

Notes on Climate Signal Detection

Gerald R. North
December 6, 2004

1 Introduction

One often wishes to see if there is a deterministic signal buried in the natural variability of a climate data stream. How can we process the data stream in such way as to estimate the strength of the signal embedded in the data and assess the confidence interval in our estimate. We start with a simple example in which there is only one signal in the stream that we wish to detect. We accomplish our goal by building a statistical model of the process.

2 Detection of a Single Signal

Our statistical model can be expressed as follows:

$$T^{data}(x) = \alpha S(x) + N(x) \quad (1)$$

The variable x refers to a point in space-time (say (x, y, t)); the term $T^{data}(x)$ refers to the data referred evaluated at the point x . The function $S(x)$ is the signal we are trying to detect. In our case it is given and not random. The function $N(x)$ is the natural variability in the background field; $N(x)$ is a random field assumed to have ensemble average zero. It can be thought of as the ‘noise’ in the problem, but note that this noise might be quite complicated, since it represents a realization of the climatic field in the absence of any signal. We have one realization of the data field T^{data} and we want to assess the strength of the signal embedded in the field α . If the strength is statistically different from zero we say we have *detected* the signal.

The first step in our procedure is to develop the natural variability into its EOFs:

$$\int \langle N(x)N(x') \rangle \psi_n(x') dx' = \lambda_n \psi_n(x) \quad (2)$$

where the angular brackets denote ensemble average. The $\psi_n(x)$ are the EOFs and they form a complete set such that any of the

fields can be expanded into them:

$$N(x) = \sum_{n=1}^{\infty} N_n \psi_n(x) \quad (3)$$

$$T^{data}(x) = \sum_{n=1}^{\infty} T_n^d \psi_n(x) \quad (4)$$

$$S(x) = \sum_{n=1}^{\infty} S_n \psi_n(x) \quad (5)$$

We can now expand each side in the EOFs and project out the components:

$$T_n^d = \alpha S_n + N_n \quad (6)$$

In this expression the N_n is a random variable along with T_n^d , while S_n is deterministic. Our task is to find the value of α which best makes the model fit the data. We also know that the N_n satisfy:

$$\langle N_n N_m \rangle = \lambda_n \delta_{nm} \quad (7)$$

where λ_n is the variance associated with EOF n . Notice that (6) is almost in the form of a standard least squares problem except that the ‘error term’ is not white noise. We can correct this by multiplying through by the *prewhitening* matrix

$$W_{mm'} = \frac{\delta_{mm'}}{\sqrt{\lambda_m}} \quad (8)$$

We then have

$$\tilde{T}_n^d = \alpha \tilde{S}_n + \tilde{N}_n \quad (9)$$

where

$$\tilde{T}_n^d = \sum_{n'=1}^{\infty} W_{nn'} T_{n'}^d, \quad \tilde{S}_n = \sum_{n'=1}^{\infty} W_{nn'} S_{n'}, \quad \tilde{N}_n = \sum_{n'=1}^{\infty} W_{nn'} N_{n'}, \quad (10)$$

Now (9) is in the form of a standard least squares problem since the error term \tilde{N}_n is white; i.e., its variance is independent of n .

We proceed through the least squares regression step by forming the total mean squared error, $E^2 = \sum_{n=1}^{\infty} (\tilde{T}_n^d - \alpha \tilde{S}_n)^2$ and finding its minimum with respect to α :

$$\frac{dE^2}{d\alpha} = -2 \sum_{n=1}^{\infty} \tilde{S}_n (\tilde{T}_n^d - \hat{\alpha} \tilde{S}_n) = 0 \quad (11)$$

where we have put a hat on α to indicate that it is the value it attains when the derivative on the *lhs* is set to zero. Solving for α yields

$$\hat{\alpha} = \frac{\sum_{n=1}^{\infty} \tilde{S}_n \tilde{T}_n^d}{\sum_{n=1}^{\infty} \tilde{S}_n^2} \quad (12)$$

$$= \frac{\sum_{nmm'} W_{nm} S_m W_{nm'} T_{m'}^d}{\sum_{nmm'} W_{nm} S_m W_{nm'} S_{m'}} \quad (13)$$

$$= \frac{\sum_{m=1}^{\infty} \left(\frac{S_m T_m^d}{\lambda_m} \right)}{\sum_{m=1}^{\infty} \left(\frac{S_m^2}{\lambda_m} \right)} \quad (14)$$

Note that α is itself a random variable from one realization of the data stream to another. This can be seen from the last equation since T_n^d is a random variable. We can determine whether the estimate of α is biased by inserting (6) into the last equation and taking the expectation value:

$$\langle \hat{\alpha} \rangle = \alpha \quad (15)$$

which tells us that this estimator of α is unbiased. This is not unexpected since we know that the standard estimators in regression analysis are unbiased.

