

Regression analysis of spectral data taken from a multicomponent solution. Find composition from spectral data. A combination of mole fraction and total concentration completely specifies the composition.

Source of data: fluorescence spectra for binary NADH-tyrosine solutions taken in my laboratory.

Instructor: Nam Sun Wang

Read data into a matrix:

ORIGIN=1

```
data := READPRN(nadh_dat)    n := rows(data)    n = 33    cols(data) = 32
```

Extract from the given data independent variable x and dependent variable y.

```
y := submatrix(data, 1, n, 1, 2) ... Columns 1-2 contain the mole fraction component #1 (NADH) and
total concentration (dependent variable). We perform regression on
each one of the dependent variables separately or simultaneously.
y_1 := data<1>
y_2 := data<2>
We take the former approach in this worksheet.
```

```
x := submatrix(data, 1, n, 3, 32)... Columns 3-32 contain 30 spectral intensities (independent variables x)
```

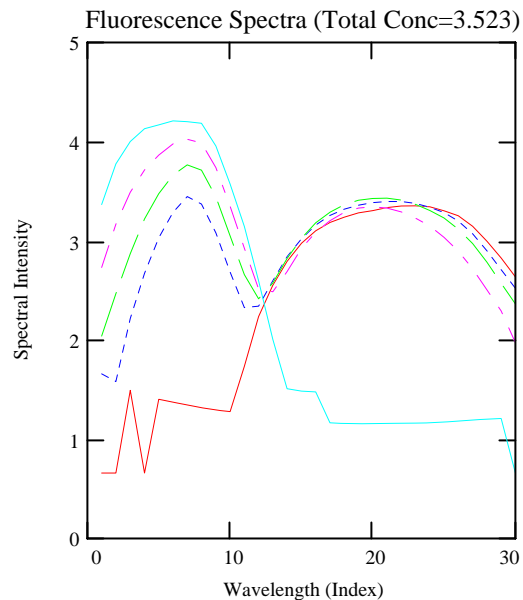
```
n = 33    i := 1 .. n    ... total number of observations = n = 33
```

```
m := 30    j := 1 .. m    ... total number of independent variables
```

The number of spectral intensities can easily be in the thousands; thus, the number of independent variables, m, may possibly be greater than the number of data points, n. This poses a problem in applying the following normal equation to find the regression coefficients: $a=(X^T \cdot X)^{-1} \cdot X^T \cdot Y$. Actually, even when $m < n$, we still face the co-linearity problem, as the supposedly independent variables are not mutually independent, but highly correlated. Thus, we need to resort to eigenvalue-eigenvector. First, let us examine the given spectral data.



