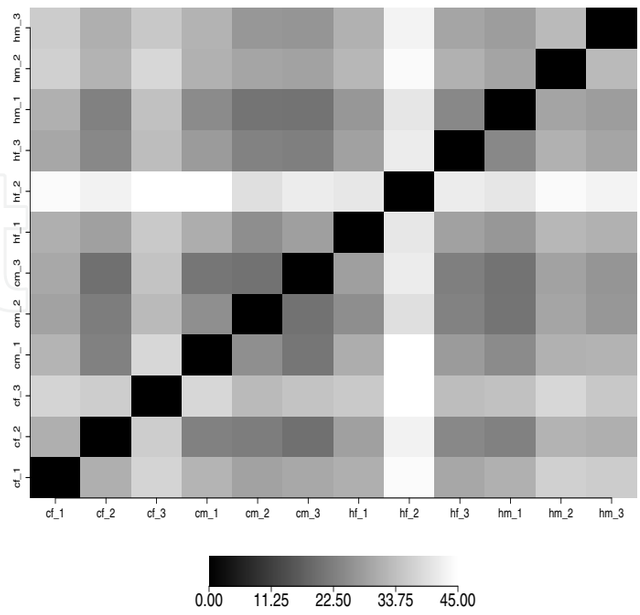


Fig. 5. The distance matrix heatmap showing the measured Euclidean distance between the samples using the protein profile for each sample. From this figure, the profile for sample hf_2 is the furthest in terms of Euclidean distance from the other samples.

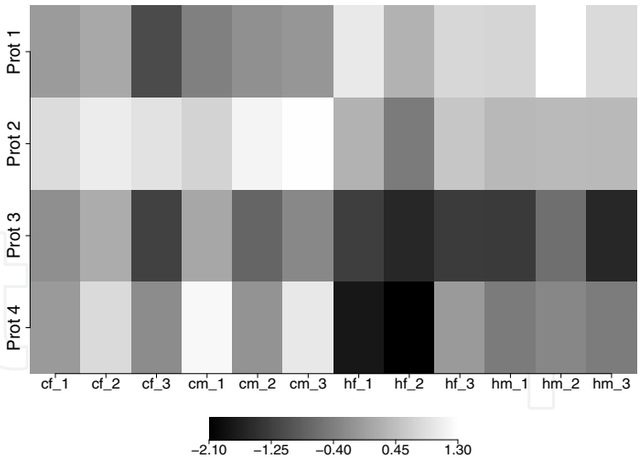


Fig. 6. Heatmap of the significant proteins as determined in case study using Equation (16) and a k-FWER error controlling scheme.

After determining the significant proteins in this study, it is reasonable to examine models for prediction. We explored using a logistic regression model as described in Section 3.2. We choose a classification model using the most significant protein ($Prot1$) as determined from

fitting the model in Equation (14). The fitted logistic regression equation is

$$\hat{r}(Prot1) = \frac{exp(-10.303 * Prot1)}{1 + exp(-10.303 * Prot1)}, \tag{17}$$

where $Prot1$ is the intensity of the most significant protein after fitting Equation (14); the profile for this protein is shown in Row 1 in Figure 6. Our logistic classifier is then specified by

$$\hat{h}(x) = \begin{cases} \text{Disease} & \text{if } \hat{r}(x) > 0.5 \\ \text{Healthy} & \text{if } \hat{r}(x) \leq 0.5. \end{cases} \tag{18}$$

Using a leave-one-out cross-validation method as described in Section 3.2.1, we obtain a sensitivity estimate of 83.3% (5/6) and a specificity estimate of 100% (6/6); the misclassified sample is cf_2. As seen in Row 1 of Figure 6, cf_2 has a value for $Prot1$ indicating a pattern more aligned with the healthy samples. Similar to Schröder et al. (2010), our conclusion from this analysis is that a urine proteomic profile as measured on antibody arrays shows promise in diagnosing pancreatic cancer. Due to the limited sample size (12 samples) of our case study, however, we caution the reader not to rely heavily on these estimates of sensitivity and specificity.

