Ph 21.5: Covariance and Principal Component Analysis (PCA)

-v20150527-

Introduction

Suppose we make a measurement for which each data sample consists of two measured quantities. A simple example would be temperature (T) and pressure (P) taken at time (t) at constant volume(V). The data set is $\{T_i, P_i|t_i\}_N$, which represents a set of N measurements. We wish to make sense of the data and determine the dependence of, say, P on T. Suppose P and T were for some reason independent of each other; then the two variables would be *uncorrelated*. (Of course we are well aware that P and V *are* correlated and we know the ideal gas law: $PV = nRT$). How might we infer the correlation from the data?

The tools for quantifying correlations between random variables is the *covariance*. For two real-valued random variables (X, Y), the covariance is defined as (under certain rather non-restrictive assumptions):

$$\mathrm{Cov}(X, Y) \equiv \sigma^2_{XY} \equiv \langle (X - \langle X\rangle)(Y - \langle Y\rangle)\rangle$$

where $\langle ...\rangle$ denotes the expectation (average) value of the quantity in brackets. For the case of P and T, we have

$$\begin{aligned}
\mathrm{Cov}(P,T) &= \langle (P - \langle P\rangle)(T - \langle T\rangle)\rangle \\
&= \langle P \times T\rangle - \langle P\rangle \times \langle T\rangle \\
&= \left(\frac{1}{N}\sum_{i=0}^{N-1} P_i T_i\right) \; - \; \left(\frac{1}{N}\sum_{i=0}^{N-1} P_i\right)\left(\frac{1}{N}\sum_{i=0}^{N-1} T_i\right)
\end{aligned}$$

The extension of this to real-valued random vectors $(\vec{X}, \vec{Y})$ is straighforward:

$$\mathrm{Cov}(\vec{X}, \vec{Y}) \equiv \sigma^2_{\vec{X}\vec{Y}} \equiv \left\langle (\vec{X} - <\vec{X}>)(\vec{Y} - <\vec{Y}>)^T \right\rangle$$

This is a matrix, resulting from the product of a one vector and the transpose of another vector, where $\vec{X}^T$ denotes the transpose of $\vec{X}$. This matrix, called the *covariance matrix*, is one of the most important quantities that arises in data analysis.

In many cases, we will simply be interested in the covariance of a single set of random variables $\{X_i\}_N = \vec{X}$:

$$\mathrm{Cov}(\vec{X}, \vec{X}) = \left\langle (\vec{X} - <\vec{X}>)(\vec{X} - <\vec{X}>)^T \right\rangle$$

For specificity, suppose $\vec{X} = \{x, y\}$, i.e. we are back to the case of two real-valued random variables. Then:

$$\mathrm{Cov}(\vec{X}, \vec{X}) \equiv \begin{pmatrix} \sigma^2_x & \sigma^2_{xy} \\ \sigma^2_{xy} & \sigma^2_y \end{pmatrix}$$

where σ^2_x is just the usual "sigma squared" variance for x. When $\sigma^2_{xy} = 0$ the errors in x and y are uncorrelated, but otherwise they are correlated.

So, covariance matrices are very useful: they provide an estimate of the variance in individual random variables and also measure whether variables are correlated. A concise summary of the covariance can be found on Wikipedia by looking up 'covariance'. A more complete and better description is provided by [5] which provides a more general Bayesian context for the definition of the covariance matrix. If you have not seen covariance matrices before (or even if you have), [5] is a good place to learn about them.

You may have noticed that there can be cases where both σ^2_x and σ^2_y can be large, but because x and y are highly correlated, measuring one variable accurately specifies the other, possibly with very little error. Such is the case, for example, when $x = a \; + \; b\, y$. The covariance matrix show immediately if this is the