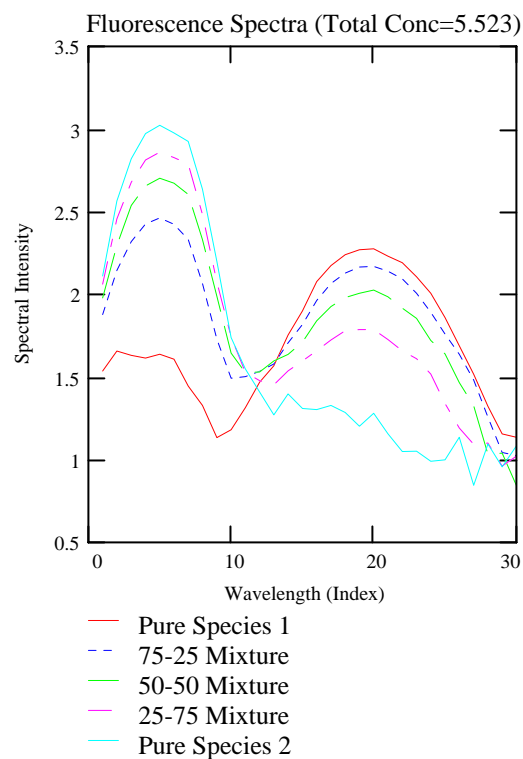
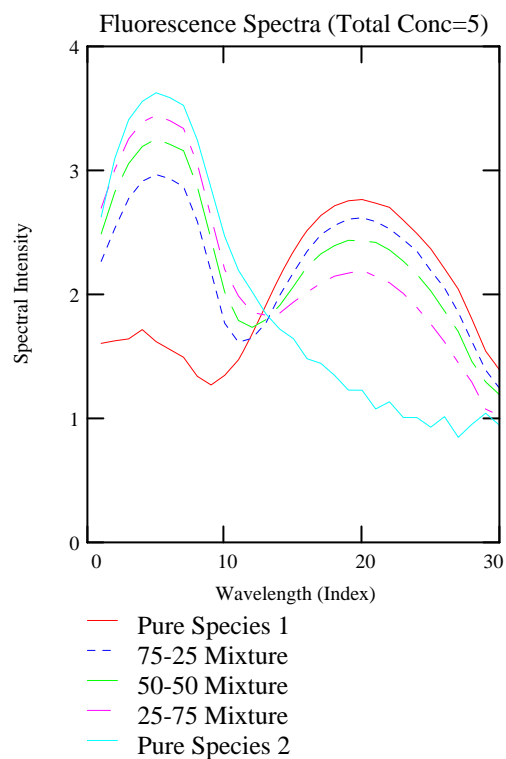
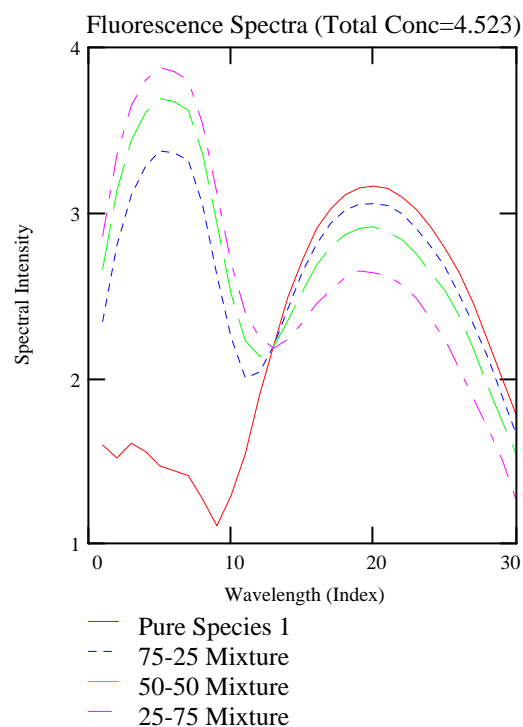
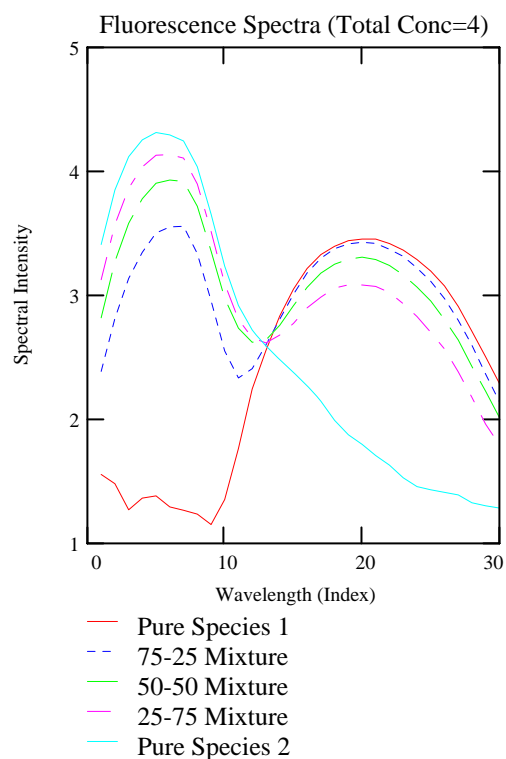
— Pure Species 1
 - - 75-25 Mixture
 — 50-50 Mixture
 - - 25-75 Mixture
 — Pure Species 2



— Pure Species 1
 - - 75-25 Mixture
 — 50-50 Mixture
 - - 25-75 Mixture
 — Pure Species 2