6. Summary

Chemical sensor data appear in a variety of contexts, from breathalyzers to carbon monoxide or smoke detectors. Given their pervasive existence in various aspects of life, it is essential that these sensors work and provide proper analysis to accurately assess chemical questions of interest. This can only happen with proper statistical insight and tools to provide accurate assessments that can lead to appropriate decision-making. Hirschfeld et al. (1984) noted the importance of chemometrics with sensor arrays, particularly the use of pattern recognition strategies and learning algorithms. As a result, calibration, quantification, and reproducibility are all attainable. In this chapter, we highlight some of the state-of-the-art statistical methods that are used to calibrate, quantify, and ultimately, benchmark modern chemical sensors. We stress that further advancements in the use and utility of chemical sensors will require input from statisticians to determine the accuracy/reproducibility of the sensors as well as their ability to make inferences on a population.

7. Appendix

In this section, we provide a brief overview of neural networks as they are commonly used for prediction with chemical sensor data (see Hashem et al. (1995)). As discussed in Wasserman (2004), neural networks often take the form,

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j \sigma \left(\alpha_0 + \alpha^T X \right) \tag{19}$$

where σ is a smooth function. When compared to models such as in Equations (14) and (16), the neural networks model is obviously more complex. As such, these models are often difficult to fit to datasets and often require large datasets and heavy computational power. Besides Wasserman (2004), other references for neural networks can be found in Bhadeshia (1999); MacKay (2003).

8. References

Adams, R. and Bischof, L. (1994). Seeded region growing, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 16(6): 641–647.

Baggerly, K., Morris, J. and Coombes, K. (2004). Reproducibility of seldi-tof protein patterns in serum: comparing datasets from different experiments, *Bioinformatics* 20(5): 777–785.

Befekadu, G., Tadesse, M., Hathout, Y. and Ressom, H. (2008). Multi-class alignment of lc-ms data using probabilistic-based mixture regression models, *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, IEEE, pp. 4094–4097.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency, *The Annals of Statistics* 29(4): 1165–1188.

Bhadeshia, H. (1999). Neural networks in materials science, *ISIJ international* 39(10): 966–979.

Bolstad, B., Irizarry, R., Åstrand, M. and Speed, T. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias, *Bioinformatics* 19(2): 185.

Cai, T., Pepe, M., Zheng, Y., Lumley, T. and Jenny, N. (2006). The sensitivity and specificity of markers for event times, *Biostatistics* 7(2): 182.

Carmel, L., Levy, S., Lancet, D. and Harel, D. (2003). A feature extraction method for chemical sensors in electronic noses, *Sensors and Actuators B: Chemical* 93(1-3): 67 – 76. Proceedings of the Ninth International Meeting on Chemical Sensors.
URL: *http://www.sciencedirect.com/science/article/pii/S0925400503002478*

Chandrasekhar, R., Miecznikowski, J., Gaile, D., Govindaraju, V., Bright, F. and Sellers, K. (2009). Xerogel package, *Chemometrics and Intelligent Laboratory Systems* 96(1): 70–74.

Cohen, J. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*, Vol. 1, Lawrence Erlbaum.

Coombes, K., Fritsche Jr, H., Clarke, C., Chen, J., Baggerly, K., Morris, J., Xiao, L., Hung, M. and Kuerer, H. (2003). Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionization, *Clinical Chemistry* 49(10): 1615.

Coombes, K., Tsavachidis, S., Morris, J., Baggerly, K., Hung, M. and Kuerer, H. (2005). Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform, *Proteomics* 5: 4107–4117.

Du, P., Kibbe, W. and Lin, S. (2006). Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching, *Bioinformatics* 22(17): 2059.

Eggins, B. (2002). *Chemical sensors and biosensors*, Wiley.

Fushiki, T., Fujisawa, H. and Eguchi, S. (2006). Identification of biomarkers from mass spectrometry data using a "common" peak approach, *BMC Bioinformatics* 7: 358.

