

CCSR Series Lectures

Fall 1995

Climate Sampling Errors

By

Samuel S. Shen

University of Tokyo

Preface

These notes were written for the lectures I gave at CCSR (the Center for Climate System Research), the University of Tokyo in the fall of 1995 which hosted me for half a year during my sabbatical leave from the University of Alberta, Canada. The purpose of this series of lectures is to introduce the basic mathematics and statistics related to the estimation of the mean square errors in sampling climate fields. The topics range from the classical sampling theorem to the modern optimal sampling network design and are listed below:

1. Exact sampling, aliasing, and basic statistics on mean square errors,
2. North-Nakamoto method for computing mean square errors,
3. Trigonometric functions, spherical harmonics and EOFs,
4. Minimal mean square errors, global warming detection and optimal network designs.

According to the above four topics, the lecture notes are divided into four chapters. Some of the materials were used in a graduate course on statistics for geophysical sciences in the University of Alberta in the January term of 1995. Various suggestions were given and many errors were corrected by the enthusiastic students in the class. I hope that the audience of this CCSR class

will also give me generous suggestions with regard to correcting mistakes and improving the presentation style.

The lectures are directed to both senior level undergraduate students and graduate students in geophysical, mathematical and statistical sciences, as well as interested research scientists. The mathematical pre-requisites of the lectures include: calculus, linear algebra, elementary ordinary and partial differential equations.

I would like to thank both my host institution (the University of Tokyo) and my home institution (the University of Alberta) for their generous and sincere support of my research and teaching activities. In particular I would like to thank Professors A. Sumi and T. Nakajima for their effort to materialize my visit to Tokyo. Finally my gratitude goes to Dr. M. Takahashi who made the arrangement for my CCSR lectures.

Sam Shen
Tokyo, Japan
September 1995
E-mail: shen@cake.math.ualberta.ca

Contents

1	Aliased Power and Sampling Theorem	1
1.1	Exact sampling	1
1.2	Fourier transform and Fourier series	2
1.2.1	Fourier transform	2
1.2.2	Fourier series	4
1.3	Aliasing	6
1.3.1	Aliasing for a signal defined on $(-\infty, \infty)$	7
1.3.2	Aliasing for periodic signals	10
2	Sampling Errors	15
2.1	Mean square errors	15
2.2	Spectrum-filter formula for satellite MSE	17
2.3	Spectral-filter formula for rain gauges	23
2.4	A simple diffusive rain model	24
2.5	Interpretations of the results	27
2.6	References	28
3	EOFs	29
3.1	Covariance matrix and data preparation	29
3.2	Empirical orthogonal functions	32
3.3	EOFs of a stochastic field	34
3.4	EOFs on a unit circle	36
3.4.1	Homogeneous real EOFs	36
3.4.2	Homogeneous complex EOFs	37
3.4.3	Inhomogeneous EOFs on a unit circle	38
3.5	EOFs on a unit sphere	38
3.5.1	Simple orthogonal bases	38
3.5.2	Spherical harmonics	39
3.5.3	EOFs for a homogeneous field on a sphere	44
3.6	T-truncation and R-truncation	45
3.6.1	Spectral truncation	45
3.6.2	Spatial resolution	46
3.7	EOFs for nonhomogeneous field	49
3.8	References	50

4 Minimal MSE Optimization	51
4.1 Numerical integration	51
4.1.1 Optimal weights only	52
4.1.2 Optimize both weights and positions	54
4.1.3 Monte Carlo method	56
4.2 Global warming	58
4.3 Spherical harmonic components	61
4.4 Homogeneous reduction	64
4.4.1 The covariance function kernel	65
4.4.2 The MSE formula	66
4.5 Spectra derived from noise forced EBM	67
4.6 Network design on unit sphere	69
4.7 References	74

Chapter 1

Aliased Power and Sampling Theorem

Our goal is to use limited amount of data for a time series X_t (to be detected or predicted) to reconstruct a realization of X_t so that the reconstructed realization is closest to the true one. We often need to do this when X_t is a continuous time series and the measured data are discrete. Let \hat{X}_t be the reconstructed time series. The mean square sampling error is defined as

$$\epsilon^2 = \langle (X_t - \hat{X}_t)^2 \rangle \quad (1.0.1)$$

In this book, we consider only the stationary time series, so the sampling error is independent of time t .

Therefore our mathematical problem is to obtain an estimated \hat{X}_t so that ϵ^2 is minimal. Then we use \hat{X}_t as an approximate realization of X_t . This is called the sampling problem, in which we need to find both \hat{X}_t and ϵ^2 . This chapter will devote to the sampling problem of single variate time series.

1.1 Exact sampling

One can imagine that the sampling error must depend on the oscillation frequency and smoothness of the function $x_t = f(t)$ (regarded as a realization of a time series X_t). For instance, for a low frequency signal x_t , the function $f(t)$ is smooth and the sampling error should be small. The extreme case is the zero frequency signal: a constant signal. The sampling error is zero by only one sampling point.

One may think that if the signal contains only limited number of frequencies, then one can choose finitely many sampling points to have zero error measurements. This is true and can be proved by linear algebra techniques.

Consider a signal $f(t)$ in $[0, 1]$ of n frequencies:

$$f(t) = \sum_{k=1}^n a_k \sin(2\pi kt). \quad (1.1.1)$$

We use N sampling points in $[0, 1]$ to sample the signal $f(t)$:

$$f(t_j) = \sum_{k=1}^n a_k \sin(2\pi kt_j), \quad j = 1, 2, \dots, N. \quad (1.1.2)$$

One can completely determine a_k ($k = 1, 2, \dots, n$) as long as the sampling points are chosen such that the rank of the matrix $[\sin(2\pi kt_j)]$ is equal to n . Therefore, for the complete determination of a_k ($k = 1, 2, \dots, n$) (i.e. zero sampling error), it is necessary to have $N \geq n$ (i.e. the number of the sampling points must not be less than that of the frequencies).

Another related question is more interesting for applications. You know that your signal has only a finite number of frequencies. But you do not have enough sampling points for getting zero error sampling. So what is this non-zero sampling error? This problem is usually called the aliasing problem and will be addressed in Section 1.3.

1.2 Fourier transform and Fourier series

The motion of a subject usually is associated with sound, colors of light, and, in general, vibration. In particular, many types of motions are oscillatory or vibrating, such as the rotation of machines, fluctuations of the climate, the cycles of human mood, and others. The physical motion may be described by a displacement function which depends on time t . The "sound" is another physical quantity usually one can readily hear or detect. This "sound" is also physical and called the spectrum of the motion. When an experienced driver hears some abnormal sound of a car, he can immediately know which part of the car is in trouble. It is also quite common that from the sound of an engine one can tell whether the engine runs fast or slow. There are numerous examples of detecting the motion (i.e. the displacement function) of a subject from sound (i.e. the spectrum). As a matter of fact, the displacement signal can be completely described by the spectral signal, and vice versa. This kind of principle can be described by the Fourier transform (from displacement to spectrum) and the inverse Fourier transform (from spectrum to displacement).

In many cases, it is easier to measure the spectra directly. By the inverse Fourier transform, one can then indirectly determine the displacement function. Thus to study the detection problem, naturally we need to understand the basics of the theory of the Fourier transform.

1.2.1 Fourier transform

We use $f(t)$ to denote the displacement function and $\hat{f}(\omega)$ for the spectrum function (called the Fourier transform). Usually, one regards t as time and ω

as frequency. There are also many cases that t plays the role of spatial coordinate and then the corresponding ω is the inverse of the wave length, called the wave number. In the formality of mathematics, it seems that the difference of wavenumber and frequency is only a matter of name; but in quantum mechanics, wave number and frequency have different physical meanings. The former represents momentum and the later measures the energy of a particle.

The transforms between $f(t)$ and $\tilde{f}(\omega)$ are defined by

$$\tilde{f}(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} dt f(t) e^{i\omega t} \quad (\text{Fourier transform}), \quad (1.2.1)$$

$$f(t) = \int_{-\infty}^{\infty} d\omega \tilde{f}(\omega) e^{-i\omega t} \quad (\text{inverse Fourier transform}). \quad (1.2.2)$$

It is not very hard for one to accept the Fourier transform formula (1.2.1), since the spectrum (i.e. the sound) should be completely determined by the displacement as long as the displacement function does not jump too violently. One may be suspicious about the validity of the inverse formula (1.2.2). The question is: can the spectrum completely determine the displacement? The answer is "yes" as long as: (i) $f(t)$ does not have sudden jumps, (ii) $f(t)$ does not have too many peaks, and (iii) $\int_{-\infty}^{\infty} |f(t)| dt < \infty$.

Mathematically the inverse formula can be easily derived. We start with

$$\begin{aligned} & \int_{-\infty}^{\infty} d\omega \tilde{f}(\omega) e^{-i\omega t} \\ &= \int_{-\infty}^{\infty} d\omega \left(\frac{1}{2\pi} \int_{-\infty}^{\infty} dt' f(t') e^{i\omega t'} \right) e^{-i\omega t} \\ &= \int_{-\infty}^{\infty} dt' f(t') D(t-t') \quad (\text{exchange the order of integration}), \end{aligned}$$

where

$$D(t-t') = \frac{1}{2\pi} \int_{-\infty}^{\infty} d\omega e^{-i\omega(t-t')}. \quad (1.2.3)$$

The right hand side of the above integral takes the Cauchy principal value:

$$\begin{aligned} D(x) &= \lim_{R \rightarrow \infty} \int_{-R}^R d\omega e^{-i\omega x} \\ &= \lim_{R \rightarrow \infty} \frac{\sin Rx}{\pi x} \quad (\text{a delta convergent sequence}) \\ &= \delta(x). \end{aligned}$$

See Fig. ~~ref~~ 1.1 for the convergence process of the above delta-convergent sequence. Another verification is to check whether

$$\lim_{R \rightarrow \infty} \int_{-\infty}^{\infty} \frac{\sin Rx}{\pi x} dx = 1.$$

Using Mathematica, we can do

`NIntegrate[Sin[30* x] / (Pi x), {x, 0.0000001, 5}]`

The answer is 0.498525 which is very close to 0.5.

Thus,

$$\int_{-\infty}^{\infty} dt' f(t')D(t-t') = \int_{-\infty}^{\infty} dt' f(t')\delta(t-t') = f(t).$$

The last equality requires that $f(t)$ is a continuous function.

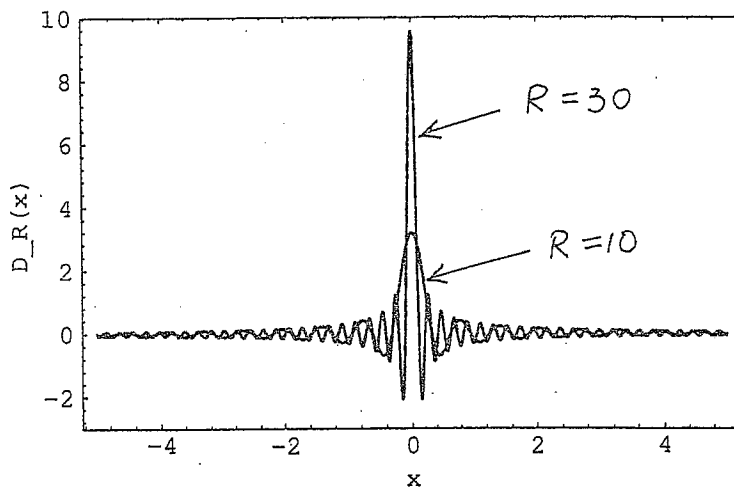


Figure 1.1: The convergence process of a delta-convergent sequence.

1.2.2 Fourier series

We just considered the Fourier transform for a function defined on $(-\infty, \infty)$. Now we consider the similar transform for periodic functions in $(-\infty, \infty)$. Since we can always change the coordinate so that the period of the function is 2π . Let us work in $(-\pi, \pi]$. The integral of the inverse Fourier transform can be written as an infinite sum:

$$f(t) = \sum_{n=-\infty}^{\infty} c_n e^{int}. \quad (1.2.4)$$

This infinite series is called the Fourier series. The Fourier coefficients c_n are determined by

$$c_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} dt f(t) e^{-int}, \quad n = 0, \pm 1, \pm 2, \dots \quad (1.2.5)$$

One may feel that the condition for (1.2.4) to hold should be pretty loose. Indeed, it is true. If a $f(t)$ is piecewise continuous, then

$$\frac{f(t-0) + f(t+0)}{2} = \sum_{-\infty}^{\infty} c_n e^{int}. \quad (1.2.6)$$

The proof of this formula is similar to the above derivation of the inverse Fourier transform. It goes as follows:

$$\begin{aligned} & \sum_{-\infty}^{\infty} c_n e^{int} \\ = & \sum_{-\infty}^{\infty} \frac{1}{2\pi} \int_{-\pi}^{\pi} dt' f(t') e^{-int'} e^{int} \\ = & \int_{-\pi}^{\pi} dt' f(t') K(t-t'), \end{aligned}$$

(exchange the order of summation and integration)

where the kernel function $K(t-t')$ is defined by

$$K(x) = \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} e^{inx}. \quad (1.2.7)$$

We also take the Cauchy principal value of this infinite series

$$\begin{aligned} K(x) &= \lim_{N \rightarrow \infty} \frac{1}{2\pi} \sum_{n=-N}^N e^{inx} \\ &= \lim_{N \rightarrow \infty} \frac{1}{2\pi} e^{-iNx} \frac{1 - e^{(2N+1)ix}}{1 - e^{ix}} \\ &= \lim_{N \rightarrow \infty} \frac{1}{2\pi} \frac{e^{-iNx} e^{-ix/2} - e^{iNx} e^{ix/2}}{e^{-ix/2} - e^{ix/2}} \\ &= \lim_{N \rightarrow \infty} \frac{1}{2\pi} \frac{\sin[(N+1/2)x]}{\sin(x/2)} \\ &= \delta(x). \end{aligned}$$

Again, please see Fig. 1.2 for the convergence process of the above limit. One may also wish to check whether

$$\int_{-2}^2 \frac{1}{2\pi} \frac{\sin(N+1/2)x}{\sin(x/2)} dx = 1.$$

Using Mathematica, we have

```
NIntegrate[(1/(2 Pi)) * Sin[(n + 0.5) x] / Sin[x/2],
{x, 0.0000001, 2}]
```

The result is 0.501663 which is very close to 1.0.

Therefore,

$$\int_{-\pi}^{\pi} dt' f(t')K(t-t') = \frac{f(t-0) + f(t+0)}{2}, \quad -\pi < t \leq \pi. \quad (1.2.8)$$

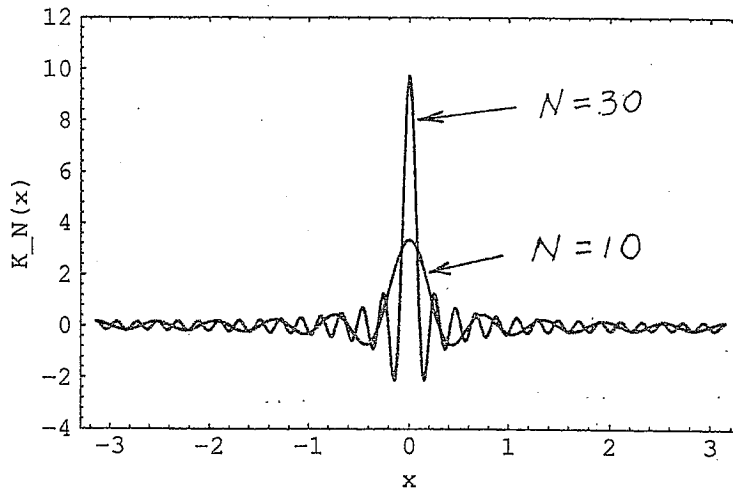


Figure 1.2: The convergence process of $K_N(x)$ to a delta function.

1.3 Aliasing

The aliasing topic is about the accuracy of computation or measurement. The knowledge we are going to learn in this section is a useful preparation for understanding the truncation of spherical harmonic series and the resolution of a spectral general circulation model (GCM), which will be discussed in Chapter 3.

From Section 1.1 and our sixth sense feeling, we may think that we perhaps do not need all the values of $f(t)$ (for every t) to determine the spectrum $\bar{f}(\omega)$. This feeling has certain truth, but strictly speaking, not correct. Then we can think further: what spectrum do we catch if we have only finite or countable (infinite) number of values of $f(t)$? Now our intuition may suggest that the spectra at the lower frequencies can be caught. This conclusion is again not completely true. Actually the spectra over the higher frequencies are aliased (or moved) to those of the lower frequencies. Therefore the spectra derived from the finite or countable (still infinite) number of values of $f(t)$ is the distorted spectra over the lower frequency regime. Finally we come to the question: can we compute the distorted part of the spectra? The answer is now yes. We can compute the aliased spectrum, which is determined by the sampling design

(i.e., where we sample the values of $f(t)$) and the spectral property of the $f(t)$ (i.e. the property of $\tilde{f}(\omega)$).

1.3.1 Aliasing for a signal defined on $(-\infty, \infty)$

As defined by (1.2.1), the Fourier transform is

$$\tilde{f}(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} dt f(t) e^{i\omega t} \quad (1.3.1)$$

If we have function values at discrete points t_n (the index n takes whatever choice you have made), then the spectrum estimated from the discretized form of the above integral is

$$\hat{\tilde{f}}(\omega) = \frac{1}{2\pi} \sum_n f(t_n) e^{i\omega t_n} w_n. \quad (1.3.2)$$

Here, the index n takes whatever choice (i.e. the sampling design) you have made, such as $n = -\infty, \dots, -1, 0, 1, \dots, \infty$, and w_n are the weights determined by certain type of numerical integration method. The above formula can be written into the form

$$\hat{\tilde{f}}(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} dt f(t) H(t) e^{i\omega t}, \quad (1.3.3)$$

where

$$H(t) = \sum_n w_n \delta(t - t_n) \quad (1.3.4)$$

is determined by the sampling design means to choose t_n and w_n .

Next we transform the above expression into the spectral form:

$$\begin{aligned} \hat{\tilde{f}}(\omega) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} dt \left(\int_{-\infty}^{\infty} d\omega' \tilde{f}(\omega') e^{-i\omega' t} H(t) e^{i\omega t} \right) \\ &= \int_{-\infty}^{\infty} d\omega' \tilde{f}(\omega') \Gamma(\omega - \omega') \\ &= (\tilde{f} * \Gamma)(\omega) \quad (\text{convolution product}), \end{aligned} \quad (1.3.5)$$

where

$$\Gamma(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} dt H(t) e^{i\omega t} \quad (1.3.6)$$

is called the designed filter. Hence the formula (1.3.5) is in the spectrum-filter form. The spectrum is of course determined by the property of the function $f(t)$. It is the property of the subject under investigation, such as the annual cycle of the climate, the 11-year cycle of sunspot activity, and the diurnal cycles of rain fall in some regions. The filter is completely determined by the sampling design. However, to detect a signal in most efficient way we have to design our filter according to the properties of the signal. Thus an optimal

filter (in a well defined sense in the case of practical applications) depends on the spectrum of the signal. We will talk about the optimal design in Chapter 4. At this moment, let us put the optimal design aside and consider the simplest sampling design which is the design of uniform sampling points and uniform weighting:

$$t_n = n(1/\omega_0) \quad \text{and} \quad w_n = \Delta t_n = 1/\omega_0 \quad (n = 0, \pm 1, \pm 2, \dots) \quad (1.3.7)$$

The quantity $\omega_0 = 1/\Delta t_n$ is called the sampling frequency and $\omega_{Nq} = 1/(2\Delta) = \omega_0/2$ is called the Nyquist frequency. This Nyquist frequency determines where the aliasing starts. We will see this soon.

The filter of the design is

$$\begin{aligned} \Gamma(\omega) &= \sum_{n=-\infty}^{\infty} e^{i\omega n(1/\omega_0)} \frac{1}{\omega_0} \\ \sum e^{i\omega n x} &= 2\pi \sum \delta(x - j) \\ &= 2\pi \frac{1}{\omega_0} \sum_{j=-\infty}^{\infty} \delta\left(\frac{1}{\omega_0}\omega - j\right) \\ &= 2\pi \sum_{j=-\infty}^{\infty} \delta(\omega - j\omega_0). \end{aligned} \quad (1.3.8)$$

Hence

$$\tilde{f}(\omega) = \bar{f}(\omega) + \Delta \bar{f}(\omega), \quad (1.3.9)$$

where

$$\Delta \bar{f}(\omega) = \sum_{j \neq 0} \bar{f}(\omega + j\omega_0) \quad (1.3.10)$$

is the aliased spectrum. So the spectrum estimated from discrete data include both true spectrum $\bar{f}(\omega)$ and the aliased spectrum $\Delta \bar{f}(\omega)$ (which is not known and supposed to be detected). It is the aliased spectrum that distorted the true spectrum. The spectra over the high frequency region are moved to the lower frequency region (see Fig. 1.3). This spectra flipping is called the aliasing phenomenon.

There are some cases that the the true spectrum of the signal is contained in a bounded interval. This type of signals are called the band limited signal. For a band limited signal, there is always a positive value ω_c such that $\bar{f}(\omega) = 0$ if $|\omega| > \omega_c$. For the uniform sampling design, the aliased spectrum is zero if $\omega_0 > 2\omega_c$. This is the famous sampling theorem in mathematics

Theorem 1.1 *If $f(t)$ is a banded limited signal, the signal can be exactly detected by uniform sampling as long as the half sampling frequency $\omega_0/2$ (i.e. the Nyquist frequency) is greater than the maximal frequency ω_c of the signal, and further*

$$f(t) = \Delta t \sum_{n=-\infty}^{\infty} f(n\Delta t) \frac{\sin((n\Delta t - t)\omega_0)}{\pi(n\Delta t - t)}, \quad (1.3.11)$$

with $\Delta t = 1/\omega_0$.

ω_0 is dependence

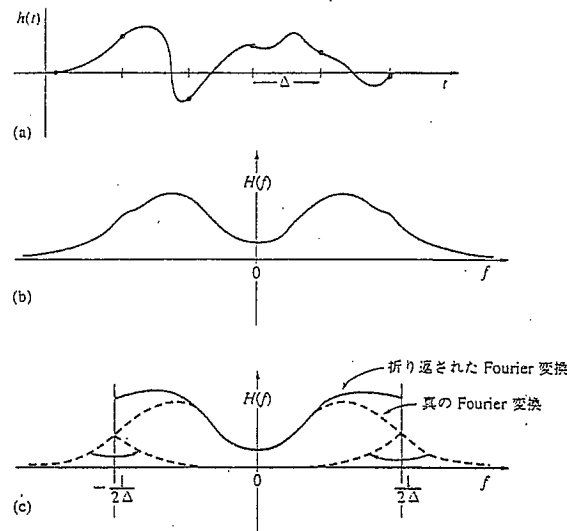


Figure 1.3: The aliasing phenomenon: spectra flipping.

