Multivariate Analysis

Homework #1 (2008) Answers

What happens to regression and PCA when you combine datasets?

Q1. Consider the basic regression set-up: given a matrix X (elements x_{mn}, whose nth column is the nth regressor) and a dataset Y (considered as a column vector y_m) find a column A (elements a_n) for which $R = \sum_m (\sum_n x_{mn} a_n - y_m)^2 = tr\left((XA-Y)^T (XA-Y)\right)$ *is minimized.*

Say that A_1 is the solution for dataset Y_1, and that A_2 is the solution for dataset Y_2 (both based on the same regressors X). Can you write a simple expression for the solution A corresponding to the combined dataset $Y_c = Y_1 + Y_2$? Why or why not? (For example, if you have an experiment with multiple subjects, and you do a regression analysis separately on each subject's data, what can you say about a regression analysis on the combined data?)

Solution.

The regression coefficients *A* are linear in the data, so adding datasets corresponds to adding regressors. Specifically (page 5 of notes),

$$A = (X^T X)^{-1} X^T Y . \qquad (1)$$

So, $A_c = (X^T X)^{-1} X^T Y_c = (X^T X)^{-1} X^T (Y_1 + Y_2) = (X^T X)^{-1} X^T Y_1 + (X^T X)^{-1} X^T Y_2 = A_1 + A_2$.

A simple consequence is that averaging together two datasets results in averaging the regressors.

An important implication of this, related to the effects of noisy measurements, is the following. We recognize that any experimental measurement *Y* is corrupted by noise. But we can reasonably hope that our experimental measurement *Y* is unbiased – i.e., that in any single experiment, the measured value of *Y* does not systematically deviate above or below its true value. Since regression is linear in the data, this means that the estimated regression coefficients are unbiased estimates of the true coefficients.

Note also that regularized and penalized regression is also linear in the data (page 7 of notes: $A = (X^T C_Y^{-1} X + C_A^{-1})^{-1} X^T C_Y^{-1} Y$), so the above analysis applies to that context.

Q2: Same as Q1, but for PCA. That is, say you have a dataset Y_1 (elements $y_{1;mr}$), for which the principal components are the matrix X_1, and a second dataset Y_2 with

principal components X_2. *Can you write a simple expression for the principal components of the combined dataset* $Y_c = Y_1 + Y_2$ *? Why or why not?*

Solution.

No, the PCA's of the combined dataset need not be simply related to the principal components of the individual ones. Abstractly, the reason is that the principal components are not linear functions of the data, but rather, they are the eigenvectors of the covariance matrix $Y^T Y$ -- and $Y_c{}^T Y_c \neq Y_1{}^T Y_1 + Y_2{}^T Y_2$.

One reason that this is important is that it means that estimates of principal components are necessarily biased – as opposed to the unbiased nature of regression estimates.

Numerical example: $Y_1 = \begin{pmatrix} 6 & 4 \\ 3 & 2 \\ 9 & 6 \end{pmatrix}$, $Y_2 = \begin{pmatrix} 4 & 8 \\ -2 & -4 \\ 3 & 6 \end{pmatrix}$. To make the situation clearcut, these have been "cooked" so that each dataset individually has only one principal component, namely

$Y_1 = \begin{pmatrix} 2 \\ 1 \\ 3 \end{pmatrix} \begin{pmatrix} 3 & 2 \end{pmatrix}$ and $Y_2 = \begin{pmatrix} 4 \\ -2 \\ 3 \end{pmatrix} \begin{pmatrix} 1 & 2 \end{pmatrix}$.

$Y_1{}^T Y_1 = \begin{pmatrix} 126 & 84 \\ 84 & 56 \end{pmatrix}$, which has only one nonzero eigenvalue $\lambda_1 = 182$ with eigenvector proportional to $(3 \quad 2)$ (check this, and that $\lambda_1 = tr(Y_1{}^T Y_1) = \sum_{i,j} \left(Y_{1;ij} \right)^2$ and that $Y_1{}^T Y$ is singular).

$Y_2{}^T Y_2 = \begin{pmatrix} 29 & 58 \\ 58 & 116 \end{pmatrix}$, which has only one nonzero eigenvalue $\lambda_2 = 145$ with eigenvector proportional to $(2 \quad 1)$ (check this, and that $\lambda_2 = tr(Y_2{}^T Y_2) = \sum_{i,j} \left(Y_{2;ij} \right)^2$ and that $Y_2{}^T Y_2$ is singular).

$Y_c = Y_1 + Y_2 = \begin{pmatrix} 10 & 12 \\ 1 & -2 \\ 12 & 12 \end{pmatrix}$. $Y_c{}^T Y_c = \begin{pmatrix} 245 & 262 \\ 262 & 292 \end{pmatrix}$. This is not singular; it has two nonzero left eigenvalues (use Matlab's eig), namely approximately 531.55 and 5.44, and corresponding eigenvalues are approximately proportional to $\begin{pmatrix} 0.674 & 0.738 \end{pmatrix}$ and $\begin{pmatrix} -0.738 & 0.674 \end{pmatrix}$ (use Matlab's eigs, and be careful about row and column conventions). Essentially, what happened is that the first component of the combined problem is the best compromise between the two datasets, and the second component contains the rest of the variance.