Gaile, D., Shepherd, L., Bruno, A., Liu, S., Morrison, C., Sucheston, L. and Miecznikowski, J. (2010). iGenomicViewer: R package for visualisation of high dimension genomic data, *International Journal of Bioinformatics Research and Applications* 6(6): 584–593.

Genovese, C. and Wasserman, L. (2004). A stochastic process approach to false discovery control, *The Annals of Statistics* 32(3): 1035–1061.

Gentleman, R. (2005). *Bioinformatics and computational biology solutions using R and Bioconductor*, Springer Verlag.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H. and Zhang, J. (2004). Bioconductor: Open software development for computational biology and bioinformatics, *Genome Biology* 5: R80. URL: *http://genomebiology.com/2004/5/10/R80*

Gold, D., Miecznikowski, J. and Liu, S. (2009). Error control variability in pathway-based microarray analysis, *Bioinformatics* 25(17): 2216–2221.

Hashem, S., Keller, P., Kouzes, R. and Kangas, L. (1995). Neural-network-based data analysis for chemical sensor arrays, *Proceedings of SPIE*, Vol. 2492, p. 33.

Hastie, T., Tibshirani, R., Friedman, J. and Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction, *The Mathematical Intelligencer* 27(2): 83–85.

Hirschfeld, T., Callis, J. B. and Kowalski, B. R. (1984). Chemical sensing in process analysis, *Science* 226(4672): 312–318. URL: *http://www.sciencemag.org/content/226/4672/312.abstract*

Jurs, P. C., Bakken, G. A. and McClelland, H. E. (2000). Computational methods for the analysis of chemical sensor array data from volatile analytes, *Chemical Reviews* 100(7): 2649–2678. URL: *http://pubs.acs.org/doi/abs/10.1021/cr9800964*

Lehmann, E. and Romano, J. (2005). Generalizations of the familywise error rate, *The Annals of Statistics* 33(3): 1138–1154.

Lin, S., Haney, R., Campa, M., Fitzgerald, M. and Patz, E. (2005). Characterising phase variations in maldi-tof data and correcting them by peak alignment, *Cancer Informatics* 1(1): 32–40.

MacKay, D. (2003). *Information theory, inference, and learning algorithms*, Cambridge Univ Pr.

Miecznikowski, J., Gold, D., Shepherd, L. and Liu, S. (2011). Deriving and comparing the distribution for the number of false positives in single step methods to control k-fwer, *Statistics & Probability Letters* (in press).

Morris, J., Coombes, K., Koomen, J., Baggerly, K. and Kobayashi, R. (2005). Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum, *Bioinformatics* 21(9): 1764.

Natale, C., Martinelli, E., Pennazza, G., Orsini, A. and Santonico, M. (2006). Data Analysis for Chemical Sensor Arrays, *Advances in Sensing with Security Applications* pp. 147–169.

Pardo, M. and Sberveglieri, G. (2007). Comparing the performance of different features in sensor arrays, *Sensors and Actuators B: Chemical* 123(1): 437 – 443. URL: *http://www.sciencedirect.com/science/article/pii/S0925400506006411*

Parkinson, H., Kapushesky, M., Kolesnikov, N., Rustici, G., Shojatalab, M., Abeygunawardena, N., Berube, H., Dylag, M., Emam, I., Farne, A. et al. (2009). Arrayexpress update-from an archive of functional genomics experiments to the atlas of gene expression, *Nucleic acids research* 37(suppl 1): D868.

Pepe, M. (2004). *The statistical evaluation of medical tests for classification and prediction*, Oxford University Press, USA.

Pepe, M., Feng, Z., Huang, Y., Longton, G., Prentice, R., Thompson, I. and Zheng, Y. (2008). Integrating the predictiveness of a marker with its performance as a classifier, *American journal of epidemiology* 167(3): 362.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. URL: *http://www.R-project.org*

Rawlings, J., Pantula, S., Dickey, D. and MyiLibrary (1998). *Applied regression analysis: a research tool*, Springer New York, NY, US.