Example The function

$$f(t) = \frac{\sin \sqrt{1+t^2}}{\sqrt{1+t^2}} \quad (1.3.12)$$

is a band limited signal whose Fourier transform is

$$\tilde{f}(\omega) = \begin{cases} \frac{1}{2} J_0(\sqrt{1-\omega^2}), & |\omega| < 1, \\ 0, & |\omega| \geq 1, \end{cases} \quad (1.3.13)$$

where J_0 is the zeroth order Bessel function. So this $f(t)$ is a band limited signal whose $\omega_c = 1$. To get exact sampling, we should choose the sampling frequency $\omega_0 \geq 2\omega_c = 2$. Let me present you two cases: $\omega_0 = 2$ and 3.

```
f[t_] := Sin[Sqrt[1. + t^2]]/Sqrt[1 + t^2];
om = 2.;
x=0.8;
del = 1.0 / om;
fx = del * Sum[f[n * del] * N[Sin[(n * del - x)*
  om]/(Pi (n* del -x))],{n, -20,20}]
```

The result is 0.748514. The exact value is computed by $f[x]$ which gives 0.748224. So the sampling result can be considered as exact.

when If we choose $\omega_0 = 3$ and do the same calculation as above, the sampling result is 0.747097, which can also be considered as exact.

The first part of the theorem is the direct result of equation (1.3.10). The second part can be derived as follows

$$f(t) = \int_{-\infty}^{\infty} d\omega \tilde{f}(\omega) e^{-i\omega t}$$

$$\begin{aligned}
&= \int_{-\omega_0}^{\omega_0} d\omega \bar{f}(\omega) e^{-i\omega t} \\
&= \int_{-\omega_0}^{\omega_0} d\omega \left(\frac{1}{2\pi} \sum_{n=-\infty}^{\infty} f(n\Delta t) e^{i\omega n\Delta t} \Delta t \right) e^{-i\omega t} \\
&= \frac{1}{2\pi} \Delta t \sum_{n=-\infty}^{\infty} f(n\Delta t) \int_{-\omega_0}^{\omega_0} d\omega e^{i(n\Delta t - t)\omega} \\
&= \Delta t \sum_{n=-\infty}^{\infty} f(n\Delta t) \frac{\sin[(n\Delta t - t)\omega_0]}{\pi(n\Delta t - t)}.
\end{aligned}$$

The aliased power is defined as

$$\epsilon^2(\omega) = |\Delta \bar{f}(\omega)|^2. \quad (1.3.14)$$

If the signal is a stationary time series, then the aliased power is the ensemble average of the above:

$$\epsilon^2(\omega) = \langle |\Delta \bar{f}(\omega)|^2 \rangle. \quad (1.3.15)$$

For a stationary time series, the spectra of different frequencies are not correlated, i.e.

$$\langle \bar{f}(\omega) \bar{f}(\omega') \rangle = \sigma^2 \delta(\omega - \omega'), \quad (1.3.16)$$

where $\sigma^2 = \langle f^2(t) \rangle = \langle |\bar{f}(0)|^2 \rangle$ is the variance of the signal. Thus, we have

$$\langle |\Delta \bar{f}(\omega)|^2 \rangle = \sum_{j \neq 0} \langle |\bar{f}(\omega + j\omega_0)|^2 \rangle. \quad (1.3.17)$$

1.3.2 Aliasing for periodic signals

In this subsection, we study the aliasing of periodic signals in $(-\infty, \infty)$. Let us consider the case of period equal to 2π . The cases of other periods can be converted into a 2π -period function by coordinate stretching.

Periodic signal is a special case of the signals we discussed in the above subsection. The relevant formulas in the above subsection might be simplified. The simplification is like the reduction from Fourier integral to Fourier series: integration to summation. Many people in the pre-computer era may not consider the change from the integration form to the summation form as a simplification because of the beauty and simplicity of the fundamental theorem of calculus. People have systematic way to calculate the definite integrals and tend to forget that the usual integral is a limit of a sum (called the Riemman sum), But for experienced researchers we all know that the summation form is a better form for a computer to do the calculation. Discretization of an integral is a dedicated skill of mathematics and it is a category of numerical analysis.

As defined by (1.2.4), the Fourier coefficients are

$$c_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} dt f(t) e^{-int}. \quad (1.3.18)$$

Then under certain conditions, we have

$$f(t) = \sum_{n=-\infty}^{\infty} c_n e^{int}. \quad (1.3.19)$$

Since $f(t)$ is a periodic function of period equal to 2π , we need only to consider the function in $(-\pi, \pi]$. Suppose we have N sampling points: $-\pi < t_1 < t_2 < \dots < t_j < \dots < t_N \leq \pi$. Now we use the function values at these points to estimate the Fourier coefficient (i.e., the spectra):

$$\hat{c}_n = \frac{1}{2\pi} \sum_{j=1}^N f(t_j) e^{-int_j} w_j, \quad (1.3.20)$$

where w_j are the weights to be chosen according to certain criterion under the constraint that

$$\sum_{j=1}^N w_j = 2\pi. \quad (1.3.21)$$

(A sampling design means to choose both t_j and w_j for $j = 1, 2, \dots, N$.)

The above formula can be written in the integral form

$$\hat{c}_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} dt f(t) H(t) e^{-int}, \quad (1.3.22)$$

where

$$H(t) = \sum_{j=1}^N \delta(t - t_j) w_j. \quad (1.3.23)$$

We would like to write \hat{c}_n into the convolution product of spectra and filter like (1.3.10):

$$\begin{aligned} \hat{c}_n &= \frac{1}{2\pi} \int_{-\pi}^{\pi} dt \left(\sum_{m=-\infty}^{\infty} c_m e^{imt} \right) H(t) e^{-int} \\ &= \sum_{m=-\infty}^{\infty} c_m \frac{1}{2\pi} \int_{-\pi}^{\pi} dt H(t) e^{i(m-n)t} \\ &= \sum_{m=-\infty}^{\infty} c_m \Gamma(m-n) \equiv (c * \Gamma)(n), \end{aligned} \quad (1.3.24)$$

where

$$\Gamma(k) = \frac{1}{2\pi} \sum_{j=1}^N e^{ikt_j} w_j \quad (1.3.25)$$

consists the filter.

The simplest design is the uniform space and uniform weight sampling: $\Delta t = 2\pi/N$, $t_j = -\pi + j\Delta t$ and $w_j = 2\pi/N$. Then the filter is

$$\Gamma(k) = \frac{(-1)^k}{N} \sum_{j=1}^N e^{ikj\Delta t} \quad (1.3.26)$$

Noticing that $\Gamma(0) = 1$, we have

$$\hat{c}_n = c_n + \Delta c_n \quad (1.3.27)$$

where the aliased spectra Δc_n are given by

$$\Delta c_n = \sum_{m \neq 0, (m-n)/N = \text{integer}} (-1)^{m-n} c_m. \quad (1.3.28)$$

If the signal is band-limited signal

$$c_n = 0 \quad \text{if} \quad |n| > n_c, \quad (1.3.29)$$

then $\Delta c_n = 0$ when $N > 2n_c$. This conclusion is similar to that in Section 1.1, where only sine waves are sampled and hence only half the sampling points are need to get the exact sampling.

With $\Delta c_n = 0$, we can further have:

$$f(t) = \frac{\Delta t}{2\pi} \sum_{j=1}^N f(-\pi + j\Delta t) \frac{\cos[(n_0 + 1/2)(t - j\Delta t)]}{\cos[(t - j\Delta t)/2]}, \quad (1.3.30)$$

where $n_0 \geq n_c$.

Example: The signal

$$f(t) = \sin t + 3 \cos 2t + \sin 4t$$

is band limited and $n_c = 4$. Let us choose $n_0 = 4$, $N = 9$ and test the above formula using Mathematica.

```
f[t_]:= Sin[t] + 3.0 Cos[2 t] + Sin[4 t];
nn = 9;
del = 2 * Pi / nn;
x = 0.8;
ex =N[ (del / (2* Pi)) * Sum[f[-Pi + j * del] *
  Cos[(4 + 0.5) (x - j * del)] / Cos[0.5* (x - j * del)],
  {j,1,nn}] ];
f[x];
```

The sampling result is $ex = 0.571383$ and the function result is $f[x] = 0.571383$. They are exactly the same.

This is the sampling theorem for periodic signals. The derivation of the above formula is as follows:

$$\begin{aligned}
 f(t) &= \sum_{n=-\infty}^{\infty} c_n e^{int} \\
 &= \sum_{n=-n_0}^{n_0} c_n e^{int} \\
 &= \sum_{n=-n_0}^{n_0} \left(\frac{1}{2\pi} \sum_{j=1}^N f(t_j) e^{-i(-\pi+j\Delta t)n \Delta t} \right) e^{int} \\
 \rightarrow &= \frac{\Delta t}{2\pi} \sum_{j=1}^N f(t_j) \sum_{n=-n_0}^{n_0} e^{in(t-j\Delta t+\pi)} \\
 &= \frac{\Delta t}{2\pi} \sum_{j=1}^N f(t_j) \frac{\cos[(n_0 + 1/2)(t - j\Delta t)]}{\cos[(t - j\Delta t)/2]}.
 \end{aligned}$$

If the signal is a stationary time series in $(-\pi, \pi]$, then the aliased power is

$$\epsilon_n^2 = \langle |\Delta c_n|^2 \rangle = \sum_{m \neq 0, (m-n)/N = \text{integer}} |c_m|^2. \quad (1.3.31)$$

Chapter 2

Sampling Errors

The sampling error refers to the difference of the true value of a climate quantity, such as monthly mean temperature and precipitation, and the value of the same quantity derived from incomplete samples. In climatology, sampling errors appear everywhere. The global average of the surface air temperature of Earth is derived from the data obtained only from finitely many (hence incomplete samples) surface stations or satellites. The rain rate of a rectangular area that includes Japan is derived from finitely many rain gauges and surface radars (cf. AMeDAS document: Automated Meteorological Data Acquisition System). In principle the true value of a climate quantity over an area in a certain time interval, such as the monthly rainfall over Japan, can never be exactly measured due to either spatial gaps or temporal gaps. However when sufficient number of instruments are deployed to sample a quantity in a small area, the value derived from the measurements (or called samples) may be regarded as the "true" one because of small errors. In this chapter, we will discuss the calculation of the sampling error, in particular for rain rates, but the methodology can also be applied to other climate quantities.

2.1 Mean square errors

Mathematically, sampling errors can be in various kinds of forms. But the two which are not strange to climatologists are the absolute error and the mean square error:

$$\epsilon = \langle |Q_{true} - Q_{sample}| \rangle \quad (\text{absolute error}) \quad (2.1.1)$$

$$\epsilon^2 = \langle (Q_{true} - Q_{sample})^2 \rangle, \quad (\text{mean square error}), \quad (2.1.2)$$

where $\langle \cdot \rangle$ signifies the ensemble average (i.e. the expectation value).

The mathematics for the mean square error (MSE) is relatively easier than that for the absolute error. In our lectures, we only talk about the computation of the MSE and the absolute error is replaced by

$$\epsilon = [\langle (Q_{true} - Q_{sample})^2 \rangle]^{1/2}. \quad (2.1.3)$$

Please notice that

$$\epsilon \neq \varepsilon \quad (2.1.4)$$

in general.

In practice the ensemble average is approximated by a weighted average:

$$\epsilon^2 = \sum_{n=1}^N w_n (Q_{true} - Q_n)^2, \quad (2.1.5)$$

where Q_n is the Q quantity derived from the n th measurement, which can be either the n th temporal measurement (for the satellite case) or the n th spatial measurement (for the case of fixed stations). The weights w_n satisfy a normalization condition:

$$\sum_{n=1}^N w_n = 1. \quad (2.1.6)$$

In the past climatological practice, it has been quite often to take the weights to be uniform:

$$w_n = \frac{1}{N}, \quad n = 1, 2, \dots, N. \quad (2.1.7)$$

However, because of the development of the mathematical theory in climatology and the power of modern computers, people tend to favor the use of optimal weights. We will talk about the optimal weighting theory in Chapter 4.

The relative error often used in climatology is given by

$$\frac{[(Q_{true} - Q_{sample})^2]^{1/2}}{|Q_{true}|}. \quad (2.1.8)$$

As we mentioned before, usually we do not know the true value Q_{true} (which is exactly the quantity we desired to measure by incomplete samples). Hence one has no way to carry out the computations according to the above formulas. But in certain cases we want to calibrate a new sampling technique. For instance, TRMM is a new technology and we need to know its sampling error. There are two ways to evaluate the sampling error for the new technology. One needs to use ground truth and the other needs to use models of the field. For the former, we can regard a very high resolution measurement as the ground truth. For instance, the rain rate over Japan derived from the AMeDAS data, whose spatial resolution is 5.0×5.0 [km], may be regarded as the truth. Then the TRMM sampling error over Japan can be computed by

$$\frac{[\sum_{n=1}^N (R_{AMeDAS} - R_{n,TRMM})^2 / N]^{1/2}}{R_{AMeDAS}} \times 100\%. \quad (2.1.9)$$

Please refer to Oki and Sumi (1994) for details.

The other way makes use of mathematical models, either a stochastic model for the field or the statistical model for fitting the data. In this case, the spectrum of the field, which is derived from the second moment of the stochastic field, is crucial in the computation. In the rest of this chapter, we will

exclusively talk about this method. We mainly discuss the spectrum-filter theory developed by North and Nakamoto (1989) for computing the mean square errors (MSE). They considered the measurement of the average rain-rate whose unit is [unit rain] [unit area]⁻¹ [unit time]⁻¹ in the space-time box $\Omega = [0, L] \times [0, L] \times [0, T]$ by using both satellite and rain gauge devices. For the rain gauge measurement, there are spatial gaps between the rain gauges and for the satellite measurement there are temporal gaps (except for stationary satellites) between different visits of a fixed area. They formulated a theory that can be used to estimate the mean square sampling errors (not including the instrumental errors). From mathematical view point, their main contribution is the derivation of the MSE in the spectrum-filter form. Of course, from climatological view point, the implications they derived from the spectrum-filter formula is perhaps more important.

2.2 Spectrum-filter formula for satellite MSE

The accurate measurement of rain rate in an area is of apparent importance in atmospheric sciences, agriculture, and civil engineering, etc. For example, from the record of the rain rate one can monitor the release rate of the regional latent heat which is an important drive of the atmospheric general circulation.

To make the theory simple, we consider measuring how much rain Φ is contained in the space-time box $\Omega = [0, L] \times [0, L] \times [0, T]$ shown in Fig. 2.1. For instance, if $L = 500$ [km] and $T = 30$ [day], the Φ is the rainfall in these 30 days over the square area of $250,000$ [km]². The average rain rate in Ω is

$$\Psi = \frac{\Phi}{TL^2}. \quad (2.2.1)$$

The rain rate field (i.e. the instant and point-to-point rain rate) which cannot be measured directly by an instrument is denoted by $\psi(\mathbf{r}, t)$ [unit rain][unit area]⁻¹ [unit time]⁻¹. This is a stochastic field whose structure, however, is quite complicated. First of all most of the time at a place it is not raining and hardly all the points in an interested area rain at the same time. The PDF of $\psi(\mathbf{r}, t)$ has a large peak at the zero rain rate and a fat tale (i.e. the heavy rain) according to Kedem et al. (1990). This property can be pretty well described by a mixed-lognormal distribution.

The stochastic property directly used in the calculation of MSE is the covariance function. Again, to make the theory simple we assume that the covariance structure is homogeneous in space and stationary in time:

$$\langle \psi(\mathbf{r}, t) \psi(\mathbf{r}', t') \rangle = \sigma^2 \rho(\xi, \tau) \quad (2.2.2)$$

where $\sigma^2 = \langle \psi^2(\mathbf{r}, t) \rangle$ is the point variance of the field. Since the field is assumed to be homogeneous in space and stationary in time, the point variance σ^2 is a constant. The function $\rho(\xi, \tau)$ is the autocorrelation function with $\xi = \mathbf{r} - \mathbf{r}'$, $\tau = t - t'$ and $\rho(0, 0) = 1$.

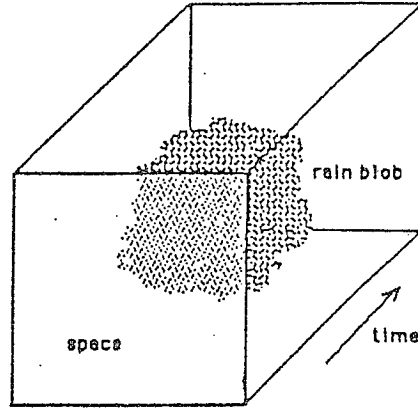


Figure 2.1: The rain clusters in the space-time box $\Omega = [0, L] \times [0, L] \times [0, T]$.

The Fourier transform pairs for $\rho(\xi, \tau)$ are defined by

$$S(\nu, f) = \int_{-\infty}^{\infty} d\xi_1 \int_{-\infty}^{\infty} d\xi_2 \int_{-\infty}^{\infty} d\tau \rho(\xi, \tau) e^{2\pi i(\xi \cdot \nu + f\tau)}, \quad (2.2.3)$$

$$\rho(\xi, \tau) = \int_{-\infty}^{\infty} d\nu_1 \int_{-\infty}^{\infty} d\nu_2 \int_{-\infty}^{\infty} df S(\nu, f) e^{-2\pi i(\xi \cdot \nu + f\tau)}, \quad (2.2.4)$$

where $\nu = (\nu_1, \nu_2)$ and $\xi = (\xi_1, \xi_2)$. The average rain rate in Ω is

$$\Psi = \frac{1}{TL^2} \int_{\Omega} d\Omega \psi(\hat{x}, t). \quad (2.2.5)$$

When using a satellite to measure the rain, it take a picture over the area $[0, L] \times [0, L]$ at a time, and then it comes back in a Δt time to take another picture. Usually Δt is designed to be close to 12 hours. The ideal case is that every time the satellite picture covers the whole area $[0, L] \times [0, L]$. This is called the flush visit. But of course this is not realistic. When a satellite is over the area $[0, L] \times [0, L]$, its picture often covers part of the area $[0, L] \times [0, L]$. Moreover, when the satellite passes the adjacent area of $S = [0, L] \times [0, L]$, its picture can also cover part of $[0, L] \times [0, L]$. Hence the area $S = [0, L] \times [0, L]$ may have partial coverage for more than twice a day (see Shin and North, 1988; North et al., 1992; and Oki and Sumi, 1994). But that is getting too complicated to be dealt with here. We still consider the ideal situation: the rain rate field is homogeneous in space and stationary in time; the visits are flush

and twice a day. So the average rain rate derived from the satellite pictures is

$$\Psi_s = \frac{1}{TL^2} \sum_{n=1}^N \int_0^L dx \int_0^L dy \psi(\hat{r}, n\Delta t) w_n, \quad (2.2.6)$$

where $\Delta t = T/N$ (see Fig. 2.2) and w_n are the weights subject to

$$\sum_{n=1}^N w_n = T. \quad (2.2.7)$$

The weights are part of the sampling design. Of course the simplest weights are the uniform ones: $w_n = T/N$. The case of optimal weights (often non-uniform weights) for satellite samplings is still a research problem.

SATELLITE SAMPLING DESIGN

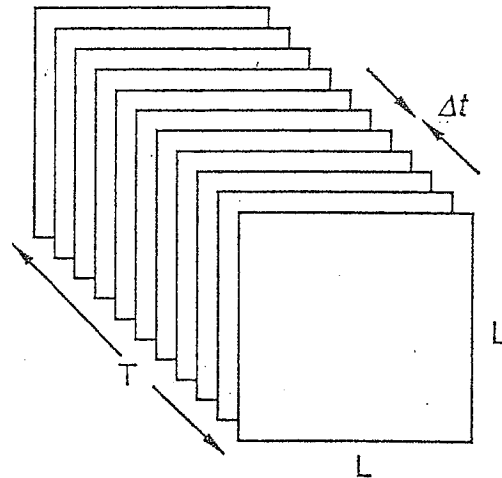


Figure 2.2: The uniform sampling design for a satellite.

We can re-write the Ψ_s formula into the following form

$$\Psi_s = \frac{1}{TL^2} \int_{\Omega} d\Omega K(t) \psi(\hat{r}, t), \quad (2.2.8)$$

where

$$K(t) = T \sum_{n=1}^N \delta(t - n\Delta t). \quad (2.2.9)$$

The difference between the true rain rate Ψ and the estimated value from the satellite Ψ_s is measured by the MSE:

$$\epsilon^2 = \langle (\Psi - \Psi_s)^2 \rangle, \quad (2.2.10)$$

which can be changed to:

$$\begin{aligned}
 \epsilon^2 &= \int_{\Omega} d\Omega \int_{\Omega'} d\Omega' \langle \psi(\hat{\mathbf{r}}, t) \psi(\hat{\mathbf{r}}', t') \rangle [1 - K(t)][1 - K(t')] \\
 &= \sigma^2 \int_{\Omega} d\Omega \int_{\Omega'} d\Omega' \int_{-\infty}^{\infty} d\nu_1 \int_{-\infty}^{\infty} d\nu_2 \int_{-\infty}^{\infty} df S(\nu, f) e^{-2\pi i(\xi \cdot \nu + f\tau)} \\
 &\quad \times [1 - K(t)][1 - K(t')] \\
 &= \sigma^2 \int_{-\infty}^{\infty} d\nu_1 \int_{-\infty}^{\infty} d\nu_2 \int_{-\infty}^{\infty} df S(\nu, f) \\
 &\quad \times \int_{\Omega} d\Omega \int_{\Omega'} d\Omega' e^{-2\pi i(\xi \cdot \nu + f\tau)} [1 - K(t)][1 - K(t')].
 \end{aligned}$$

The second part of the last expression is the filter we desired to find. It is denoted by H_s^2 . After some straightforward but tedious manipulations, we can get

$$H_s(\nu, f) = G(\nu_1 L) G(\nu_2 L) G(fT) \times \left[1 - \frac{1}{G(f\Delta t)} \right], \quad (2.2.11)$$

and

$$G(x) = \frac{\sin \pi x}{\pi x}. \quad (2.2.12)$$

Finally we have a compact and nice spectrum-filter formula:

$$\epsilon^2 = \sigma^2 \int_{-\infty}^{\infty} d\nu_1 \int_{-\infty}^{\infty} d\nu_2 \int_{-\infty}^{\infty} df S(\nu, f) H_s^2(\nu, f). \quad (2.2.13)$$

Hence for given design parameters: L, T, N , the filter can be constructed. If the spectrum $S(\nu, f)$ is known, then in principle we can calculate the sampling error by the above formula using a numerical integration method. The spectrum $S(\nu, f)$, in general, can be determined by: (i) data, (ii) empirical formulas, and (3) climate models. The data approach is not realistic here since we have assumed the homogeneity, but an observation data set is often inhomogeneous. The empirical formula approach attracts lots of criticism lately because of its difficulty for validation. North and Nakamoto took the third approach: deriving the spectrum from a mathematical model. We will talk about this model in Section 2.4.

