Appendix A

Statistical notes

This appendix is an attempt to describe the main features of the some statistical techniques used in this study. Some of them (e.g. variance, confidence interval) which are largely used in sciences that use samples, measurements, and all sort of numerical data are described here again, as being a support to the following "more complex" techniques. The complete understanding of the mathematical definitions seems to be the only antidote to the many confusions that can arise, for example, in discussion of Empirical Orthogonal Functions (EOF) and Single Value Decomposition (SVD).

It is not the intention to present a complete description of those statistical analyses using demonstrations and theorems, but a summary capable to evidence their main differences, where they can exist.

Nomenclature, symbols, and their respective meanings, which are used in this description are shown in the following list, or described as the text flows on.

$\mathbf{X}$	= {X[m,n]}	data matrix, consisting of set the set of all data vectors, one vector per column ($m = 1 \ldots M, n = 1 \ldots N$)
N	...	the number of columns vectors in the data set
M	...	the number of elements in each column vector
L	...	the number of dimensions in the dimensionally reduced subspace, $1 \leq L \leq M$
$\mathbf{C}$	= {C[m,k]}	covariance matrix
$\mathbf{R}$	= {R[m,k]}	correlation matrix
$\mathbf{V}$	= {V[m,k]}	matrix consisting of the set of all eigenvectors of $\mathbf{C}$, one eigenvector per column
$\mathbf{D}$	= {D[m,k]}	diagonal matrix consisting of the set of all eigenvalues of $\mathbf{C}$ along its principal diagonal, and 0 for all other elements
$\mathbf{W}$	= {W[m,k]}	matrix of basis vectors, one vector per column, where each basis vector is one of the eigenvectors of $\mathbf{C}$, and where the vectors in $\mathbf{W}$ are a sub-set of those in $\mathbf{V}$

A.1 Variance

The variance of a random variable is a measure of its statistical dispersion, indicating how far from the expected value its values typically are. The variance of a random variable is the square of its standard deviation.

If $\mu = \mathrm{E}(x)$ is the expected value (mean) of the random variable x, then the variance is

$$var(x) = \mathrm{E}((x - \mu)^2) \tag{A.1}$$

That is, it is the expected value of the square of the deviation of x from its own mean - it is the *mean squared deviation*. The variation of random variable x is typically designated as $var(x)$, σ_x^2, or simply σ^2.

If a distribution does not have an expected value, it does not have a variance either. The converse is not true: there are distributions for which the expected values exists, but the variance does not.

A.2 Confidence intervals (CI)

A confidence interval is an interval in which a measurement or trial falls corresponding to a given probability. Usually, the confidence interval of interest is symmetrically placed around the mean.

If independent samples are taken repeatedly from the same population, and a confidence interval calculated for each sample, then a certain percentage (confidence level) of the intervals will include the unknown population parameter. If this parameter is the mean, the width of the confidence interval gives us some idea about how uncertain we are about the mean.

Confidence intervals are more informative than the simple results of hypothesis tests (where we decide "reject" or "do not reject") since they provide a range of plausible values for the unknown parameter.

For a normal distribution, the probability that a measurement falls within n standard deviations ($n\sigma$) of the mean μ (i.e., within the interval $[\mu - n\sigma, \mu + n\sigma]$) is given by

$$P(\mu - n\sigma < x < \mu + n\sigma) \equiv \frac{1}{\sigma\sqrt{2\pi}} \int_{\mu-n\sigma}^{\mu+n\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \tag{A.2}$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{\mu}^{\mu+n\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \tag{A.3}$$

Let $u \equiv \frac{x-\mu}{\sqrt{2}\sigma}$, so $du = \frac{dx}{\sqrt{2}\sigma}$. Then,

$$P(\mu - n\sigma < x < \mu + n\sigma) = \frac{2}{\sigma\sqrt{2\pi}}\sqrt{2}\sigma \int_0^{x/\sqrt{2}} e^{-u^2} du \tag{A.4}$$

$$= \frac{2}{\pi} \int_0^{x/\sqrt{2}} e^{-u^2} du = \mathrm{erf}\left(\frac{n}{\sqrt{2}}\right) \tag{A.5}$$

where $\mathrm{erf}(x)$ is the called erf function. The following table summarizes the probabilities $P(\mu - x_n < x < \mu + x_n)$ that measurements from a normal distribution fall within $[\mu - x_n, \mu + x_n]$ for $x_n = n\sigma$ with small values of n.

x_n	$P(\mu - x_n < x < \mu + x_n)$
σ	0.6826895
2σ	0.9544887
3σ	0.9973002
4σ	0.9999366
5σ	0.9999994

Conversely, to find the probability-P confidence interval centered about the mean for a normal distribution in units of σ, solve equation (A.5) for n to obtain

$$n = \sqrt{2}\,\mathrm{erf}^{-1}(P) \tag{A.6}$$

where erf^{-1} is the inverse erf function.

The following table then gives the values of x_P such that $[\mu - x_P, \mu + x_P]$ is the probability-P confidence interval for a few representatives values of P.

