Dennis L. Hartmann and Elizabeth A. Barnes

Data Analysis for Atmospheric, Oceanic and Climate Science

July 24, 2025

Springer

Contents

Chapter 1
Introduction

This book is intended for graduate students and professionals working in atmospheric and oceanic sciences or closely related fields.

It has been prepared by Professord Dennis L. Hartmann and Elizabeth A. Barnes with the idea of publishing it, but we here make it available for use by all.

It is very helpful if the reader has a background that includes elementary statistics, probability and matrix algebra, but an attempt is made to provide some basic background in these subjects, so that the book is largely self-contained.

The focus of the book is on using some standard methods for practical purposes intelligently, rather than providing a deep and complete theoretical analysis of these methods. Examples are shown of appropriate uses of the techniques in atmospheric and oceanic sciences.

The choice of topics reflects the experience of the authors in doing research in the atmospheric and oceanic sciences over an extended period. Some of the examples presented are of historical interest, while many of the techniques are very applicable today and in the future.

Most of the figures and examples in this book were done using Python and Jupyter Notebooks. The notebooks we used can be found in the GitHub page for this book.

Chapter 2
Basic Statistical Techniques

2.1 Basic statistical quantities: means and other moments

2.1.1 The Mean

The sample *mean* of a set of N values, x_i, where $i = 1, 2, 3...N$ is given by

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{2.1}$$

The mean is the *first moment* about zero and should be distinguished from the *median*, which is the value in the center of the population (or the average of the two middle values if N is even).

The sample mean $\overline{x}$ is an *unbiased estimate* of the true population mean μ. An unbiased estimate implies that if we draw an infinite number of samples from the same underlying distribution, then the mean of all of the sample means will be equal to the underlying distribution's population mean μ.

In Practice.

- The median is a very useful quantity when the distribution of your dataset is not symmetric or contains outliers. For example, in **Fig. 2.1**, the median may be considered more representative of the data.

2.1.2 The Variance

The sample *variance* of a set of values, x_i, is given by

$$\overline{x'^2} = \frac{1}{N-1}\sum_{i=1}^{N} (x_i - \overline{x})^2 \tag{2.2}$$

where the prime denotes departures from the mean. The variance is the *second moment* about the mean. The division by $N-1$ instead of the expected N is required for an unbiased estimate of the variance. An explanation for why this is can be found in any standard statistics textbook, but it basically boils down to the fact that the sample mean is itself an estimate and comes with its own uncertainties which gives the sample variance a low bias without the $N-1$ correction.

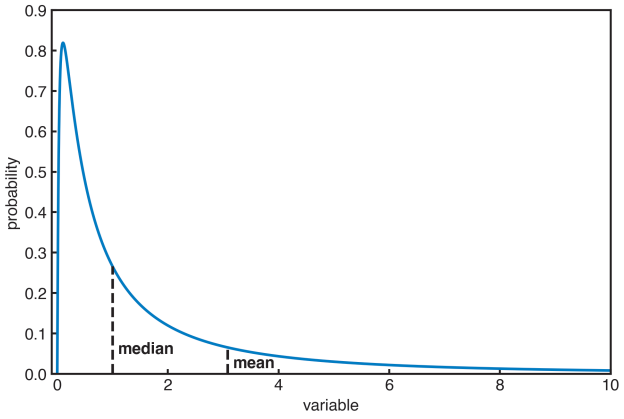

Figure 2.1 Comparison of the mean vs the median for a highly skewed distribution.

2.1.3 The Standard Deviation

The *standard deviation* is the square root of the variance and is often denoted as σ. The sample standard deviation of a set of values, x_i, is similarly defined as

$$s = \sqrt{\overline{x'^2}} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2} \tag{2.3}$$

2.1.4 Higher Moments

We can define an arbitrary moment about the mean as

$$m_r = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^r \tag{2.4}$$

so that m_2 is the variance, m_3 is the *skewness*, and m_4 is the *kurtosis*. Written in this way, note that m_2 is actually a biased estimate of the variance due to the division by N rather than $N-1$.

These m_r moments can be standardized (non-dimensionalized) by defining

$$a_r = \frac{m_r}{\sigma^r} \tag{2.5}$$

where σ is the standard deviation. The first two standardized moments are zero and 1, but the third and fourth are the coefficients of skewness and kurtosis, which give information about the shape of the distribution.

The *coefficient of skewness*, a_3 indicates the degree of asymmetry of the distribution about the mean. If $a_3 > 0$ then the distribution is said to be *skewed to the right* and has a longer tail on the positive side. If $a_3 < 0$ then the distribution is said to be *skewed to the left* and has a longer tail on the negative side. **Fig. 2.2** shows examples of a positively and negatively skewed distribution.