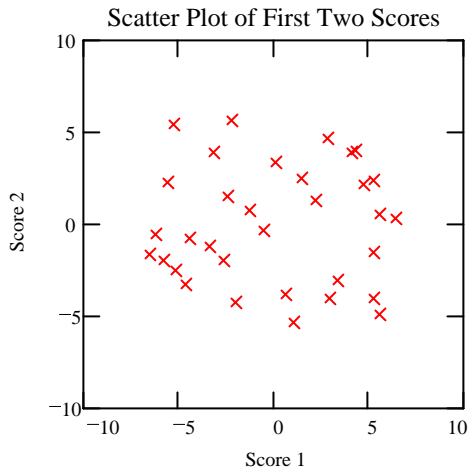
Mole Fraction Total Conc.

| i | y_{1_i} | y_{2_i} |
|----|-----------|-----------|
| 1 | 0.75 | 3.523 |
| 2 | 0 | 3 |
| 3 | 0.5 | 5.523 |
| 4 | 0.25 | 5 |
| 5 | 0 | 6 |
| 6 | 0.25 | 3.523 |
| 7 | 0.75 | 5.523 |
| 8 | 0.5 | 4 |
| 9 | 0.75 | 5 |
| 10 | 1 | 4.523 |
| 11 | 1 | 3 |
| 12 | 0 | 4 |
| 13 | 0.5 | 3 |
| 14 | 0.25 | 6 |
| 15 | 1 | 5.523 |
| 16 | 1 | 3.523 |
| 17 | 0.25 | 5.523 |
| 18 | 0 | 5 |
| 19 | 0.5 | 3.523 |
| 20 | 0 | 5.523 |
| 21 | 0.75 | 6 |
| 22 | 1 | 5 |
| 23 | 0.75 | 3 |
| 24 | 0.25 | 3 |
| 25 | 1 | 4 |
| 26 | 0.5 | 5 |
| 27 | 0.75 | 4 |
| 28 | 0.5 | 4.523 |
| 29 | 0.25 | 4.523 |
| 30 | 0.75 | 4.523 |
| 31 | 0.5 | 6 |
| 32 | 0 | 3.523 |
| 33 | 0.25 | 4 |



Express X in the eigenvector direction. This quantity is commonly referred to as the **score** vector(s). Essentially, we change the basis (or coordinate system). The scores are just another way of expressing X, except that we do so along the eigenvector directions. In other words, we express the original X with a new coordinate system. The new coordinates (which are the eigenvectors) are commonly referred to as the **loading** vectors. The figure above shows the first three loading vectors. Thus, the scores are conceptually "equivalent" to the values of the independent variable X, and the loadings are the combination of the independent variable X.

$$\text{score} := X \cdot V$$



$$\lambda_1 = 567.07 \quad \dots \text{variance of score}^{<1>}$$

$$\lambda_2 = 325.94 \quad \dots \text{variance of score}^{<2>}$$

Linear Representation. We wish to express Y (the given dependent vector) as a linear combination of the scores (independent vectors).

$$Y = a_1 \cdot \text{score}^{<1>} + a_2 \cdot \text{score}^{<2>} + \dots + a_m \cdot \text{score}^{<m>} + \text{error} = \sum_{j=1}^m a_j \cdot \text{score}^{<j>} + \text{error}$$

Regression Coefficients. Since the scores are orthogonal vectors, we can apply the normal equation sequentially in an element-wise fashion to calculate the coefficients.

$$a_{1j} := \frac{(\text{score}^{<j>} \cdot Y_1)}{(\text{score}^{<j>} \cdot \text{score}^{<j>)} \quad a_{2j} := \frac{(\text{score}^{<j>} \cdot Y_2)}{(\text{score}^{<j>} \cdot \text{score}^{<j>)}$$

Regression Equation. We trace our regression steps backward to arrive at the regression equation. (Any one of the following equivalent expressions will do.)

