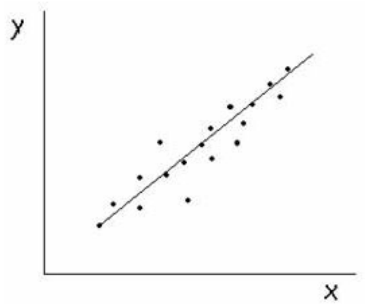
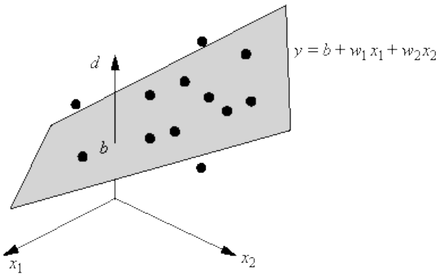
May 9, 2014

Using Ideas from Multiple Linear Regression to Construct an Image Compression Matrix

Irma Stevens

Since Ryan Livingston provided a description of how PCA can be used to construct decent Image Compression Matrices, I will try not to be redundant and instead focus on how I gained an intuitive sense of the problem by comparing it to Multiple Linear Regression. To start, both methods attempt to solve the least squares problem by considering the relationship (i.e. the correlation) between different variables (i.e. image vectors). This is difficult to visualize in multiple dimensions, but the idea can be seen below with one and two variables being considered. The regression line below minimizes the sum of the squares of the vertical distances of the given points. The regression plane follows a similar process.





Images from:
http://www.mu-sigma.com/analytics/thought_leadership/cafe-cerebral-linear-regression.html
http://www.nd.com/NSBook/NEURAL%20AND%20ADAPTIVE%20SYSTEMS17_Regression_for_Multiple_Varia.html

The correlation matrix provides the setup for PCA (through SVD) to do its job by essentially ranking the top vectors to include in the compression matrices. Although the construction of this correlation matrix (multiplying the image vector matrix by its transpose) deviates from the methods often employed in statistical linear model building, the purpose of the matrix is similar. Essentially, the vectors with the strongest correlations will be chosen.

This is important to keep in mind when choosing the images used to form the cloud of vectors. As is the case with multiple linear regression modeling, while adding more variables to a set will not harm your results, the key is really finding the most useful variables to include in the model. As my Statistics professor stressed, if you want to create a model to predict the winning percentage of a baseball team, do not choose only offensive statistics (e.g. hitting, running, steals, etc.) to do so. If this is done, the model will run into the issue multicollinearity, or the redundancy of variables (due to the high correlation between the variables). To avoid this issue when constructing an image compression matrix, it is important to choose images that vary in terms of its features. For example, I chose some images that were images of dark animals against a light background, others that were light animals against a light background, etc. This helps to obtain a broader range of varying image vectors for the PCA process to utilize.

On the other hand, it is also possible to exploit this idea of correlation matrices by purposefully choosing several images that display certain characteristics that will potentially be common in the dataset. For instance, since I was given that the image I would need to compress was going to be an animal, I purposefully chose several pictures of animals with detailed stripes, spots, whiskers, and fur in order to get them to produce high enough correlations to end up in my final image compression matrices.