Principal component analysis (PCA)

2.6 Scaling the PCs and eigenvectors [Book, Sect. 2.1.6]

Various options for scaling the PCs $\{a_j(t)\}$ and the eigenvectors $\{\mathbf{e}_j\}$. One can introduce an arbitrary scale factor α,

$$a'_j = \frac{1}{\alpha} a_j, \quad \mathbf{e}'_j = \alpha \mathbf{e}_j \,, \tag{1}$$

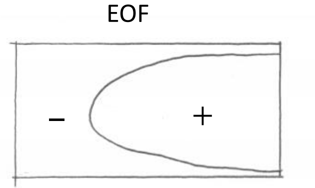
so that

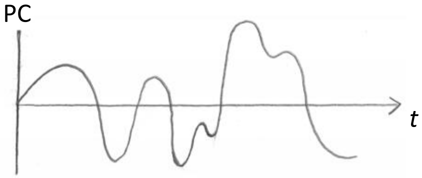
$$\mathbf{y} - \overline{\mathbf{y}} = \sum_j a_j \mathbf{e}_j = \sum_j a'_j \mathbf{e}'_j \,. \tag{2}$$

Thus $a_j(t)$ and $\mathbf{e}_j$ are defined only up to an arbitrary scale factor.

With $\alpha = -1$, one reverses the sign of both $a_j(t)$ and $\mathbf{e}_j$, which is often done to make them more interpretable.

Q3: Suppose the first PCA mode has the following EOF spatial pattern and PC time series:





You want the EOF to have positive anomalies on the left side instead of on the right side. What would you do to achieve this and what would the new PC look like?

Our choice for the scaling has so far been

$$\mathbf{e}_i^{\mathrm{T}}\mathbf{e}_j = \delta_{ij}\,, \tag{3}$$

which was the choice of Lorenz (1956).

Q4: Show that with $\mathbf{y} - \overline{\mathbf{y}} = \sum_{j=1}^{m} a_j \mathbf{e}_j$ and $\mathbf{e}_i^{\mathrm{T}} \mathbf{e}_j = \delta_{ij}$, the variance of the data $\mathbf{y}$ is contained in $\{a_j(t)\}$, with

$$\mathrm{var}(\mathbf{y}) = \mathrm{E}\left[\|\mathbf{y} - \overline{\mathbf{y}}\|^2\right] = \mathrm{E}\left[\sum_{j=1}^{m} a_j^2\right] . \tag{4}$$

Another common choice is Hotelling's original choice

$$a_j' = \frac{1}{\sqrt{\lambda_j}} a_j, \quad \mathbf{e}_j' = \sqrt{\lambda_j}\mathbf{e}_j \,, \tag{5}$$

whence

$$\mathrm{var}(\mathbf{y}) = \sum_{j=1}^{m} \lambda_j = \sum_{j=1}^{m} \|\mathbf{e}_j'\|^2 \,, \tag{6}$$

$$\mathrm{cov}(a'_i, a'_j) = \delta_{ij}\,. \tag{7}$$

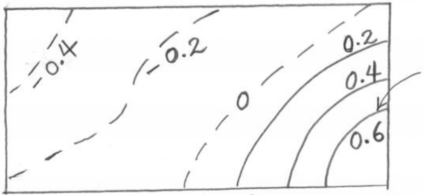
The variance of the original data is now contained in $\{\mathbf{e}_j\}$.

Regardless of the arbitrary scale factor, the PCA eigenvectors are orthogonal and the PCs are uncorrelated.

If PCA is performed on the *standardized* variables $\tilde{y}_l$, one can show that the correlation

$$\rho(a'_j(t), \tilde{y}_l(t)) = e'_{jl}\,, \tag{8}$$

the lth element of $\mathbf{e}'_j$ (Jolliffe, 2002, p.25).



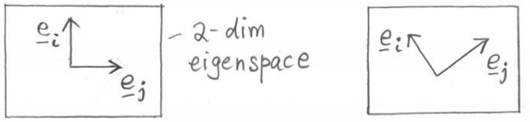
$a'_j(t)$ correlated at $\rho = 0.6$ with y'_ℓ at grid pt. ℓ.

Contour map of $\underline{e}'_j$

Hence the lth element of $\mathbf{e}'_j$ conveniently provides the correlation between the PC a'_j and the standardized variable $\tilde{y}_l$, which is a reason why Hotelling's scaling (5) is also widely used.

