Multivariate Analysis

Homework #1 (2008)

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 $n$th column is the $n$th regressor) and a dataset $Y$ (considered as a column vector $y_m$) find a column $A$ (elements $a_n$) for which

$$R = \sum_{m} (\sum_{n} x_{mn} a_n - y_m)^2 = \text{tr}((XA-Y)(XA-Y))$$

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?)

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?