The *coefficient of kurtosis* (Greek word for curved or arching), a_4, indicates the degree to which the distribution is spread about the mean value or the length of the tails. The kurtosis can be thought of as the "tailedness" of the distribution, and is typically compared with the kurtosis of the Normal distribution which has $a_4 = 3$. Thus, distributions with excess kurtosis ($a_4 > 3$) are very peaked about the mean with long tails and are called *leptokurtic* (Greek for *leptos*, meaning small or narrow) and distributions with $a_4 < 3$ are very flat about the mean with short tails and are called *platykurtic* (Greek *platys*, meaning broad or flat).

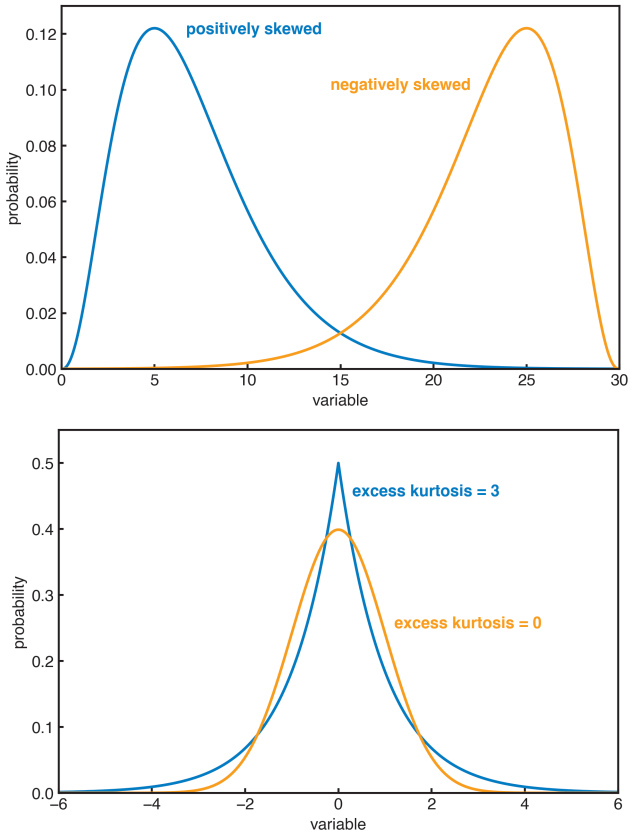

Figure 2.2 Examples of distributions with different skewness and kurtosis.

In Practice.

- In many software packages, the calculated kurtosis is actually the *excess kurtosis*, that is, the kurtosis minus 3 ($\mathfrak{a}_4 - 3$) since 3 is the kurtosis of the Normal distribution. Thus, platykurtic distributions will have negative kurtosis, and leptokurtic positive kurtosis.

2.2 Probability Concepts and Theorems

2.2.1 Unions and Intersections of Probability - Venn Diagram

The probability of some event E happening is written as $\boldsymbol{Pr}(\mathsf{E})$. For example, E could be that you roll a die (a die is a cube with a different number, 1 through 6, on each side) and get a 2. If the die is fair, then

$$\boldsymbol{Pr}(\mathsf{E}) = \frac{1}{6}. \tag{2.6}$$

The probability of E not happening

$$\boldsymbol{Pr}\left(\widetilde{\mathsf{E}}\right) = 1 - \boldsymbol{Pr}(\mathsf{E}) \tag{2.7}$$

where $\widetilde{\mathsf{E}}$ is the event that E *does not* happen. In the case of rolling the fair die

$$\boldsymbol{Pr}\left(\widetilde{\mathsf{E}}\right) = 1 - \frac{1}{6} = \frac{5}{6}. \tag{2.8}$$

The probability that either or both of two events, E_1 and E_2, will occur is called the *union* of the two probabilities and is given by

$$\boldsymbol{Pr}(\mathsf{E}_1 \cup \mathsf{E}_2) = \boldsymbol{Pr}(\mathsf{E}_1) + \boldsymbol{Pr}(\mathsf{E}_2) - \boldsymbol{Pr}(\mathsf{E}_1 \cap \mathsf{E}_2) \tag{2.9}$$

where $\boldsymbol{Pr}(\mathsf{E}_1 \cap \mathsf{E}_2)$ is the probability that both events will occur, and is called the *intersection*. It is the overlap between the two probabilities and must be subtracted from the sum. This is easily seen via a Venn diagram in **Fig. 2.3**. The area inside the two event circles indicates the probability of the two events. The intersection between them gets counted twice when you add the two areas and so must be subtracted to calculate the union of the probabilities. If the two events are *mutually exclusive* (i.e. the circles do not overlap), then no intersection occurs.

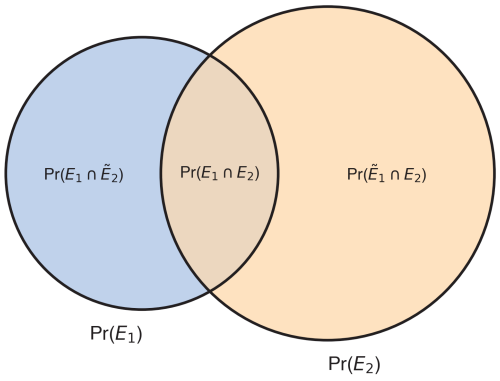

Figure 2.3 Venn Diagram illustrating the intersection of two probabilities.

