Appendix — Principal Components Analysis

The data initially taken from the experimental setup in this thesis are a long stream of resistances read from each sensor over time. From this data, for each exposure to an analyte from each sensor, we extract a single descriptor:

$$\frac{\Delta R_{\text{max}}}{R_{\text{b}}} \tag{5.1}$$

where R_{b} is the baseline resistance of the sensor prior to exposure, and ΔR_{max} is the maximum change in steady-state resistance (Figure 5.1). This $\Delta R_{\text{max}}/R_{\text{b}}$ value is the partial differential resistance response of one sensor to one exposure of an analyte.

This yields $\mathbf{R} = \{r_{ij}\}$, an $m \times n$ matrix of sensor values, where n is the number of sensors, m is the number of exposures, and r_{ij} represents the response of the jth sensor to the ith exposure of analyte, as shown in Equation 5.1. This leaves the problem of having data in n-space, which is difficult to interpret and visualize. Principal components analysis (PCA) is a multivariate statistical technique employed to reduce the dimensionality of the data, and make it more amenable to interpretation.[1] This is a common method used in pattern analysis, and has been extensively used and reviewed in the sensor array literature.[2] This is the primary method for analyte discrimination used in this thesis.

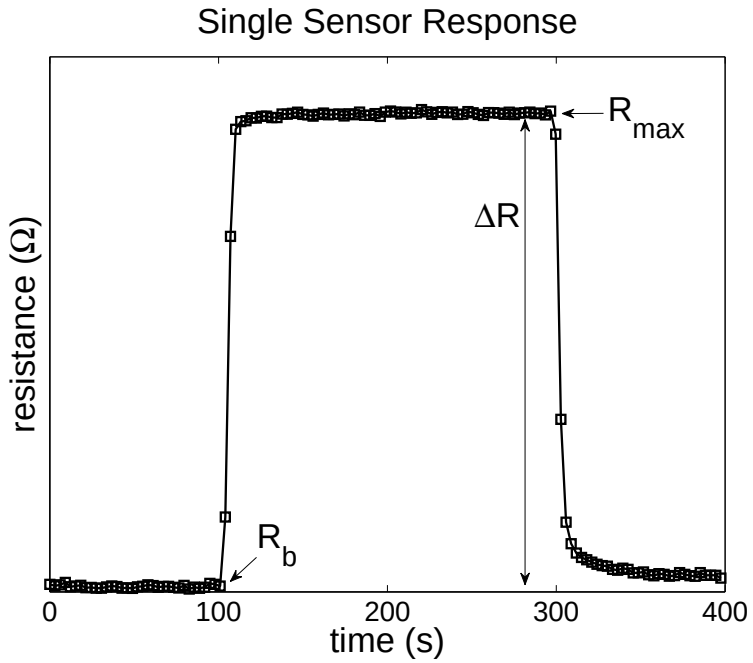

Figure 5.1: Response of a poly(ethylene oxide)/carbon black composite sensor to a 200 second exposure to 2 ppth of chloroform vapor, at an overall flow rate of 2.5 L min^{-1}.

The matrix $\mathbf{R}$ is first preprocessed such that each column in the matrix is normalized and autoscaled (i.e., centered about the mean and defined to have unit standard deviation, resulting in a final matrix $\mathbf{D} = \{d_{ij}\}$. First the r_{ij} values are normalized, creating the matrix $\mathbf{Q} = \{q_{ij}\}$ which helps correct for differences in solvent vapor pressure.

$$q_{ij} = \frac{r_{ij}}{\sum_j r_{ij}} \tag{5.2}$$

These normalized values are then autoscaled, such that they are both mean-centered and set to have a standard deviation of unity.

$$d_{ij} = \frac{q_{ij} - \overline{q}_j}{\sigma_j} \tag{5.3}$$

Here, $\overline{q}_j$ and σ_j represent the mean and standard deviation of each sensor j to all analytes presented to it. This matrix $\mathbf{D} = \{d_{ij}\}$ is then diagonalized (i.e., multiplied by its transpose) to obtain a correlation matrix $\mathbf{M}$.

$$\mathbf{M} = \mathbf{D^T} \cdot \mathbf{D} \tag{5.4}$$

The eigenvalues and eigenvector matrix of $\mathbf{M}$ are then obtained. The n eigenvectors of the eigenvector matrix $\mathbf{V}$ are mutually orthogonal. We multiply this $n \times n$ matrix $\mathbf{V}$ by the data matrix $\mathbf{D}$ to obtain our matrix of principal components, $\mathbf{P}$, an $m \times n$ matrix, in which each row is still associated with a particular analyte exposure, and each column is now a principal component of the data, in which the maximal variance between the members of the original data set is found in the first principal component, the maximal remaining variance found in the second component, and so on. The corresponding eigenvalues of $\mathbf{M}$ tell us how much of the total variance is to be found in each principal component.

$$\mathbf{P} = \mathbf{DV} \tag{5.5}$$

The maximal amount of variance is now front loaded into the first few principal components, allowing us to much more easily visualize the information in the data in just two or three dimensions, rather than the full n-dimensionality of the original sensor set.

5.1 Bibliography

[1] Duda, R.; Hart, P.; Stork, D. *Pattern Classification;* Wiley: New York, 2nd ed., 2001.

[2] Jurs, P.; Bakken, G.; McClelland, H. *Chem. Rev.* **2000**, *100*(7), 2649–2678.