What is somewhat surprising is that if we take a single term and use it to estimate α we would use the estimator:

$$\hat{\alpha}_m = \frac{T_m}{S_m} \quad (16)$$

By the same procedure as before we find that this is an unbiased estimator. After a bit of work we can establish:

$$\text{var}(\hat{\alpha}) = \frac{\lambda_m}{S_m^2} \quad (17)$$

This last is reasonable, since if the component of $S(x)$ along $\psi_m(x)$ is small the error variance in $\hat{\alpha}$ will be large. Likewise if λ_m is large for that component the error variance will be large.

2.1 Estimators with a Finite Number of Terms

We now have an unbiased estimator for α in (16). We could improve our estimate by a single EOF component by adding up a finite number of them in say a set \mathcal{M} consisting of M terms and dividing by

the number in the sum:

$$\hat{\alpha}_{subopt} = \frac{1}{M} \sum_{m \in \mathcal{M}} \frac{T_m}{S_m} \quad (18)$$

But an optimal estimator based upon the individual unbiased estimators (16) consists of weighting the individual terms by the inverse of their variances:

$$\alpha_{opt} = \sum_{n \in \mathcal{M}} \beta_n \frac{T_n}{S_n}; \quad \sum_{n \in \mathcal{M}} \beta_n = 1 \quad (19)$$

The β_n turn out to be just:

$$\beta_n = \frac{\frac{S_n^2}{\lambda_n}}{\sum_{n \in \mathcal{M}} \frac{S_n^2}{\lambda_n}} \quad (20)$$

when this is substituted into (19) we find that we are back to the original regression estimator (14) in the case where $M \rightarrow \infty$. We see that this optimal estimator works for a finite sum, and it is unbiased and has the least error variance for the set \mathcal{M} .

The sum $I_{\mathcal{M}} = \sum_{n \in \mathcal{M}} \frac{S_n^2}{\lambda_n}$ has an interesting interpretation. Each term is an expression of the signal squared to the noise variance for an individual mode. This sum is the error variance for the optimal estimator of the coefficient α . With each term the error variance grows but the signal to noise ratio diminishes. Consider the product:

$$\left(\sum_n \frac{S_n^2}{\lambda_n} \right) \left(\sum_n \lambda_n \right) \geq \sum_n S_n^2 \quad (21)$$

where inequality is Schwartz's inequality. This means the sum $I_{\mathcal{M}}$ is an upper bound on the signal to noise ratio squared.

2.2 Statistically Independent Estimators

A simpler example might aid in understanding the results obtained above. Consider first a pair of thermometers that are to be used in estimating the temperature of a bath whose actual temperature is T_0 . Each of the estimators is unbiased:

$$\langle \hat{T}_1 \rangle = \langle \hat{T}_2 \rangle = T_0 \quad (22)$$

Our model of the temperature estimators is:

$$\hat{T}_1 = T_0 + \epsilon_1 \quad (23)$$

$$\hat{T}_2 = T_0 + \epsilon_2 \quad (24)$$

with the gaussian random error terms $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2 \delta_{ij})$. This last says the error terms are uncorrelated. The estimators \hat{T}_1 and \hat{T}_2 are statistically independent estimators, but the variances of the errors might be quite different from one another.

The question we want to answer is how can we employ some linear combination of the two instruments in such way as to minimize the error in the estimate of the temperature of the bath. If we have two measurements, even though one might have a large error, there should be some means of making use of even this poor datum. We set up our problem as

$$\hat{T} = \alpha \hat{T}_1 + (1 - \alpha) \hat{T}_2 \quad (25)$$

We wish to know the value of α such that the overall error is minimal. Note that the coefficient of \hat{T}_2 is such as to preserve the zero bias, thus leaving only one unknown. Our task is to express the mean squared error in terms of α and then minimize it.

$$E^2 = \langle (T_0 - \hat{T})^2 \rangle \quad (26)$$

$$= \langle (\alpha \epsilon_1 + (1 - \alpha) \epsilon_2)^2 \rangle \quad (27)$$

$$= \alpha^2 \sigma_1^2 + (1 - \alpha)^2 \sigma_2^2 \quad (28)$$

Now we can set the derivative of E^2 to zero and find:

$$\alpha_{opt} = \frac{1}{\sigma_1^2} \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} \right)^{-1} \quad (29)$$

The optimal estimator is then:

$$\hat{T}_{opt} = \left(\frac{\hat{T}_1}{\sigma_1^2} + \frac{\hat{T}_2}{\sigma_2^2} \right) \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} \right)^{-1} \quad (30)$$

The second factor is a normalization merely to maintain the zero bias. Note that the individual estimators are weighted the inverse of their error variances. Estimators with large errors will be weighted less than those with smaller errors.