Another important concept is *conditional probability*. We write the probability that E_2 will occur given that E_1 has occurred as

$$\boldsymbol{Pr}(\mathsf{E}_2|\mathsf{E}_1) = \frac{\boldsymbol{Pr}(\mathsf{E}_1 \cap \mathsf{E}_2)}{\boldsymbol{Pr}(\mathsf{E}_1)} \tag{2.10}$$

Moving terms around, one can also obtain a formula for the probability that both events will occur, and this is called the multiplicative law of probability,

$$\boldsymbol{Pr}(\mathsf{E}_1 \cap \mathsf{E}_2) = \boldsymbol{Pr}(\mathsf{E}_2|\mathsf{E}_1)\, \boldsymbol{Pr}(\mathsf{E}_1) = \boldsymbol{Pr}(\mathsf{E}_1|\mathsf{E}_2)\, \boldsymbol{Pr}(\mathsf{E}_2)\,. \tag{2.11}$$

If E_1 and E_2 are *independent events*, that is, their probabilities do not depend on one another, then

$$\boldsymbol{Pr}(\mathsf{E}_1|\mathsf{E}_2) = \boldsymbol{Pr}(\mathsf{E}_1) \tag{2.12}$$

$$\boldsymbol{Pr}(\mathsf{E}_2|\mathsf{E}_1) = \boldsymbol{Pr}(\mathsf{E}_2) \tag{2.13}$$

and so

$$\boldsymbol{Pr}(\mathsf{E}_1 \cap \mathsf{E}_2) = \boldsymbol{Pr}(\mathsf{E}_1)\, \boldsymbol{Pr}(\mathsf{E}_2) \tag{2.14}$$

(2.14) is the definition of statistical *independence*.

Worked Example 2.1.
If the probability of getting heads on a coin flip is 0.5, and one coin flip is independent of every other coin flip, then, using (2.14), the probability of getting N heads in a row is 0.5^{N}.

Worked Example 2.2.
The probability of it raining on Monday is 60%. But, you know from looking at historical records that the probability of it raining the day after it rains is 80% (it is more likely than not to rain the day after it rains). So, whether it rains on Tuesday is dependent on whether it rains on Monday. What is the probability it will rain Monday and Tuesday?

$$\mathsf{M} = \text{event that it rains Monday} \tag{2.15}$$
$$\mathsf{T} = \text{event that it rains Tuesday} \tag{2.16}$$
$$\boldsymbol{Pr}(\mathsf{M} \cap \mathsf{T}) = \boldsymbol{Pr}(\mathsf{T}|\mathsf{M}) \cdot \boldsymbol{Pr}(\mathsf{M}) = 0.8 \cdot 0.6 = 48\% \tag{2.17}$$

2.2.2 Bayes Theorem

Theorem 2.1 (Bayes Theorem). *Let* $\mathsf{E}_\mathsf{i}, \mathsf{i} = 1, 2, 3...\mathsf{N}$ *be a set of N events, each with positive probability, such that* E *includes all possibilities in a set* S *and the events are mutually exclusive. Then, for any event* B *defined on* S*, with* $\boldsymbol{Pr}(\mathsf{B}) > 0$*,*

$$\boldsymbol{Pr}(\mathsf{E}_\mathsf{j}|\mathsf{B}) = \frac{\boldsymbol{Pr}(\mathsf{B}|\mathsf{E}_\mathsf{j})\,\boldsymbol{Pr}(\mathsf{E}_\mathsf{j})}{\sum_{\mathsf{i}=1}^{\mathsf{N}} \boldsymbol{Pr}(\mathsf{B}|\mathsf{E}_\mathsf{i})\,\boldsymbol{Pr}(\mathsf{E}_\mathsf{i})} \tag{2.18}$$

Bayes Theorem may at first appear quite complicated, but in fact, we have already discussed all of the pieces that go into its derivation. We start with the conditional probability of an event E given that an event B has occurred:

$$\boldsymbol{Pr}(\mathsf{E}|\mathsf{B}) = \frac{\boldsymbol{Pr}(\mathsf{E} \cap \mathsf{B})}{\boldsymbol{Pr}(\mathsf{B})}. \tag{2.19}$$

This can be rearranged as

$$\boldsymbol{Pr}(\mathsf{E} \cap \mathsf{B}) = \boldsymbol{Pr}(\mathsf{E}|\mathsf{B})\,\boldsymbol{Pr}(\mathsf{B})\,. \tag{2.20}$$

If the E_i cover all possible outcomes, with a little thought one can see that the following must be true:

$$\boldsymbol{Pr}(\mathsf{B}) = \sum_{\mathsf{i}=1}^{\mathsf{N}} \boldsymbol{Pr}(\mathsf{B}|\mathsf{E}_\mathsf{i})\,\boldsymbol{Pr}(\mathsf{E}_\mathsf{i})\,. \tag{2.21}$$

Plugging (2.21) into the denominator of (2.19) gives us (2.18).