$$Y = \text{score} \cdot a = X \cdot V \cdot a = X \cdot (V^{<1>} \quad V^{<2>} \quad \dots \quad V^{<m>}) \cdot a = (X \cdot V^{<1>} \quad X \cdot V^{<2>} \quad \dots \quad X \cdot V^{<m>}) \cdot a \\ = a_1 \cdot X \cdot V^{<1>} + a_2 \cdot X \cdot V^{<2>} + \dots + a_m \cdot X \cdot V^{<m>}$$

$$Y = \text{score} \cdot a = (\text{score}^{<1>} \quad \text{score}^{<2>} \quad \dots \quad \text{score}^{<m>}) \cdot a = a_1 \cdot \text{score}^{<1>} + a_2 \cdot \text{score}^{<2>} + \dots + a_m \cdot \text{score}^{<m>}$$

$$Y = a_1 \cdot \sum_{j=1}^m X^{<j>} \cdot (V^{<1>})_j + a_2 \cdot \sum_{j=1}^m X^{<j>} \cdot (V^{<2>})_j + \dots + a_m \cdot \sum_{j=1}^m X^{<j>} \cdot (V^{<m>})_j$$

$$Y = \left[\sum_{i=1}^m \left[a_i \cdot \sum_{j=1}^m X^{<j>} \cdot (V^{<i>})_j \right] \right] \quad \text{where} \quad X^{<j>} = \left(\frac{x^{<j>} - x \text{ mean}_j}{x \text{ std}_j} \right)$$

$$y = X \cdot V \cdot a \quad Y = \left(\frac{y - y \text{ mean}}{y \text{ std}} \right)$$

Sum of squared error (sse). Examine sse to determine which terms (factors) we should retain for regression.

Initial sse (total variations in y) before regression.

$$\begin{aligned} \text{sse}_{10} &:= Y_1 \cdot Y_1 & \text{sse}_{10} &= 33 & \leftarrow \text{compare} \rightarrow & n = 33 \\ \text{sse}_{20} &:= Y_2 \cdot Y_2 & \text{sse}_{20} &= 33 \end{aligned}$$

Note that since we normalized each data point in Y (rather than the length of the whole vector Y) with respect to the standard deviation at the beginning of this worksheet, the elements of Y are scattered around 1, i.e., the average of Y_i^2 is 1. Consequently, the variance of Y, which is the sum of variances of individual elements of Y, $\sum_i Y_i^2$, is the number of data points n.

Error is reduced as we include each additional term. $e_0 = Y$ $e_j = e_{j-1} - \text{score}^{<j>} \cdot a_j$ $\text{sse} = e \cdot e$

Here comes the calculation for sse and R^2 ...

$$\begin{aligned} \text{sse}_{1_j} &:= \left(Y_1 - \sum_{k=1}^j \text{score}^{<k>} \cdot a_{1_k} \right) \cdot \left(Y_1 - \sum_{k=1}^j \text{score}^{<k>} \cdot a_{1_k} \right) & r^2_{1_j} &:= \frac{\text{sse}_{10} - \text{sse}_{1_j}}{\text{sse}_{10}} \\ \text{sse}_{2_j} &:= \left(Y_2 - \sum_{k=1}^j \text{score}^{<k>} \cdot a_{2_k} \right) \cdot \left(Y_2 - \sum_{k=1}^j \text{score}^{<k>} \cdot a_{2_k} \right) & r^2_{2_j} &:= \frac{\text{sse}_{20} - \text{sse}_{2_j}}{\text{sse}_{20}} \end{aligned}$$

Incremental R^2 as we include each additional term.

$$\begin{aligned} \Delta r^2_{1_j} &:= \frac{(\text{score}^{<j>} \cdot a_{1_j}) \cdot (\text{score}^{<j>} \cdot a_{1_j})}{\text{sse}_{10}} \\ \Delta r^2_{2_j} &:= \frac{(\text{score}^{<j>} \cdot a_{2_j}) \cdot (\text{score}^{<j>} \cdot a_{2_j})}{\text{sse}_{20}} \end{aligned}$$

sse and R^2 for the 1st independent variable Y_1 (mole fraction)