P	x_P
0.800	1.28155σ
0.900	1.64485σ
0.950	1.95996σ
0.990	2.57583σ
0.995	2.80703σ
0.999	3.29053σ

A.3 Empirical Orthogonal Functions (EOF)

In statistics and signal processing, the method of Empirical Orthogonal Functions (EOF) is a decomposition of a signal or data set in terms of orthogonal basis functions which are determined from the data. The k_{th} basis function is chosen to be orthogonal to the basis functions from the first through $k-1$, and to minimize the residual variance. That is, the basis functions are chosen to be different from each other, and to account for as much variance as possible. Thus this method has much in common with the method of kriging in geostatistics, and Gaussian process models.

The method of EOF is similar in spirit to harmonic analysis, but harmonic analysis typically uses predetermined orthogonal functions, for example, sine and cosine functions at fixed frequencies. In some cases the two methods may yield essentially the same results.

The basis functions are typically found by computing the eigenvectors of the covariance matrix of the data set. This is the same as performing Principal Components Analysis (PCA) on the data.

Assuming zero empirical mean (the empirical mean of the distribution has been subtracted away from the data set), the principal component $\mathbf{w}_1$ of the dataset $\mathbf{x}$ can be defined as:

$$\mathbf{w}_1 = argmax\mathbb{E}\{(\mathbf{w}^t\mathbf{x})^2\} \qquad \text{with } \|\mathbf{w}\| = 1 \tag{A.7}$$

where $\mathbb{E}$ is the expected value operator; for the first $k-1$ components, the k_{th} component can be found by subtracting the first $k-1$ principal components from $\mathbf{x}$:

$$\hat{\mathbf{x}}_{k-1} = \mathbf{x} - \sum_{i=1}^{k-1} \mathbf{w}_i\mathbf{w}_i^t\mathbf{x} \tag{A.8}$$

and by substituting this as the new dataset to find a principal component in

$$\mathbf{w}_k = argmax\mathbb{E}\{(\mathbf{w}^t\hat{\mathbf{x}}_{k-1})^2\} \qquad \text{with } \|\mathbf{w}\| = 1 \tag{A.9}$$

The EOF transform is therefore equivalent to finding the singular value decomposition of the data matrix $\mathbf{X}$,

$$\mathbf{X} = \mathbf{W}\sum\mathbf{V}^t \tag{A.10}$$

and then obtaining the reduced-space data matrix $\mathbf{Y}$ by projecting $\mathbf{X}$ down into the reduced space defined by only first L singular vectors, $\mathbf{W}_L$:

$$\mathbf{Y} = \mathbf{W}_L^t\mathbf{X} = \sum{}_L\mathbf{V}_L^t \tag{A.11}$$

The matrix $\mathbf{W}$ of singular vectors of $\mathbf{X}$ is equivalently also the matrix $\mathbf{W}$ of eigenvectors of the matrix of observed covariances $\mathbf{C} = \mathbf{X}\mathbf{X}^t$,

$$\mathbf{X}\mathbf{X}^t = \mathbf{W}\sum{}^2\mathbf{W}^t \tag{A.12}$$

By finding the eigenvalues and eigenvectors of the covariance matrix, the eigenvectors with the largest eigenvalues correspond to the dimensions that have the strongest correlation in the dataset.

A.4 Single Value Decomposition (SVD)

The Single Value Decomposition (SVD) method can be thought as a generalization to rectangular matrices of the diagonalization of the square symmetric matrix (like in EOF analysis).

The SVD of the cross-covariance matrix yields two spatially sets of singular vectors (spatial patterns analogous to the eigenvectors or EOF's, but one for each variable) and a set of singular values associated with each pair of vectors (analogous to the eigenvalues). Each pair of spatial patterns describe a fraction of the Square Covariance (SC) between the two variables. The first pair of patterns describes the largest fraction of the SC and each succeeding pair describes a maximum fraction of the SC that is unexplicated by the previous pair. The correlation value (r) between the k_{th} expansion coefficient of the two variables indicates how strongly related the k_{th} coupled patterns are.

As SVD is an statistical analysis of the two fields, the symbols and respective meanings already described, here, will be applied to matrices $\mathbf{X}$ (={X[m,n]} and $\mathbf{Y}$ (={Y[m,n]}.

Assuming the time mean of the the matrices has been removed, the covariance matrix is formed by

$$\mathbf{C} = \mathbf{X}^t\mathbf{Y} \tag{A.13}$$

If the matrices $\mathbf{X}$ and $\mathbf{Y}$ were normalized by their respectives standard variation, the result would be the cross-correlation matrix, rather than cross-covariance matrix.

The single value decomposition is performed on $\mathbf{C}$,

$$\mathbf{C} = \mathbf{U}\mathbf{D}\mathbf{V}^t \tag{A.14}$$

where the singular vectors of $\mathbf{X}$ are the column of $\mathbf{U}$, the singular vectors of $\mathbf{Y}$ are the column of $\mathbf{V}$, and $\mathbf{D}$ is the diagonal matrix.

Like EOF's, these patterns represent standing oscillations in the data fields.

Each mode of variability oscillates in time - the expansion coefficients - are calculated by,

$$\mathbf{A} = \mathbf{X}\mathbf{U} \tag{A.15}$$

$$\mathbf{B} = \mathbf{Y}\mathbf{V} \tag{A.16}$$

where, the columns of $\mathbf{A}$ and $\mathbf{B}$ contain the expansion coefficients of each mode. Since $\mathbf{U}$ and $\mathbf{V}$ are both orthogonal, the reconstruction of the data matrices can be done using $\mathbf{X} = \mathbf{A}\mathbf{U}^t$ and $\mathbf{Y} = \mathbf{B}\mathbf{V}^t$.

If $l_k = L(k,k)$ is the k_{th} singular value ($L = \sum l(k,k)$), the Fraction of Square Covariance (SCF) explained by the corresponding singular vector $\vec{u}_k$ and $\vec{v}_k$ is given by

$$SCF_k = \frac{l_k^2}{\sum l_k^2} \tag{A.17}$$

The computing of the SCF for each singular value allows to decide how many we want to keep.

A.5 Wavelets

In signal analysis, there are a number of different functions one can perform on that signal in order to translate it into different forms that are more suitable for different applications. The most popular function is the Fast Fourier Transform (FFT) that converts a signal from time *versus* amplitude to frequency *versus* amplitude. This transform is useful for many applications, but it is not based in time. To combat this problem, mathematicians came up with the Short Term Fourier Transform (STFT) which can convert a signal to frequency *versus* time. Unfortunately, this transform also has its shortcomings mostly that it cannot get decent resolutions for both high and low frequencies at the same time.

So how can a signal be converted and manipulated while keeping resolution across the entire signal and still be based in time? This is where wavelets come into play. Wavelets are finite windows through which the signal can be viewed. In order to move the window about the length of the signal, the wavelets can be translated about time in addition to being compressed and widened.

Wavelet transforms are broadly classified into the Discrete Wavelet Transform (DWT) and the Continuous Wavelet Transform (CWT). The principal difference between the two is the continuous transform operates over every possible scale and translation whereas the discrete uses a specific subset of all scale and translation values.

All wavelet transforms may be considered to be forms of time-frequency representation and are, therefore, related to the subject of harmonic analysis. Almost all practically useful discrete wavelet transforms make use of filterbanks containing finite impulse response filters. The wavelets forming a CWT are subject to Heisenberg's uncertainty principle and, equivalently, discrete wavelet bases may be considered in the context of other forms of the uncertainty principle.

The CWT is formally written as:

$$\gamma(s,\tau) = \int f(t)\Psi^*_{s,\tau}(t)dt \tag{A.18}$$

where * denotes complex conjugation. This equation shows how a function $f(t)$ is decomposed into a set of basis functions $\Psi_{s,\tau}(t)$, called the wavelets. The variables s and τ, scale and translation, are the new dimensions after the wavelet transform. For completeness sake the next equation gives the inverse wavelet transform

$$f(t) = \iint \gamma_{s,\tau}\Psi_{s,\tau}(t)d\tau ds \tag{A.19}$$

The wavelets are generated from a single basic wavelet $\Psi(t)$, the so-called *mother wavelet*, by scaling and translation:

$$\Psi_{s,\tau}(t) = \frac{1}{\sqrt{s}}\Psi\left(\frac{t-\tau}{s}\right) \tag{A.20}$$

In A.20 s is the scale factor, τ is the translation factor and the factor $s^{-1/2}$ is for energy normalization across the different scales. For the CWT, the pair (s,τ) varies over the full half-plane $\mathbb{R}_+ \times \mathbb{R}$; for the discrete WT this pair varies over a discrete subset of it, which is also called *affine group*.

It is important to note that in A.18, A.19 and A.20 the wavelet basis functions are not specified. This is a difference between the wavelet transform and the Fourier transform, or other transforms. The theory of wavelet transforms deals with the general properties of the wavelets and wavelet transforms only.

The most important properties of wavelets are the admissibility and the regularity conditions and these are the properties which gave wavelets their name. It can be that square integrable functions $\psi(t)$ satisfying the admissibility condition,

$$\int \frac{|\Psi(\omega)|^2}{|\omega|}d\omega < +\infty \tag{A.21}$$

In A.21 $\Psi(\omega)$ stands for the Fourier transform of $\psi(t)$.

Being in this space ensures that one can formulate the conditions of zero mean and square norm one:

$$\int \psi(t)dt = 0 \tag{A.22}$$

is the condition for zero mean, and

$$\int |\psi(t)|^2 \, dt = 1 \tag{A.23}$$

is the condition for square norm one. In other words, $\psi(t)$ must be a wave.

For $\psi(t)$ to be a wavelet for the continuous wavelet transform, the mother wavelet must satisfy an admissibility criterion in order to get a stably invertible transform. This condition - the admissibility criterion - is related to the application which the wavelet technique will be used for.