Prof. Sergio A. Alvarez
Fulton Hall 410–B
Computer Science Department
Boston College
Chestnut Hill, MA 02467 USA

http://www.cs.bc.edu/∼alvarez/
alvarez@cs.bc.edu
voice: (617) 552-4333
fax: (617) 552-6790

Summary statistics for multivariate data

For data described by a single numerical attribute, the mean and standard deviation provide good summary statistics of the overall data distribution. For multivariate data, however, merely calculating the mean and standard deviation of the individual attributes gives a limited picture as this fails to identify relationships among the attributes. We discuss pairwise linear correlation and the covariance matrix, which provide tools for the multivariate case.

Notation

We consider a dataset X consisting of m instances, each described by the values of n numerical attributes. Thus, we can concretely view X as a matrix of size $m \times n$, with one row for each instance and one column for each attribute. Using Matlab notation, we let $X(i, j)$ denote the element of X that appears in row i, column j, let $X(i, :)$ denote the i-th row of X, and let $X(:, j)$ denote the j-th column of X.

Correlation

One measure of the agreement or disagreement between two attributes is the Pearson correlation, defined as follows for two vectors x and y of length m, each corresponding to the values of a preselected attribute over the set of all data instances (e.g., a column of the data matrix X):

$$corr(x, y) = \frac{\sum_{k=1}^{m}(x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{m}(x_k - \bar{x})^2}\sqrt{\sum_{k=1}^{m}(y_k - \bar{y})^2}},$$

where the bars indicate the means of the corresponding attributes. The correlation value varies between -1 and $+1$, with -1 indicating a linearly decreasing relationship between the two variables, 0 indicating complete independence between the variables, and $+1$ indicating a linearly increasing dependence (see Fig. 1). Note that the correlation only models *linear* dependence between attributes. It is possible for two attributes to be functionally related in a simple way (e.g., $y = \sin(x)$) while the correlation equals 0. Thus, the correlation cannot be interpreted as a measure of dependence in the general sense. Also, correlation does not provide summary statistical information about the individual attributes, and hence constitutes an incomplete description of multivariate data.

The covariance matrix

There is an object that captures the spread of multivariate data in all of its dimensions. It combines the pairwise correlations of all attribute pairs, together with the variances of the individual attributes. This object is the *covariance matrix*, which for an $m \times n$ data matrix X (m instances,