2.7 Degeneracy of eigenvalues [Book, Sect. 2.1.7]

A degenerate case arises when $\lambda_i = \lambda_j$, $(i \neq j)$. When two eigenvalues are equal, their eigenspace is 2-D, i.e. a plane in which any two orthogonal vectors can be chosen as the eigenvectors, i.e. the eigenvectors are not unique.



If l eigenvalues are equal, l non-unique orthogonal vectors can be chosen in the l-dimensional eigenspace.

E.g. a propagating plane wave,

$$h(x, y, t) = A\cos(ky - \omega t)\,, \tag{9}$$

which can be expressed in terms of two standing waves:

$$h = A\cos(ky)\cos(\omega t) + A\sin(ky)\sin(\omega t)\,. \tag{10}$$

If we perform PCA on $h(x, y, t)$, we get two modes with equal eigenvalues.

Q5: If we apply PCA to the propagating plane wave (9), what do the two EOF spatial patterns look like and what do the corresponding PCs look like?

As (10) is a PCA decomposition, with the two modes having the same amplitude A, hence the eigenvalues $\lambda_1 = \lambda_2$, and the case is degenerate.

Thus propagating waves in the data leads to degeneracy in the eigenvalues.

If one finds eigenvalues of very similar magnitudes from a PCA analysis, that implies near degeneracy and there may be propagating waves in the data.

In reality, noise in the data usually precludes $\lambda_1 = \lambda_2$ exactly.

When $\lambda_1 \approx \lambda_2$, the near degeneracy causes the eigenvectors to be rather poorly defined (i.e. very sensitive to noise in the data) (North et al., 1982).

2.8 A smaller covariance matrix [Book, Sect. 2.1.8]

Let the data matrix be

$$\mathbf{Y} = \begin{bmatrix} y_{11} & \cdots & y_{1n} \\ \cdots & \cdots & \cdots \\ y_{m1} & \cdots & y_{mn} \end{bmatrix}, \tag{11}$$

where m is the number of spatial points and n the number of time points. The columns of this matrix are simply the vectors $\mathbf{y}(t_1), \mathbf{y}(t_2), \ldots, \mathbf{y}(t_n)$. Assuming

$$\frac{1}{n} \sum_{i=1}^{n} y_{ji} = 0 \,, \tag{12}$$

i.e. the temporal mean has been removed, then

$$\mathbf{C} = \frac{1}{n} \mathbf{Y}\mathbf{Y}^{\mathrm{T}} \tag{13}$$

is an $m \times m$ matrix.

The theory of singular value decomposition (SVD) (Book Sect. 2.1.10) tells us that the nonzero eigenvalues of $\mathbf{Y}\mathbf{Y}^{\mathrm{T}}$ (an $m \times m$

matrix) are exactly the nonzero eigenvalues of $\mathbf{Y}^{\mathrm{T}}\mathbf{Y}$ (an $n \times n$ matrix).

The size of the two matrices are often very different. E.g. for global $5^\circ \times 5^\circ$ monthly sea level pressure data collected over 50 years, total number of spatial grid points is $m = 2592$ while number of time points is $n = 600$. Obviously, much easier to solve the eigen problem for the 600×600 matrix than that for the 2592×2592 matrix.

Hence, when $n < m$, considerable computational savings can be gained by first finding the eigenvalues $\{\lambda_j\}$ and eigenvectors $\{\mathbf{v}_j\}$ for the alternative covariance matrix

$$\mathbf{C}' = \frac{1}{n}\mathbf{Y}^{\mathrm{T}}\mathbf{Y}\,, \tag{14}$$

i.e.

$$\frac{1}{n}\mathbf{Y}^{\mathrm{T}}\mathbf{Y}\mathbf{v}_j = \lambda_j \mathbf{v}_j \,. \tag{15}$$

Since

$$\begin{aligned} \lambda_j \mathbf{Y}\mathbf{v}_j &= \mathbf{Y}\lambda_j \mathbf{v}_j = \mathbf{Y}\frac{1}{n}\mathbf{Y}^{\mathrm{T}}\mathbf{Y}\mathbf{v}_j \,, \\ (\frac{1}{n}\mathbf{Y}\mathbf{Y}^{\mathrm{T}})(\mathbf{Y}\mathbf{v}_j) &= \lambda_j(\mathbf{Y}\mathbf{v}_j) \,. \end{aligned} \tag{16}$$

This equation is easily seen to be of the form

$$\mathbf{C}\mathbf{e}_j = \lambda_j \mathbf{e}_j \,, \tag{17}$$

with

$$\mathbf{e}_j = \mathbf{Y}\mathbf{v}_j \,, \tag{18}$$

which means $\mathbf{e}_j$ is an eigenvector for $\mathbf{C}$.