When L, N, T are large, the filter H_s^2 can be approximated by a considerably simpler expression. The numerical integration of (2.2.13) becomes not necessary. Please bear in mind that the significance of the approximation is not the simple computational convenience, but the exposition of the properties of the rain physics and the sampling design.

When L is large, we have

$$G^2(\nu_l L) = \left(\frac{\sin \pi \nu_l L}{\pi \nu_l L} \right)^2 \sim \frac{1}{L} \delta(\nu_l), \quad l = 1, 2. \quad (2.2.14)$$

Fig. 2.3 shows the convergence process of the above as L gets larger. The figure was generated by the following Mathematica commands:

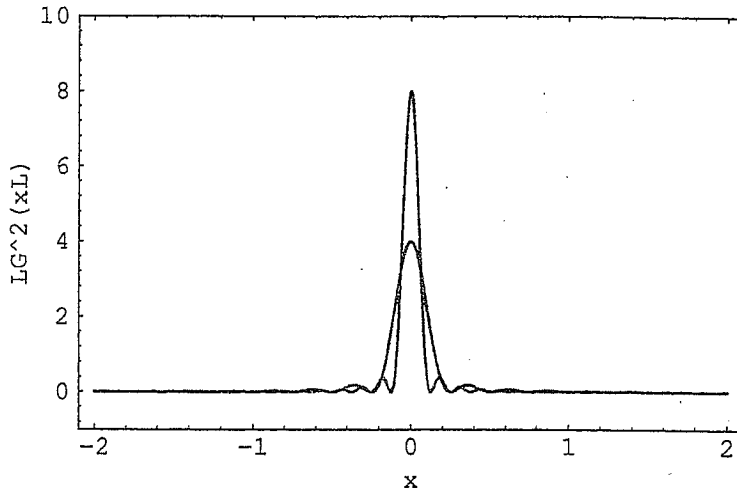


Figure 2.3: The convergence process of a delta-convergent sequence: $LG^2(xL) \sim \delta(x)$ as $L \rightarrow \infty$.

```
l=4;
Plot[1 (Sin[Pi x l] / (Pi x l))^2, {x,-2,2}, PlotRange->{-1,10},
PlotPoints->90, FrameLabel->{"x", "LG^2(xL)", "", ""},
PlotRegion->{{0.2,0.8}, {0.2,0.8}},
Frame->True, Axes->False]
```

Similarly,

$$G^2(fT) \times \left[1 - \frac{1}{G(f\Delta t)}\right]^2 \quad (2.2.15)$$

$$= \frac{\sin^2(\pi fT)}{N^2 \sin^2(\pi f\Delta t)} [1 - G(\pi f\Delta t)]^2$$

$$\sim \frac{1}{T} \sum_{n \neq 0} \delta\left(f - \frac{n}{\Delta t}\right), \quad (2.2.16)$$

since

$$\lim_{N \rightarrow \infty} \frac{\Delta t \sin^2(N\pi\Delta t f)}{N \sin^2(\pi\Delta t f)} = \sum_{n=-\infty}^{\infty} \delta\left(f - \frac{n}{\Delta t}\right). \quad (2.2.17)$$

See Fig. 2.4 for the convergence process of the last limit. We call this limit the Dirac comb, which is also generated by Mathematica:

```
del = 0.3;
n=12;
Plot[ del * (Sin[n Pi del x])^2 / ( n (Sin[Pi del x])^2),
```

```
{x,-11, 11}, PlotRange->{0,5},
PlotPoints->120, FrameLabel->{"f", "Dirac comb", "", ""},
PlotRegion->{{0.2,0.8}, {0.2,0.8}},
Frame->True, Axes->False]
```

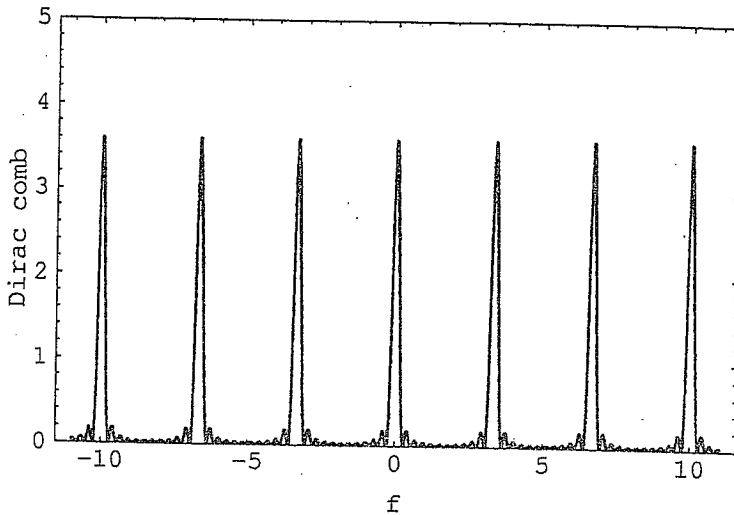


Figure 2.4: The convergence process of the Dirac comb.

Therefore, when L, T and N are large, the filter H_s can be approximated by

$$H_s^2(\nu, f) \sim \frac{1}{L^2 T} \Delta(\nu_1) \Delta(\nu_2) \sum_{n \neq 0} \delta\left(f - \frac{n}{\Delta t}\right). \quad (2.2.18)$$

It follows that the MSE formula (2.2.13) can be approximated by

$$\epsilon^2 = \frac{\sigma^2}{L^2 T} \sum_{n \neq 0} S(0, n/\Delta t). \quad (2.2.19)$$

Thus the sampling error due to the temporal gaps start to build up from the sampling frequency $1/\Delta t$ and continues with the multiples of $1/\Delta t$. The zero frequency part is exactly sampled since it associates with the uniform fluctuation pattern.

Although the above formula being quite simple, when the spectrum formula is known, the formula can be further simplified and makes the properties of the physics and the sampling design more transparent. We will see this in Section 2.4.

2.3 Spectral-filter formula for rain gauges

We take the simplest design of the rain gauges: $M \times M$ gauges uniformly distributed over $[0, L] \times [0, L]$. The gap between each pair of the nearest neighborhood gauges is $\Delta l = L/M$. See Fig. 2.5 for the design.

These gauges can make the continuous measurement of rain in time but leave gaps in space. The average rain rate in $\Omega = [0, L] \times [0, L] \times [0, T]$ derived from the gauge data is

$$\Psi_g = \frac{1}{TM^2} \sum_{n_1=1}^M \sum_{n_2=1}^M \int_0^T dt \psi(n_1 \Delta l, n_2 \Delta l, t). \quad (2.3.1)$$

This can be written in integral form

$$\Psi_g = \frac{1}{TL^2} \int_{\Omega} d\Omega K(\bar{r}) \psi(\bar{r}), \quad (2.3.2)$$

where

$$K(\bar{r}) = L^2 \sum_{n_1=1}^M \sum_{n_2=1}^M \delta(\xi_1 - n_1 \Delta l) \times \delta(\xi_2 - n_2 \Delta l). \quad (2.3.3)$$

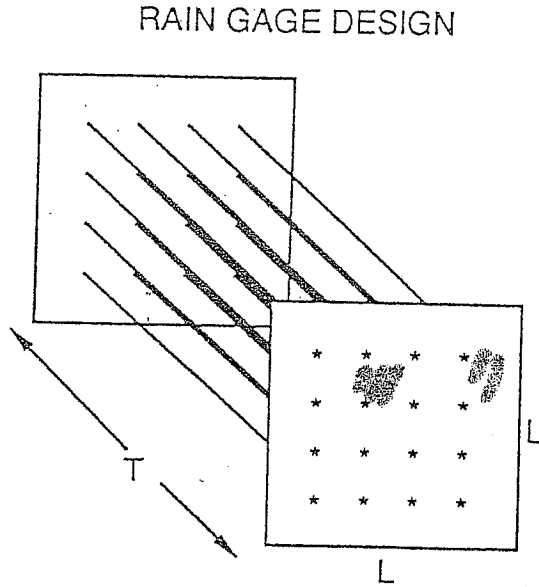


Figure 2.5: The uniform sampling design for rain gauges in the space-time box: $\Omega = [0, L] \times [0, L] \times [0, T]$.

The difference between the true rain rate Ψ and the measured Ψ_g is evaluated by the MSE:

$$\epsilon^2 = \langle (\Psi - \Psi_g)^2 \rangle, \quad (2.3.4)$$

which can be re-written as

$$\begin{aligned}
 \epsilon^2 &= \int_{\Omega} d\Omega \int_{\Omega'} d\Omega' \langle \psi(\hat{\mathbf{r}}, t) \psi(\hat{\mathbf{r}}', t') \rangle [1 - K(\hat{\mathbf{r}}, t)] [1 - K(\hat{\mathbf{r}}', t')] \\
 &= \sigma^2 \int_{\Omega} d\Omega \int_{\Omega'} d\Omega' \int_{-\infty}^{\infty} d\nu_1 \int_{-\infty}^{\infty} d\nu_2 \int_{-\infty}^{\infty} df S(\nu, f) e^{-2\pi i(\xi \cdot \nu + f\tau)} \\
 &\quad \times [1 - K(\hat{\mathbf{r}})] [1 - K(\hat{\mathbf{r}}')] \\
 &= \sigma^2 \int_{-\infty}^{\infty} d\nu_1 \int_{-\infty}^{\infty} d\nu_2 \int_{-\infty}^{\infty} df S(\nu, f) \\
 &\quad \times \int_{\Omega} d\Omega \int_{\Omega'} d\Omega' e^{-2\pi i(\xi \cdot \nu + f\tau)} \times [1 - K(\hat{\mathbf{r}})] [1 - K(\hat{\mathbf{r}}')].
 \end{aligned}$$

The second part of the last expression is the filter we desired to find. It is denoted by H_g^2 . Like the satellite case, after some straightforward but tedious manipulations, we can get

$$H_g(\nu, f) = G(fT)G(\nu_1 L)G(\nu_2 L) \left[1 - \frac{1}{G(\nu_1 \Delta l)G(\nu_2 \Delta l)} \right]. \quad (2.3.5)$$

We also have the spectrum-filter formula:

$$\epsilon^2 = \sigma^2 \int_{-\infty}^{\infty} d\nu_1 \int_{-\infty}^{\infty} d\nu_2 \int_{-\infty}^{\infty} df S(\nu, f) H_g^2(\nu, f). \quad (2.3.6)$$

The rain gauge filter H_g^2 also has an approximation which is similar to that of H_s^2 , when M, L and T are large:

$$H_g^2(\nu, f) \sim \frac{1}{TL^2} \delta(f) \times \sum_{n_1, n_2 \neq 0} \delta\left(\nu_1 - \frac{n_1}{\Delta l}\right) \times \delta\left(\nu_2 - \frac{n_2}{\Delta l}\right). \quad (2.3.7)$$

It follows that the MSE formula (2.3.6) can be approximated by

$$\epsilon^2 = \frac{\sigma^2}{L^2 T} \sum_{n_1, n_2 \neq 0} S\left(\frac{n_1}{\Delta l}, \frac{n_2}{\Delta l}, 0\right). \quad (2.3.8)$$

Hence the sampling error accumulates from the sampling wavenumber $1/\Delta l$. The zero wavenumber part of the spectral power is exactly sampled because of uniform fluctuation pattern.

2.4 A simple diffusive rain model

Here we regard that the stochastic rain rate field is generated and destroyed by a random and uncorrelated source in space and time (i.e. white noise); the field of rain rates is modified by the exponential decay of the whole field at an intrinsic time scale τ_0 ; and finally the rain rates are displaced from one point to

another by a simple down gradient diffusion process. These mechanisms lead to the following mathematical model:

$$\tau_0 \frac{\partial \psi}{\partial t} - \lambda_0 \nabla^2 \psi + \psi = F(\hat{\mathbf{r}}, t). \quad (2.4.1)$$

Here, λ_0 is the diffusion length scale, and F is the "red" noise forcing (i.e., the high frequency part of the F is set to be zero).

With this kind of assumption, the stochastic process of ψ is Gaussian. Here, again we are in a non-realistic situation. As remarked at earlier in this chapter, the realistic PDF of ψ is more like a mixed-lognormal distribution. Since our model (2.4.1) is simpler and has some deviation from a more realistic model, we obviously ought to have some gain by doing this. The advantage of assuming this model is two folds. First, the model does have two intrinsic scales: time scale τ_0 and space scale λ_0 . These are the most important scales in all the rain models available. Hence the model does carry some true physical property of rain. Further one can also adjust the cut-off frequency of the "red" noise forcing to fit whatever rain field one desires to model. Secondly, the model allows us to find explicit spectrum formula. Using it, we can write our MSE formulas in very simple forms, which clearly expose the properties of the rain field and the sampling design. Consequently, our understanding of the rain physics and sampling designs can be improved and extended to more complicated and more realistic cases.

The Fourier transform pair for ψ is

$$\bar{\psi}(\nu, f) = \int d\hat{\mathbf{r}}^2 dt \psi(\hat{\mathbf{r}}, t) e^{i2\pi\hat{\mathbf{r}} \cdot \nu}, \quad (2.4.2)$$

$$\psi(\hat{\mathbf{r}}, t) = \int d\nu^2 df \bar{\psi}(\nu, f) e^{-i2\pi\hat{\mathbf{r}} \cdot \nu}, \quad (2.4.3)$$

where the integration limits of the triple integral are all from $-\infty$ to ∞ .

We now take the Fourier transform of the model equation (2.4.1). The transform leads to

$$\bar{\psi}(\nu, f) = \frac{\bar{F}(\nu, f)}{2\pi i \tau_0 f + (1 + 4\pi^2 \lambda_0^2 \nu^2)}, \quad (2.4.4)$$

where $\nu = |\nu|$.

Since the ψ field is assumed to be homogeneous, the spectra of ψ should be uncorrelated at different frequencies and wavenumbers:

$$\langle \bar{\psi}(\nu, f) \bar{\psi}^*(\nu', f') \rangle = \sigma^2 S(\nu, f) \delta(\nu - \nu') \times \delta(f - f'), \quad (2.4.5)$$

where * signifies for complex conjugate. This assertion can be verified in the following way:

$$\frac{\langle |\bar{F}(\nu, f)|^2 \rangle}{|2\pi i \tau_0 f + (1 + 4\pi^2 \lambda_0^2 \nu^2)|^2} \delta(\nu - \nu') \times \delta(f - f')$$

$$\begin{aligned}
&= \langle \bar{\psi}(\nu, f) \bar{\psi}^*(\nu', f') \rangle \\
&= \left\langle \int d\hat{\mathbf{r}}^2 dt \psi(\hat{\mathbf{r}}, t) e^{i2\pi(\nu \cdot \hat{\mathbf{r}} + ft)} \int d\hat{\mathbf{r}}'^2 dt' \psi(\hat{\mathbf{r}}', t') e^{-i2\pi(\nu' \cdot \hat{\mathbf{r}}' + f' t')} \right\rangle \\
&= \int d\xi d\tau \sigma^2 \rho(\xi, \tau) e^{i2\pi(\xi \cdot \hat{\mathbf{r}} + f\tau)} \times \int d\hat{\mathbf{r}}'^2 dt' e^{i2\pi[(\nu - \nu') \cdot \hat{\mathbf{r}}' + (f - f')t']} \\
&= \sigma^2 S(\nu, f) \delta(\nu - \nu') \times \delta(f - f').
\end{aligned}$$

Hence,

$$S(\nu, f) = \langle |\psi(\nu, f)|^2 \rangle = \frac{\langle |\bar{F}(\nu, f)|^2 \rangle / \sigma^2}{4\pi^2 \tau_0^2 f^2 + (1 + 4\pi^2 \lambda_0^2 \nu^2)^2}. \quad (2.4.6)$$

To satisfy the normalization condition

$$\rho(0, 0) = \int_{-\infty}^{\infty} d\nu^2 df S(\nu, f) = 1, \quad (2.4.7)$$

the forcing noise must be cut off at a critical wavenumber ν_c , i.e.,

$$\langle |\bar{F}(\nu, f)|^2 \rangle = \begin{cases} \alpha \sigma^2 & \text{if } \nu < \nu_c, \\ 0 & \text{otherwise.} \end{cases} \quad (2.4.8)$$

This α is given by

$$\alpha = \frac{8\pi\tau_0\lambda_0^2}{\ln(1 + 4\pi^2\lambda_0^2\nu_c^2)}. \quad (2.4.9)$$

Using the spectrum formula (2.4.6) and the summation formula

$$\sum_{n=1}^{\infty} \frac{1}{1 + a^2 n^2} = \frac{1}{2} \left[\frac{\pi}{a} \coth\left(\frac{\pi}{a}\right) - 1 \right], \quad a \neq 0, \quad (2.4.10)$$

we can further simplify the MSE formula (2.2.19) for satellite:

$$\begin{aligned}
\epsilon^2 &= \frac{\sigma^2 \alpha}{L^2 T} \frac{1}{2} \sum_{n=1}^{\infty} \frac{1}{1 + 4\pi^2 \left(\frac{\tau_0}{\Delta t}\right)^2 n^2} \\
&= \frac{\sigma^2 \alpha}{L^2 T} \frac{1}{2} \left[\frac{\pi}{2\pi(\tau_0/\Delta t)} \coth\left(\frac{\pi}{2\pi(\tau_0/\Delta t)}\right) - 1 \right] \\
&= \frac{\sigma^2 \alpha}{TL^2} \left[\frac{\Delta t}{2\tau_0} \coth\left(\frac{\Delta t}{2\tau_0}\right) - 1 \right]. \quad (2.4.11)
\end{aligned}$$

In the above, the summation formula (2.4.10) can either be found from a mathematics handbook or derived from the the Fourier series of the coth function.

When $\Delta t/\tau_0$ is small, the Laurent expansion of coth function can be used to further reduce the MSE formula (2.4.11). The Laurent expansion is

$$\coth x = \frac{1}{x} + \frac{1}{3}x - \frac{1}{45}x^3 + \dots, \quad 0 < |x| < \pi. \quad (2.4.12)$$

When $\Delta t/\tau_0$ is small, after throwing out the $(\Delta t/\tau_0)^4$ and the higher order terms and retaining the $(\Delta t/\tau_0)^4$ term, we have

$$\epsilon^2 = \frac{\sigma^2 \alpha}{2\tau_0 L^2} \frac{1}{6} \frac{\Delta t}{T} \frac{\Delta t}{\tau_0}. \quad (2.4.13)$$

We can perform similar calculations for the case of rain gauges, but it appears difficult to use some summation formulas to get a compact form like (2.4.13). In any event, the simplest form we can further get from the MSE formula (2.3.8) is

$$\epsilon^2 = \frac{4\sigma^2 \alpha}{TL^2} \sum_{n_1, n_2 \neq 1}^{\infty} \frac{1}{[1 + 4\pi^2(\lambda_0/\Delta l)^2(n_1^2 + n_2^2)]^2}. \quad (2.4.14)$$

2.5 Interpretations of the results

In the final formulas (2.4.11) and (2.4.14), the quantity α is not yet known, although we know that it reflects the fluctuations of the noise forcing. One way to measure it is to use the ensemble average of the all the square of all the satellite pictures:

$$\sigma_A^2 = \left\langle \int_{[0,L] \times [0,L]} dA \psi^2(\bar{r}, t) \right\rangle. \quad (2.5.1)$$

Since the rain process is assumed to be stationary, the above quantity is independent of t . Similar to the derivation of the spectrum-filter formula, we can also derive the spectral representation of σ_A^2 :

$$\sigma_A^2 = \sigma^2 \int d\nu^2 df G^2(\nu_1 L) G^2(\nu_2 L) S(\nu, f). \quad (2.5.2)$$

When L is large, we have

$$G^2(\nu_1 L) \sim \frac{1}{L} \delta(\nu_1). \quad (2.5.3)$$

Using this asymptotic approximation and the spectrum formula (2.4.6), we have

$$\sigma_A^2 \approx \sigma^2 \frac{\alpha}{2\tau_0 L^2}. \quad (2.5.4)$$

The percentage sampling errors are therefore:

$$\frac{\epsilon_s}{\sigma_A} = \left\{ \frac{2\tau_0}{T} \left[\frac{\Delta t}{2\tau_0} \coth \left(\frac{\Delta t}{2\tau_0} \right) - 1 \right] \right\}^{1/2}, \quad (2.5.5)$$

for the satellite case, and

$$\frac{\epsilon_g}{\sigma_A} = \left\{ \frac{8\tau_0}{T} \sum_{n_1, n_2 \neq 1}^{\infty} \frac{1}{[1 + 4\pi^2(\lambda_0/\Delta l)^2(n_1^2 + n_2^2)]^2} \right\}^{1/2} \quad (2.5.6)$$

$$\epsilon = \frac{1}{N} \sum_n^N \sigma_A(\text{sat})$$

for the case of rain gauges.

For the GATE [GARP (Global Atmospheric Research Program) Atlantic Tropical Experiment] data and TRMM satellite, $\Delta t \approx 12$ [hour] and $\tau_0 \approx 12$ [hour], the MSE formula (2.5.5) gives that

$$\frac{\epsilon_s}{\sigma_A} = 0.0523 = 5.23\%. \quad (2.5.7)$$

This error (5%) is too small compared with the more acceptable result of 10%. The reason is that we have assumed that the satellite makes flush visits. In the real situation, a satellite makes only partial coverage of the square $[0, L] \times [0, L]$. Let us look an example on TRMM: altitude 350 [km], inclination 35° , nominal swath width 600 [km]. For a grid box of 500 [km] by 500 [km] on the equator, on average every visit of the TRMM satellite covers only 44% of the square $[0, L] \times [0, L]$ where $L = 500$ [km]. Therefore the actual sampling error is more than double the error computed from the flush visit case. More detailed computation shows that the sampling error for TRMM is 11.2% (North et al., 1993). This is a quite realistic number. But if one takes into account of the seasonal variation, diurnal cycle, spatial inhomogeneity, and perhaps other properties of the rain rate field, the sampling error can be significantly larger than this figure. It may get so large that it reaches 20% or even more for a single TRMM satellite (see Oki and Sumi, 1994).

2.6 References

- Kedem, B., L. Chiu, and G.R. North, 1990: Estimation of mean rainrate: Application to satellite observations. *J. Geophys. Res.*, **95**, 1965-1972.
- North, G.R., and S. Nakamoto, 1989: Formalism for comparing rain estimation designs. *J. Atmos. Oceanic Tech.*, **6**, 985-992.
- Oki, R., and A. Sumi, 1994: Sampling simulation of TRMM rainfall estimation using radar-AMeDAS composites. *J. Appl. Meteo.*, **33**, 1597-1608.
- North, G.R., S. S. Shen, and R. Upson, 1993: Sampling errors in rainfall estimates by multiple satellites. *J. Appl. Meteo.*, **32**, 399-410.

Chapter 3

EOFs

The climate may be regarded as a superposition of a fixed state and fluctuations about the fixed state. The fixed state is called the climatology and the fluctuations are called the anomalies. For example, when we say that the monthly average temperature of Tokyo in September is 20°C, this 20°C is the climatology. The climatology can be a good reference for us to know the climate of an area. But for an individual year, the climate may deviate quite a bit from the climatology. The deviation is described by anomalies. It is these anomalies which have direct impact to the change of the weather. People have been interested to know the patterns of the anomaly field. These patterns may be regarded as a basis functions in an infinite dimensional space. Weather can be projected to these bases. Therefore, these basis functions are helpful for statistical weather forecasting.

These basis functions are the empirical orthogonal functions (EOFs). They are the eigenvectors of the covariance matrix of the anomaly field. In this chapter, we will discuss the basic theory on computing EOFs from observation data, simple model and GCM output. We will also point out some possible troubles in computing EOFs, and unfortunately these troubles, which might be very serious, are often ignored by most of the EOF fans.

3.1 Covariance matrix and data preparation

The monthly average temperature anomalies at N stations in Kabayamaka Province are denoted by $\{X_1, X_2, \dots, X_N\}$. They are RVs whose expectation values are equal to zero. The matrix

$$C = \left[\langle X_i X_j \rangle \right]_{i,j=1}^N \quad (3.1.1)$$

is called the covariance matrix. Here $\langle \cdot \rangle$ still denotes the ensemble average.

In practice, the first question is how we derive the anomaly data from the raw data and how we compute the covariance matrix C . The climatology is a "definite" signal buried in the raw data (although there is no definition of

climatology which yields the unique output). To get the anomaly, we can delete the climatology and some definite variance from the raw data, i.e.,

$$\text{anomaly} = (\text{raw data}) \text{ DELETES } (\text{definite information}). \quad (3.1.2)$$

The climatology, as a definite part of the climate, may include: mean (the most important part), seasonal cycle, land-ocean contrast, trend (e.g., heat island effect), etc. We present an example for preparing the anomaly of annual mean temperature from the "raw" data for a station. (Here the "raw" data is not exactly raw. The instrument errors and human recording errors are assumed being removed, and the monthly average of the readings has already been computed.)

1. **Seasonal cycle:** Up to now, most climatologists regard the 30 years mean between 1951 and 1980 as the climatology. The raw data is denoted by $Z(\alpha, \beta)$ where α and β denote year and month respectively. For instance, if the data set is from 1910 to 1994, then $\alpha = 1910, 1911, \dots, 1994$ and $\beta = \text{Jan, Feb, Mar, } \dots, \text{Dec}$ (or simply $1, 2, 3, \dots, 12$). The β -th month mean temperature (as the climatology) is

$$\theta_S(\beta) = \frac{1}{30} \sum_{\alpha=1951}^{1980} Z(\alpha, \beta). \quad (3.1.3)$$

This gives the seasonal cycle of the data (see Fig. 3.1). Then the annual mean temperature (also as the climatology) is

$$\theta_A = \frac{1}{12} \sum_{\beta=1}^{12} \theta_S(\beta). \quad (3.1.4)$$

We now first remove the seasonal cycle to get the monthly anomaly:

$$\theta_M(\alpha, \beta) = Z(\alpha, \beta) - \theta_S(\beta). \quad (3.1.5)$$

2. **Trend:** Due to anthropologic forcings, the above processed anomaly may include a trend, say, a warming trend. As for what type of the trend it is (linear trend, polynomial trend, or some other curves), it depends on each individual station. One way to get such a trend is to use the moving average:

$$\theta_{trend}(\alpha, \beta) = \frac{1}{2 \times Lng} \sum_{j=-Lng}^{Lng} \theta_M(\alpha + j, \beta). \quad (3.1.6)$$

This is a moving average of length $2 \times Lng$ and the index α runs from $1910 + Lng$ to $1994 - Lng$. The function $\theta_{trend}(\alpha, \beta)$ is a smoothed curve of $\theta_M(\alpha, \beta)$ for a fixed β and may be regarded as the trend.

Or one can assume a linear trend or other types of known functions. Then using the least square method to fit the function to get the trend.

With the trend, one can process the above anomaly $\theta_M(\alpha, \beta)$ again:

$$\theta_{MT}(\alpha, \beta) = \theta_M(\alpha, \beta) - \theta_{trend}(\alpha, \beta). \quad (3.1.7)$$

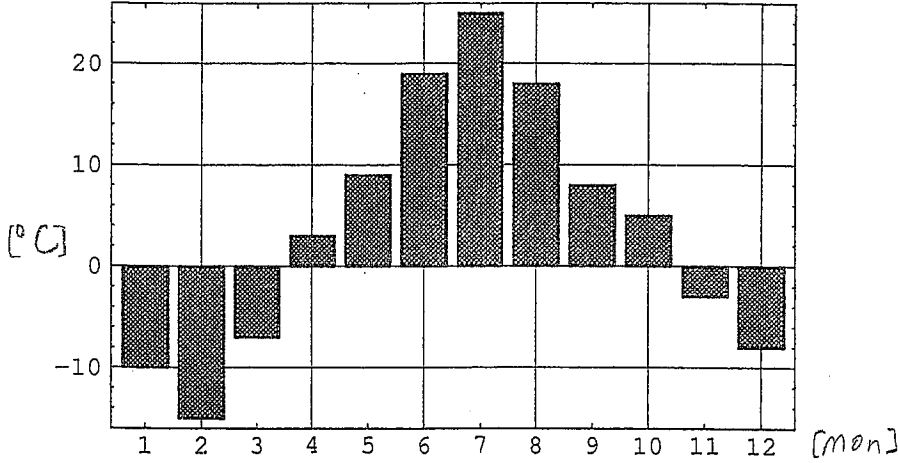


Figure 3.1: Climatology of the monthly average temperature.

This $\theta_{MT}(\alpha, \beta)$ is the anomaly with both seasonal cycle and trend removed.

3. Variance: We know that the variance of the temperature over land is much larger than that over ocean, and the variance over the higher latitude is larger than that over the lower latitude due to the land-ocean distribution and the inclination angle of the earth's process around sun. Therefore, this land-ocean contrast may also be regarded as a definite signal. Since the variance is a property of second moment statistics, we do not call it climatology. So we compute the variance:

$$\sigma_V^2(\beta) = \frac{1}{85} \sum_{\alpha=1910}^{1994} \theta_{MT}^2(\alpha, \beta). \quad (3.1.8)$$

When we do this computation, we have assumed that the time series $\theta_{MT}^2(\alpha, \beta)$ is stationary in α and the relevant stochastic process is ergodic. The variance is removed by:

$$T_S(\alpha, \beta) = \frac{\theta_{MT}(\alpha, \beta)}{\sigma_V(\beta)}. \quad (3.1.9)$$

This is called the standardized anomaly.

Finally we can compute the annual mean anomaly:

$$T(\alpha) = \frac{1}{12} \sum_{\beta=1}^{12} T_S(\alpha, \beta). \quad (3.1.10)$$

In summary, the standardized anomaly of the annual mean temperature can be computed by:

$$T(\alpha) = \frac{1}{12} \sum_{\beta=1}^{12} \frac{Z(\alpha, \beta) - \theta_S(\beta) - \theta_{trend}(\alpha, \beta)}{\sigma_V(\beta)}. \quad (3.1.11)$$

With the prepared anomaly data, we can talk about the computation of the covariance matrix. Let $T_j(\alpha)$ be the anomaly data for j th station and α th year for N stations. The length of the data stream is M_{yr} years. We assume the relevant time series are stationary and the relevant stochastic processes are ergodic (ergodicity = ensemble average is equivalent to the time average). Then the covariance matrix can be computed by

$$C = \frac{1}{M_{yr}} \sum_{\alpha=1}^{M_{yr}} T_i(\alpha)T_j(\alpha). \quad (3.1.12)$$

Ideally, this matrix is symmetric and positive definite.

Notice that the rank of this matrix is not larger than $\min(M_{yr}, N)$. For a given total number of stations N , if $M_{yr} < N$, then the covariance matrix necessarily does not have full rank. Consequently, its last a few eigenvalues must be zero, its determinant must vanish and the covariance matrix is not positive definite! Thus when the data stream is short, our replacement of the ensemble average by the time average can cause a problem. As for how serious is this problem, it is still a research topic and no definite answer has been given yet.

There are some preliminary research results on this short data stream problem based upon the perturbation theory. The idea is that the time averaged covariance matrix is regarded as a "small" perturbation of the true ensemble averaged covariance matrix. Then the mathematical question is that if the matrix has a small perturbation, what are responses of the eigenvalues and eigenvectors (North et al., 1984; Penland and Sardeshmukh, 1995). However, nobody knows what is the exact definition of the "small" perturbation.

The research on errors in computing EOFs (the eigenvectors of the covariance matrix) is challenging yet important. It is also an important research topic of modern mathematics. There are some pure mathematicians on linear operator theory who also worry about this problem. It is quite often that an infinite dimensional operator is represented by an infinite long matrix. When approximated by a finite matrix, how does it affect the characteristics of the operator: eigenvalues and eigenvectors? Therefore, before we know the answers to this "perturbation" problem, I would like to suggest the EOF fans be careful with the interpretation of their results.

3.2 Empirical orthogonal functions

We learned how to prepare the anomaly data. Suppose that the annual mean temperature anomalies at N stations in Kabayamaka Province have been obtained and are denoted by $\{X_1, X_2, \dots, X_N\}$. They are RVs whose expectation values are equal to zero. The matrix

$$C = [\langle X_i X_j \rangle]_{i,j=1}^N \quad (3.2.1)$$

is called the covariance matrix. This is a symmetric matrix and is usually positive definite. Its eigenvalue problem is

$$\mathbf{C} \mathbf{e} = \lambda \mathbf{e} \quad (3.2.2)$$

where \mathbf{e} is called the eigenvector and λ is called the eigenvalue. The solutions of this eigenvalue problem are called the eigenpairs $(\lambda_n, \mathbf{e}_n)$, $n = 1, 2, \dots, N$. The eigenvalues can be ordered $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_N$. The eigenvector corresponding to the largest eigenvalue is called the gravest mode (or the gravest eigenvector). When people say "the first a few modes", they refer to the order $\lambda_1 \geq \lambda_2 \geq \dots$.

The eigenvalues are usually different and the eigenvectors are usually normalized in the sense that $\mathbf{e}_n \cdot \mathbf{e}_n = 1$ for $n = 1, 2, \dots, N$. The eigenvectors corresponding to different eigenvalues are orthogonal:

$$\mathbf{e}_i \cdot \mathbf{e}_j = 0 \quad \text{when } i \neq j. \quad (3.2.3)$$

Hence, $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_N\}$ forms an orthonormal basis of an N -dimensional Euclidean space. This set of eigenvectors are called the empirical orthogonal functions (EOFs).

There are cases that two or three eigenvalues are the same. Hence for one eigenvalue, there might be two or more eigenvectors. Namely, this eigenspace is of more than one dimension. One can orthogonalize the eigenvectors in this eigenspace. So we still have a set of N orthonormal vectors.

From linear algebra, we know that for a symmetric matrix of no repeated eigenvalues, there is a similarity transform \mathbf{B} such that this symmetric matrix can be diagonalized by \mathbf{B} . Namely,

$$\mathbf{B} \mathbf{C} \mathbf{B}^{-1} = \{\lambda_i \delta_{ij}\}_{i,j=1}^N \quad (3.2.4)$$

where δ_{ij} is the Kronecker delta. The trace is invariant under this transform:

$$\sum_{i=1}^N \langle X_i^2 \rangle = \sum_{i=1}^N \lambda_i, \quad (3.2.5)$$

i.e., the sum of the eigenvalues is equal to the sum of the variances of all the RVs.

This leads us to exploit the meaning of the eigenvalues one step further. As we know that in a mechanical vibration system or a quantum mechanical system, the eigenvector represents the vibration pattern (or mode) and the corresponding eigenvalue measures the energy level of this pattern. Similar explanation can be made for EOFs.

Let $\mathbf{x} = \{x_1, x_2, \dots, x_N\}$ be observations of the RVs $\{X_1, X_2, \dots, X_N\}$. This observation can be expressed in terms of the EOFs:

$$\mathbf{x} = \sum_{n=1}^N c_n \mathbf{e}_n \quad (3.2.6)$$

where

$$c_n = \mathbf{x} \cdot \mathbf{e}_n, \quad n = 1, 2, \dots, N. \quad (3.2.7)$$

Then the variance of c_n is

$$\langle c_n^2 \rangle = \mathbf{e}_n^T \langle \mathbf{x} \mathbf{x}^T \rangle \mathbf{e}_n. \quad (3.2.8)$$

Since $\langle \mathbf{x} \mathbf{x}^T \rangle$ is the covariant matrix, the right hand side is equal to λ_n , i.e.

$$\langle c_n^2 \rangle = \lambda_n. \quad (3.2.9)$$

Therefore, the eigenvalue is the variance of observations at the corresponding mode. The percentage variance explained by a given mode \mathbf{e}_n is

$$\left(\lambda_n / \sum_{j=1}^N \lambda_j \right) \times 100\%. \quad (3.2.10)$$

(This is similar to the relative energy level in a mechanical system.) Thus the equation (3.2.5) not only gives the total energy $E = \sum_{i=1}^N \langle X_i^2 \rangle$ but also separates the energy according to the importance of the modes.

If an observation is expressed in terms of a sum of EOFs expansions, the signals (i.e. the variances) are in the superposition form. The gravest mode reflects the strongest signal. The weaker signals correspond to higher modes. But it is very often that people are only interested in the first a few modes. There are many reasons for this. First of all, the eigenvalues usually decrease really fast. It happens very often that the first ten eigenvalues explain more than 90% of the total variance, i.e. $\sum_{n=1}^{10} \lambda_n / \sum_{j=1}^N \lambda_j \times 100\% \geq 90\%$. The second reason is that the higher modes are associated with complicated patterns and small length (or time) scales. The observation network is not dense enough to resolve these smaller scales. Hence the higher modes, even though being computed, are far away from the real situation. Usually it is suggested to use a new observation network for detecting the signals of smaller scales. In this new network, the pattern of the smaller scales become the gravest modes. The third reason is that there can be large numerical errors in computing the higher modes, including both eigenvalues and eigenfunctions.

But, life is not easy. For some climate fields, their eigenvalues decay slowly. One sometimes has to include many (say, more than 40) eigenmodes. Further, some of the eigenvalues can be so close to each other that the eigenspace becomes two dimensional. Then it is quite hard to identify the eigenspaces when the resolution of a data set is not sufficiently fine. The relevant topics are worries of the modern statistical climatology and are under intensive investigations.

3.3 EOFs of a stochastic field

Consider a stochastic field $\Theta(\mathbf{x}, t)$ defined on a spatial domain Ω and a temporal domain $[-T/2, T/2]$. Still the expectation value is assumed zero. Hence the

signal of the field is its variance. The covariance function is defined as

$$K(\mathbf{x}, \mathbf{x}', t, t') = \langle \Theta(\mathbf{x}, t) \Theta(\mathbf{x}', t') \rangle. \quad (3.3.1)$$

We say that the stochastic field is stationary (in time) if $K(\mathbf{x}, \mathbf{x}', t, t') = K(\mathbf{x}, \mathbf{x}', |t - t'|) < \infty$ for any t and t' . We say that the stochastic field is homogeneous (in space) if $K(\mathbf{x}, \mathbf{x}', t, t') = K(|\mathbf{x} - \mathbf{x}'|, t, t') < \infty$. In physics and chemistry, this property is often referred to as the "isotropic" property since this property implies that the covariance is not only independent of the rigid shift but also independent of the rigid rotation. We have seen in Chapter 2 that if one assumes both stationarity and homogeneity, he can greatly simplify his analysis. However, for a highly nonhomogeneous field, the homogeneous approximation may yield misleading results, and one should retain the inhomogeneous property and use EOFs. Nevertheless, the stationary assumption is often retained for the following two reasons. First of all, the non-stationary characteristics in many cases are not prevailing. Secondly, there are no effective and systematic mathematical tools that deal with the non-stationarity.

In our book, we consider only two cases: (i) homogeneous and stationary fields, and (ii) nonhomogeneous and stationary fields.

Let $\tau = t - t'$. Then $K(\mathbf{x}, \mathbf{x}', \tau)$ may be expressed in terms of Fourier series

$$K(\mathbf{x}, \mathbf{x}', \tau) = \sum_{n=-\infty}^{\infty} K_n(\mathbf{x}, \mathbf{x}') e^{in(\pi/T)\tau}. \quad (3.3.2)$$

Here $K_n(\mathbf{x}, \mathbf{x}')$ is the covariance at the frequency n . Sometimes, one wants to look at the covariance of the field in a frequency window $N_1 \leq n \leq N_2$. This band covariance function is defined as

$$K_w(\mathbf{x}, \mathbf{x}') = \frac{1}{N_2 - N_1} \sum_{n=N_1}^{N_2} K_n(\mathbf{x}, \mathbf{x}'). \quad (3.3.3)$$

The EOFs of the a stochastic field are defined according to the band covariance function:

$$\int_{\Omega} K(\mathbf{x}, \mathbf{x}') \psi_n(\mathbf{x}') d\mathbf{x}' = \lambda_n \psi_n(\mathbf{x}), \quad n = 1, 2, 3, \dots \quad (3.3.4)$$

Here we omitted the subscript w of K for simplicity. Of course, $\psi_n(\mathbf{x})$ are the EOFs (eigenfunctions) and the λ_n are the variances (eigenvalues), and $K(\mathbf{x}, \mathbf{x}')$ is also called the kernel of the above integral equation. Similar to the matrix case, the eigenvalues are also ordered: $\lambda_1 \geq \lambda_2 \geq \dots$.

In statistics, EOFs are also called Karhunen-Loeve basis functions. They are apparently simple mathematics. But, it is a kind of surprise that the EOFs were not used for engineering analysis and natural science until 1940s.

EOFs have the following properties

$$\int_{\Omega} \psi_m(\mathbf{x}) \psi_n(\mathbf{x}) d\mathbf{x} = \delta_{mn}, \quad (\text{orthonormal property}), \quad (3.3.5)$$

$$\sum_{n=1}^{\infty} \psi_n(\mathbf{x}) \psi_n(\mathbf{x}') = \delta(\mathbf{x} - \mathbf{x}') \quad (\text{completeness property}). \quad (3.3.6)$$

These two properties imply

$$K(\mathbf{x}, \mathbf{x}') = \sum_{n=1}^{\infty} \lambda_n \psi_n(\mathbf{x}) \psi_n(\mathbf{x}'). \quad (3.3.7)$$

Mathematicians call this $K(\mathbf{x}, \mathbf{x}')$ a Hilbert-Schmidt kernel.

3.4 EOFs on a unit circle

3.4.1 Homogeneous real EOFs

Consider a homogeneous stochastic field on a unit circle (Fig. 3.2). The covariance function depends only on the open angle between the two points in question:

$$\langle \Theta(\mathbf{x}) \Theta(\mathbf{x}') \rangle = K(\theta) \quad (3.4.1)$$

where θ is the angle between \mathbf{x} and \mathbf{x}' (Fig. 3.2). Let $\mathbf{x} = \exp[i\phi]$ and $\mathbf{x}' = \exp[i\phi']$. Then $\theta = \phi - \phi'$. The EOFs are defined by

$$\int_0^{2\pi} K(\phi - \phi') \psi_n(\phi') d\phi' = \lambda_n \psi_n(\phi). \quad (3.4.2)$$

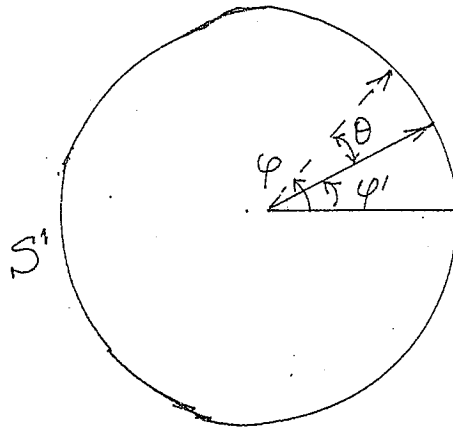


Figure 3.2: Open angle and homogeneous field on the unit circle.

The covariance function in this case is certainly a periodic one with its period at most equal to 2π . Hence it can be expressed in terms of Fourier cosine series (since K is an even function)

$$K(\phi - \phi') = \sum_{j=0}^{\infty} K_j \cos[j(\phi - \phi')]. \quad (3.4.3)$$

The Fourier coefficients are

$$K_j = \frac{1}{\pi} \int_0^{2\pi} K(\theta) \cos(j\theta) d\theta \quad \text{for } j = 1, 2, 3, \dots \quad (3.4.4)$$

and

$$K_0 = \frac{1}{2\pi} \int_0^{2\pi} K(\theta) d\theta. \quad (3.4.5)$$

We substitute this into (3.4.2) and find that the solution of (3.4.2) (i.e., the eigenvalues and EOFs) as follows. When $n = 0$,

$$\lambda_0 = \sqrt{2\pi} K_0, \quad \psi_0(\phi) = \frac{1}{\sqrt{2\pi}}. \quad (3.4.6)$$

When $n \geq 1$, each eigenvalue corresponds to two eigenfunctions:

$$\lambda_n = \pi K_n \quad (\text{eigenvalue}) \quad (3.4.7)$$

$$\psi_n^+(\phi) = \frac{1}{\sqrt{\pi}} \cos[n\phi], \quad \psi_n^-(\phi) = \frac{1}{\sqrt{\pi}} \sin[n\phi], \quad n = 1, 2, \dots \quad (3.4.8)$$

The above EOF result can be derived by a simpler way. The Fourier expansion (3.4.3) of K can be written in terms of the form of a Hilbert-Schmidt kernel:

$$K(\phi - \phi') = (\sqrt{2\pi} K_0) \frac{1}{\sqrt{2\pi}} + \sum_{n=1}^{\infty} \left[(\pi K_n) \frac{\sin(n\phi)}{\sqrt{\pi}} \frac{\sin(n\phi')}{\sqrt{\pi}} + (\pi K_n) \frac{\cos(n\phi)}{\sqrt{\pi}} \frac{\cos(n\phi')}{\sqrt{\pi}} \right]. \quad (3.4.9)$$

Since

$$\frac{1}{\sqrt{2\pi}} \quad \text{and} \quad \left\{ \frac{\cos(n\phi)}{\sqrt{\pi}}, \frac{\sin(n\phi)}{\sqrt{\pi}} \right\}_{n=1}^{\infty}$$

form a complete orthonormal set for functions of period 2π , they must be the EOFs for our covariant function $K(\phi - \phi')$.

With EOFs, any function on the unit circle can be expressed in terms of the EOFs expansions, which are exactly the Fourier series expansions.

3.4.2 Homogeneous complex EOFs

The symmetric kernel K can be complex (called Hermitian kernel) and the eigenfunctions can also be complex valued, although the eigenvalues must be real and positive. The complex EOFs can be defined in the following way:

$$\int_{\Omega} K(\mathbf{x}, \mathbf{x}') \psi_n^*(\mathbf{x}') d\mathbf{x}' = \lambda_n \psi_n(\mathbf{x}), \quad (3.4.10)$$

$$\int_{\Omega} \psi_m(\mathbf{x}) \psi_n^*(\mathbf{x}) d\mathbf{x} = \delta_{mn}, \quad (3.4.11)$$

$$\sum_{n=1}^{\infty} \psi_n(\mathbf{x}) \psi_n^*(\mathbf{x}') = \delta(\mathbf{x} - \mathbf{x}'), \quad (3.4.12)$$

where * still signifies the complex conjugate.