Ritchie, M., Silver, J., Oshlack, A., Holmes, M., Diyagama, D., Holloway, A. and Smyth, G. (2007). A comparison of background correction methods for two-colour microarrays, *Bioinformatics* 23(20): 2700.

Robins, P., Rapley, V. and Thomas, P. (2005). A probabilistic chemical sensor model for data fusion, *Information Fusion, 2005 8th International Conference on*, Vol. 2, IEEE, pp. 7–pp.

Schröder, C., Jacob, A., Tonack, S., Radon, T., Sill, M., Zucknick, M., Rüffer, S., Costello, E., Neoptolemos, J., Crnogorac-Jurcevic, T. et al. (2010). Dual-color proteomic profiling of complex samples with a microarray of 810 cancer-related antibodies, *Molecular & Cellular Proteomics* 9(6): 1271.

Sellers, K., Miecznikowski, J., Viswanathan, S., Minden, J. and Eddy, W. (2007). Lights, Camera, Action! Systematic variation in 2-D difference gel electrophoresis images, *Electrophoresis* 28(18): 3324–3332.

Shao, Y. and Tseng, C. (2007). Sample size calculation with dependence adjustment for fdr-control in microarray studies, *Statistics in medicine* 26(23): 4219–4237.

Smyth, G. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments, *Statistical applications in genetics and molecular biology* 3(1): 3.

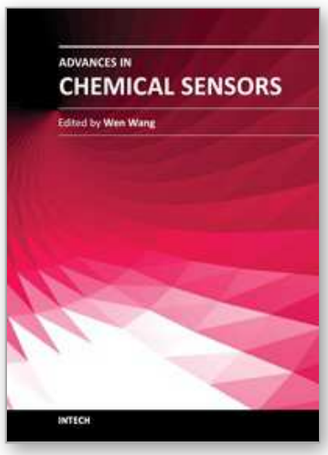
Smyth, G. and Speed, T. (2003). Normalization of cdna microarray data, *Methods* 31(4): 265–273.

Storey, J. (2002). A direct approach to false discovery rates, *Journal of the Royal Statistical Society. Series B, Statistical Methodology* pp. 479–498.

Wang, M., Perera, A. and Gutierrez-Osuna, R. (2004). Principal discriminants analysis for small-sample-size problems: application to chemical sensing, *Sensors, 2004. Proceedings of IEEE*, IEEE, pp. 591–594.

Wasserman, L. (2004). *All of statistics: a concise course in statistical inference*, Springer Verlag.

Yang, Y., Dudoit, S., Luu, P., Lin, D., Peng, V., Ngai, J. and Speed, T. (2002). Normalization for cdna microarray data: a robust composite method addressing single and multiple slide systematic variation, *Nucleic acids research* 30(4): e15.



Advances in Chemical Sensors

Edited by Prof. Wen Wang

ISBN 978-953-307-792-5
Hard cover, 358 pages
Publisher InTech
Published online 20, January, 2012
Published in print edition January, 2012

The chemical sensor plays an essential role in the fields of environmental conservation and monitoring, disaster and disease prevention, and industrial analysis. A typical chemical sensor is a device that transforms chemical information in a selective and reversible way, ranging from the concentration of a specific sample component to total composition analysis, into an analytically useful signal. Much research work has been performed to achieve a chemical sensor with such excellent qualities as quick response, low cost, small size, superior sensitivity, good reversibility and selectivity, and excellent detection limit. This book introduces the latest advances on chemical sensors. It consists of 15 chapters composed by the researchers active in the field of chemical sensors, and is divided into 5 sections according to the classification following the principles of signal transducer. This collection of up-to-date information and the latest research progress on chemical sensor will provide valuable references and learning materials for all those working in the field of chemical sensors.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Jeffrey C. Miecznikowski and Kimberly F. Sellers (2012). Statistical Analysis of Chemical Sensor Data, Advances in Chemical Sensors, Prof. Wen Wang (Ed.), ISBN: 978-953-307-792-5, InTech, Available from: http://www.intechopen.com/books/advances-in-chemical-sensors/statistical-analysis-of-chemical-sensor-data

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen