

Chapter 2. Background on Orthogonal Functions and Covariance

The purpose of this chapter is to present some basic mathematics and statistics that will be used heavily in subsequent chapters. The organization of the material and the emphasis on some important details peculiar to the geophysical discipline should help the reader.

2.1 Orthogonal Functions

Two functions f and g are defined to be orthogonal on a domain S if

$$\int_S f(s) g(s) ds = 0 \quad (2.1)$$

where s is space, 1, 2 or more dimensions, and the integral is taken over S . Since we will work with data observed at discrete points and with large sets of orthogonal functions we redefine and extend a discrete version of (2.1) as

$$\begin{aligned} \sum_{s=1}^{n_s} e_k(s) e_m(s) &= 0 && \text{for } k \neq m \\ &= \text{positive} && \text{for } k = m \\ &= 1 \text{ for } && k = m; \text{ orthonormal} \end{aligned} \quad (2.2)$$

The summation is over $s = 1$ to n_s , the number of (observed) points in space. The $e_k(s)$ are basis functions, orthogonal to all $e_m(s)$, $k \neq m$. The functions $e_k(s)$ are said to be orthonormal when the rhs in (2.2) is either zero or unity. For non-equal area data, such as on a latitude/longitude grid, or for irregular station data, the summation involves a weighting factor $w(s)$, i.e.

$$\sum_s w(s) e_k(s) e_m(s), \quad 1 \leq s \leq n_s$$

where $w(s)$ is related to the size of the area each data point represents. Except when noted otherwise, $w(s)$ will be left off throughout the book for simplicity. However this detail is not always trivial.

The $e_m(s)$ can be thought of as vectors consisting of n_s components. In that context the type of product in (2.2) is often referred to as inner or dot product $\mathbf{e}_k \cdot \mathbf{e}_m$.

A major convenience of a set of orthogonal functions $e_m(s)$ satisfying Eq(2.2) is functional representation of data, that is to say, for any discrete $f(s)$, say a map of Mean Sea Level Pressure on a grid, we can write:

$$f(s) = [f] + \sum_{m=1}^M \alpha_m e_m(s) \quad 1 \leq s \leq n_s \quad (2.3)$$

where $[f]$ is the spatial mean, α_m is the expansion coefficient, and M is at most n_s-1 . In the context of (2.3) it is clear why the $e_m(s)$ are called a basis. The equal sign in (2.3) only applies when the basis is complete.

For now, let's consider $[f]$ to be zero (or removed from the data as per $f^* = f - [f]$, then drop the $*$). We now define the spatial variance (SV) as:

$$SV = \sum_s f^2(s) / n_s \quad (2.4)$$

and note that SV, for orthonormal $e_m(s)$, can also be written as

$$SV = \sum_{m=1}^M \alpha_m^2 \quad (2.5)$$

which establishes the link between variance in physical (2.4) and spectral space (2.5). The equal sign in (2.5) only applies when M is the required minimum value, which could be as high as n_s-1 . As per (2.5) each basis function 'explains' a non-overlapping part of the variance, or (put another way) contributes an independent piece of information contained in $f(s)$. α_m^2 is the classical Fourier power spectrum if the e 's are sin/cos on the domain. In that context (2.5) is known as Parseval's theorem, and counting the sin/cos pair as one mode, M is at most $n_s/2$.

When e is known, then α_m can be easily calculated as:

$$\alpha_m = \sum_s f(s) e_m(s) \quad 1 \leq m \leq M \quad (2.6)$$

i.e one finds the projection coefficients α_m by simply projecting the data $f(s)$ onto the e 's on the points where they coincide. (2.6) is only valid when the e 's have unit length (orthonormal). If not divide rhs of (2.6) by $\sum e_m^2(s)$. Since orthogonal functions do not compete for the same variance, (2.6) can be evaluated for each m separately, and in any order. The ordering $m = 1$ to M is quite arbitrary. When sin/cos is used the low m values correspond to the largest spatial scales. But

ordering by amplitude (or equivalently: explained variance) makes a lot of sense too. If one were to truncate to N functions ($N < M$) the explained variance (EV) is given by (2.5), but summed only over $m = 1$ to N , so ordering orthogonal functions by EV is natural and advantageous for many purposes.

When the α_m and $e_m(s)$ are both known, the data in physical space can be retrieved via (2.3). When the α_m and $f(s)$ are known, a hypothetical situation, $e_m(s)$ cannot be retrieved in general.

It is necessary to reflect where the physical units reside in (2.3)-(2.6). If $f(s)$ is say pressure, in millibar (mb), the α_m assume the unit mb, and the SV has the unit mb^2 . The $e_m(s)$ are dimensionless, and it is convenient for $e_m(s)$ to have unit length, such that the numerical values of α_m and SV make sense in physical units mb and mb^2 respectively.

The above was written for unspecified orthogonal functions. There is an infinity of orthogonal functions. Analytical orthogonal functions include the $\sin ms$, $\cos ms$ pair (or $\cos m(s - \epsilon)$) as the most famous of all - in this case (2.6) is known as the Fourier transform. Legendre polynomials, and spherical harmonics (a combination of \sin/\cos in the east west direction and Legendre in the north south direction) are also widely used in meteorology, starting with Baer and Platzman(1961). But the list includes Bessel, Hermite, Chebyshev, Laguerre functions etc etc. Why prefer one function over the other? There are issues of taste, preference, accuracy, theory, scaling, tradition, ...convenience. We mention here specifically the issue of 'efficiency'. For practical reasons one may have to truncate, in (2.3), to much less than M . If only N orthogonal functions are allowed ($N < M$) it matters which orthogonal functions will explain the most variance. The remainder is relegated to unresolved scales, unexplained variance and truncation error. {A different type of efficiency has to do with the speed by which transforms like (2.3) and (2.6), can be executed on a computer.}

One can easily imagine non-analytical orthogonal functions. Examples include zeros at all points in space except one - this makes for a set of orthogonal functions equal to n_s . An

advantage of analytical functions is that there is theory and a wealth of information. Moreover, analytical functions suggest values and meaning in between the data points. This makes differentiation and interpolation easy. Nevertheless, ever since Lorenz(1956) non-analytical empirical orthogonal functions (no more than a set of numbers on a grid) are highly popular in Meteorology as a device to “let the data speak”. Moreover, these empirical orthogonal functions (EOF) are the most efficient in explaining variance for a given data set.

We have written the above, (2.1) thru (2.6), for space s . One can trivially replace space s by time t ($1 \leq t \leq n_t$) and keep the exact same equations showing t instead of s . However, if one has a space-time data set, $f(s, t)$, as one typically does, the situation becomes quickly more involved. For instance, choosing orthogonal functions, as before, in space, (2.3) can be written as:

$$f(s, t) = [f(s, t)] + \sum_{m=1}^M \alpha_m(t) e_m(s) \quad 1 \leq s \leq n_s \quad 1 \leq t \leq n_t \quad (2.7)$$

where the projection coefficients and the space mean are a function of time.

While choosing orthogonal functions in time would lead to:

$$f(s, t) = \langle f(s, t) \rangle + \sum_{m=1}^M \alpha_m(s) e_m(t) \quad 1 \leq s \leq n_s \quad 1 \leq t \leq n_t \quad (2.7a)$$

where the time mean, and the projection coefficients are a function of space. Time mean is denoted by $\langle \rangle$.

Eq (2.7) and (2.7a) lead, in general, to drastically different looks of the same data set, sometimes referred to as T-mode and S-mode analysis. There is, however, one unique set of functions for which this space-time ambiguity can be removed: EOFs, i.e $\alpha_m(t)$ in (2.7) is the same as $e_m(t)$ in (2.7a), and $\alpha_m(s)$ in (2.7a) is the same as $e_m(s)$ in (2.7). We phrase this as follows: For EOFs one can reverse (interchange) the roles of time and space. This does, however, require a careful treatment of the space-time mean, see 2.3.

2.2 Correlation and Covariance

Here we discuss elementary statistics in one dimension first (time) and use as a not-so-arbitrary example of two times series, $D(t)$ the seasonal mean pressure at Darwin in Australia near a center of action of a phenomenon called ENSO, and seasonal mean temperature $T(t)$ at some far away location in mid-latitude: $T(t)$, $1 \leq t \leq n_t$, where t is a year index, 1948..2005 say; $n_t=58$). One can define the time mean of D as

$$\langle D \rangle = \sum_{t=1}^{n_t} D(t) / n_t, \quad (2.8)$$

and T similarly has a time average $\langle T \rangle$. We now formulate departures from the mean, often called anomalies:

$$\begin{aligned} D'(t) &= D(t) - \langle D \rangle \\ &\text{for all } t \quad (2.9) \\ T'(t) &= T(t) - \langle T \rangle \end{aligned}$$

The covariance between D and T is given by

$$\text{cov}_{DT} = \sum D'(t) T'(t) / n_t \quad (2.10)$$

The physical units of covariance in this example are (mb °C). The variance is given by

$$\text{var}_D = \sum D'(t) D'(t) / n_t, \quad (2.11)$$

and similarly for var_T . The standard deviation is $\text{sd}_D = \sqrt{\text{var}_D}$, and the correlation between D and T is:

$$\rho = \text{cov}_{DT} / (\text{sd}_D \cdot \text{sd}_T) \quad (2.12)$$

The correlation is a non-dimensional quantity, $-1 \leq \rho \leq 1$.

Calling D the predictor, and T the predictand, there is a regression line $T'_{\text{fcst}} = b D'$ which, over $t = 1, n_t$, explains ρ^2 % of the variance in T . The subscript 'fcst' designates a forecast for T given D . The regression coefficient b is given by

$$b = \rho \text{sd}_T / \text{sd}_D \quad (2.13)$$

The correlation has been used widely in tele-connection studies to gauge relationships or

‘connection’ between far away points. The suggestion of a predictive capability is more explicit when $D(t)$ and $T(t)$, while both time series of length n_t , are offset in time, D leading T . If D and T are the same variable at the same location, but offset in time, the above describes the first steps of an auto-regressive forecast system. Note also that the correlation is used frequently for verification of forecasts against observations.

In many texts the route to (2.12) is taken via ‘standardized’ variables, i.e. using (2.8), (2.9) and (2.11) we define:

$$D''(t) = (D(t) - \langle D \rangle) / sd_D \quad (2.9a)$$

which have no physical units. Given these standardized anomalies, correlation and covariance become the same and are given by

$$\rho = cov_{D'',T''} = \sum_t D''(t) T''(t) / n_t \quad (2.12a)$$

So, the correlation does not change wrt (2.12), but the covariance does change relative to (2.10) and loses its physical units.

It should be trivial to replace time t by space s in (2.8) - (2.12) and define covariance or correlation in space in analogous fashion. Extending to both space and time, and using general notation, we have a data set $f(s, t)$ with a mean removed. The covariance in time between two points s_i and s_j is given by

$$q_{ij} = \sum_t f(s_i, t) f(s_j, t) / n_t \quad (2.14)$$

while the covariance in space between two times t_i and t_j is given by:

$$q_{ij}^a = \sum_s f(s, t_i) f(s, t_j) / n_s \quad (2.14a)^1$$

q_{ij} and q_{ij}^a are the elements of the two renditions of the all-important covariance matrices Q and Q^a - the superscript a stands for alternative. The n_s by n_s matrix Q measures ‘teleconnection’ between any two points s_i and s_j while the n_t by n_t matrix Q^a measures the similarity of two maps at any

¹ When a non-equal area grid is used the expression has to be adjusted in a calculation as follows:

$$q_{ij}^a = \sum_s w(s) f(s, t_i) f(s, t_j) / W \quad (2.14a)$$

where $w(s)$ represents the size of the area for each data point in space, and W is the sum (in space) of all $w(s)$. On the common lat-lon grid the weight is $\cos(\text{latitude})$.

two times t_i and t_j , a measure of analogy. These two features (teleconnection and analogy) are totally unrelated at first thought, but under the right definitions the eigenvalues of Q and Q^a are actually the same, such that the role of space and time can be thought of as reversible. One issue to be particularly careful about is the mean value that is removed from $f(s,t)$. This has strong repercussions in both sections 2.1 and 2.2.

2.3 Issues about removal of “the mean”.

An important but murky issue is that of forming anomalies. The general idea is one of splitting a datum into a part that is easy to know (some mean value we are supposed to know), and the remainder, or anomaly, which deals with variability around that mean and is a worthy target for prediction efforts. All attention is subsequently given to the anomaly. Should it be $f'(t) = f(t) - \langle f \rangle$ as in Eq (2.9) or $f'(t) = f(t) - \{f\}$, where $\{f\}$ is a reference value, not necessarily the sample time mean. This is a matter of judgement. Examples where this question arises:

- a) When the mean is known theoretically. (There may be no need to calculate a flawed mean from a limited sample)
- b) When forecasts are made of the type: “warmer than normal”, w.r.t. a normal which is based on past data by necessity at the time of issuance of the forecast.
- c) The widely used anomaly correlation in verification, see inset/appendix.
- d) In EOF calculations (a somewhat hidden problem)
- e) When the climatology is smoothed across calendar months, resulting in non-zero time mean anomalies at certain times of the year.

While no absolute truth and guidelines exist we here take the point of view that removal of a reference value, acting as an approximate time mean, is often the right course of action.

The removal of a space mean is not recommended. And $\sum_s f(s) \neq 0$ and $\sum_t f(t) \neq 0$ are acceptable. Removal of a calculated space mean is problematic. On planet earth, with its widely varying climate removing a space mean first makes little sense as it creates, for example,

anomalies warmer(colder) than average in equatorial(polar) areas.

Given a space-time data set $f(s, t)$ we will thus follow this practice.

1) Remove at each point in space a reference value $\{ f(s) \}$, i.e form anomalies as follows:

$$f'(s, t) = f(s, t) - \{ f(s) \} \quad (2.15)$$

In some cases and examples the $\{ \}$ reference will be the sample time mean, such that anomalies do sum up to zero over time and f' is strictly centered. But we do not impose such a requirement.

2) We do NOT remove the spatial mean of f' .

Under this practice we evaluate (2.14) and (2.14a). We furthermore note that under the above working definition the space-time variance is given by

$$STV = \sum_{s,t} f'^2(s,t) / (n_s n_t) \quad (2.16)$$

Exactly the same total variance to be divided among orthogonal functions (EOF or otherwise) calculated from either Q or Q^a . We acknowledge that some authors would take an additional space mean out when working with Q^a (because they feel they should require the sum of the anomalies in space to be zero). This, however, modifies the STV, and all information derived from Q^a would change. We do not recommend taking the space mean out. Especially on small domains, taking out the space mean of f' takes away much of the signal of interest. We thus work with anomalies that do not necessarily sum up to exactly zero in either time or space domain. This also means that variance and standard deviation, as defined in (2.11), (2.12) and (2.16) are augmented by the (usually small) offset from zero mean. When calculating EOFs the domain means get absorbed into one or more modes.

2.4 Concluding remarks

We have approached the representation of a data set $f(s, t) = [f(s, t)] + \sum \alpha_m(t) e_m(s)$ both by classical mathematical analysis theory and by basic statistical concepts that will allow calculation of orthogonal functions from a data set (mainly in Chapter 5). It may be a good idea to reflect on the commonality of section 2.1 and 2.2 and the juxtaposition of terminology. For

example, the inner product used to determine orthogonality (2.1) is the same as the measure for covariance (or correlation) in (2.10). Indeed a zero correlation is a sure sign of two orthogonal time series or two orthogonal maps. In both sections 2.1 and 2.2 we mentioned the notion of explained variance (EV), once for orthogonal functions, once for regression. We invited the reader to follow the physical units, and numerically the numbers should make sense when the basis is orthonormal. This task becomes difficult because the use of EOFs allows basis functions to be orthogonal in time and space simultaneously - both are basis function and projection coefficients all at the same time. We emphasized the ‘reversibility’ of time and space in some of the calculations. We finally spent some paragraphs on removing a mean, which, while seemingly a detail, can cause large differences in interpretation.

Inset: The anomaly correlation.

One of the most famous correlations in meteorology is the anomaly correlation used for verification. It has been in use at least since Miyakoda(1972). Imagine we have, as a function of latitude and longitude, a 500 mb height field $Z(s)$. The field is given on a grid, cosine weights not shown. How to correlate two 500 mb height maps?, like for instance Z_{fcst} and Z_o , a set of paired forecast/observed maps, a challenge faced each day by operational weather forecast centers. The core issue is forming anomalies or splitting Z into a component we are supposed to know (no reward for forecasting it right) and the much tougher remainder. A definitely ‘wrong way’² would be to form anomalies by $Z^*(s) = Z(s) - [Z]$ where $[Z] = \sum Z(s) / n_s$, a space mean. Removal of a calculated space mean is problematic. On planet earth with its widely varying climate removing a space mean makes little sense as it creates, for example, anomalies warmer/colder than average in

²There is no absolutely right or wrong in these issues. In a 2D homogeneous turbulence experiment time and space means would be expected to be the same, so taking out the space mean may be quite acceptable.

equatorial/polar areas. Even a terrible forecast would have a high anomaly correlation.

A ‘better way’¹ is to form anomalies is via $Z'(s) = Z(s) - Z_{\text{climo}}(s)$ and likewise $Z'_{\text{fcst}}(s) = Z_{\text{fcst}}(s) - Z_{\text{climo}}(s)$ where $Z_{\text{climo}}(s)$ is based on a long multi-year observed data set $Z(\lambda, \phi, \text{pressure level, day of the year, hour of the day})$. The anomaly correlation is then given by

$$AC = \frac{\sum Z'_{\text{fcst}}(s) Z'_o(s) / n_s}{[\sum Z'_{\text{fcst}}(s) Z'_{\text{fcst}}(s) / n_s \cdot \sum Z'_o(s) Z'_o(s) / n_s]^{1/2}} \quad (2.17)$$

where summation is over space. $-1 \leq AC \leq 1$

Among the debatable issues: should $\sum Z'(s)$ be (made) zero?? There is no reason to do that, especially in verification, but some people feel that way. The removal of the space mean removes a potentially important component of the forecast from the verification process. This is especially true on small domains that may be dominated by anomalies of one sign. Note that $-1 \leq AC \leq 1$, like a regular correlation, even though the spatial means are not exactly zero and the two terms in the denominator are augmented versions of the classical notion variance.

Eq(2.17) is for a single pair of maps. When we have a large set of paired maps we sum the three terms in (2.17) over the whole set, then execute the multiplication and division.

End-end-end