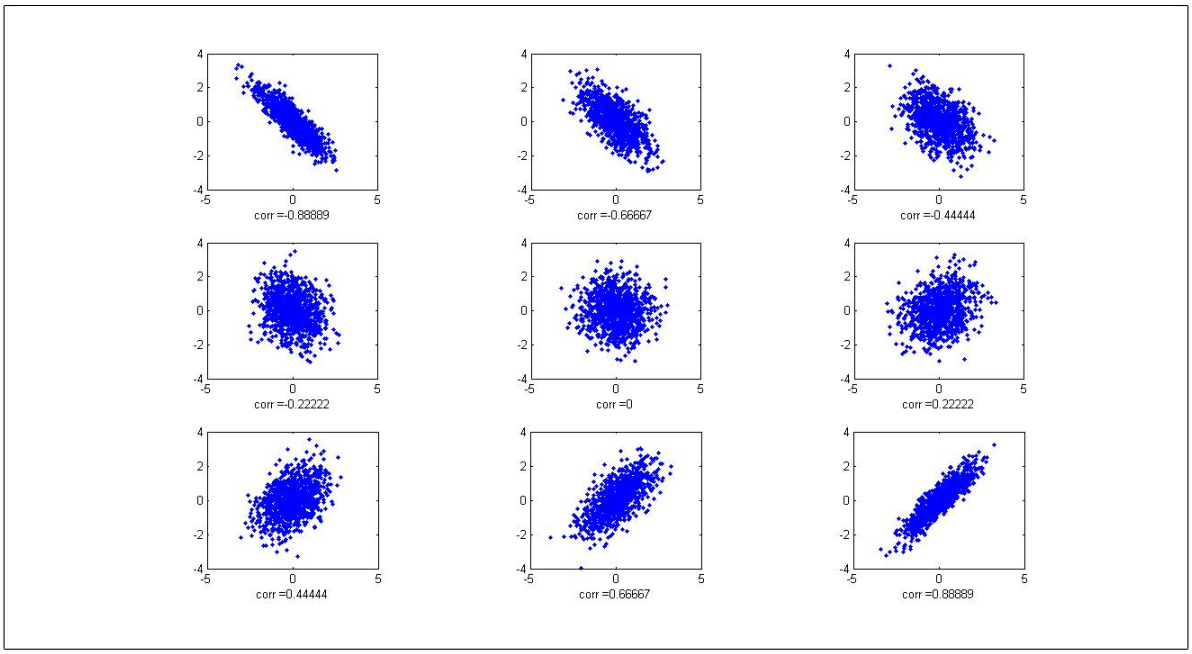

Figure 1: Examples of datasets with various correlation values

each described by n attributes) is defined as the $m \times m$ matrix with the following elements:

$$C(i,j) = \frac{1}{m-1}\sum_{k=1}^{m}(X(k,i) - \overline{X(:,i)})(X(k,j) - \overline{X(:,j)})$$

Notice that, for $i \neq j$:

$$C(i,j) = corr(X(:,i), X(:,j))\sqrt{var(X(:,i))var(X(:,j))},$$

while for $i = j$:

$$C(i,i) = var(X(:,i))$$

In particular, the pairwise correlations may be recovered from the entries of the covariance matrix:

$$corr(X(:,i), X(:,j)) = \frac{C(i,j)}{\sqrt{C(i,i)C(j,j)}}$$

Notice also that the covariance matrix may be expressed succintly in matrix product terms:

$$cov(X) = \frac{1}{m-1}X'X,$$

where X' is the transpose of X.

Using the covariance matrix to construct a decorrelating frame

The covariance matrix $C \in \mathrm{R}^{m\times m}$ of a dataset $X \in \mathrm{R}^{m\times n}$ encodes geometric information about the spread of the data in R^n. We show how the covariance matrix may be used to construct a

new system of coordinates (attributes) that are completely uncorrelated relative to the dataset. Geometrically, in the new frame the data will be symmetrically distributed in all directions, with unit variances for all attributes and zero correlations between all pairs of attributes.

We first state the following important fact without proof.

Choleski decomposition of a symmetric positive semi–definite matrix. If $A \in \mathrm{R}^{p\times p}$ satisfies $A' = A$ and $x'Ax \geq 0$ for all column vectors $x \in \mathrm{R}^p$, then there is an upper triangular matrix $R \in \mathrm{R}^{p\times p}$ such that $A = R'R$. The matrices R and R' are called the Choleski factors of A.

The Choleski decomposition may be used to decorrelate the set of attributes of a dataset X, as follows. Let C be the covariance matrix of X. Assuming that C is invertible (it will be, unless the matrix X has rank less than n, which would happen only if the data points are constrained to some subspace of dimension lower than n within R^n), let R be the upper triangular Choleski factor of the inverse matrix C^{-1}. Define a new dataset Y by $Y = XR'$. Notice that this amounts to a transformation of coordinates: the new dataset Y contains exactly the same number of instances (rows) as X, but the new coordinates of each instance (row) in Y are linear combinations of the old coordinates of that instance in X, according to the columns of R'. I claim that the covariance matrix of Y is the $n \times n$ identity matrix, which has 1's along the main diagonal and 0's everywhere else. Before proving this claim, notice that it implies the assertion that the new coordinates are completely uncorrelated, as the diagonal entries of the correlation matrix of Y are the variances of the individual attributes and the off–diagonal entries are scaled versions of the pairwise correlations of different attributes.

An example of the results obtained through the decorrelation procedure described above is shown in Fig.2.