In Practice.

- In general, Bayes Theorem takes information about the $\boldsymbol{Pr}(A|\mathsf{B})$ and turns it into information about the $\boldsymbol{Pr}(\mathsf{B}|A)$.

Worked Example 2.3.
You recently started measuring daily precipitation in Argentina to study extreme precipitation events in the area. Past experience at the site indicates that 5% of the days exhibit what you consider dangerous amounts of precipitation (e.g. lead to landslides, crop damage, etc.).

You are testing a new rain gauge that measures daily precipitation totals and then logs it into a computer. Unfortunately, the particular gauge in question has some reliability problems. Your gauge indicates extreme precipitation on only 95% of the days that extreme downpours actually occur. Furthermore, your gauge also incorrectly indicates extreme precipitation on 10% of the days when the actual precipitation was below what you consider extreme.

What is the probability that a day for which the gauge indicated extreme precipitation did not have extreme precipitation?

.....

If we let E denote the event of extreme precipitation, and M denote the event where the gauge flags extreme precipitation, then we want to know $\boldsymbol{Pr}\left(\widetilde{\mathsf{E}}|\mathsf{M}\right)$. In this case, Bayes Theorem takes the form of

$$\boldsymbol{Pr}\left(\widetilde{\mathsf{E}}|\mathsf{M}\right) = \frac{\boldsymbol{Pr}\left(\mathsf{M}|\widetilde{\mathsf{E}}\right)\boldsymbol{Pr}\left(\widetilde{\mathsf{E}}\right)}{\boldsymbol{Pr}\left(\mathsf{M}|\widetilde{\mathsf{E}}\right)\boldsymbol{Pr}\left(\widetilde{\mathsf{E}}\right) + \boldsymbol{Pr}(\mathsf{M}|\mathsf{E})\,\boldsymbol{Pr}(\mathsf{E})} \tag{2.22}$$

$$= \frac{0.1 \cdot 0.95}{0.1 \cdot 0.95 + 0.95 \cdot 0.05} \approx 0.67 \tag{2.23}$$

Thus, a Bayesian would conclude that there is a 67% chance that the gauge is wrong and that extreme precipitation did not occur.

2.2.3 Probability philosophies: frequentist vs Bayesian views

While there are a wide range of philosophies on the meaning of probability, two general philosophies are discussed most frequently: the frequentist viewpoint, and the Bayesian viewpoint.

A *frequentist* approach takes the following form: If you have some large number of opportunities for an event to occur, then the number of times that event actually occurs, divided by the number of opportunities for it to occur is the probability. The probability varies between zero and one. The frequentist view has a solid foundation in the *Weak Law of Large Numbers* which states that if you have an event E that occurs N_E times in N trials, then $\mathsf{N}_\mathsf{E}/\mathsf{N}$ converges to the probability of event E occurring as the number of trials goes to infinity.

An alternative philosophy is attributed to Rev. Thomas Bayes (1701-1761), who figured that in many cases one is unlikely to have a large enough sample with which to measure the frequency of occurrence, and so, one must take a more liberal view. *Bayesian* inference is given that name for its frequent use of Bayes Theorem, which it uses to take into account *a priori* information, that may not be derivable from a frequentist point of view.

While the frequentist viewpoint is often found in the scientific literature in association with hypothesis testing and p-values, many recent articles have come out arguing against this approach due its broad misuse and the prevalence of "p-hacking" (e.g. Nuzzo, 2014; Goodman, 2001). Both the frequentist and the Bayesian approaches can be valid and useful, if done carefully and objectively. Bayesian analysis can be useful if you only have a small sample and you have prior information that you feel is reliable. New data can then be added to improve the estimate of probabilities. A weakness might be that this prior information could be subjective, and the methods of Bayesian analysis are a bit more complex. The Frequentist approach is simple to apply and works well if a large amount of data is available. Which approach to choose may depend on

the problem at hand. In all cases one must be alert to the possibilities of errors in the logic or application of statistical tests.

Worked Example 2.4.
Sometimes, the frequentist approach and the Bayesian approach can result in different conclusions, as demonstrated by returning to our previous example of the faulty rain gauge.

Frequentist Approach: A frequentist would conclude that the probability that extreme precipitation did not occur is 10%, since this is the probability that the gauge incorrectly flags extreme precipitation when none actually occurred.

Bayesian Approach: The Bayesian approach would take into account the background rate of extreme precipitation and plug everything into Bayes Theorem. Taking this Bayesian approach, we previously concluded that there is a 67% chance that the gauge is wrong and that extreme precipitation did not occur.

The reason the two approaches result in such wildly different answers is that the Bayesian approach took into account information that the frequentist approach did not. Namely, the frequency with which extreme precipitation actually occurs.

