

# Thoughts on von Storch and Zweiss, ch 13: Empirical Orthogonal Functions

Jonathan Rougier, Durham University, UK

Inception 4 October, 2006; last update October 10, 2006.

The thing that concerns me about this chapter is that it is too statistical, which might seem an odd thing for a Statistician to say. The starting point is an random  $p$ -vector with given mean vector  $\mu$  and variance matrix,  $\Sigma$ . But such things are not a natural consequence of climate physics, or of climate models. What we typically start with is collections of quantities,  $X = (x_1, \dots, x_n)$ , where the generic quantity  $x$  is a  $p$ -vector with components that might have some relationship to each other. For example,  $x$  might be mean monthly sea-surface temperatures along a transect across the Atlantic, and the  $i$ s might index time, measured in months, since 1950. Or  $x$  might be the complete climate state vector in pre-industrial steady state after a 4000 year spin-up of some climate model, and the  $i$ s represent different initial conditions. Or, somewhere between these two cases, we might spin up the model to a pre-industrial steady state, and then sample the climate state vector once every hundred years, so the  $i$ s index the model-dates 4100, 4200, 4300, and so on.

If we judged that the  $x_i$ s could be treated as exchangeable, then we can do some interesting Statistics. In this case, for example, the estimated mean vector  $m$  and the estimated variance matrix  $S$  can function as sufficient statistics: we throw away  $X$  and focus on  $m$  as though it was  $\mu$ , and  $S$  as

though it was  $\Sigma$ . Possibly in the second example this would be appropriate, if the initial conditions are randomly sampled from some specified distribution. But definitely not in the first example, and the third is also questionable. Therefore, to me, the primitive object is the matrix  $X$ . It is a matter of judgement whether the rows of  $X$  are exchangeable, and in many climate applications this seems difficult to justify. Where we judge that the rows of  $X$  are not exchangeable, we have to focus on a bigger question: not “What is the structure in the random vector  $x$ , as summarised by  $m$  and  $S$ ?”, but “What is the structure in the matrix  $X$ ?” where, in general, we expect to see an interrelationship between rows and columns.

So, for me, EOFs are about the Singular Value Decomposition (SVD). Just a reminder, the SVD says that we can find an orthogonal  $n \times n$  matrix  $U$  and an orthogonal  $p \times p$  matrix  $V$  such that

$$U^T X V = \text{diag}(\sigma_1, \dots, \sigma_k) \in \mathbb{R}^{n \times p} \quad \text{where } k \triangleq \min\{n, p\} \quad (1)$$

and  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_k \geq 0$ ; the diagonal matrix will need to be padded with zeros unless  $n = p$ . I like the proof of the SVD found in Golub and Van Loan (1989, sec. 2.5), which uses inequalities based on vector and matrix norms. Note that  $\|X\|_2 = \sigma_1$ .

The  $V$  are a set of eigenvectors for  $X^T X$ . If  $X$  is column-centred and the rows of  $X$  are exchangeable, then it follows that  $V$  are also the eigenvectors of  $S$ , the sample variance matrix of  $x$ . But if the rows of  $X$  are not exchangeable, then it does not even make sense to talk about the variance of  $x$ , because there is no clearly defined generic quantity for which all of the  $x_i$  are equally good candidates.

If we let  $r \triangleq \max\{i : \sigma_i > 0\}$ , then  $\text{rank } X = r$ , by construction, and we can write  $X$  as

$$X = \sum_{i=1}^r \sigma_i B^i, \quad \text{where } B^i \triangleq u_{(i)} v_{(i)}^T \quad (2)$$

and  $u_{(i)}$  is column  $i$  of  $U$ , and likewise for  $v_{(i)}$ . Here each  $B^i$  is a rank-one  $n \times p$  matrix, so we see that the SVD decomposes  $X$  into the sum of  $r$  rank-one matrices. Obviously, if some of the later  $\sigma_i$ s are small then we might reasonably truncate before  $r$ , to give the rank- $s$  approximation