| j | λ_j | sse 1_j | $r^2 1_j$ | $\Delta r^2 1_j$ |
|----|-------------|-----------|-----------|------------------|
| 1 | 567.070 | 25.054 | 0.241 | 0.241 |
| 2 | 325.940 | 9.303 | 0.718 | 0.477 |
| 3 | 57.636 | 6.283 | 0.810 | 0.092 |
| 4 | 19.823 | 6.273 | 0.810 | 0.000 |
| 5 | 11.562 | 5.541 | 0.832 | 0.022 |
| 6 | 4.367 | 4.951 | 0.850 | 0.018 |
| 7 | 1.515 | 4.734 | 0.857 | 0.007 |
| 8 | 0.525 | 4.549 | 0.862 | 0.006 |
| 9 | 0.453 | 4.208 | 0.872 | 0.010 |
| 10 | 0.318 | 4.196 | 0.873 | 0.000 |
| 11 | 0.196 | 3.359 | 0.898 | 0.025 |
| 12 | 0.148 | 3.326 | 0.899 | 0.001 |
| 13 | 0.107 | 3.320 | 0.899 | 0.000 |
| 14 | 0.074 | 3.320 | 0.899 | 0.000 |
| 15 | 0.059 | 3.032 | 0.908 | 0.009 |
| 16 | 0.055 | 2.861 | 0.913 | 0.005 |
| 17 | 0.043 | 2.570 | 0.922 | 0.009 |
| 18 | 0.036 | 2.207 | 0.933 | 0.011 |
| 19 | 0.023 | 2.111 | 0.936 | 0.003 |
| 20 | 0.018 | 0.956 | 0.971 | 0.035 |
| 21 | 0.013 | 0.929 | 0.972 | 0.001 |
| 22 | 0.008 | 0.927 | 0.972 | 0.000 |
| 23 | 0.004 | 0.867 | 0.974 | 0.002 |
| 24 | 0.004 | 0.740 | 0.978 | 0.004 |
| 25 | 0.002 | 0.732 | 0.978 | 0.000 |
| 26 | 0.002 | 0.701 | 0.979 | 0.001 |
| 27 | 0.001 | 0.333 | 0.990 | 0.011 |
| 28 | 0.000 | 0.166 | 0.995 | 0.005 |
| 29 | 0.000 | 0.163 | 0.995 | 0.000 |
| 30 | 0.000 | 0.155 | 0.995 | 0.000 |

← These factors capture at least 1% of variations in Y_1 .

←

←

←

←

←

←

←

←

←

sse and R^2 for the 2nd independent variable Y_2 (total concentration)

| j | λ_j | sse $_{2_j}$ | $r^2_{2_j}$ | $\Delta r^2_{2_j}$ |
|----|-------------|--------------|-------------|--------------------|
| 1 | 567.070 | 15.736 | 0.523 | 0.523 |
| 2 | 325.940 | 13.324 | 0.596 | 0.073 |
| 3 | 57.636 | 2.653 | 0.920 | 0.323 |
| 4 | 19.823 | 2.606 | 0.921 | 0.001 |
| 5 | 11.562 | 1.513 | 0.954 | 0.033 |
| 6 | 4.367 | 1.310 | 0.960 | 0.006 |
| 7 | 1.515 | 0.982 | 0.970 | 0.010 |
| 8 | 0.525 | 0.970 | 0.971 | 0.000 |
| 9 | 0.453 | 0.962 | 0.971 | 0.000 |
| 10 | 0.318 | 0.900 | 0.973 | 0.002 |
| 11 | 0.196 | 0.586 | 0.982 | 0.009 |
| 12 | 0.148 | 0.583 | 0.982 | 0.000 |
| 13 | 0.107 | 0.580 | 0.982 | 0.000 |
| 14 | 0.074 | 0.566 | 0.983 | 0.000 |
| 15 | 0.059 | 0.546 | 0.983 | 0.001 |
| 16 | 0.055 | 0.410 | 0.988 | 0.004 |
| 17 | 0.043 | 0.410 | 0.988 | 0.000 |
| 18 | 0.036 | 0.375 | 0.989 | 0.001 |
| 19 | 0.023 | 0.264 | 0.992 | 0.003 |
| 20 | 0.018 | 0.263 | 0.992 | 0.000 |
| 21 | 0.013 | 0.262 | 0.992 | 0.000 |
| 22 | 0.008 | 0.253 | 0.992 | 0.000 |
| 23 | 0.004 | 0.245 | 0.993 | 0.000 |
| 24 | 0.004 | 0.203 | 0.994 | 0.001 |
| 25 | 0.002 | 0.174 | 0.995 | 0.001 |
| 26 | 0.002 | 0.073 | 0.998 | 0.003 |
| 27 | 0.001 | 0.025 | 0.999 | 0.001 |
| 28 | 0.000 | 0.021 | 0.999 | 0.000 |
| 29 | 0.000 | 0.020 | 0.999 | 0.000 |
| 30 | 0.000 | 0.003 | 1.000 | 0.000 |

← These factors capture at least 1% of variations in Y_2 .

←

←

←

←

←

We need less number of factors to describe Y_2 (total concentration) than Y_1 (mole fraction) because spectral intensities correlates more directly with total concentration than mole fraction. In general, the more complex the behavior we try to describe, the more factors we will need. Be sure not to employ more factors than we need. With a sufficient number of factors, we can always drive sse to 0 exactly and R^2 to 1 exactly. In regression, whether or not we utilize eigenvalues and eigenvectors, we pursue a good fit, not an exact fit, of the given data with as few factors as possible. In other words, do not overfit and capture noises in the given data instead of the genuine underlying trends. We leave out the noise and random fluctuations from our regression model.

Note that the factors that have higher variations in X (i.e., the ones with larger eigenvalues) do not necessarily capture more variations in Y . The arrows in the above tables point to the ten factors that capture at least 1% of variations in Y_1 and the six most important factors that capture the majority of variations in Y_2 . Furthermore, the set of factors that explain Y_1 and those that explain Y_2 well are not necessarily identical. It means that some factors are more closely related to mole fraction, while others are more closely related to total concentration.

For simplicity, we re-sort the eigenvalues based on incremental R² values, Δr². First, we define a function that rearranges a vector x based on descending order of a second vector y.

$$\text{resortvec}(x, y) := \text{reverse}\left(\text{csort}(\text{augment}(y, x), 1)^{\langle 2 \rangle}\right)$$

With this function, we now re-order the eigenvalues and eigenvectors.

$$\lambda_1 := \text{resortvec}(\lambda, \Delta r_{2_1}) \quad V1^{\langle j \rangle} := \text{eigenvec}(X^T \cdot X, \lambda_{1_j}) \quad \text{score} := X \cdot V1 \quad a_{1_j} := \frac{(\text{score}^{\langle j \rangle} \cdot Y_1)}{(\text{score}^{\langle j \rangle} \cdot \text{score}^{\langle j \rangle})}$$

$$\lambda_2 := \text{resortvec}(\lambda, \Delta r_{2_2}) \quad V2^{\langle j \rangle} := \text{eigenvec}(X^T \cdot X, \lambda_{2_j}) \quad \text{score} := X \cdot V2 \quad a_{2_j} := \frac{(\text{score}^{\langle j \rangle} \cdot Y_2)}{(\text{score}^{\langle j \rangle} \cdot \text{score}^{\langle j \rangle})}$$

Regression equation in semi-vector notation where *each data point is given as a column vector of x* (which is different from how each point is given in the original data table as a row vector) because Mathcad works with column vectors, not row vectors. The following regression includes the factors in decreasing order of Δr².

$$y_{1.\text{regress}}(x, \text{factor}) := \left[\sum_{k=1}^{\text{factor}} \left[a_{1_k} \cdot \sum_{j=1}^m \frac{x_j - x_{\text{mean}_j}}{x_{\text{std}_j}} \cdot (V1^{\langle k \rangle})_j \right] \right] \cdot y_{1.\text{std}} + y_{1.\text{mean}}$$

Repeat for the 2nd dependent variable (total concentration).

$$y_{2.\text{regress}}(x, \text{factor}) := \left[\sum_{k=1}^{\text{factor}} \left[a_{2_k} \cdot \sum_{j=1}^m \frac{x_j - x_{\text{mean}_j}}{x_{\text{std}_j}} \cdot (V2^{\langle k \rangle})_j \right] \right] \cdot y_{2.\text{std}} + y_{2.\text{mean}}$$

Let us perform prediction on the given data.

$$y_{1.\text{pred}_1} := y_{1.\text{regress}}\left[\left(x^T\right)^{\langle i \rangle}, 10\right] \quad \dots \text{ 10 factors suffice for mole fraction.} \quad yy_1 := 0, 0.01 \dots 1$$

$$y_{2.\text{pred}_1} := y_{2.\text{regress}}\left[\left(x^T\right)^{\langle i \rangle}, 6\right] \quad \dots \text{ 6 factors suffice for total concentration} \quad yy_2 := 2, 2.1 \dots 7$$