Summary: solving the eigen problem for the smaller matrix $\mathbf{C}'$ yields the eigenvalues $\{\lambda_j\}$ and eigenvectors $\{\mathbf{v}_j\}$. The eigenvectors $\{\mathbf{e}_j\}$ for the bigger matrix $\mathbf{C}$ are then obtained from (18).

2.9 Temporal and spatial mean removal [Book, Sect. 2.1.9]

Given a data matrix $\mathbf{Y}$ as in (11), what type of mean are we trying to remove from the data?

We have removed the *temporal* mean, i.e. the average of the jth row, from each datum y_{ji}.

We could instead have removed the *spatial* mean, i.e. the average of the ith column, from each datum y_{ji}.

Which type of mean should be removed is very dependent on the type of data one has. For most applications, one removes the temporal mean.

For satellite sensed sea surface temperature data, the precision is much better than the accuracy. Also, the subsequent satellite image may be collected by a different satellite, which would have different systematic errors. So more appropriate to subtract the *spatial* mean of an image from each pixel (as was done in Fang and Hsieh, 1993).

Also possible to remove both the temporal and spatial means, by subtracting the average of the jth row and then the average of the ith column from each datum y_{ji}.

2.10 Singular value decomposition [Book, Sect. 2.1.10]

Instead of solving the eigen problem of the data covariance matrix $\mathbf{C}$, a computationally more efficient way to perform PCA is via *singular value decomposition* (SVD) of the $m \times n$ data matrix $\mathbf{Y}$ given by (11) (Kelly, 1988).

Without loss of generality, we can assume $m \geq n$, then the SVD theorem (Strang, 2005) says that

$$\mathbf{Y} = \mathbf{E}\mathbf{S}\mathbf{F}^{\mathrm{T}} = \underbrace{\overset{\mathbf{E}}{\boxed{\begin{array}{c:c} \mathbf{E}' & \\ m \times n & 0 \end{array}}}}_{m \times m} \underbrace{\overset{\mathbf{S}}{\boxed{\begin{array}{c} \mathbf{S}' \\ n \times n \\ \hdashline 0 \end{array}}}}_{m \times n} \underbrace{\overset{\mathbf{F}^{\mathrm{T}}}{\boxed{}}}_{n \times n} . \tag{19}$$

The $m \times m$ matrix $\mathbf{E}$ contains an $m \times n$ submatrix $\mathbf{E}'$— and if $m > n$, some zero column vectors.

The $m \times n$ matrix $\mathbf{S}$ contains the diagonal $n \times n$ submatrix $\mathbf{S}'$, and possibly some zero row vectors.
$\mathbf{F}^{\mathrm{T}}$ is an $n \times n$ matrix.
(If $m < n$, one can apply the above arguments to the transpose of the data matrix).

$\mathbf{E}$ and $\mathbf{F}$ are *orthonormal matrices*, i.e.

$$\mathbf{E}^{\mathrm{T}}\mathbf{E} = \mathbf{I}\,, \qquad \mathbf{F}^{\mathrm{T}}\mathbf{F} = \mathbf{I}\,, \tag{20}$$

where $\mathbf{I}$ is the identity matrix. The leftmost n columns of $\mathbf{E}$ contain the n *left singular vectors*, and the columns of $\mathbf{F}$ the n *right singular vectors*, while the diagonal elements of $\mathbf{S}'$ are the *singular values*.

Using (19) and (20),

$$\mathbf{C} = \frac{1}{n}\mathbf{Y}\mathbf{Y}^{\mathrm{T}} = \frac{1}{n}\mathbf{E}\mathbf{S}\mathbf{S}^{\mathrm{T}}\mathbf{E}^{\mathrm{T}}\,. \tag{21}$$

The matrix

$$\mathbf{S}\mathbf{S}^{\mathrm{T}} \equiv \mathbf{\Lambda} \tag{22}$$

is diagonal and zero everywhere, except in the upper left $n \times n$ corner, containing $\mathbf{S}'^2$.
Right multiply Eq.(21) by $n\mathbf{E}$, and using (22) and (20) give

$$n\mathbf{C}\mathbf{E} = \mathbf{E}\mathbf{\Lambda}\,, \tag{23}$$

where $\mathbf{\Lambda}$ contains the eigenvalues for the matrix $n\mathbf{C}$.