The problem is easily generalized to the case of M thermometers with the i th instrument having error variance σ_i^2 . In this case we write:

$$\hat{T} = \sum_{i=1}^M \alpha_i \hat{T}_i \quad (31)$$

and the constraint to insure no bias is

$$\sum_{i=1}^M \alpha_i = 1 \quad (32)$$

We proceed as before with the use of a Lagrange multiplier, etc. The final result is

$$\hat{T}_{opt} = \frac{\sum_{i=1}^M \frac{\hat{T}_i}{\sigma_i^2}}{\sum_{j=1}^M \frac{1}{\sigma_j^2}} \quad (33)$$

We have solved the problem of how to linearly combine M independent measurements in such way as to minimize the error. If the instruments had some covariance between them, it would be possible to find linear combinations of them that would be independent (the PCs). This should shed some light on why it is important to go to independent coordinates in the estimation problem.

3 Several Signals

The problem of more than one simultaneous signal comes up frequently. This model can be formulated as:

$$T^{data}(x) = \sum_{s=1}^{n_s} \alpha_s S_s(x) + N(x) \quad (34)$$

In this case we have n_s different signals with strengths α_s . As before we expand into EOFs of the noise field, $N(x)$.

$$T_n^d = \sum_{s=1}^{n_s} \alpha_s S_{sn} + N_n \quad (35)$$

Again, T_n^d and N_n are random, while S_{sn} is a given deterministic shape in space-time. We need to prewhiten the terms in the equation

by multiplying through by the matrix $W_{nm} = \frac{\delta_{nm}}{\sqrt{\lambda_n}}$. The sum of squares of errors is given by

$$E^2 = \sum_n \left(\tilde{T}_n^d - \sum_s \alpha_s \tilde{S}_{sn} \right)^2 \quad (36)$$

where the variables with tilde are defined as in the previous section. We seek to minimize E^2 with respect to the $\alpha_s, s = 1, \dots, n_s$ simultaneously.

$$\frac{\partial E^2}{\partial \alpha_s} = -2 \sum_n \tilde{S}_{sn} \left(\tilde{T}_n^d - \sum_{s'} \hat{\alpha}_{s'} \tilde{S}_{s'n} \right) = 0 \quad (37)$$

Rearranging:

$$\sum_{s'} \left(\sum_n \tilde{S}_{sn} \tilde{S}_{s'n} \right) \hat{\alpha}_{s'} = \sum_n \tilde{S}_{sn} \tilde{T}_n^d, \quad s = 1, \dots, n_s \quad (38)$$

These last are a set of linear equations for the $\alpha_{s'}, s' = 1, \dots, n_s$.

There are several important issues we have to contend with. A major one involves the properties of the symmetric matrix \mathcal{G} :

$$\mathcal{G} \equiv \sum_n \tilde{S}_{sn} \tilde{S}_{s'n} \quad (39)$$

The rank of \mathcal{G} is the number of its nonzero eigenvalues. The rank will be $r_{\mathcal{G}} \leq M$, where M is the number of EOF modes in the sum over n . It follows that to determine all n_s of the α_s we need to have $M \geq n_s$. So if there are four signals whose strengths are to be estimated, we need at least four EOF modes in the estimation process.

Besides the rank of \mathcal{G} we have to worry about *collinearity*. Here the issue is whether two or more columns (or rows) are nearly linearly independent of one another. Physically, this means that two of the signals have nearly the same shapes in x . Collinearity makes discrimination between these two signals very difficult, and this will show up in the next issue to be discussed.

The estimators $\hat{\alpha}_s$ are random variables since the T_n^d are random variables from one realization of the data stream to another. This problem is simple in the single signal problem since $\hat{\alpha}$ is normally

distributed about its true value. But in the case $n_s > 1$ there will be correlations between the different estimators $\hat{\alpha}_s$ and this leads to a more involved analysis as to whether some of the $\hat{\alpha}_s$ are significantly different from zero.