2.3 Probability Distributions

The probability that a randomly selected value of a random variable x falls between the limits a and b is

$$\boldsymbol{Pr}(a \leqslant x \leqslant b) = \int_a^b f(x)dx \tag{2.24}$$

This expression defines the *probability density function (PDF)*, $f(x)$, in the continuous case. Note that the probability that x is exactly equal to some value c is exactly zero.

To be a probability density function, $f(x)$ must satisfy the following criteria:

$$\int_{-\infty}^{\infty} f(x)dx = 1 \tag{2.25}$$

$$f(x) \geqslant 0 \text{ for all } x \tag{2.26}$$

The moments about the mean of the distribution can be obtained directly from the probability density function using the following formula,

$$m_r = \int_{-\infty}^{\infty} (x-\mu)^r f(x)dx, \tag{2.27}$$

where μ is the true, population mean.

The *cumulative distribution function (CDF)*, $F(x)$, is defined as the probability that a random variable assumes a value less than x,

$$F(x) = \int_{-\infty}^{x} f(t)dt. \tag{2.28}$$

The probability density function and the cumulative density function are linked via the fundamental theorem of calculus and it is straightforward to show that

$$\frac{dF}{dx} = f(x), \tag{2.29}$$

and

$$\boldsymbol{Pr}(a \leqslant x \leqslant b) = \int_a^b f(x)dx = F(b) - F(a). \tag{2.30}$$

In Practice.

- The probability density function and cumulative density function of a finite data set can be approximated by the smoothed histogram of the data. One common method for smoothing is called the *kernel density estimation*, an example of which is given in **Fig. 2.4**.

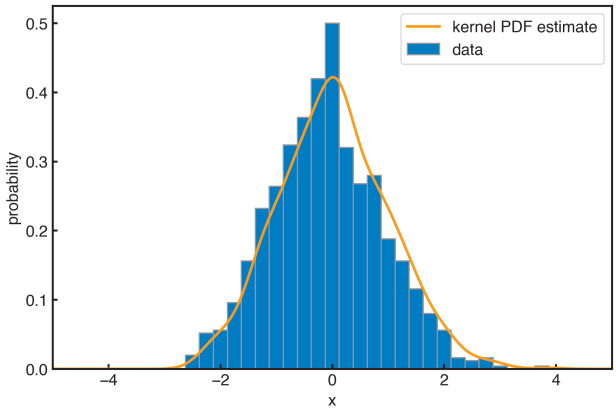

Figure 2.4 Histogram of a data set (blue) along with the kernel estimated probability density function (orange).

2.3.1 The Normal Distribution

The Normal (Gaussian) distribution is one of the most important in nature. Most observables are distributed normally about their means, or can be transformed in such a way that they become normally distributed. Because of this tendency for things to be normally distributed, the most common statistical tests assume normality. Thus, it is very important to verify that your random variable of interest is normally distributed before using common Gaussian statistics.

The probability density function for a normally distributed random variable x is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{2.31}$$

The associated cumulative distribution function is

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(t-\mu)^2}{2\sigma^2}} dt \tag{2.32}$$

It is often useful to write the Normal distribution functions in terms of *standardized* random variables, that is, a random variable with mean of 0 and standard deviation of 1. Letting z denote such a standardized random variable,

$$z = \frac{x - \mu}{\sigma}. \tag{2.33}$$

The probability density function and cumulative density function for a Normally distributed, standardized random variable z then simplifies to

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} \tag{2.34}$$

$$F(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}} dt \tag{2.35}$$

The probability that a standardized, normally distributed random variable z falls within ± 1, ± 2 and ± 3 standard deviations of its mean is given by

$$\boldsymbol{Pr}(-1 \leqslant z \leqslant 1) = \int_{-1}^{1} f(z)dz = 68.27\% \tag{2.36}$$

$$\boldsymbol{Pr}(-2 \leqslant z \leqslant 2) = \int_{-2}^{2} f(z)dz = 95.45\% \tag{2.37}$$

$$\boldsymbol{Pr}(-3 \leqslant z \leqslant 3) = \int_{-3}^{3} f(z)dz = 99.73\% \tag{2.38}$$

These probabilities can also be visualized as the area under the Gaussian $f(x)$ curve, as shown in **Fig. 2.5**. There is only a 4.55% probability that a normally distributed variable will fall more than 2 standard deviations away from its mean. This is a *two-tailed* probability. The probability that a normal variable will exceed its mean by more than 2 standard deviations is only half of that, 2.275%, since the normal distribution is symmetric. This is a *one-tailed* probability.

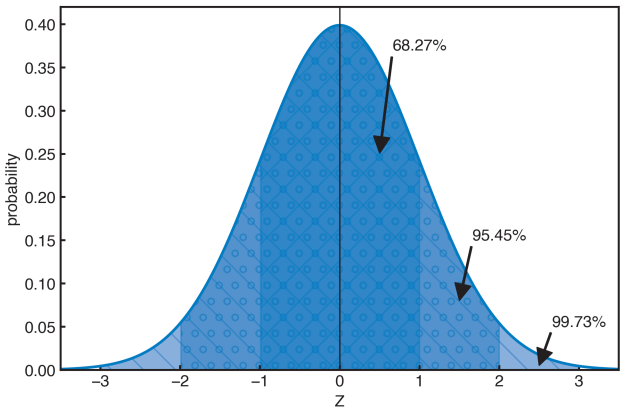

Figure 2.5 Probability density function of z and the probability that z falls within $\pm 1\sigma$, $\pm 2\sigma$ and $\pm 3\sigma$ (area under the curve).

In Practice.

- Standardizing your data using (2.33) comes in handy for comparing particular values with others who may have normally distributed data with different means and standard deviations, or those unfamiliar with the units of your data. For example, if I measured the ozone at a remote site and told you the measurements were normally distributed and today's level was 100 parts per billion (ppb), you may not know what to think. But, if I told you the standardized level was $z = 4\sigma$, you would know that ozone was extremely high today.
- Your data does not need to be Normally distributed to standardize it following (2.33), it is merely a unit conversion, like going from Celsius to Fahrenheit. When this is done, the resulting values can still be interpreted as the number of standard deviations about the sample mean. If the data is not normally distributed, however, the probabilities given in (2.36)-(2.38) and **Fig. 2.5** will not be applicable.

2.3.2 Other Common Distributions

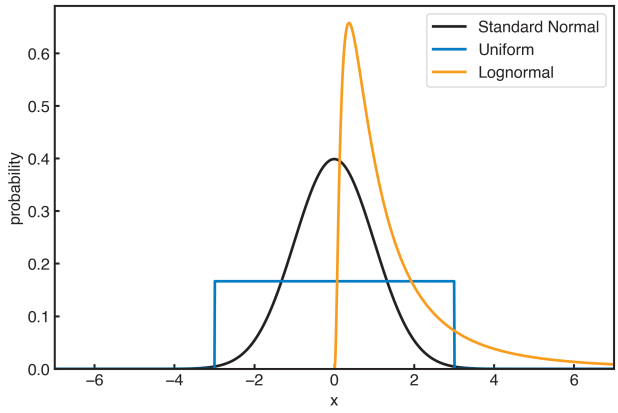

Figure 2.6 The probability density functions of three well-known distributions.

Uniform Distribution

The *uniform distribution* describes a random variable that is equally likely to take any value in the closed interval $[a, b]$. Its probability density function is plotted in **Fig. 2.6** and given by

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leqslant x \leqslant b, \\ 0 & \text{for } x < a \text{ or } x > b \end{cases} \tag{2.39}$$

The cumulative distribution function is

$$F(x) = \begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } a \leqslant x \leqslant b, \\ 1 & \text{for } x > b \end{cases} \tag{2.40}$$

Lognormal Distribution

A positive random variable x has a *lognormal distribution* if the natural logarithm ($\log$) of x is normally distributed. Put another way, x is lognormally distributed if $Y = \log x$ is normal. To determine the probability density function, it is straight-forward to plug $\log x$ into (2.31), perform a change of variables, and show that the lognormal probability density function is

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{\frac{-(\log x - \mu)^2}{2\sigma^2}}, \quad x > 0 \tag{2.41}$$

The cumulative density function is more complicated and requires the complementary error function to be written in full and so we will not do so here. An example lognormal probability density function is shown in **Fig. 2.6**.

Gamma Distribution

The Gamma Distribution is a two-parameter family of distributions. It is included here since it can fit positive-definite highly skewed distributions such as that of precipitation or wind speed. The two parameters are a shape factor, $a > 0$, and a scale factor, $b > 0$. The pdf for the Gamma Distribution is given by

$$f(x) = \frac{1}{\Gamma(a) b^a} x^{(a-1)} e^{-x/b} \tag{2.42}$$

The probability density functions for the gamma distribution with six sets of parameters are shown in **Fig. 2.7**. A small shape and large scale factor give a highly skewed distribution peaking near zero, which can be a good fit to variables like precipitation. A large shape and small scale parameter gives a distribution that is peaked at a non-zero value and less positively skewed.

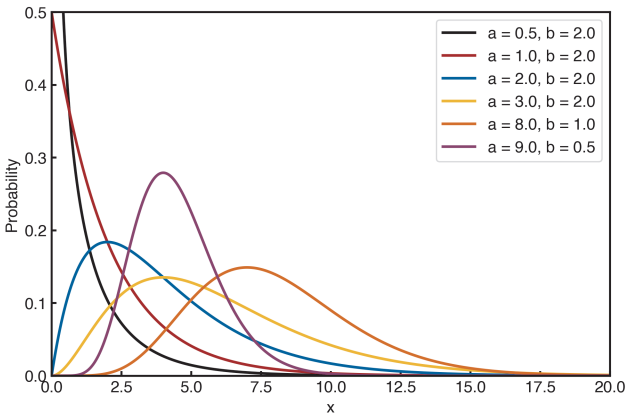

Figure 2.7 The probability density functions for gamma distributions with the shape and scale factors indicated in the legend.

2.4 Central Limit Theorem

Theorem 2.2 (Central Limit Theorem). *The arithmetic mean of a sufficiently large number of iterates of independent random variables, each with a well-defined mean and variance, will be approximately normally distributed, with standard deviation* $\sigma/\sqrt{N}$*, where* N *is the size of each sample.*

In simpler terms, the Central Limit Theorem says that no matter the underlying distribution of your data, if you take a large enough sample of your data, and compute its average, then take another sample and take its average, then another, etc., the distribution of these sample means will be normal with mean equal to the mean of the random variable and standard deviation of $\sigma/\sqrt{N}$, where σ is the standard deviation of the underlying distribution of the random variable. This concept is absolutely fundamental to much of the statistics that we do in the physical sciences, and significantly simplifies the statistics we must master. Let's look at two examples.

Worked Example 2.5.
Suppose we know that our data is normally distributed with $\mu = 0$ and $\sigma = 1$ (standard normal distribution). What is the distribution of sample means for sample sizes of $N = 25$? $N = 100$? $N = 200$?

.....

The Central Limit Theorem says that for a "sufficiently large number", that is, for sufficiently large N, the distribution will be normal. But what is "sufficiently large"? It turns out that if the underlying data is normal, the Central Limit Theorem applies for any $N \geqslant 1$. Thus, the sample mean will have a normal distribution with the same mean as the underlying distribution,$\mu = 0$, and standard deviation:

$$\sigma_{N=25} = \sigma/\sqrt{N} = 1/\sqrt{25} = 0.2 \tag{2.43}$$

$$\sigma_{N=100} = \sigma/\sqrt{N} = 1/\sqrt{100} = 0.1 \tag{2.44}$$

$$\sigma_{N=200} = \sigma/\sqrt{N} = 1/\sqrt{200} = 0.07 \tag{2.45}$$

To visualize this result, **Fig. 2.8** displays the distribution of 10000 sample means of values drawn from a standard normal distribution. The dashed gray curves denote the theoretical distribution given in (2.43) - (2.45).

Worked Example 2.6.
In the previous example, the underlying distribution was normal. However, the Central Limit Theorem applies to *all underlying distributions* as long as N is large enough.

Fig. 2.9 shows the distributions of 10000 sample means of length $N = 25, 100, 200$ drawn from the three distributions plotted in **Fig. 2.6**. As in **Fig. 2.8**, the dashed gray curves denote the theoretical normal distribution predicted by the Central Limit Theorem. As N increases, the theoretical estimate and the actual distribution agree more and more. Note how the lognormal distribution of sample means still does not agree completely with the theoretical estimate, and this is also the distribution that is most skewed (looks the least like a Gaussian).

2.5 Testing for Significance

Many geophysical variables are approximately normally distributed, furthermore, as we discussed in Section 2.4, if you take a large enough sample, the sample mean of *any* variable is normally distributed. Thus, we can often use the theoretical normal probability distribution to calculate the probability of measuring a certain value. We have so far covered how to determine the probability of drawing a value x_i within a range of values, but what about comparing a sample's mean to some other value? For example, instead of asking "what is the probability that this summer's average temperature will be greater than 80°F", we might instead want to ask "was this summer's average temperature significantly warmer than that of the summer of 1950?" As in this example, many research questions revolve around determining whether two means are different from

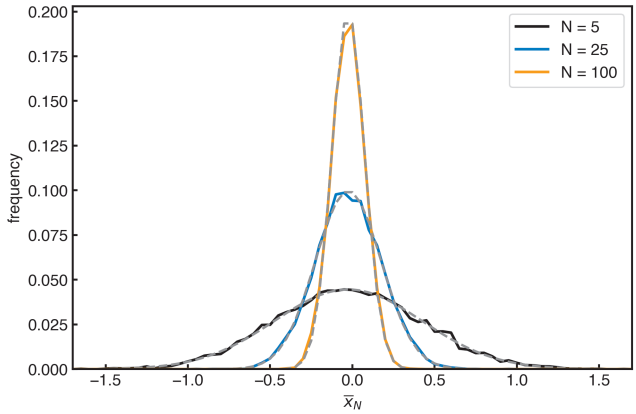

Figure 2.8 Distribution of 10000 sample means drawn from a standard normal distribution for sample sizes of $N = 25, 100, 200$. Dashed gray lines denote the distributions predicted by theory.

one another. To do this we need to know our data's true population mean and population standard deviation *a priori*. Unfortunately, the best that we are likely to have are the sample mean $\bar{x}$ and the sample standard deviation s based on a sample of finite length N.

If we know our data are normally distributed, and N is large enough, then we can use $\bar{x}$ and s to compute the z-statistic. If N is not sufficiently large, we need to use the Student-t distribution (see Section 2.5.1), which is appropriate for small sample sizes.