Instead of solving the eigen problem (23), we use SVD to get $\mathbf{E}$ from $\mathbf{Y}$ by (19). Eq.(23) implies that there are only n eigenvalues in $\mathbf{\Lambda}$ from $\mathbf{S}'^2$, and the eigenvalues $=$ (singular values)^2. As (23) and (17) are equivalent except for the constant n, the eigenvalues in $\mathbf{\Lambda}$ are simply $n\lambda_j$, with λ_j the eigenvalues from (17).

Similarly, for the other covariance matrix

$$\mathbf{C}' = \frac{1}{n}\mathbf{Y}^{\mathrm{T}}\mathbf{Y}\,, \tag{24}$$

$$\mathbf{C}' = \frac{1}{n}\mathbf{F}\mathbf{S}'^{2}\mathbf{F}^{\mathrm{T}}\,, \tag{25}$$

$$n\mathbf{C}'\mathbf{F} = \mathbf{F}\mathbf{S}'^{2}\,. \tag{26}$$

Hence the eigen problem (26) has the same eigenvalues as (23).

The PCA decomposition

$$\mathbf{y}(t) = \sum_{j} \mathbf{e}_j a_j(t)\,, \tag{27}$$

is equivalent to the matrix form

$$\mathbf{Y} = \mathbf{E}\mathbf{A}^{\mathrm{T}} = \sum_j \mathbf{e}_j \mathbf{a}_j^{\mathrm{T}} \,, \tag{28}$$

where the eigenvector $\mathbf{e}_j$ is the jth column in the matrix $\mathbf{E}$, and the PC $a_j(t)$ is the vector $\mathbf{a}_j$, the jth column in the matrix $\mathbf{A}$. Eqs.(19) and (28) yield

$$\mathbf{A}^{\mathrm{T}} = \mathbf{S}\mathbf{F}^{\mathrm{T}} \,. \tag{29}$$

From the SVD (19), we get the eigenvectors $\mathbf{e}_j$ from $\mathbf{E}$, and the PCs $a_j(t)$ from $\mathbf{A}$ in (29).

Can also left multiply (28) by $\mathbf{E}^{\mathrm{T}}$, and invoke (20) to get

$$\mathbf{A}^{\mathrm{T}} = \mathbf{E}^{\mathrm{T}}\mathbf{Y} \,. \tag{30}$$

Kelly (1988) pointed out that the SVD approach to PCA is at least twice as fast as the eigen approach.

2.11 Missing data [Book, Sect. 2.1.11]

Missing data produce gaps in data records. If the gaps are small, one can interpolate the missing values using neighbouring data. If the gaps are not small, then instead of

$$\mathbf{C} = \frac{1}{n}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\,, \tag{31}$$

(assuming the means have been removed from the data), one computes

$$c_{kl} = \frac{1}{n'}\sum_i{}' y_{ki}\, y_{il} \tag{32}$$

where the prime denotes that the summation is only over i with neither y_{ki} nor y_{il} missing— with a total of n' terms in the summation. The eigenvectors $\mathbf{e}_j$ can then be obtained from this new covariance matrix.

The principal components a_j cannot be computed from

$$a_j(t_l) = \sum_i e_{ji}\, y_{il}\;, \tag{33}$$

as some values of y_{il} are missing. Instead a_j is estimated (von Storch and Zwiers, 1999, Sect.13.2.8) as a least squares solution to minimizing $\mathrm{E}[\|\mathbf{y} - \sum a_j \mathbf{e}_j\|^2]$, i.e.

$$a_j(t_l) = \frac{\sum_i' e_{ji}\, y_{il}}{\sum_i' |e_{ji}|^2}\;, \tag{34}$$

where for a given value of l, the superscript prime means that the summations are only over i for which y_{il} is not missing.