For the case of a homogeneous field on a unit circle, we have the complex Fourier expansion:

$$K(\phi - \phi') = \sum_{n=-\infty}^{\infty} K_n e^{in(\phi - \phi')}, \quad (3.4.13)$$

where

$$K_n = \frac{1}{2\pi} \int_0^{2\pi} K(s) e^{-in\phi'} ds. \quad (3.4.14)$$

Hence,

$$K(\phi - \phi') = \sum_{n=-\infty}^{\infty} K_n e^{in\phi} (e^{in\phi'})^*. \quad (3.4.15)$$

This is a Hilbert-Schmidt kernel whose eigenpairs are

$$\{K_n, e^{in\phi}\}, \quad n = 0, \pm 1, \pm 2, \dots \quad (3.4.16)$$

The eigenspaces are two-dimensional for $n \neq 0$ since $K_{-n} = K_n$.

3.4.3 Inhomogeneous EOFs on a unit circle

If the field is not homogeneous, the trigonometric functions are no longer EOFs, rather each EOF is a linear combination of the trigonometric functions (i.e., Fourier series):

$$\psi_n(\phi) = \sum_{m=-\infty}^{\infty} \psi_{nm} e^{im\phi}. \quad (3.4.17)$$

One can solve linear algebraic equations for ψ_{nm} to find the EOF $\psi_n(\phi)$. For our climatology applications, there is no need to get into too much detail of the EOFs on the unit circle, instead later in this chapter, we will talk about computing EOFs for more complicated case: spherical harmonics on a unit sphere.

3.5 EOFs on a unit sphere

3.5.1 Simple orthogonal bases

The simplest orthonormal basis for functions over a domain is the one generated by the eigenvalue problem of the the simplest second order differential operator, which, mathematically speaking, is the simplest possible non-trivial self-adjoint operator.

Example 1. Orthonormal basis over $[-\pi, \pi]$.

The solutions of the eigenvalue problem

$$\frac{d^2}{dx^2} u = \lambda u, \quad (3.5.1)$$

$$u(x = -\pi) = u(x = \pi) = 0 \quad (3.5.2)$$

are:

$$\left\{ -n^2, \exp[inx]/\sqrt{2\pi} \right\}_{n=-\infty}^{\infty}. \quad (3.5.3)$$

Notice that $\left\{ \exp[inx]/\sqrt{2\pi} \right\}_{n=-\infty}^{\infty}$ is the regular Fourier orthonormal basis over $[-\pi, \pi]$.

Example 2. Orthonormal basis over $[-\pi, \pi] \times [-\pi, \pi]$.

The solutions of the eigenvalue problem

$$\left(\frac{\partial^2}{\partial^2 x} + \frac{\partial^2}{\partial^2 y} \right) u = \lambda u, \quad (3.5.4)$$

$$u(x = -\pi, y) = u(x = \pi, y) = 0, \quad (3.5.5)$$

$$u(x, y = -\pi) = u(x, y = \pi) = 0 \quad (3.5.6)$$

are:

$$\left\{ -(m^2 + n^2), \exp[i(mx + ny)]/(2\pi) \right\}_{m,n=-\infty}^{\infty}. \quad (3.5.7)$$

Notice that

$$\left\{ \exp[i(mx + ny)]/2\pi \right\}_{m,n=-\infty}^{\infty}$$

is the regular two-dimensional Fourier orthonormal basis over $[-\pi, \pi] \times [-\pi, \pi]$.

Example 3. Orthonormal basis over a unit circle. (Mathematically the unit circle is denoted by S^1 .)

Because the circle has a 2π - period (mathematically denoted by $\text{mod}(2\pi)$) the solutions of

$$\frac{d^2}{d\theta^2} u = \lambda u, \quad (3.5.8)$$

must be

$$\left\{ -n^2, \exp[in\theta]/\sqrt{2\pi} \right\}_{n=-\infty}^{\infty}. \quad (3.5.9)$$

Notice that $\left\{ \exp[in\theta]/\sqrt{2\pi} \right\}_{n=-\infty}^{\infty}$ is the regular Fourier basis over $[-\pi, \pi]$.

The functions $\exp[in\theta]$ are called harmonics.

3.5.2 Spherical harmonics

Following the same philosophy, we try to find the simplest possible basis for functions on a unit sphere. (Mathemicians like to denote the unit sphere in 3-dimensional space by S^2 .) This basis consists of spherical harmonics (or normalized spherical harmonic functions).

The spherical coordinates are defined by (θ, ϕ) : ϕ being the latitude and θ being the longitude (Fig.3.3). Then the unit vectors are

$$\begin{aligned} \hat{n} &= (\cos \phi \cos \theta, \cos \phi \sin \theta, \sin \phi), \\ \hat{n}' &= (\cos \phi' \cos \theta', \cos \phi' \sin \theta', \sin \phi'). \end{aligned}$$

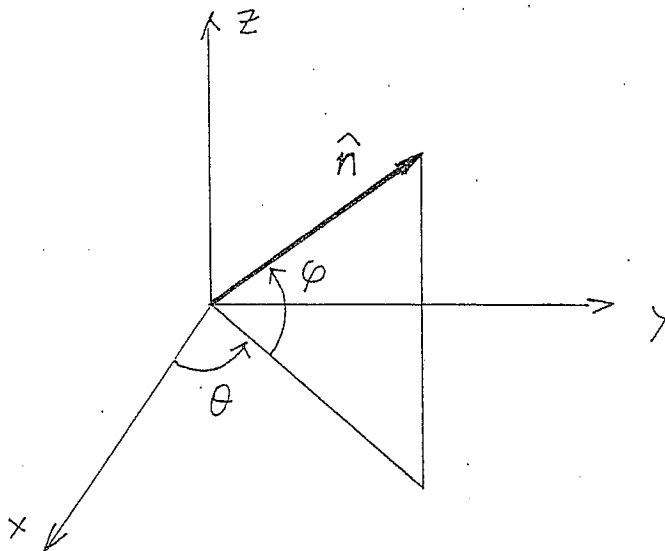


Figure 3.3: A unit vector on the unit sphere.

On the sphere, the simplest self-adjoint operator is the Laplace operator (in spherical coordinates). In terms of the spherical coordinates (θ, ϕ) , the eigenvalue problem for the Laplace operator on the sphere is

$$\left[\frac{1}{\cos \phi} \frac{\partial}{\partial \phi} \left(\cos \phi \frac{\partial}{\partial \phi} \right) + \frac{1}{\cos^2 \phi} \frac{\partial^2}{\partial \theta^2} \right] u = \lambda u, \quad \text{mod}(2\pi \times \pi). \quad (3.5.10)$$

The method of the separation of variables can be used:

$$u = \Theta(\theta)\Phi(\phi). \quad (3.5.11)$$

Then the above eigenvalue problem can be decomposed into two ordinary differential equations

$$\frac{d^2 \Theta}{d\theta^2} = -m^2 \Theta, \quad (3.5.12)$$

$$\frac{1}{\cos \phi} \frac{d}{d\phi} \left(\cos \phi \frac{d\Phi}{d\phi} \right) + \frac{-m^2}{\cos^2 \phi} \Phi = \lambda \Phi. \quad (3.5.13)$$

To satisfy the 2π -period condition for θ , we must have m being an integer according to (3.5.12). The general solution of (3.5.12) is:

$$\Theta = \exp[im\theta]. \quad (3.5.14)$$

As for (3.5.13), we let $x = \sin \theta$ be the latitude height from the the equator plane. Then the ODE becomes

$$\frac{d}{dx} \left[(1-x^2) \frac{d\Phi}{dx} \right] + \left[-\lambda - \frac{m^2}{1-x^2} \right] \Phi = 0. \quad (3.5.15)$$

This form is the standard ODE: associated Legendre equation as long as

$$-\lambda = l(l+1), \quad (3.5.16)$$

The solutions of the Legendre equation are the associated Legendre functions $P_l^m(x)$ (when m is an integer). If we require that l be a positive integer and $|m| \leq l$, then $P_l^m(\sin \phi)$ is a periodic function of period equal to 2π . By the uniqueness of the eigenspace of equation (3.5.13), $P_l^m(\sin \phi)$ are the solutions we desired to find. Therefore we adopt

$$\frac{d}{dx} \left[(1-x^2) \frac{d\Phi}{dx} \right] + \left[l(l+1) - \frac{m^2}{1-x^2} \right] \Phi = 0, \quad l = 0, 1, 2, \dots, \quad |m| \leq l. \quad (3.5.17)$$

When $m = 0$, the associated Legendre functions become Legendre polynomials $P_l(x)$ which are given by the Rodrigues' formula:

$$P_l(x) = \frac{1}{2^l l!} \frac{d^l}{dx^l} (x^2 - 1)^l, \quad l = 0, 1, 2, \dots, \quad |x| \leq 1. \quad (3.5.18)$$

(Of course, the derivatives of polynomials are polynomials.) The first nine Legendre polynomials are

$$\begin{aligned} P_0(x) &= 1, \\ P_1(x) &= x, \\ P_2(x) &= \frac{1}{2}(3x^2 - 1), \\ P_3(x) &= \frac{1}{2}(5x^3 - 3x), \\ P_4(x) &= \frac{1}{8}(3 - 30x^2 + 35x^4), \\ P_5(x) &= \frac{1}{8}(15x - 70x^3 + 63x^5), \\ P_6(x) &= \frac{1}{16}(-5 + 105x^2 - 315x^4 + 231x^6), \\ P_7(x) &= \frac{1}{16}(-35x + 315x^3 - 693x^5 + 429x^7), \\ P_8(x) &= \frac{1}{128}(35 - 1260x^2 + 6930x^4 - 12012x^6 + 6435x^8), \\ P_9(x) &= \frac{1}{128}(315x - 4620x^3 + 18018x^5 - 25740x^7 + 12155x^9). \end{aligned}$$

Legendre polynomials are a set of orthogonal polynomials

$$\int_{-1}^1 P_l(x) P_{l'}(x) dx = \frac{2}{2l+1} \delta_{ll'}, \quad (3.5.19)$$

and

$$P_l(1) = 1. \quad (3.5.20)$$

$P_l(x)$ has l zeros in $(-1, 1)$.

The associated Legendre functions are defined by

$$P_l^m(x) = (-1)^m (1-x^2)^{m/2} \frac{d^m}{dx^m} P_l(x). \quad (3.5.21)$$

(Again, this derivative gives only polynomials.) This is also a set of orthogonal functions with respect to the index l :

$$\int_{-1}^1 P_l^m(x) P_{l'}^m(x) dx = \frac{2}{2l+1} \frac{(l+m)!}{(l-m)!} \delta_{ll'}. \quad (3.5.22)$$

$P_l^m(x)$ also has $l - |m|$ zeros in $(-1, 1)$ and $P_l^m(\pm 1) = 0$.

In Mathematica, you can type `LegendreP[n,x]`. The machine will give you the expression of $P_n(x)$. To get the associated Legendre function, you type `LegendreP[l,m,x]`. For plotting, you can type `Plot[LegendreP[6,x], x,-1,1]` for the graph of $P_6(x)$.

The spherical harmonics $Y_{lm}(\theta, \phi)$ must be proportional to

$$P_l^m(\sin \phi) \exp[im\theta]. \quad (3.5.23)$$

The normalization condition

$$\int_{4\pi} Y_{lm}(\theta, \phi) Y_{l'm'}^*(\theta, \phi) d\Omega = \delta_{ll'} \delta_{mm'} \quad (3.5.24)$$

determines that

$$Y_{lm}(\theta, \phi) = \sqrt{\frac{2l+1}{4\pi} \frac{(l+m)!}{(l-m)!}} P_l^m(\sin \phi) \exp[im\theta]. \quad (3.5.25)$$

In the above, * signifies the complex conjugate, i.e. $(a + ib)^* = a - ib$ for real number a and b . And

$$Y_{lm}^*(\theta, \phi) = (-1)^m Y_{l,-m}(\theta, \phi). \quad (3.5.26)$$

$Y_{l0}(\theta, \phi)$ has l zeros along a longitude line (not including the two pole points) and $Y_{lm}(\theta, \phi)$ $2m$ zeros on a latitude circle. We say that it has l north-south waves and $2m$ east-west waves. So the wave patterns become more complicated when l and m are large. For a fixed l , when $m = 0$, there is no wave in east-west direction. So, Y_{l0} signifies a longitude average (also called zonal average in meteorology) of a physical quantity.

Using Mathematica, you can type `SphericalHarmonicY[2,1,phi,theta]` to get the expression of $Y_{21}(\theta, \phi)$. But be careful with the coordinates Mathematica uses. Mathematica uses θ to denote latitude and ϕ for longitude. That is why I phi, theta order in my Mathematica command. In most of mathematics books, authors use ϕ to denote the longitude and their θ is counted

from the north pole (i.e. $\theta = 0$ at the north pole). According to our coordinates, the first several spherical harmonics are:

$$\begin{aligned}
 Y_{00} &= \frac{1}{\sqrt{4\pi}}, \quad (\text{no wave}), \\
 Y_{10} &= \sqrt{\frac{3}{4\pi}} \sin \theta, \\
 Y_{11} &= -\sqrt{\frac{3}{8\pi}} \cos \phi e^{i\theta}, \\
 Y_{20} &= \frac{1}{4} \sqrt{\frac{5}{\pi}} (-1 + 3 \cos^2 \phi), \\
 Y_{21} &= -3 \sqrt{\frac{5}{24\pi}} \cos \phi e^{i\theta}, \\
 Y_{22} &= 3 \sqrt{\frac{5}{96\pi}} \sin^2 \phi e^{2i\theta}, \\
 Y_{30} &= \frac{1}{4} \sqrt{\frac{7}{\pi}} (-3 \cos \phi + 5 \cos^3 \phi), \\
 Y_{31} &= \frac{1}{8} \sqrt{\frac{21}{\pi}} (1 - 5 \cos^2 \phi) \sin \phi e^{i\theta}, \\
 Y_{32} &= 15 \sqrt{\frac{7}{480\pi}} \cos \phi \sin^2 \phi e^{2i\theta}, \\
 Y_{33} &= -\frac{5}{8} \sqrt{\frac{7}{5\pi}} \sin^3 \phi e^{3i\theta}.
 \end{aligned}$$

The spherical harmonics form a complete basis. The completeness means that

$$\sum_{l=0}^{\infty} \sum_{m=-l}^l Y_{lm}(\theta, \phi) Y_{lm}^*(\theta', \phi') = \delta(\theta - \theta') \delta(\sin \phi - \sin \phi'). \quad (3.5.27)$$

With this completeness property, you can do series expansion

$$v(\hat{\mathbf{n}}) = \sum_{l=0}^{\infty} \sum_{m=-l}^l v_{lm} Y_{lm}(\hat{\mathbf{n}}) \quad (3.5.28)$$

where the expansion coefficients are

$$v_{lm} = \int_{4\pi} \Omega v(\hat{\mathbf{n}}) Y_{lm}^*(\hat{\mathbf{n}}). \quad (3.5.29)$$

These coefficients v_{lm} are also called the spectra and $|v_{lm}|^2$ are called the power spectra.

The spherical harmonic series expansion of $P_l(\hat{\mathbf{n}} \cdot \hat{\mathbf{n}}')$ with respect to $\hat{\mathbf{n}}$ results in only finitely many terms. This is the Addition Theorem:

$$P_l(\hat{\mathbf{n}} \cdot \hat{\mathbf{n}}') = \frac{4\pi}{2l+1} \sum_{m=-l}^l Y_{lm}(\hat{\mathbf{n}}) Y_{lm}^*(\hat{\mathbf{n}}'). \quad (3.5.30)$$

This theorem can be regarded as the 3-D sphere extension of the cosine law on a unit circle in 2-D space:

$$\cos(a - b) = \cos a \cos b + \sin a \sin b. \quad (3.5.31)$$

The Addition Theorem will be used in the next subsection.

We may summarize the most useful properties of the spherical harmonics:

$$\nabla^2 Y_{lm}(\hat{n}) = -l(l+1)Y_{lm}(\hat{n}), \quad (3.5.32)$$

$$\int_{4\pi} d\Omega Y_{lm}(\theta, \phi) Y_{l'm'}^*(\theta, \phi) = \delta_{ll'} \delta_{mm'}. \quad (3.5.33)$$

3.5.3 EOFs for a homogeneous field on a sphere

On the unit sphere in 3-D, the EOFs are defined by

$$\int_{4\pi} K(\hat{n}, \hat{n}') \psi_n(\hat{n}) d\Omega = \lambda_n \psi_n(\hat{n}) \quad (3.5.34)$$

where 4π denotes the integration domain and also signifies that the total solid angle of the sphere is 4π .

Let us first look at the homogeneous case:

$$K(\hat{n}, \hat{n}') = K(\hat{n} \cdot \hat{n}'). \quad (3.5.35)$$

Let $x = \hat{n} \cdot \hat{n}'$. Then $K(x)$ is a function defined on the interval $[-1, 1]$ and can be expressed in terms of Legendre polynomials:

$$K(x) = \sum_{n=0}^{\infty} K_n P_n(x). \quad (3.5.36)$$

Now we can use the Addition Theorem for spherical harmonics:

$$P_l(\hat{n} \cdot \hat{n}') = \frac{4\pi}{2l+1} \sum_{m=-l}^l Y_{lm}(\phi, \theta) Y_{lm}^*(\phi', \theta'). \quad (3.5.37)$$

Hence,

$$K(x) = \sum_{l=0}^{\infty} K_l \frac{4\pi}{2l+1} \sum_{m=-l}^l Y_{lm}(\phi, \theta) Y_{lm}^*(\phi', \theta'). \quad (3.5.38)$$

This expression is already in the Hilbert-Schmidt form. As we know that the spherical harmonics are orthonormal and complete, they must be the EOFs. For each eigenvalue

$$\frac{4\pi}{2l+1} K_l, \quad (3.5.39)$$

there are $2l+1$ eigenfunctions

$$Y_{lm}(\phi, \theta) \quad m = -l, -l+1, \dots, l-1, l. \quad (3.5.40)$$

3.6 T-truncation and R-truncation

3.6.1 Spectral truncation

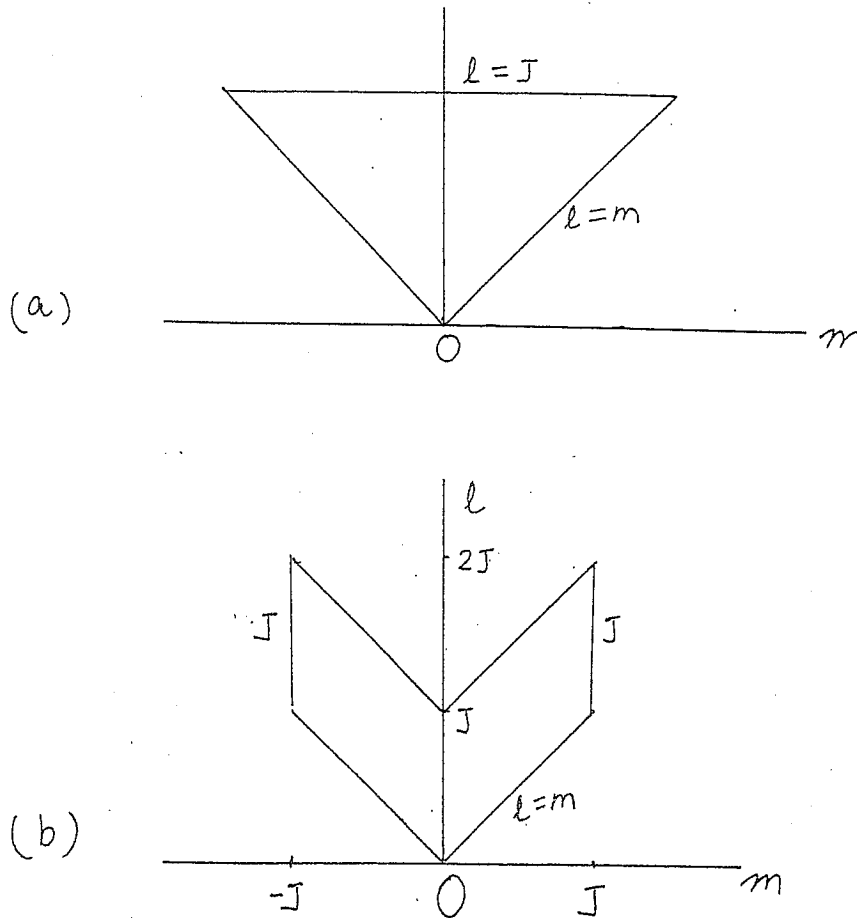


Figure 3.4: Truncation method of spherical harmonic functions: (a) T-Truncation, and (b) R-truncation.

For practical machine computation, the spherical harmonic expansion

$$v(\hat{n}) = \sum_{l=0}^{\infty} \sum_{m=-l}^l v_{lm} Y_{lm}(\hat{n}) \quad (3.6.1)$$

must be truncated at a certain level of m and l . Usually people use two types of truncations: T-truncation (triangular truncation) and R-truncation (rhomboidal truncation) (Fig. 3.4). Both of these truncations are widely used in the spectral GCM models.

T-truncation expansion is

$$v(\hat{\mathbf{n}}) = \sum_{l=0}^N \sum_{m=-l}^l v_{lm} Y_{lm}(\hat{\mathbf{n}}) \quad (3.6.2)$$

This is called the TN truncation. The indices are hence in a triangular region (Fig. 3.4a) T42 truncation means that $N=42$.

Parallelogram expansion is

$$v(\hat{\mathbf{n}}) = \sum_{m=-M}^M \sum_{l=|m|}^{|m|+J} v_{lm} Y_{lm}(u, v, n) \quad (3.6.3)$$

The indices in the above summations are in two parallelograms. Usually people take $M = J$. The parallelograms become rhomboidals. The truncation is now called the rhomboidal truncation RJ. The R21 truncation means that

$$v(\hat{\mathbf{n}}) = \sum_{m=-21}^{21} \sum_{l=|m|}^{|m|+21} v_{lm} Y_{lm}(\hat{\mathbf{n}}) \quad (3.6.4)$$

The indices are in two rhomboidal regions (Fig. 3.4b).

Here I would like to recommend every graduate student interested in climate modeling to take a look at the book by Washington and Parkinson (1986) (in particular, pp. 200-204).

3.6.2 Spatial resolution

In this subsection we discuss the basics of the transforms between the grid point data and the spectral data. We will examine the criteria by which the spatial resolutions are determined for various spectral truncations.

