X.1 Principal component analysis

• Dimension reduction methods seek to distill information-rich high-dimension data to a few key dimensions that carry the majority of the desired information.
• In PCA, the algorithm is guided by identifying a few directions that contain the majority of the overall variance within a data set, since these dimensions likely carry the greatest information content within the signal.
• Note – PCA implicitly assumes that the S/N > 1, such that the variations in the measured data due to changes in the signals are greater than variations due to noise.
Limitations of PCA

- Often, we wish to discriminate between classes using known databases of information (e.g., molecular identification based on compiled Raman or MS data, determination of the origin of complex samples, chemical fingerprinting, etc.).
- The outcome of the measurement is often: i) assignment of an unknown into a class, ii) and assessment of the confidence of that assignment.
- PCA is “unsupervised”, in that the algorithm does not utilize class information in dimension reduction. The entire data set, including both samples of known and unknown origin, are treated equally in the analysis.
- The directions of maximum variance in the data do not necessarily correspond to the directions of maximum class discrimination.
The 5 key steps in PCA

1. Arrange the data as a matrix $D$, with each “spectrum” as a column in $D$.

2. Zero-center the data row by row by subtraction of the mean for each row to generate $D_{zc}$.

3. Calculate the variance/covariance matrix, $\alpha$, of the data by $\alpha = D_{zc} \cdot D_{zc}^T$.

4. Determine the eigenvectors and eigenvalues of $\alpha$.

5. Select the key eigenvectors based on the eigenvalues, which describe the variance of the data projected along that eigenvector. Transform the data to the principal coordinate system.
• Consider the following experiment, in which a sample set of solutions contain mixtures of both A and B. The goal of the experiment is to separate the samples into A-rich groups and B-rich groups. However, neither the identity nor the spectra for A and B are known.

• Although we cannot obtain them directly from the experiments (without some work), we will assume that the noise-free pure component spectra for A and B are shown below.
Example: two-component mixtures

• Each spectrum with M data points generated from combinations of A and B together with noise represents a point described by the position vector $\mathbf{x}$ in an M-dimensional space. The complete set of spectra represent a set of data points in this high-dimensional surface.
• To illustrate the dimension problem, let’s consider just two wavelengths from the spectrum and consider the trends in absorbance values from a data set containing varying concentrations of A and B together with measurement noise.
• Both A and B contribute to absorbance at both $\lambda_1$ and $\lambda_2$. 

![Pure component spectra of "A" and "B"](image1)

![Graph](image2)
Example: two-component mixtures

- In this two-dimensional case considering just these two wavelengths, the principal axis containing the greatest spread in signal and the greatest ability to separate between the two data sets is not parallel to either wavelength axis.
- The axis containing the greatest spread in signal is a combination of the two axis, shown by the green line.
- The blue line is the only other dimension perpendicular to the green axis, and contains mostly noise.

\[
\begin{bmatrix}
(D^T)_{600}^T
\end{bmatrix}
\begin{bmatrix}
(D^T)_{300}^T
\end{bmatrix}
\]
Example: two-component mixtures

To find these PCA axes mathematically, we follow the steps in slide 2:

- subtract the mean from all data points, D, where each column of D is a “spectrum”.
- calculate $\alpha$.
- find the eigenvectors and eigenvalues of $\alpha$.
- use the eigenvalues to pick the important eigenvectors.
- project each data vector on to the appropriate eigenvectors.

$$DT := D^T$$

$$DT zc^{(j)} := DT^{(j)} - \text{mean}(DT^{(j)})$$

$$Dzc := DT zc^T$$

$$D2zc := \text{augment}(DT zc^{(600)}, DT zc^{(300)})^T$$

$$\alpha2 := D2zc D2zc^T$$

$$\text{EV2} := \text{eigenvects}(\alpha2)$$

$$\text{PC2} := \text{EV2} D2zc$$

Pool the total data to subtract the mean spectrum.

Just select the 2-element spectra to be used in this simple example.

Each column of EV2 corresponds to a two-element eigenvector with the data variance as the eigenvalue.
\[ \alpha = D_{zc} \cdot D_{zc}^T \]

Explicitly evaluate \( D_{zc}D_{zc}^T \) for wavelengths \( a \) and \( b \) (2D) with \( N \) measurements.

Variance about the mean for \( D_a \) and \( D_b \).

\[
\alpha_{aa} = \sum_{i=0}^{N-1} [(D_a)_i - \mu_a][D_a)_i - \mu_a]
\]

\[
\alpha_{bb} = \sum_{i=0}^{N-1} [(D_b)_i - \mu_b]^2
\]

\[
\alpha_{ab} = \sum_{i=0}^{N-1} [(D_a)_i - \mu_a][D_b)_i - \mu_b]
\]

\[
\alpha_{ba} = \alpha_{ab}
\]

Thus, \( \alpha \) is dubbed the variance/covariance matrix.
Example: two-component mixtures

• If we consider PC0 to be unimportant (containing mostly noise) with the majority of the information content contained within PC1, we can reduce the dimensionality of the problem to 1.

• Hopefully, this correlates to the amount of “A” and “B” in each sample

• The same can be done using all wavelength channels of the data instead of just 2, which will extract a wider variance with less relative noise.
Example: two-component mixtures

To find these PCA axes mathematically, we follow the steps in slide 2:

• subtract the mean from all data points, D, where each column of D is a “spectrum”.
• calculate $\alpha$,
• find the eigenvectors and eigenvalues of $\alpha$,
• use the eigenvalues to pick the important eigenvectors,
• project each data vector on to the appropriate eigenvectors.

\[
\begin{align*}
D^{T} & := D^{T} \\
D^{T} z^{(j)} & := D^{T} z^{(j)} - \text{mean}\left(D^{T} z^{(j)}\right) \\
Dzc & := D^{T} z^{C}^{T} \\
\alpha & := Dzc Dzc^{T} \\
evals & := \text{eigenvals}\left(\alpha\right) \quad \text{evecs} := \text{eigenvcs}\left(\alpha\right) \\
\text{Dim} & := 3 \\
\text{transform} & := \begin{cases} 
\text{for } n \in 0..\text{rows(evecs)} - 1 \\
\text{for } m \in 0..\text{Dim} - 1 \\
\text{out}_{n,m} \leftarrow \text{evecs}_{n,\text{cols(evecs)} - 1 - m} \\
\text{out} \end{cases} \\
\text{PC} & := \text{transform}^{T} \cdot D
\end{align*}
\]