PCA is also used to fill missing data. Suppose the data record can be divided into two parts, $\mathbf{Y}$ which contains no missing values, and $\tilde{\mathbf{Y}}$ which contains missing values. From (28), PCA applied to $\mathbf{Y}$ yields $\mathbf{E}$, which contains the eigenvectors $\mathbf{e}_j$. The PCs for $\tilde{\mathbf{Y}}$ are then computed from (34)

$$\tilde{a}_j(t_l) = \frac{\sum_i' e_{ji}\, \tilde{y}_{il}}{\sum_i' |e_{ji}|^2} \;. \tag{35}$$

The missing values in $\tilde{\mathbf{Y}}$ are filled in $\tilde{\mathbf{Y}}'$, where

$$\tilde{\mathbf{Y}}' = \mathbf{E}\tilde{\mathbf{A}}^{\mathrm{T}} \,, \tag{36}$$

where the jth column of $\tilde{\mathbf{A}}$ is given by $\tilde{a}_j(t_l)$. More sophisticated interpolation by PCA in Kaplan et al. (2000).

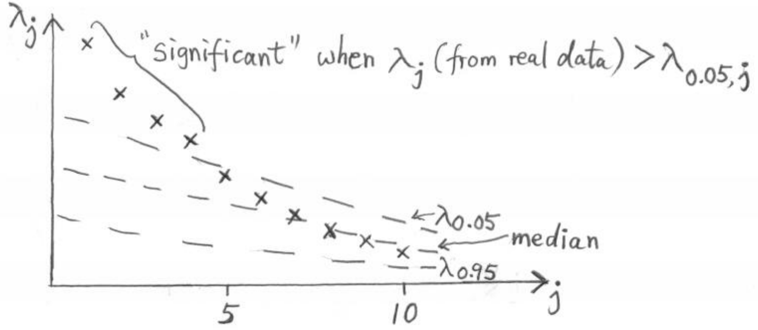
2.12 Significance tests [Book, Sect.2.1.12]

The higher PCA modes basically contain noise. How many modes to retain?
Some ‘rules of thumb’:

One approach is to plot the eigenvalues λ_j as a function of the mode number j. Hopefully, one finds an abrupt transition from large eigenvalues to small eigenvalues around mode number k. Keep the first k modes.

Computationally more involved is the Monte Carlo test (Preisendorfer, 1988): Set up random data matrices $\mathbf{R}_l$ $(l = 1, \ldots, L)$, of the same size as the data matrix $\mathbf{Y}$. The random elements are normally distributed, with the variance of the random data matching the variance of the actual data.

Do PCA on each of the random matrices, yielding eigenvalues $\lambda_j^{(l)}$. Assume for each l, the set of eigenvalues are sorted in descending order. For each j, one examines the distribution of the L values of $\lambda_j^{(l)}$, and finds the level $\lambda_{0.05}$, which is exceeded only by 5% of the $\lambda_j^{(l)}$ values.



Eigenvalues λ_j from **Y** lying above this $\lambda_{0.05}$ level are "significant".

If data have strong autocorrelation, dimension of $\mathbf{R}_l$ should be reduced, with the effective sample size n_{eff} replacing sample size n.

Monte Carlo method performs PCA on L matrices with L about 100–1000 $\Rightarrow$ costly for large data matrices.
Hence asymptotic methods based on central limit theorem are often used with large data matrices — see Mardia et al. (1979, pp.230-237) and Preisendorfer (1988, pp.204-206).

References

Fang, W. and Hsieh, W. W. (1993). Summer sea Surface temperature variability off Vancouver Island from satellite data. *Journal of Geophysical Research*, 98(C8):14391–14400.

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer, New York.

Kaplan, A., Kushnir, Y., and Cane, M. A. (2000). Reduced space optimal interpolation of historical marine sea level pressure: 1854-1992. *Journal of Climate*, 13(16):2987–3002.

Kelly, K. (1988). Comment on "Empirical orthogonal function analysis of advanced very high resolution radiometer surface temperature patterns in Santa Barbara Channel" by G.S.E. Lagerloef and R.L. Bernstein. *Journal of Geophysical Research*, 93(C12):15,743–15,754.

Lorenz, E. N. (1956). Empirical orthogonal functions and statistical weather prediction. Sci. rep. no. 1, Statistical Forecasting Project, Dept. of Meteorology, Mass. Inst. Tech.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis*. Academic Pr., London.

North, G. R., Bell, T. L., Cahalan, R. F., and Moeng, F. J. (1982). Sampling errors in the estimation of empirical orthogonal functions. *Monthly Weather Review*, 110:699–706.

Preisendorfer, R. W. (1988). *Principal Component Analysis in Meteorology and Oceanography*. Elsevier, New York.

Strang, G. (2005). *Linear Algebra and Its Applications*. Cole Brooks.

von Storch, H. and Zwiers, F. W. (1999). *Statistical Analysis in Climate Research*. Cambridge Univ. Pr., Cambridge.