Let us illustrate our analysis by using the TJ truncation and $M \times N$ grid points. Usually both M and N are even integers. The anomaly field is denoted by $Z(\hat{\mathbf{n}})$. The transforms between the spectral space and the physical space (i.e., the grid points) are:

$$\hat{Z}_{lm} = \frac{2\pi}{N} \sum_{p=1}^N \sum_{q=1}^M w_q Z(\hat{\mathbf{n}}_{pq}) Y_{lm}^*(\hat{\mathbf{n}}_{pq}) \quad (\text{from grids to spectra}) \quad (3.6.5)$$

$$\hat{Z}_{pq} = \sum_{l=0}^J \sum_{m=-l}^l Z_{lm} Y_{lm}(\hat{\mathbf{n}}_{pq}) \quad (\text{from spectra to grids}), \quad (3.6.6)$$

where w_q are the Gaussian weights. As usual, we take the Gaussian grids: The latitude circles are determined by the zeros of the Legendre polynomial of order $M/2$ and the longitude circles are uniformly distributed with separation angle $\Delta\theta = 2\pi/N$. When M is large enough (say, greater than 14), the distribution of the zeros of the Legendre polynomial $P_{M/2}$ is approximately uniform. For

the GCMs used in most of research centers in the world, it is required that the integration of the vorticity equation be exact with all the modes in the TJ truncation. This condition requires the number of the grid points be sufficiently large:

$$N \geq 3J + 1, \quad M \geq (3J + 1)/2. \quad (3.6.7)$$

For example, If $J = 9$, then the grid points are: 28×14 ; and If $J = 21$, the grid points are: 64×32 . The above condition can be easily derived. The vorticity equation is in the form:

$$\frac{\partial \psi}{\partial t} = A(\partial \psi). \quad (3.6.8)$$

The expansion of ψ is:

$$\psi(\hat{\mathbf{r}}, t) = \sum \psi_{lm}(t) Y_{lm}(\hat{\mathbf{r}}). \quad (3.6.9)$$

Then the vorticity equation becomes

$$\frac{d\psi_{lm}}{dt} = \int_{4\pi} d\Omega A(\psi) Y_{lm}^*(\hat{\mathbf{r}}). \quad (3.6.10)$$

The nonlinearity of the operator A is of second order. The highest order of the trigonometric polynomials in TJ truncation is J . Hence the highest order of the trigonometric polynomials of $A(\psi) Y_{lm}^*(\hat{\mathbf{r}})$ is $3J$. Along a longitude circle, since the integration method is Gaussian, so N points can render exact integration for $2N - 1$ order polynomial. Here we have $3J$ order polynomial, so we $(3J + 1)/2$ points can exactly integrate the waves on a longitude circle. This is M value. The weights on a latitude circle are uniform and the points are also uniformly distributed. Thus we need $3J + 1$ points to numerically integrate a $3J$ order polynomial. This is N value.

We depict the correspondence between the spectral space and the grids (also called the physical space) as follows:

$$TJ \text{ (spectra)} \leftrightarrow (3J + 1) \times (3J + 1)/2 \text{ (grids)}. \quad (3.6.11)$$

But one does not have to satisfy this no-aliasing condition. As a matter of fact, it has been pointed out that this condition of no-aliasing is not optimal (Chen, 1993). The spectra of the TJ truncation can be described by $J(J + 1)$ real numbers. The corresponding number of the grid points is $(3J + 1)^2/2$. When J is large, the ratio of $(3J + 1)^2/2$ and $J(J + 1)$ is 4.5. Namely to catch the same amount of information, in the grid point space one needs to use 4.5 times data compared with the spectral space if one chooses to use the transforms given by (3.6.5) and (3.6.6) if one requires (3.6.7) to hold.

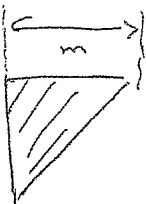
The total number of grid points is: $N_{net} = M \times N$. The covariance matrix on the grid points and that in the spectral space are respectively

$$[C_{ij}] = \frac{1}{M_{yr}} \sum_{\alpha=1}^{M_{yr}} Z_i(\alpha) Z_j(\alpha), \quad (3.6.12)$$

Handwritten notes and diagram:

$$\psi_{lm}^* = \int \psi_{lm} Y_{lm}^*$$

$$= (-1)^m \psi_{lm}^*$$

$$\psi_{lm} = (-1)^m \psi_{lm}^*$$


and

$$[\tilde{C}_{lm'l'm'}] = \frac{1}{M_{yr}} \sum_{\alpha=1}^{M_{yr}} Z_{lm}(\alpha) Z_{l'm'}^*(\alpha). \quad (3.6.13)$$

The covariance matrix $[C_{ij}]$ is a real symmetric $N_{net} \times N_{net}$ matrix. The covariance matrix in the spectral space $[\tilde{C}_{lm'l'm'}]$ is a Hermitian matrix of order $N_{spt} = (J+1)(J+2)/2$, in which there are $J(J+1)/2$ complex entries and $J+1$ real entries. As mentioned before, N_{net} is 4.5 times N_{spt} .

The eigenvalue problems in the physical space and the spectral space are defined respectively by

$$\sum_{j=1}^{N_{net}} C_{ij}(E_j)_k = \lambda_k (E_i)_k, \quad (3.6.14)$$

and

$$\sum_{l'=0}^J \sum_{m'=-l'}^{l'} \tilde{C}_{lm'l'm'} (\tilde{E}_{l'm'})_k = \bar{\lambda}_k (\tilde{E}_{lm})_k. \quad (3.6.15)$$

The transforms between the eigenvectors in the physical space and the spectral space are:

$$(\hat{E}_i)_k = \sum_{l,m} (\tilde{E}_{lm})_k Y_{lm}(\hat{n}_i), \quad (3.6.16)$$

and

$$(\hat{\tilde{E}}_{l'm'})_k = \sum_{j=1}^{N_{net}} w_j (E_j)_k Y_{l'm'}^*(\hat{n}_j). \quad (3.6.17)$$

The differences between the quantities computed in the spectral space and the physical space are denoted by

$$Z_i(\alpha) = \hat{Z}_i(\alpha) + z_i(\alpha), \quad (3.6.18)$$

$$(\tilde{E}_{lm})_k = (\hat{\tilde{E}}_{lm})_k + \tilde{e}_{lm}, \quad (3.6.19)$$

$$(E_i)_k = (\hat{E}_i)_k + (e_i)_k. \quad (3.6.20)$$

Then one can derive the difference between the eigenvalues computed from the spectral space and those from the physical space:

$$\begin{aligned} \bar{\lambda}_k - \lambda_k = & \bar{\lambda}_k \sum_{i=1}^{N_{net}} (e_i)_k (E_i)_k \\ & + \sum_{i,j=1}^{N_{net}} \frac{1}{M_{yr}} \sum_{\alpha=1}^{M_{yr}} z_i(\alpha) Z_j(\alpha) (E_i)_k (E_j)_k \\ & + \sum_{l,m} \frac{1}{M_{yr}} \sum_{\alpha=1}^{M_{yr}} \sum_{i=1}^{N_{net}} z_i(\alpha) (E_i)_k T_{lm}^*(\alpha) \tilde{e}_{lm}. \end{aligned} \quad (3.6.21)$$

If z_i , e_i and \bar{e}_{lm} are of order ϵ (a small positive number). Then the third part of the above formula is of order ϵ^2 and can be ignored compared with ϵ . Hence $\bar{\lambda}_k - \lambda_k = O(\epsilon)$. This is a theorem in computational mathematics. It says that when using Rayleigh-Ritz method to estimate the eigenvalues in an eigen-system, the resulted error is of the same order as that of the guessed pattern.

3.7 EOFs for nonhomogeneous field

The key step that leads to the results in the subsection 3.4.2 is the addition theorem (which is equivalent to $\cos(a-b) = \cos a \cos b + \sin a \sin b$ used in the unit circle case). For a nonhomogeneous field, this addition theorem cannot be applied.

Although for a nonhomogeneous field, the EOFs are not the spherical harmonics, they can still be expanded in terms of spherical harmonics.

$$\psi_n(\hat{\mathbf{n}}) = \sum_{l=0}^{\infty} \sum_{m=-l}^l \psi_{n,lm} Y_{lm}(\hat{\mathbf{n}}). \quad (3.7.1)$$

The question of computing the EOFs becomes the problem of computing the coefficients $\psi_{n,lm}$.

Recall that the covariance function is

$$K(\hat{\mathbf{n}}, \hat{\mathbf{n}}', |t - t'|) = \langle \Theta(\hat{\mathbf{n}}, t) \Theta(\hat{\mathbf{n}}', t') \rangle. \quad (3.7.2)$$

The expansion of the Θ field is

$$\Theta(\hat{\mathbf{n}}, t) = \sum_{l=0}^{\infty} \sum_{m=-l}^l \Theta_{lm}(t) Y_{lm}(\hat{\mathbf{n}}). \quad (3.7.3)$$

Then

$$K(\hat{\mathbf{n}}, \hat{\mathbf{n}}', |t - t'|) = \sum_{l,m} \sum_{l',m'} \langle \Theta_{lm}(t) \Theta_{l'm'}(t') \rangle Y_{lm}(\hat{\mathbf{n}}) Y_{l'm'}^*(\hat{\mathbf{n}}'). \quad (3.7.4)$$

The Fourier transform of $\Theta_{lm}(t)$ is denoted by $\tilde{\Theta}_{lm}(f)$. Since we assume that the time series is stationary, we have

$$\langle \tilde{\Theta}_{lm}(f) \tilde{\Theta}_{l'm'}(f') \rangle = \bar{K}_{lm'l'm'}(f) \delta(f - f'). \quad (3.7.5)$$

Consider the integral of the frequency dependent covariance function over the frequency window $[f_1, f_2]$:

$$\bar{K}(\hat{\mathbf{n}}, \hat{\mathbf{n}}') = \sum_{l,m} \sum_{l',m'} \frac{1}{f_2 - f_1} \int_{f_1}^{f_2} df \langle \tilde{\Theta}_{lm}(f) \tilde{\Theta}_{l'm'}(f') \rangle Y_{lm}(\hat{\mathbf{n}}) Y_{l'm'}^*(\hat{\mathbf{n}}'). \quad (3.7.6)$$

Then the eigenvalue problem for the covariance function

$$\int_{4\pi} d\Omega' \bar{K}(\hat{\mathbf{n}}, \hat{\mathbf{n}}') \psi_n(\hat{\mathbf{n}}') = \lambda_n \psi_n(\hat{\mathbf{n}}), \quad (n = 1, 2, 3, \dots) \quad (3.7.7)$$

can be projected to the basis functions of spherical harmonics $Y_{lm}(\hat{\mathbf{n}})$. Then we have a matrix eigenvalue problem

$$\sum_{l'=1}^{\infty} \sum_{m'=-l'}^{l'} \langle \bar{\Theta}_{lm} \bar{\Theta}_{l'm'} \rangle \psi_{n,l'm'} = \lambda_n \psi_{n,lm}, \quad (n = 1, 2, 3, \dots). \quad (3.7.8)$$

In this case, the eigenvalues are the variances of the modes over a frequency range $[f_1, f_2]$. If one considers only one-year cycle, then he needs to set f (with a proper unit) so that it corresponds to one-year cycle. No integrations are needed in this case.

3.8 References

- Chen, X.-S., 1993: The aliased and the dealiased spectral models of the shallow-water equations. *Mon. Wea. Rev.*, **121**, 834-852.
- North, G.R., T.L. Bell, R.F. Cahalan, and F.J. Moeng, 1982: Sampling errors in the estimation of empirical orthogonal functions. *Mon. Wea. Rev.*, **100**, 699-706.
- Penland, C., and P.D. Sardeshmukh, 1995: Error and sensitivity analysis of geophysical eigensystems. *J. Clim.*, **8**, 1988-1998.
- Washington, W., and C. L. Parkinson, 1986: An introduction of three dimensional climate modeling, Univ Sci Books, California.

Chapter 4

Minimal MSE Optimization

In the last three chapters we learned that due to the spatial and/or temporal gaps in a sampling process there exists an error when one derives an average of a climate quantity from observational data. For example, there are errors in deriving the averaged hourly rainfall from 830 rain gauges deployed by Tokyo Metropolitan government, the monthly rainfall from the TRMM satellite, and global average annual mean surface air temperature from historical stations, etc. These errors can be assessed by the mean square errors (MSE).

The ideal situation is that the MSE is zero. Of course, in practice no sampling process can yield zero error. Hence our goal is to minimize the MSE. Thus we have two missions here. One is to do data analysis for already collected data. For this data analysis problem, the best we can do is to weight each data entry differently so that the MSE become minimal. The other mission is to design a future sampling process and perform analysis for simulations. For this sampling design problem, we have more freedoms. We can not only choose different positions for surface stations and flight orbits for air-borne (or satellite-borne) instruments, but also assign a different weight for each data entry to minimize the MSE. In this chapter we will use a few concrete examples to show how to obtain the aforementioned minimization.

4.1 Numerical integration

As we understand that the average value of a function over an interval or a region is the definite integral of the function divided by the length of the interval or the size of the region, and the numerical integration is to use the data at discrete points to approximate the definite integral. Thus an average value of a climate quantity derived from discrete samplings is similar to the definite integral derived from numerical integration. The difference is mainly in that a climate quantity has stochastic fluctuations and the function of being

numerically integrated is a deterministic subject. Since most of the numerical integration techniques are linear operations, the order of the ensemble average operation can often be exchanged with the numerical integration. Thus, the ideas we discuss here on numerical integration (or called quadrature) are useful for us to derive optimal averaging method for climate quantities presented later.

4.1.1 Optimal weights only

We consider the following definite integral:

$$I = \int_{-1}^1 f(x) dx. \quad (4.1.1)$$

The known data are:

$$\{x_j, f_j = f(x_j)\}, \quad j = 1, 2, \dots, N,$$

with

$$-1 \leq x_1 < x_2 < \dots < x_{N-1} < x_N \leq 1.$$

See Fig. 4.1.

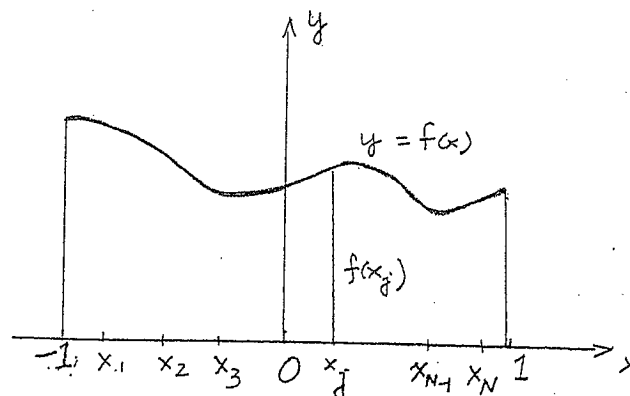


Figure 4.1: Grid points for a numerical integration on $[-1, 1]$.

The numerical integration is given by

$$\hat{I} = \sum_{j=1}^N w_j f_j. \quad (4.1.2)$$

The weights w_j satisfy

$$\sum_{j=1}^N w_j = 2 \quad (4.1.3)$$

and can be optimized such that \hat{I} is closest to I . So we minimize

$$\epsilon^2 = (I - \hat{I})^2. \quad (4.1.4)$$

According to the standard technique we learned in calculus to find extreme values subject to constraints, we construct a Lagrangian:

$$J[w_1, \dots, w_N] = \epsilon^2(w_1, \dots, w_N) - 2\Lambda \left(\sum_{j=1}^N w_j - 2 \right), \quad (4.1.5)$$

where -2Λ is the Lagrange multiplier. The conditions

$$\frac{\partial J}{\partial w_i} = 0 \quad \text{and} \quad \frac{\partial J}{\partial \Lambda} = 0$$

yield

$$(\hat{I} - I)f_i = \Lambda, \quad i = 1, 2, \dots, N, \quad (4.1.6)$$

$$\sum_{j=1}^N w_j = 2. \quad (4.1.7)$$

Now we see that the only possible solutions are: either $f_i = \text{constant}$ or $\hat{I} = I$. When $f_i = \text{constant}$, we have $\hat{I} = I$ anyway. Therefore, our optimization procedure can only work for the zero-error-quadrature. This zero-error-quadrature has a unique solution and can be derived easily from Lagrange interpolation formula.

When we have data

$$\{x_j, f_j = f(x_j)\}, \quad j = 1, 2, \dots, N,$$

the Lagrange interpolation formula for $f(x)$ is

$$L(x) = \sum_{i=1}^N \prod_{j=1, j \neq i}^N \frac{x - x_j}{x_i - x_j} f_i. \quad (4.1.8)$$

This is a polynomial of order $N - 1$. At the data points, the approximate function $L(x)$ and the original function have the same value: $L(x_j) = f_j$, $j = 1, 2, \dots, N$. If $f(x)$ has N -th derivative, then the difference between $f(x)$ and $L(x)$ can be expressed in a nice form:

$$f(x) - L(x) = \frac{f^{(N)}(\xi)}{n!} \prod_{j=1}^N (x - x_j) \quad (4.1.9)$$

where ξ is a function of x and $\xi \in [-1, 1]$.

If $f(x)$ is a polynomial of order $N - 1$, then $f(x) = L(x)$ since $f^{(N)}(x) = 0$. The integration of $L(x)$ is

$$\int_{-1}^1 L(x) dx = \sum_{i=1}^N \left(\int_{-1}^1 dx \prod_{j=1, j \neq i}^N \frac{x - x_j}{x_i - x_j} \right) f_i. \quad (4.1.10)$$

Hence if we choose

$$w_i = \int_{-1}^1 dx \prod_{j=1, j \neq i}^N \frac{x - x_j}{x_i - x_j}, \quad (4.1.11)$$

we have

$$\hat{I} = I \quad (4.1.12)$$

as long as $f(x)$ is a polynomial of order not higher than $N - 1$.

This claim itself is more or less trivial since a polynomial of order $N - 1$ has N coefficients. When one has function values at N distinguished points, the polynomial function is uniquely determined and hence its numerical integration based upon these known data at the N distinguished points should be exact.

Now the question is: can we optimize both the weights and the positions of the points so that we can use data at N points to get zero error quadrature for higher order polynomials? The answer is yes. Using Legendre polynomials, one can get zero error quadrature for polynomials of order up to $2N - 1$. This is the Gaussian quadrature method and will be discussed in the next subsection.

4.1.2 Optimize both weights and positions

Again we approximate

$$I = \int_{-1}^1 f(x) dx \quad (4.1.13)$$

by

$$\hat{I} = \sum_{j=1}^N w_j f(x_j). \quad (4.1.14)$$

What was done in the last section has an apparent shortcoming since it does not prevent the points being cluttered together and leaving a large interval with no sampling points. This is obviously a bad sampling design. To improve it, unlike the last section where only the weights w_j are optimized, we here optimize both the weights w_j and the position x_j . Because of large number of freedoms, there can be many optimization schemes that optimize both w_j and x_j . The Gaussian quadrature method is an optimal scheme that can integrate exactly any polynomials of order $2N - 1$ with N data points given. We may state this conclusion as a theorem.

Theorem 4.1 *If $f(x)$ is a polynomial of order $2N - 1$, and x_j ($j = 1, 2, \dots, N$) are the zeros of the Legendre polynomial $P_N(x)$ of order N , then*

$$\int_{-1}^1 f(x) dx = \sum_{j=1}^N w_j f(x_j) \quad (4.1.15)$$

with weights

$$w_i = \int_{-1}^1 dx \prod_{j=1, j \neq i}^N \frac{x - x_j}{x_i - x_j}, \quad j = 1, 2, \dots, N. \quad (4.1.16)$$

Example: Let us consider 5 points integration over $[-1, 1]$. The five zeros of $P_5(x)$ are

$$x: \quad 0, \pm 0.538469, \pm 0.90618.$$

The weights are:

$$w: \quad 0.566889, 0.47863, 0.23693.$$

One can either find the above data from a mathematics handbook or by using Mathematica.

For

$$I = \int_0^1 x^8 dx,$$

the exact value is $1/9 = 0.11111$. The Gaussian integration value by using 3 points (only half of the interval $[-1, 1]$) is:

$$\hat{I} \approx 0.5 \times 0^8 \times 0.566889 + 0.538469^8 \times 0.47863 + 0.90618^8 \times 0.23693 = 0.111113.$$

This, as we know, should be exactly equal to I by the claim of the theorem since $2 \times 5 - 1 = 9 > 8$.

We also see that even in the situation that it does not render exact value, the Gaussian integration still yields rather accurate approximation with only few integration points. Look at

$$I = \int_0^1 e^{-x^2} dx.$$

The "true" value is 0.746824 (obtained by many points numerical integration). If we use the 3 points Gaussian integration, we have:

$$\begin{aligned} \hat{I} \approx & 0.5 \times e^{-0^2} \times 0.566889 + e^{-0.538469^2} \times 0.47863 \\ & + e^{-0.90618^2} \times 0.23693 = 0.745834. \end{aligned}$$

This is a very impressive accuracy: with only three points our error is as small as 0.13%!

The weights and points defined above are called the Gaussian weights and Gaussian points respectively. The formula (4.1.15) is called the Gaussian quadrature formula. The derivation of the Gaussian quadrature formula is quite easy with the preparation we have had and it is shown below.

Since $f(x)$ is a polynomial of order $2N - 1$, we can write it in the form

$$f(x) = Q(x)P_N(x) + R(x), \quad (4.1.17)$$

where both $Q(x)$ and $R(x)$ are polynomials of order less or equal to $N - 1$. Hence, they are orthogonal to $P_N(x)$, i.e.,

$$\int_{-1}^1 Q(x)P_N(x) dx = 0. \quad (4.1.18)$$

Thus,

$$I = \int_{-1}^1 f(x) dx = \int_{-1}^1 R(x) dx. \quad (4.1.19)$$

This last integral can have an exact numerical integration when choosing the optimal weights, for $R(x)$ is a polynomial of order not larger than $N - 1$. So

$$\begin{aligned} I &= \sum_{j=1}^N w_j R(x_j) \\ &= \sum_{j=1}^N w_j [Q(x_j)P_N(x_j) + R(x_j)] \\ &= \sum_{j=1}^N w_j f(x_j). \end{aligned} \quad (4.1.20)$$

If the integrand is not a polynomial, the Gaussian method can still be used to compute an approximate value of I , but certainly it is not the exact value of I . In this method, we have further dissatisfaction: the weights and the points, although optimal in the sense that it can integrate $2N - 1$ order polynomials exactly, are determined *a priori* and are irrelevant to the structure of the function. One method that improves this shortcoming is the numerical integration based on the spline interpolation formulas. The theory of spline interpolation leads to a method whose weights are dependent on the structure of the function. We are not in the position to study the detailed theory of spline interpolation, but this idea of weights depending on the integrand function is very useful and will be described in next section for finding the global average temperature.

4.1.3 Monte Carlo method

The theory in the above subsection is interesting and useful (particularly in spectral GCM models), but it may not be very convenient, for one has to find the zeros of a Legendre polynomial and the weights (although they are available in mathematical handbooks). A lazy-boy may prefer another method which is to gamble: sample lots of points and add them together. You do not care optimization, you do not care the function structure, and you care nothing. Only thing you do is to sample MANY MANY points. This method is obviously useless without modern computers. But with the high speed computers as we have now, it is a very handy tool. Its procedure is as below.

