KLT applications and a Fortran code

*Stephane Dumas*

KLT is a modern data analysis tool—a black box that is widely used but poorly understood. It is also known as Principal Component Analysis (PCA), Singular Value Decomposition (SVD), Singular Spectral Analysis (SSA), and Pisarenko's Method. Applications of the Karhunen–Loève Transform (KLT) are numerous in several fields of research, as this short list illustrates:

- In biology, it is used to classify bacteria by means of surface-enhanced Raman spectroscopy.
- In genetics, it is used to analyze and classify genes.
- In geology, it is used to study earthquakes. Geophysicists have used SSA to analyze a wide variety of time series such as solar oscillations, precipitation, stream flow and sea surface temperature, the chemical constituents of ice cores, global temperature, magnetosphere dynamics, and suspended sediment concentration in an estuary.
- In mathematics, it is used to study chaos theory.
- In climatology, it is used to study the fluctuation of temperature over area and time.
- In astrophysics, KLT methods are used in the search for acoustic oscillations of the Sun. They are also used to classify variable stars (3,200 stars according to 51 features).

1 Stephane Dumas is a Canadian physicist working for the Canadian government in the field of modeling and simulation. He has worked with simulated integrated optical devices for a few years. His expertise covers such subjects as physics, mathematics, astronomy, SETI, cryptography, the theory of information/communication/data compression. He is currently the President of the Quebec Center of the Royal Astronomical Society of Canada (RASC) and is a member of the Society of Amateur Radio Astronomers (SARA) and of the Planetary Society. He is also the SETI League Coordinator for east Canada. At the moment he is doing a master’s degree in physics/astrobiology (see [http://www3.sympatico.ca/stephane_dumas/CETI/biographies.html](http://www3.sympatico.ca/stephane_dumas/CETI/biographies.html)).
KLT’s capacity to predict and fill gaps in missing data will not be discussed in this chapter.

28.1 THE EIGENPROBLEM

The core of KLT is its capacity to solve the eigenproblem. An $N \times N$ matrix $A$ is said to have an eigenvector $\mathbf{x}$ and a corresponding eigenvalue $\lambda$ if

$$A\mathbf{x} = \lambda \mathbf{x}$$  \hspace{1cm} (28.1)

which can be solved if and only if

$$\det(A - \lambda I) = \begin{vmatrix} A_{11} - \lambda_1 & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} - \lambda_2 & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} - \lambda_n \end{vmatrix} = 0$$  \hspace{1cm} (28.2)

Suppose that we have a series of $\mathbf{x}$ composed of $p$ random variables and that we are interested in analyzing the structure of the covariance or correlation between $p$ variables. If $p$ is small or the structure is simple, the solution may be easy. However, this is not often the case and we end up with a large $p$ and a complex structure. Working in the eigenspace corresponding to this dataset gives us the capability to manipulate a smaller dataset. Figure 28.1 illustrates this for $p = 2$.

Matrix $A$ is the covariance, or correlation, matrix of the dataset described by the vectors $\mathbf{x}_i$.

28.1.1 Correlation vs. covariance

Variance is a measure of how far a set of numbers is spread out. The covariance of two datasets, $X$ and $Y$, is given in equation (28.3)

$$\sigma_{XY}^2 = \sum_{i=1}^{N} (X_i - \overline{X})(Y_i - \overline{Y}) / N$$  \hspace{1cm} (28.3)

where $X_i$ and $Y_i$ are the $i$-th element of their respective vector, and $\overline{X}$ and $\overline{Y}$ are the expected value of each vector (or mean value).

Covariance is sensitive to the scale of measurement adopted and is not a good measure of coherence—being multiplied by any scale factor introduced in one variable or another. For example, if $X$ and $Y$ are measured in inches, covariance will be 144 times what it would be if they were measured in feet. Furthermore, covariance provides little information on data if the variance of some variables is too large compared with others.

To obtain a measure of coherence that does not have this defect, the correlation coefficient is used (equation 28.4):

$$\rho_{X,Y} = \frac{E[(X - \overline{X})(Y - \overline{Y})]}{\sigma_X \sigma_Y}$$  \hspace{1cm} (28.4)

Correlation is linked to convolution. The correlation of two functions, denoted