$$\hat{X}_s \triangleq \sum_{i=1}^s \sigma_i B^i \approx X. \quad (3)$$

If we want to know how good this approximation is, we cannot in general refer to the proportion of *variance* explained, since the notion of variance is not pertinent unless the rows of  $X$  are exchangeable. We can, however, by sleight-of-hand, refer to the proportion of *variation* explained, where  $X$  is column-centred and variation is defined as  $\text{tr } X^T X$ , possibly normalised by  $n^{-1}$ . Then

$$\text{pp}^{\text{n}} \text{ variation explained by } \hat{X}_s = \frac{\sum_{i=1}^s \sigma_i^2}{\sum_{i=1}^r \sigma_i^2}. \quad (4)$$

In fact,  $\hat{X}_s$  is the best rank- $s$  approximation to  $X$  in the matrix 2-norm:

$$\min_{\hat{X} : \text{rank } \hat{X} = s} \|X - \hat{X}\|_2 = \|X - \hat{X}_s\|_2 = \sigma_{s+1} \quad (5)$$

Golub and Van Loan (1989, p. 73).

A lot of effort has been devoted to the issue of how to choose a good value for  $s \leq r$  in (3). It seems clear to me that there can be no hard-and-fast rules about this, *especially* if the rows of  $X$  are not exchangeable. In the latter case we must be suspicious of methods based on the sampling behaviour of the eigenvalues, since sampling theory is not appropriate here. Peres-Neto *et al.* (2005) provide a useful recent study on selecting  $s$ .

Two PhD students<sup>1</sup> and I have been experimenting with a simple permutation test, a development of **Rnd-Lambda**, in the reference above. Suppose we want to know whether to keep the first component, i.e., should we have

---

<sup>1</sup>Danny Williamson and David Randell; the sequential approach described here arose out of a conversation with Michael Goldstein.

$s \geq 1$ . We can compute  $\sigma_1$  from  $X$ . We can also shake-up  $X$ , by randomly permuting the values in each column. We can do this a number of times, and collect a sample of  $\sigma_1$  values. If our true  $\sigma_1$  is obviously bigger than the sampled values, then there is some structure in  $X$  that is being destroyed by the random permutation. So we ought to set  $s \geq 1$ .

Given that we are keeping the first component, we subtract it from  $X$ , i.e., replace  $X$  with  $X - \sigma_1 B^1$ , and then repeat the whole process. This is our twist, to apply the method sequentially, subtracting off the components of  $X$  as we decide to keep them, so that once they are ‘in’, they no longer affect the decision in any way. But there’s one catch: if we have subtracted off the first  $s$  components then SVD of the remainder,  $\hat{X}_{-s}$  say, is difficult to compare directly with that of  $X$ , because the variation in  $\hat{X}_{-s}$  is less than that in  $X$ . Therefore we scale  $\hat{X}_{-s}$  by dividing by the scalar quantity

$$\sqrt{1 - \text{pp}^n \text{ variation explained by } \hat{X}_s}, \quad (6)$$

see (4), which makes the two variations the same. I’m not completely sure about this last step: there may be a better way of re-scaling this.

The three of us have also discussed another idea: relaxing the orthogonality of the SVD to see if we can recover some interesting physics. I’m *very* suspicious of first reducing the dimension and then looking for the physics, e.g., by rotating, unless the dimension is totally overwhelming, and we leave out stuff that is obviously just noise (something that is very hard to assess if the rows of  $X$  are not exchangeable). It seems to me much better to look for the physics, and then leave out stuff that doesn’t look very interesting. We’re still thinking about how we might implement rotations, for example in a block-diagonal structure. In this case we ought to be able to represent  $X$  not as, say, the sum of 7 rank-one matrices, but as the sum of a rank-three matrix and a rank-four matrix. But this is still work-in-progress.

## References

- G.H. Golub and C.F. Van Loan, 1989. *Matrix Computations*. Baltimore: John Hopkins University Press, 2nd edition.
- Pedro R. Peres-Neto, Donald A. Jackson, and Keith M. Somers, 2005. How many Principal Components? stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics and Data Analysis*, **49**, 974–997.