1. Generate N (a very large number, say, more than 1000) independent uniformly distributed random points in $[-1, 1]$.
2. Compute

$$\hat{I} = \frac{2}{N} \sum_{j=1}^N f(x_j). \quad (4.1.21)$$

This \hat{I} may be considered as an approximation of I . Here we have assumed that the function $f(x)$ is known (which may be very complicated), otherwise we cannot compute $f(x_j)$. From above formula we see that the weights are uniform: $w_j = 2/N$. So Monte Carlo method may be regarded as a uniformly positioned (in probability sense) and uniformly weighted numerical integration scheme. The power of this scheme is not very obvious in computing the simple integral in $[-1, 1]$, and it can be surprisingly powerful in estimate high dimensional multiple integrals:

$$\int \dots \int f(x) d\Omega \approx \frac{\text{Vol}(\Omega)}{N} \sum_{j=1}^N f(x_j), \quad (4.1.22)$$

where $\text{Vol}(\Omega)$ is the volume of the integration domain.

The error resulted from the Monte Carlo integration is random since the positions are random. Its expectation value satisfies

$$\epsilon^2 = \langle (I - \hat{I})^2 \rangle = \gamma \frac{1}{N}, \quad (4.1.23)$$

for a constant γ .

Let us look an example of Monte Carlo integration by Mathematica. Consider

$$I = \int_{-1}^1 x^2 dx. \quad (4.1.24)$$

Its exact value is $2/3 = 0.666667$. Using Monte Carlo method, one does the following:

1. Generate random $N = 1000$ points:

```
t = Table[Random[Real, {-1,1}], {1000}] .
```

2. Compute:

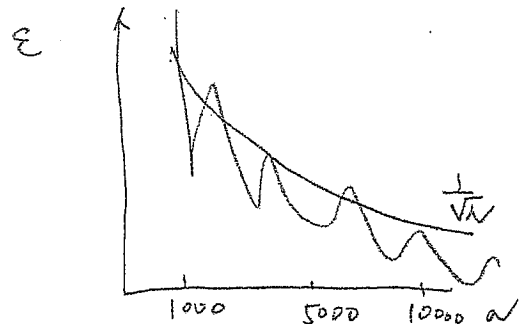
```
(2/1000) Sum[(t[[i]])^2, {i,1000}].
```

It gives result : 0.64919.

If one choose $N = 10000$, the result is more accurate. One of our tests shows the result: 0.667956.

Please notice that the positions are random. Hence, even using the same number of points, the results can be different for different experiments.

The use of Monte Carlo method is now so popular that it reaches almost every branch of science since it is really a lazy-boy tool and it is very effective. So as long as one can convert the problem into an averaging problem, Monte Carlo method can help you.



4.2 Global warming

In this section we mainly discuss the optimal weighting method introduced by Shen et al. (1994) to the detection of climate change. The weights here depend on the inhomogeneous structure of the climate field (to be exact, the surface air temperature field).

We use $T(\hat{\mathbf{r}}, t)$ to denote the annual mean temperature anomaly. Here the unit for t is [year]. The station data $T(\hat{\mathbf{r}}_j, t)$ ($j = 1, 2, \dots, N_{net}$) is prepared according to the method described in Section 3.1. Our goal is to compute the global average of $T(\hat{\mathbf{r}}, t)$ by using the data $T(\hat{\mathbf{r}}_j, t)$ ($j = 1, 2, \dots, N_{net}$) for a quite small N_{net} , say, less than 100. It is well known that there are several schools which computed the global average annual mean temperature for the last 130 years or so. The typical ones are those of Jones et al. (1986) (using 1873 stations), Hansen and Lebedeff (1987) (using 2685 stations) and Vinnikov et al. (1990) (using 566 stations). We took a subset of Jones data and consider only the period 1891-1990.

We will make use of the EOFs. The covariance function (or called auto-covariance function) of the $T(\hat{\mathbf{r}}, t)$ field is

$$\rho(\hat{\mathbf{r}}, \hat{\mathbf{r}}') = \langle T(\hat{\mathbf{r}}, t)T(\hat{\mathbf{r}}', t) \rangle. \quad (4.2.1)$$

We adopt the assumption that the temperature time series is stationary. This assumption may be justified in the sense that the changes of the temperature anomaly (both mean and variance) are not large compared with the standard deviation of the anomaly. Of course, strictly speaking the temperature time series is not stationary (particularly due to the reason that there is supposedly a trend caused by anthropological forcing we desire to detect).

The EOFs $\psi_n(\hat{\mathbf{r}})$ are defined by

$$\int_{4\pi} d\Omega \rho(\hat{\mathbf{r}}, \hat{\mathbf{r}}') \psi_n(\hat{\mathbf{r}}') = \lambda_n \psi_n(\hat{\mathbf{r}}), \quad n = 1, 2, \dots \quad (4.2.2)$$

EOFs have the following properties:

$$\rho(\hat{\mathbf{r}}, \hat{\mathbf{r}}') = \sum_{n=1}^{\infty} \lambda_n \psi_n(\hat{\mathbf{r}}) \psi_n(\hat{\mathbf{r}}'), \quad (4.2.3)$$

$$\int_{4\pi} d\Omega \psi_m(\hat{\mathbf{r}}) \psi_n(\hat{\mathbf{r}}) = \delta_{mn} \quad (4.2.4)$$

$$\sum_{n=1}^{\infty} \psi_n(\hat{\mathbf{r}}) \psi_n(\hat{\mathbf{r}}') = \delta(\hat{\mathbf{r}} - \hat{\mathbf{r}}'). \quad (4.2.5)$$

The last formula is a spectral expression of the spatial white noise.

Now let us look at the global average of $T(\hat{\mathbf{r}}, t)$ given by

$$\bar{T}(t) = \frac{1}{4\pi} \int_{4\pi} d\Omega T(\hat{\mathbf{r}}, t). \quad (4.2.6)$$

This integral is estimated by

$$\hat{T}(t) = \sum_{j=1}^{N_{net}} w_j T(\hat{\mathbf{r}}_j, t), \quad (4.2.7)$$

where the weights satisfy a normalization condition:

$$\sum_{j=1}^{N_{net}} w_j = 1. \quad (4.2.8)$$

The mean square error (MSE) is

$$\epsilon^2 = \langle (\bar{T} - \hat{T})^2 \rangle. \quad (4.2.9)$$

The expansion of the above becomes

$$\epsilon^2 = \langle (\bar{T})^2 \rangle - 2 \sum_{j=1}^{N_{net}} w_j \langle \bar{T} T(\hat{\mathbf{r}}_j, t) \rangle + \sum_{i,j=1}^{N_{net}} w_i w_j \langle T(\hat{\mathbf{r}}_i, t) T(\hat{\mathbf{r}}_j, t) \rangle. \quad (4.2.10)$$

Again due to the assumption that the temperature time series is stationary, the above ensemble average is independent of time. We will minimize this MSE under the constraint (4.2.8). The Lagrange multiplier method is used. Define a function:

$$J[w_1, \dots, w_{N_{net}}] = \epsilon^2(w_1, \dots, w_{N_{net}}) + 2\Lambda \left[\sum_{j=1}^{N_{net}} w_j - 1 \right]. \quad (4.2.11)$$

The extreme value conditions

$$\frac{\partial J}{\partial w_i} = 0, \quad i = 1, 2, \dots, N_{net},$$

the MSE expression (4.2.10) and the constraint (4.2.8) lead to $N_{net} + 1$ linear algebraic equations for the weights $w_1, \dots, w_{N_{net}}$ and the Lagrange multiplier Λ :

$$\sum_{j=1}^{N_{net}} w_j \rho(\hat{\mathbf{r}}_i, \hat{\mathbf{r}}_j) + \Lambda = \bar{\rho}(\hat{\mathbf{r}}_i), \quad i = 1, 2, \dots, N_{net}, \quad (4.2.12)$$

$$\sum_{j=1}^{N_{net}} w_j = 1. \quad (4.2.13)$$

Here

$$\rho(\hat{\mathbf{r}}_i, \hat{\mathbf{r}}_j) = \langle T(\hat{\mathbf{r}}_i, t) T(\hat{\mathbf{r}}_j, t) \rangle \quad (4.2.14)$$

is the auto-covariance matrix (or simply called covariance matrix) and

$$\bar{\rho}(\hat{\mathbf{r}}_i) = \frac{1}{4\pi} \int_{4\pi} d\Omega \rho(\hat{\mathbf{r}}, \hat{\mathbf{r}}_i) \quad (4.2.15)$$

is the global average of the covariance function around the point $\hat{\mathbf{r}}_i$. From the EOF expression of ρ we have

$$\bar{\rho}(\hat{\mathbf{r}}_i) = \sum_{n=1}^{\infty} \lambda_n \psi_n(\hat{\mathbf{r}}_i) \bar{\psi}_n, \quad (4.2.16)$$

where

$$\bar{\psi}_n = \frac{1}{4\pi} \int_{4\pi} d\Omega \psi_n(\hat{\mathbf{r}}) \quad (4.2.17)$$

is the global average of the EOF $\psi_n(\hat{\mathbf{r}})$. This EOF $\psi_n(\hat{\mathbf{r}})$ has a spherical harmonic expansion

$$\psi_n(\hat{\mathbf{r}}) = \sum_{l=0}^{\infty} \sum_{m=-l}^l \psi_{n,lm} Y_{lm}(\hat{\mathbf{r}}). \quad (4.2.18)$$

Then

$$\bar{\psi} = \sqrt{4\pi} \psi_{n,00}. \quad (4.2.19)$$

Therefore, as long as we have EOFs expressed in terms of spherical harmonics (regarded as a data bank like the data set of Jone et al), we can solve the $N_{net}+1$ equations (4.2.12) and (4.2.13) to find the optimal weights $w_1, \dots, w_{N_{net}}$. The minimal MSE can be written in a nice form

$$\epsilon_{opt}^2 = \sum_{n=1}^{\infty} \lambda_n \left| \bar{\psi}_n - \sum_{i=1}^{N_{net}} w_i \psi_n(\hat{\mathbf{r}}_i) \right|^2. \quad (4.2.20)$$

This formula implies that the sampling errors for the global average temperature is the sum of the sampling errors of the EOFs weighted by the variance (λ_n) of each EOF component. For relatively dense networks (say, more than 50 well distributed stations), the convergence of the above series is very fast. The first five modes would be enough to give satisfactory results (See Fig. 4.2). This fast convergence is partly due to the decrease of λ_n and partly due to the accurate sampling of the the lower order EOFs.

After obtaining the optimal weights, we can compute the (estimated) global average $\hat{T}(t)$ for each year. Two examples are presented here. One is a network of 4×4 uniformly distributed stations and the other is the Angell-Koshover (A-K) network of 63 stations. The 4×4 network stations are on the node points of $67.5S, 22.5S, 22.5N, 67.5N$ latitude circles and $45E, 135E, 45W, 135W$ longitude circles. The A-K network was designated in 1958 by WMO to measure the temperature in stratosphere by radiosonde (see Fig. 4.3 for the positions of the stations).

The global average temperature averaged by uniform weights

$$\bar{T}_u = \frac{4\pi}{N_{net}} \sum_{j=1}^{N_{net}} T(\hat{\mathbf{r}}_j, t) \quad (4.2.21)$$

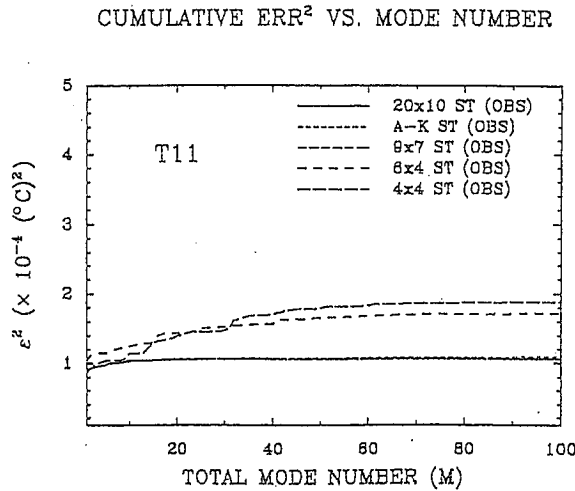


Figure 4.2: MSE and its convergence according to the number of EOF modes.

and optimal weights

$$\bar{T}_o = \sum_{j=1}^{N_{net}} w_j T(\hat{r}_j, t) \quad (4.2.22)$$

are shown in Fig. 4.4 for the networks of 16 stations and 63 stations. The minimal sampling errors (the square root of the MSE) for the two networks are

$$\epsilon_{4 \times 4} = 0.014^\circ\text{C} \quad \text{and} \quad \epsilon_{A-K} = 0.010^\circ\text{C}. \quad (4.2.23)$$

The above results show that using about 60 well distributed stations, one can obtain a very accurate global average temperature. This claim, although remarkably important and interesting, is based upon that the known EOFs are correct. So there is a problem here. If one does not have the computed EOFs in data bank but just the data from 60 stations, can one construct EOFs with reasonable accuracy and also get the satisfactory global average? This problem is still open.

4.3 Spherical harmonic components

In the above section, we discussed how to estimate the global average. As we know that the change of the global temperature and its impact is not uniform, there is a rather uneven pattern of the global change. This requires one to consider the evolution of the pattern. As we have done before, it is better to decompose the pattern in term of known sub-patterns: spheric harmonics or EOFs. The component of the lowest order spheric harmonic function is actually equal of the global average times $\sqrt{4\pi}$.

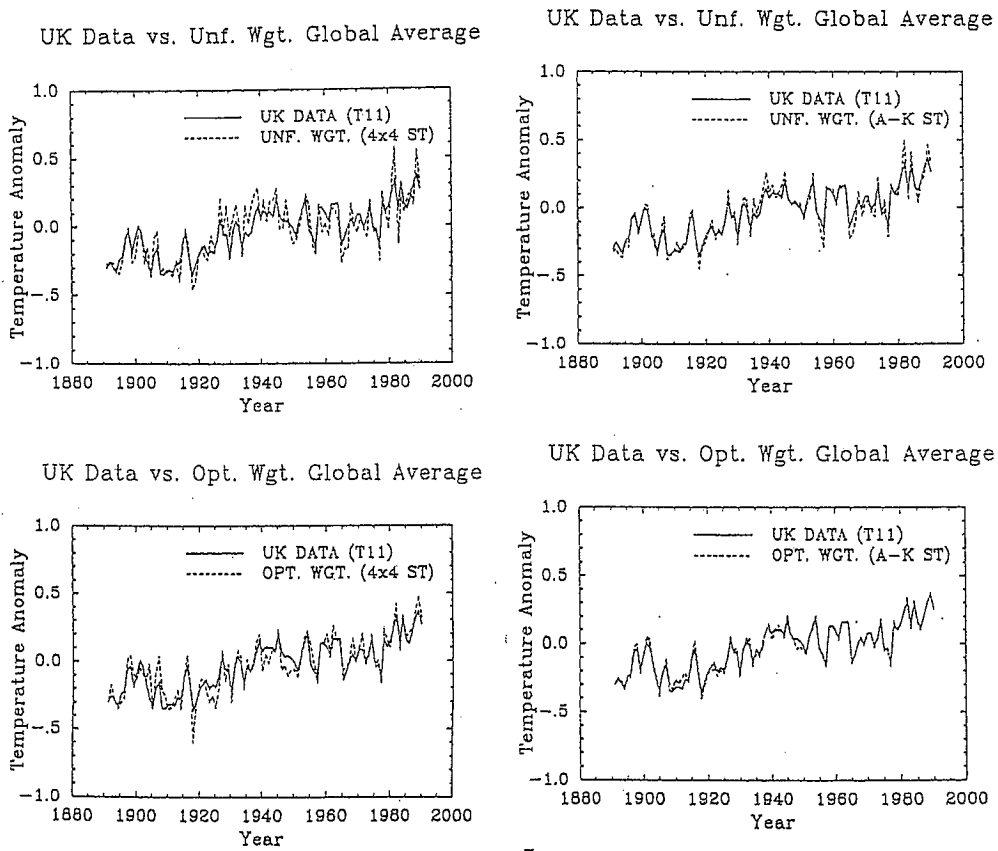


Figure 4.4: The reconstructed time series of the average surface air temperature by using 16 stations (a) and 63 station (b)

Let me describe the mathematics of the problem now. Again use $T(\hat{\mathbf{r}}, t)$ denote the annual mean temperature anomaly field. The spheric harmonic expansion of it is

$$T(\hat{\mathbf{r}}, t) = \sum_{l=0}^{\infty} \sum_{m=-l}^l T_{lm}(t) Y_{lm}(\hat{\mathbf{r}}), \quad (4.3.1)$$

where T_{lm} are called the spherical harmonic components defined by

$$T_{lm}(t) = \int_{4\pi} d\Omega T(\hat{\mathbf{r}}, t) Y_{lm}^*(\hat{\mathbf{r}}). \quad (4.3.2)$$

Our problem is to use the data: $T(\hat{\mathbf{r}}_j, t)$ ($j = 1, 2, \dots, N_{net}$) to estimate the $T_{lm}(t)$:

$$\hat{T}_{lm}(t) = \sum_{j=1}^{N_{net}} w_j^{(lm)} T(\hat{\mathbf{r}}_j, t) Y_{lm}(\hat{\mathbf{r}}_j). \quad (4.3.3)$$

Here we regard the weights independent of time and applicable all the time. The constraint put on the weights is

$$\sum_{j=1}^{N_{net}} w_j^{(lm)} = 4\pi. \quad (4.3.4)$$

Following the same procedure in the above section, consider the MSE:

$$\epsilon_{(lm)}^2 = \langle |T_{lm} - \hat{T}_{lm}|^2 \rangle. \quad (4.3.5)$$

Due to the assumption that the temperature time series is stationary, the above MSE is time independent. We minimize the MSE under the constraint condition (4.3.4) by the method of Lagrange multiplier. The extreme value conditions lead to the $N_{net} + 1$ linear equations for the weights and the Lagrange multiplier. With the weights, we can of course compute the estimate of $T_{lm}(t)$.

The minimal MSE in this analysis can finally be expressed in terms of a nice form:

$$(\epsilon_{lm}^2)_{opt} = \sum_{n=1}^{\infty} \lambda_n \left| \psi_{lm}^n - \sum_{j=1}^{N_{net}} w_j^{(lm)} \psi_n(\hat{\mathbf{r}}_j) Y_{lm}^*(\hat{\mathbf{r}}_j) \right|^2. \quad (4.3.6)$$

Here, $\psi_n(\hat{\mathbf{r}})$ are EOFs and λ_n are the corresponding eigenvalues of the covariance function.

The above is only a sketch of the idea and the computation procedures. The actual computations are quite tedious. For details, please read Kim et al. (1995).

4.4 Homogeneous reduction

If the climate field is homogeneous, then the EOFs on a sphere are the spherical harmonic functions as we pointed out in Sub-section 3.5.3. Then the computation procedures are simpler than those presented in the last two sections. The

obtained results sometimes have their significance in climatology. But of course one gets to be careful not over interpreting the results since the real climate field is often far away from being homogeneous.

4.4.1 The covariance function kernel

The nature of mean square error implies that only the first two moments of the measured field are relevant to the MSE. The temperature field under our consideration is its spatial anomaly. Hence, its first moment vanishes, i.e.

$$\langle T(\hat{\mathbf{n}}) \rangle = 0$$

where $\langle \cdot \rangle$ denotes the ensemble average. The second moment is described by its covariance function, which can be regarded as a symmetric kernel of an integral operator:

$$K(\hat{\mathbf{n}}, \hat{\mathbf{n}}') = \langle T(\hat{\mathbf{n}})T(\hat{\mathbf{n}}') \rangle. \quad (4.4.1)$$

By definition, when we say that a field is homogeneous, we mean that

$$K(\hat{\mathbf{n}}, \hat{\mathbf{n}}') = K(|\hat{\mathbf{n}} - \hat{\mathbf{n}}'|),$$

or

$$\langle T(\hat{\mathbf{n}})T(\hat{\mathbf{n}}') \rangle = \sigma^2 \rho(\hat{\mathbf{n}} \cdot \hat{\mathbf{n}}') = \sigma^2 \rho(x) \quad (4.4.2)$$

where $\sigma^2 = \langle T^2(\hat{\mathbf{n}}) \rangle$ is the low frequency point-variance of the temperature field at point $\hat{\mathbf{n}}$. Note the $x = (\hat{\mathbf{n}} \cdot \hat{\mathbf{n}}')$ is the cosine of the opening angle between the directions (stations) $\hat{\mathbf{n}}$ and $\hat{\mathbf{n}}'$. The correlation function $\rho(x)$ is dimensionless and normalized by $\rho(x=1) = 1$.

An important consequence of the homogeneity assumption is that the spectra of the covariance field consist only of the coefficients of the Fourier-Legendre series of the function $\rho(x)$:

$$\rho_n = \frac{1}{2} \int_{-1}^1 dx \rho(x) P_n(x), \quad (n = 0, 1, 2, 3, \dots). \quad (4.4.3)$$

Correspondingly, the correlation function $\rho(x)$ is expressed in a series sum of Legendre polynomials:

$$\rho(x) = \sum_{n=0}^{\infty} (2n+1) \rho_n P_n(x). \quad (4.4.4)$$

Now we apply the addition theorem for Legendre polynomials:

$$P_n(\hat{\mathbf{n}} \cdot \hat{\mathbf{n}}') = \frac{4\pi}{2n+1} \sum_{k=-n}^n Y_{nk}(\hat{\mathbf{n}}) Y_{nk}^*(\hat{\mathbf{n}}') \quad (\text{Addition theorem}). \quad (4.4.5)$$

The covariance function of a homogeneous field can now be written as

$$\langle T(\hat{\mathbf{n}})T(\hat{\mathbf{n}}') \rangle = \sum_{n=0}^{\infty} \sum_{k=-n}^n 4\pi \sigma^2 \rho_n Y_{nk}(\hat{\mathbf{n}}) Y_{nk}^*(\hat{\mathbf{n}}'). \quad (4.4.6)$$

Hence, the spherical harmonics $Y_{n,k}(\hat{\mathbf{n}})$ are now the eigenfunctions (EOFs), and one eigenvalue $4\pi\sigma^2\rho_n$ corresponds to $2n+1$ different eigenfunctions $Y_{n,k}(\hat{\mathbf{n}})$, ($k = -n, \dots, n-1, n$).