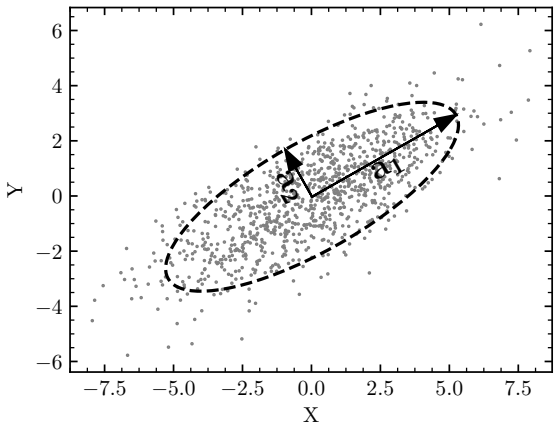

Figure 1: Geometric interpretation of diagonalizing the covariance matrix in 2D: it is essentially fitting an ellipse (in general, an N-dimensional ellipsoid) to the distribution of your data. If the eigenvalues of the covariance matrix are (λ_1, λ_2), then the lengths of the ellipsoid's axes are proportional to $\sqrt{\lambda_1}$ and $\sqrt{\lambda_2}$, and the axis vectors $\mathbf{a_1}$ and $\mathbf{a_2}$ are oriented along the covariance matrix's eigenvectors. Here the ellipsoid plotted is the '2σ' contour of the underlying 2D Gaussian distribution of the data.

case. You may then wonder if we can use covariance matrices to determine underlying correlations in a set of measured variables, even when we do not know the underlying dependencies of the parameters. The answer is 'yes' (partially), and this is the subject of *principal component analysis*, or PCA.

Principal Component Analysis (PCA)

Entire books have been written about PCA, and the closely related subject if *independent component analysis*, or ICA (e.g. [2] or [1]). We will not explore ICA in this assignment. Basically, PCA is a special case of ICA in which only second-order statistics (Gausian behavior) is considered. A decent tutorial on PCA is provided by [4].

PCA is simply described as "diagonalizing the covariance matrix". What does diagonalizing a matrix mean in this context? It simply means that we need to find a non-trivial linear combination of our original variables such that the covariance matrix is diagonal. When this is done, the resulting variables are uncorrelated, i.e. independent.

How do we compute the linear combinations of the original variables, called the 'principal components'? We provide a prescription here, taken from [4], leaving the justification of the prescription to, e.g., [4]:

1. Organize the data as an $m \times n$ matrix, $\mathbf{X}$,where m is the number of measurement types and n is the number of samples.
2. Subtract off the mean from each measurement type to produce a matrix, $\mathbf{X'} = \mathbf{X} - < \mathbf{X} >$.
3. Calculate the eigenvectors of the covariance of the matrix $\mathbf{X'}$, i.e. $\mathbf{C_{X'}} = \frac{1}{n}\mathbf{X'}\mathbf{X'}^T$.

What is an eigenvector? An eigenvector of a matrix, $\mathbf{A}$ is a vector $\vec{x}$ such that

$$\mathbf{A}\vec{x} \; = \; \lambda\vec{x}$$

i.e. multiplication of the eigenvector of a matrix by the matrix returns the eigenvector times a constant, called the 'eigenvalue'. If an eigenvalue computed in PCA is large, then the corresponding eigenvector (principal component) is important in describing the underlying data dependencies. If the eigenvalue is near zero, that principal component is relatively unimportant and the data depend primarily on fewer components

than there are measurement types. A geometric interpretation of the eigenvalues and eigenvectors of the covariance matrix is illustrated in Figure 1.

We cannot go into a full discussion of eigenvectors and matrices here. Read [4] or [5] for additional explanation. For our purposes here, we merely need the machinery of eigenvectors and covariance matrices in order to find linear combinations of our original variables that are uncorrelated. Fortunately, such machinery is available as python modules.

The Assignment

1. Read/skim at least sections I-V of [4].
2. Write some general python code to compute principle components of a set of measurements. Use the modules: `numpy.cov` and `numpy.linalg.eig` (or one of the variants of eig). Done correctly using numpy modules, this code can require very few lines.
3. Try your code on a simple simulated data set. Start by generating a data set x_i, y_i where x_i is linearly dependent on y_i. Include errors in the measurement of x_i and y_i. What are the principal components?
4. Now simulate a higher dimensional data set. You can invent your own problem, or use the three camera problem (with errors) discussed in [4].

References

[1] A. Hyvarinen, J. Karhunen, and E. Oja, Independent Component Analysis (2001).

[2] I.T. Jolliffe, Principal Component Analysis (2002)

[3] W.H. Press et al., Numerical Recipes, various editions beginning in 1988.

[4] J. Shiens, A Tutorial on Principal Component Analysis, `http://arxiv.org/pdf/1404.1100.pdf` (2009).

[5] D.S. Sivia, Data Analysis: A Bayesian Tutorial (2006)