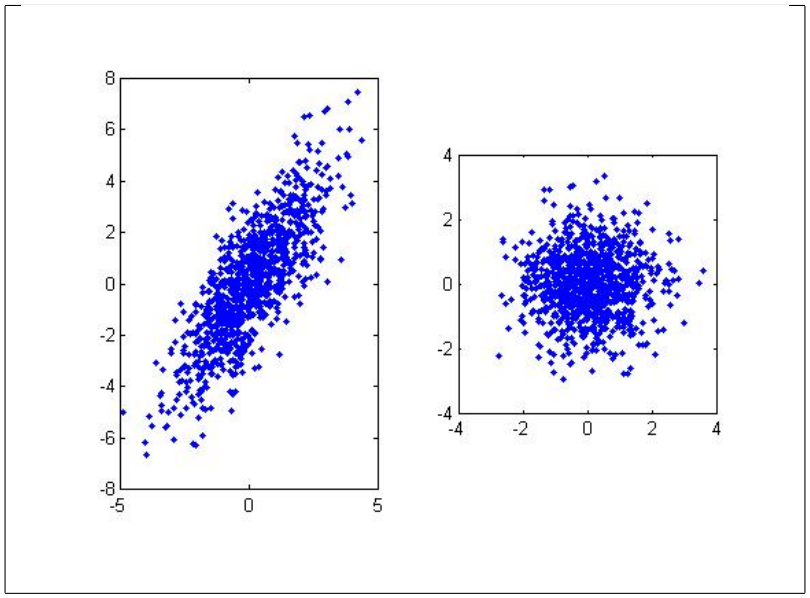

Figure 2: Correlated dataset (left) and its decorrelated version (right)

Proof that the decorrelation procedure works correctly. We now show that $cov(Y) = I$. Start from the matrix product definition of the covariance matrix:

$$cov(Y) = \frac{1}{m-1} Y'Y$$

Substitute the definition of Y:

$$cov(Y) = \frac{1}{m-1}(XR')'(XR') = \frac{1}{m-1} RX'XR' = Rcov(X)R'$$

Thus, if we right–multiply the covariance matrix by R, we obtain:

$$cov(Y)R = Rcov(X)R'R = R,$$

since $R'R$ is the inverse of $cov(X)$ by the Choleski decomposition. The latter equation shows that all of the vectors that form the columns of R are invariant under left multiplication by $cov(Y)$, or, in other words, they are eigenvectors of $cov(Y)$ with eigenvalue 1. But these columns form a complete basis of R^n since R is non-singular (since $cov(X)$ is non-singular). Therefore, any vector $v \in \mathrm{R}^n$ may be expressed as a linear combination of the columns of R, and it follows that $cov(Y)v = v$ for all such vectors. This shows that $cov(Y) = I$, as claimed.

The Mahalanobis distance

The decorrelation procedure described above yields in particular an appealing distance metric in the space of instances. For correlated data X (non–diagonal covariance matrix), the Euclidean metric in the original attribute space is not a good choice because it assigns distances symmetrically in all directions. This accounts for neither different variances among individual attributes, nor for correlations between pairs of attributes. How to address these issues? Well, since decorrelation completely symmetrizes the data distribution, one can simply decorrelate the data first, and then use the standard Euclidean distance metric in the decorrelated space. The resulting distance metric may be described as follows for row vectors x and y, where R is the upper triangular Choleski factor of the inverse $cov(X)^{-1}$, which satisfies $X'X = cov(X)^{-1}$:

$$d(x, y) = \|xR' - yR'\| = \|(x - y)R'\|$$

Since Euclidean length of a row vector z may be expressed in terms of a dot product:

$$\|z\| = \sqrt{zz'},$$

we may rewrite the symmetrized metric $d(x, y)$ as follows:

$$d(x, y) = \sqrt{((x - y)R')((x - y)R')'} = \sqrt{(x - y)R'R(x - y)'} = \sqrt{(x - y)cov(X)^{-1}(x - y)'}$$

This distance metric $d(x, y)$ is known as the Mahalanobis distance for the dataset X. As described above, the Mahalanobis distance is a normalized distance function that is tuned to the particular distribution of the dataset X.