4.4.2 The MSE formula

We use the data from N stations at points $\hat{\mathbf{n}}_1, \hat{\mathbf{n}}_2, \dots, \hat{\mathbf{n}}_N$ to estimate the spherical harmonic components T_{lm} . The linear estimator is

$$\hat{T}_{lm} = \sum_{j=1}^N w_j^{(lm)} T(\hat{\mathbf{n}}_j) Y_{lm}^*(\hat{\mathbf{n}}_j). \quad (4.4.7)$$

This is the Riemann sum of the integral (2). The surface of the unit sphere is partitioned into N sub-regions and the weight $w_j^{(lm)}$ is the area of the j th sub-region ($j = 1, 2, \dots, N$). Hence the weights $w_j^{(lm)}$ ($j = 1, 2, \dots, N$) are real-valued and satisfy the normalization condition

$$\sum_{j=1}^N w_j^{(lm)} = 4\pi. \quad (4.4.8)$$

The MSE for estimating T_{lm} is defined as

$$\epsilon_{(lm)}^2 = \langle |T_{lm} - \hat{T}_{lm}|^2 \rangle. \quad (4.4.9)$$

This can be re-written into:

$$\begin{aligned} \epsilon_{(lm)}^2 &= \left\langle \left| \int_{4\pi} d\Omega T(\hat{\mathbf{n}}) [1 - w^{(lm)}(\hat{\mathbf{n}})] Y^*(\hat{\mathbf{n}}) \right|^2 \right\rangle \\ &= \int_{4\pi} d\Omega \int_{4\pi} d\Omega' \langle T(\hat{\mathbf{n}}) T(\hat{\mathbf{n}}') \rangle [1 - w^{(lm)}(\hat{\mathbf{n}})] \\ &\quad \times [1 - w^{(lm)}(\hat{\mathbf{n}}')] Y_{lm}^*(\hat{\mathbf{n}}) Y_{lm}(\hat{\mathbf{n}}') \\ &= \sum_{n=1}^{\infty} \lambda_n |\psi_{n,lm} - \hat{\psi}_{n,lm}|^2 \end{aligned} \quad (4.4.10)$$

where

$$w^{(lm)}(\hat{\mathbf{n}}) = \sum_{j=1}^N w_j^{(lm)} \delta(\hat{\mathbf{n}} - \hat{\mathbf{n}}_j), \quad (4.4.11)$$

$$\psi_{n,lm} = \int_{4\pi} d\Omega \psi_n(\hat{\mathbf{n}}) Y_{lm}^*(\hat{\mathbf{n}}), \quad (4.4.12)$$

$$\hat{\psi}_{n,lm} = \sum_{j=1}^N w_j^{(lm)} \psi_n(\hat{\mathbf{n}}_j) Y_{lm}^*(\hat{\mathbf{n}}_j), \quad (4.4.13)$$

and ψ_n are the EOFs.

If the field is homogeneous, then the spherical harmonics are the EOFs as pointed in the above sub-section

$$\lambda_n \leftrightarrow 4\pi\sigma^2\rho_n, \quad \psi_n(\hat{\mathbf{n}}) \leftrightarrow Y_{nk}(\hat{\mathbf{n}}), \quad k = -n, \dots, n-1, n.$$

Then the MSE formula (4.4.10) becomes

$$\epsilon_{(lm)}^2 = \sum_{n=0}^{\infty} \sum_{k=-n}^n 4\pi\sigma^2\rho_n |\delta_{nl}\delta_{km} - \sum_{j=1}^N w_j^{(lm)} Y_{nk}(\hat{\mathbf{n}}_j) Y_{lm}^*(\hat{\mathbf{n}}_j)|^2. \quad (4.4.14)$$

Using the addition theorem again, one can reduce the MSE formula (4.4.14) to an easy-to-compute form:

$$\begin{aligned} \frac{\epsilon_{(lm)}^2}{4\pi\sigma^2} &= \sum_{n=0}^{\infty} (2n+1)\rho_n \frac{1}{4\pi} \sum_{i,j=1}^N w_i^{(lm)} w_j^{(lm)} P_n(\hat{\mathbf{n}}_i \cdot \hat{\mathbf{n}}_j) Y_{lm}^*(\hat{\mathbf{n}}_i) Y_{lm}(\hat{\mathbf{n}}_j) \\ &\quad + \rho_l \left(1 - 2 \sum_{j=1}^N w_j^{(lm)} |Y_{lm}(\hat{\mathbf{n}}_j)|^2 \right). \end{aligned} \quad (4.4.15)$$

Thus, the sampling error is explicitly expressed in terms of a series sum of spectral components whose coefficients are functions of the positions and weights of the stations. And the spectra ρ_n ($n = 0, 1, 2, \dots$) can be obtained from a homogeneous climate model.

4.5 Spectra derived from noise forced EBM

Here we consider the case of a simple climate model which is a white noise forced linear energy balance model (EBM) given by

$$\tau_0 \frac{\partial}{\partial t} T(\hat{\mathbf{n}}, t) - \lambda_0^2 \nabla^2 T(\hat{\mathbf{n}}, t) + T(\hat{\mathbf{n}}, t) = F(\hat{\mathbf{n}}, t) \quad (4.5.1)$$

where $T(\hat{\mathbf{n}}, t)$ is the local departure of the temperature from its climatology; τ_0 is an inherent time scale and λ_0 is an inherent length scale. We are interested only in the low frequency limit (annual average or two-year average) of the climate process. With this limit the time dependent term in the above model drops out. Hence we simply consider the time independent model. The unit of the length scale λ_0 is the Earth radius. The forcing function is a spatial white noise, i.e.,

$$\langle F(\hat{\mathbf{n}}) F(\hat{\mathbf{n}}') \rangle = \sigma_F^2 \delta(\hat{\mathbf{n}} - \hat{\mathbf{n}}'), \quad (4.5.2)$$

where δ is the Dirac delta function. The time independent noise forced EBM is

$$-\lambda_0^2 \nabla^2 T(\hat{\mathbf{n}}) + T(\hat{\mathbf{n}}) = F(\hat{\mathbf{n}}). \quad (4.5.3)$$

The spherical harmonic expansions for $T(\hat{\mathbf{n}})$ and $F(\hat{\mathbf{n}})$ are

$$T(\hat{\mathbf{n}}) = \sum_{l=0}^{\infty} \sum_{m=-l}^l T_{lm} Y_{lm}(\hat{\mathbf{n}}), \quad (4.5.4)$$

$$F(\hat{\mathbf{n}}) = \sum_{l=0}^{\infty} \sum_{m=-l}^l F_{lm} Y_{lm}(\hat{\mathbf{n}}). \quad (4.5.5)$$

Substituting these two expressions into the model equation (4.5.3), we can obtain

$$T_{lm} = \frac{F_{lm}}{1 + \lambda_0^2 l(l+1)}. \quad (4.5.6)$$

Substituting (4.5.4) into the left hand of (4.4.2) and using the Fourier-Legendre expansion (4.4.4) and the addition theorem for the spherical harmonic functions (4.4.5), we have

$$\begin{aligned} & \sum_{l=0}^{\infty} \sum_{m=-l}^l \sum_{l'=0}^{\infty} \sum_{m'=-l'}^{l'} \langle T_{lm} T_{l'm'}^* \rangle Y_{lm}(\hat{\mathbf{n}}) Y_{l'm'}^*(\hat{\mathbf{n}}') \\ &= \sigma^2 \sum_{n=0}^{\infty} \rho_n 4\pi \sum_{k=-n}^n Y_{nk}(\hat{\mathbf{n}}) Y_{nk}^*(\hat{\mathbf{n}}'). \end{aligned} \quad (4.5.7)$$

This equality and equation (4.5.6) imply that

$$\rho_n = \frac{\langle |F_{nm}|^2 \rangle / (4\pi\sigma^2)}{[1 + \lambda_0^2 n(n+1)]^2}, \quad n = 0, 1, 2, \dots \quad (4.5.8)$$

Here $\langle |F_{nm}|^2 \rangle$ can be found from the white noise assumption (4.5.2) and the expansion (4.5.5):

$$\sum_{l=0}^{\infty} \sum_{m=-l}^l \sum_{l'=0}^{\infty} \sum_{m'=-l'}^{l'} \langle F_{lm} F_{l'm'}^* \rangle Y_{lm}(\hat{\mathbf{n}}) Y_{l'm'}^*(\hat{\mathbf{n}}') \quad (4.5.9)$$

$$= \sigma_F^2 \sum_{n=0}^{\infty} \sum_{k=-n}^n Y_{nk}(\hat{\mathbf{n}}) Y_{nk}^*(\hat{\mathbf{n}}'). \quad (4.5.10)$$

This implies that

$$\langle F_{lm} F_{l'm'}^* \rangle = \sigma_F^2 \delta_{ll'} \delta_{mm'}. \quad (4.5.11)$$

When $n = 0$ in (4.5.8), we have

$$\rho_0 = \frac{\sigma_F^2}{4\pi\sigma^2}. \quad (4.5.12)$$

This ρ_0 can be determined by the normalization condition $\rho(x=1) = 1$, i.e.,

$$\sum_{n=0}^{\infty} (2n+1) \frac{\rho_0}{[1 + \lambda_0^2 n(n+1)]^2} P_n(1) = 1. \quad (4.5.13)$$

Noting that $P_n(1) = 1$ ($n = 0, 1, 2, \dots$), we have

$$\rho_0 = \frac{1}{\sum_{n=0}^{\infty} (2n+1)/[1 + \lambda_0^2 n(n+1)]^2}. \quad (4.5.14)$$

Hence ρ_0 is only a function of the length scale λ_0 . The larger the λ_0 is, the more variance is explained by the spectral component ρ_0 . Fig. 4.5 shows the relationship between ρ_0 and λ_0 , which, of course, is a monotonically increasing function. The value of λ_0 is determined by the length scale of the anomaly field. For the annual mean field, EBM length scale is about 2000 km. If we take the radius of the earth to be 6367 km, λ_0 takes the value: $2000/6367 = 0.3141$. The corresponding ρ_0 is 0.0954. Thus, the zeroth order spectral component ρ_0 explains about 10% of the low frequency point variance of the surface air temperature.

With the above preparation, one can compute the weights and MSEs. The details are described in Shen et al. (1995).

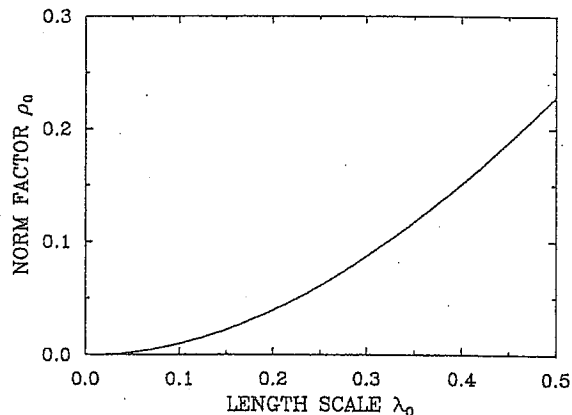


Figure 4.5: The length scale of a noise driven linear EBM.

4.6 Network design on unit sphere

For the future network design, there can be many methods and criteria. Here we present one which is of more interest in mathematics and grid design for global climate models rather than in climate data analysis. However, the idea used here might have some interesting implications for the future methods on climate data analysis.

The question is where to put n points on a unit sphere in three dimensional space so that the sum of their mutual distances takes the maximal value? This

is a very old and yet still difficult mathematical problem (in the sense of rigorous mathematics). We are going to solve this problem by sequential procedures.

Let p_1, p_2, \dots, p_{n-1} be $n-1$ points fixed on the unit sphere. We ask the question: where do we put the n th point so that the sum of their mutual distances becomes maximum? Let x be a point which can be moved on the sphere. Then, the sum of the mutual distances among the point x and the first $(n-1)$ points is a function of x :

$$S(x) = \sum_{i=1}^{n-1} |x - p_i|. \quad (4.6.1)$$

This function reaches its maximum at p_n which must be a critical point of the function $S(x)$. Hence the directional derivative of $S(x)$ in any tangential direction of the sphere at p_n is equal to zero. Equivalently, the gradient of $S(x)$ must be normal to the tangential plane and hence is in the radial direction. Therefore, there is a scalar α such that

$$\nabla S(p_n) = \alpha p_n. \quad (4.6.2)$$

Since $(x - p_i) \cdot x > 0$ ($i = 1, 2, \dots, n-1$), we have

$$\nabla S(x) \cdot x = \left(\sum_{i=1}^{n-1} \frac{x - p_i}{|x - p_i|} \right) \cdot x > 0. \quad (4.6.3)$$

Consequently, the scalar α in equation (4.6.2) is equal to

$$\alpha = |\nabla S(p_n)|. \quad (4.6.4)$$

From equation (4.6.2) the n th point p_n is the solution of the following equation:

$$p_n = \left(\sum_{i=1}^{n-1} \frac{p_n - p_i}{|p_n - p_i|} \right) / \left| \sum_{i=1}^{n-1} \frac{p_n - p_i}{|p_n - p_i|} \right|. \quad (4.6.5)$$

We solve this nonlinear equation for p_n by iteration procedure

$$p_n^{(k)} = \left(\sum_{i=1}^{n-1} \frac{p_n^{(k-1)} - p_i}{|p_n^{(k-1)} - p_i|} \right) / \left| \sum_{i=1}^{n-1} \frac{p_n^{(k-1)} - p_i}{|p_n^{(k-1)} - p_i|} \right|. \quad (4.6.6)$$

The initial configuration of the n points are put onto the sphere according to random and uniform distribution. The method is to generate random uniform distribution of point on an interval $[-1, 1]$ and then on a square $[-1, 1] \times [-1, 1]$. These points are finally projected to the unit sphere according to the uniform random distribution criterion.

We perform the first iteration for every point before we proceed to the second iteration, then the second iteration for every point before the third iteration, and similarly the third, fourth and k th iteration until the monotonically increasing sequence

$$D^{(k)}(n) = (1/2) \sum_{i,j=1}^n |p_i^{(k)} - p_j^{(k)}| \quad (4.6.7)$$

converges, i.e.

$$\lim_{k \rightarrow \infty} D^{(k)}(n) = D(n). \quad (4.6.8)$$

At the same time, the position sequences $p_j^{(k)}$ ($j = 1, 2, \dots, n$) converge to the equilibrium positions:

$$\lim_{k \rightarrow \infty} p_j^{(k)} = r_j, \quad j = 1, 2, \dots, n. \quad (4.6.9)$$

I developed a numerical software package to carry out the above iterations and plottings. In our numerical package, without loss of generality, we set the first point on the north pole, i.e.

$$p_1^{(k)} = (0, 0, 1). \quad (4.6.10)$$

The numerical tests we carried out seem to suggest that for $n > 4$ there are always a pair of antipodal points. The proof of this assumption is not yet available. To prevent from the free spin of the sphere, we also set a point on the Prime Meridian (i.e. zero longitude line). Our numerical tests also suggest that the following iteration scheme

$$p_j^{(k)} = \left(\sum_{i=1}^{n-1} \frac{-p_n^{(k-1)} + p_i}{|p_j^{(k-1)} - p_i|} \right) / \left| \sum_{i=1}^{n-1} \frac{p_n^{(k-1)} - p_i}{|p_n^{(k-1)} - p_i|} \right| \quad (4.6.11)$$

for $j = 2, 3, \dots, n-1$, which is the result of changing the sign on the right hand side of (4.6.6), converges faster. When $p_j^{(k-1)}$ is very close to its limit as k is sufficiently large, $p_j^{(k)}$ is almost equal to $-p_j^{(k-1)}$, a 180° rotation of $p_j^{(k-1)}$ around the axis connecting the north pole and the south pole. And the point on the south pole is

$$\lim_{k \rightarrow \infty} p_n^{(k)} = (0, 0, -1). \quad (4.6.12)$$

The precision of the solution of eq. (4.6.5) and the cut-off number for iterations depend on the what one needs and vary with respect to the total number of points n . But we observed that in the first a few steps of iterations, the convergence of $D^{(k)}(n)$ is very fast. The distance found after the first five steps of iteration, i.e. $D^{(5)}(n)$, is usually less than 1% away from the accurate solution $D(n)$. But the positions of the points may be far away from the equilibrium. In other words one may have to move the points for rather large distances in order to improve the sum of the mutual distances by merely 1%.

Interested readers may use anonymous ftp to get the package from
 cake.math.ualberta.ca

in the directory /pub/point, or write to me. Using the example of $n = 6$, one can run the package spherept.m in the following way:

```
In[1]:= n=6;
In[2]:= <<spherept.m      (load the package)
In[3]:= init             (generate the initial configuration)
Initial dist=19.9696     (the result from the initial configuration)
```

```

In[4]:= itera[5]          (instruct the computer to do five iterations)
k=1, dist=20.18004099665794, n=6
k=2, dist=21.78175758641795, n=6
k=3, dist=22.42284677829108, n=6
k=4, dist=22.43318413850058, n=6
k=5, dist=22.75807186171074, n=6
# of points=6, dist=22.75807186171074
Total # of iterations=5
In[5]:= plt              (plot the six points on the longitude-latitude grids)
In[6]:= itera[100]       (instruct the computer to do 100 more iterations)
In[7]:= plt              (plot the new result)
In[8]:= x                (show the Cartesian coordinates of the points)
Out[8]= {{0., 0., 1.}, {0.999038, -0.0438603, -0.000109174},
> {0.0438601, 0.999038, 0.000107343},
> {-0.0438604, -0.999038, 0.000107395},
-6 -8
> {-0.999038, 0.0438603, -0.000105564}, {-1.34913 10 , 3.95669 10 , -1.}}
In[9]:= tf                (show the longitude-latitude of the points)
Out[9]= {{0., 90.}, {0., -0.00625524}, {90., 0.00615029}, {-90., 0.0061533},
> {-180., -0.00604836}, {0, -89.9999}}

```

Table 1 shows the sum of the mutual distances among the n points. The formula $n \cot(\pi/2n)$ is the sum of the mutual distances of the vertices of the regular n -gon inscribed in a unit circle in two dimensional space E^2 . The formula $2n^2/3 - 1/2$ is an approximation of $D(n)$. The second formula is rather accurate, particularly when n is large. Hence it can be regarded as an asymptotic approximation of $D(n)$ as $n \rightarrow \infty$.

TABLE 1

n	11	12	13	14	15
Computer search	79.2746	94.5829	111.1704	129.1173	148.4005
$n \cot(\pi/2n)$	76.5067	91.1409	107.0646	124.2534	142.7155
$2n^2/3 - 1/2$	80.1667	95.5000	112.1667	130.1667	149.5000

n	16	17	18	19	20
Computer search	168.9781	190.9711	214.2610	238.8718	264.8362
$n \cot(\pi/2n)$	162.4507	183.4592	205.7409	229.2959	254.1241
$2n^2/3 - 1/2$	170.1667	192.1667	215.5000	240.1667	266.1667

One may expect that for $n = 4, 6, 8, 12$ and 20 , the solutions should be regular polyhedra. This is indeed true for $n = 4, 6, 8$ and 12 . But, to our

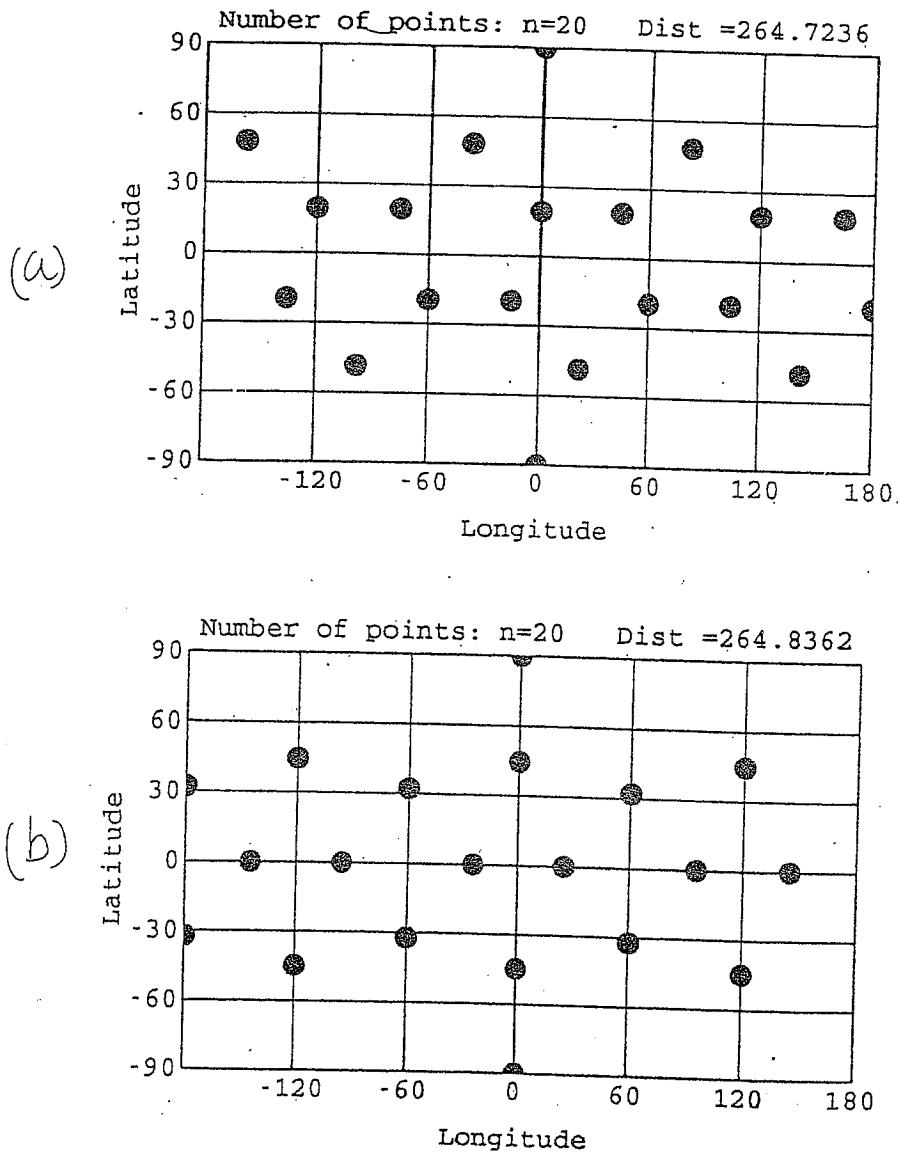


Figure 4.6: The points of dodecahedron (a) and maximal mutual distance (b).

surprise, this is not true for $n = 20$. For a dodecahedron inscribed in a unit sphere, the sum of the mutual distances among the vertices is equal to 264.7236, which is smaller than $D(20) = 264.8362$. As remarked earlier and depicted by Fig. 4.6, although this difference is small, the corresponding point positions are very much different. See Fig. 4.6 for the difference between the two sets of points: the vertices of dodecahedron (Fig. 4.6a) and the points which have the maximal sum of the mutual distances (Fig. 4.6b).

4.7 References

- Kim, K.Y., G.R. North, and S.S. Shen, 1995: Optimal estimation of spherical harmonic components from a sample with spatially nonuniform covariance statistics. *J. Climate*, in press.
- Shen, S.S., G.R. North, and K.Y. Kim, 1994: Spectral approach to optimal estimation of the global average temperature. *J. Climate*, 7, 1999-2007.
- Shen, S.S., G.R. North, and K.Y. Kim, 1995: Optimal estimation of the spherical harmonic components of the surface air temperature. *Environmetrics*, in press.