The standard variable used to compare a sample mean to the true mean is:

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{N}}} \tag{2.46}$$

where we have used the Central Limit Theorem to replace $\sigma_{\bar{x}}$ with $\sigma/\sqrt{N}$. The z-statistic is thus the number of standard errors that the sample mean deviates from the true mean. If the variable is normally distributed about its mean, then z can be converted into a probability statement.

(2.46) needs to be altered only slightly to provide a significance test for differences between two sample means:

$$z = \frac{(\bar{x_1} - \bar{x_2}) - \Delta_{1,2}}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}} \tag{2.47}$$

Here, the sample sizes are allowed to be different, and $\Delta_{1,2}$ is the hypothesized difference between the two means, which is often zero in practice.

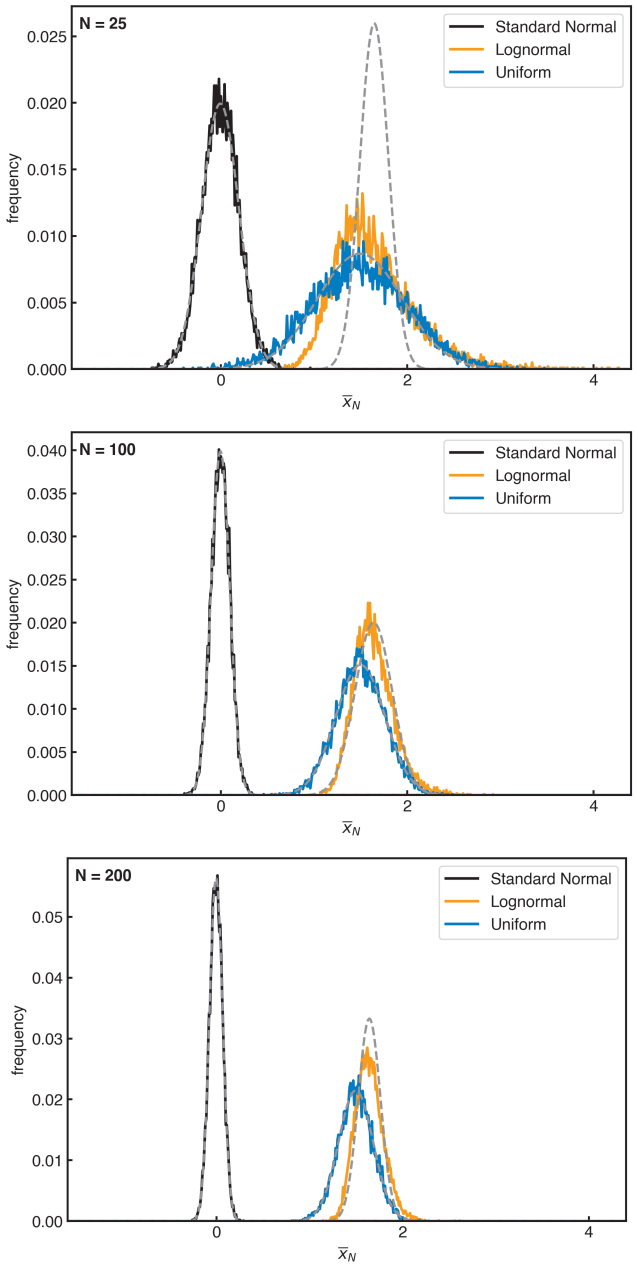

Figure 2.9 Distribution of 10000 sample means drawn from a three distributions for sample sizes of $\mathrm{N} = 25, 100, 200$.

2.5.1 Small sampling theory: the t*-statistic*

When the sample size is smaller than about 30 we cannot use the z-statistic to compare sample means, even if the underlying distribution is normally distributed. Instead, we must use the Student's t distribution to compare sample means, or the chi-squared distribution when comparing sample variances. The key difference between the z-statistic and the t-statistic is that the z-statistic requires knowledge of the population standard deviation σ while the t-statistic uses the sample standard deviation s. When the sample size is smaller than 30, s is biased low as an estimate of σ and thus, we use the t-statistic to account for this.

The Student's t-distribution is derived in exact analogy with the z-statistic:

$$t = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{N-1}}} = \frac{\overline{x} - \mu}{\frac{\hat{s}}{\sqrt{N}}} \tag{2.48}$$

$$\hat{s} = s\sqrt{\frac{N}{N-1}} \tag{2.49}$$

If we draw a sample of size N of independent values from a normally distributed population with mean μ, t (as defined by (2.48)) is distributed with the following probability density:

$$f(t) = \frac{f_0(\nu)}{\left(1 + \frac{t^2}{\nu}\right)^{\left(\frac{\nu+1}{2}\right)}}, \tag{2.50}$$

where $f_0(\nu)$ is chosen as a normalization factor to make $\int_{-\infty}^{\infty} f(t)dt = 1$ and $\nu = N - 1$ is the *number of degrees of freedom*. The degrees of freedom is defined as the number of independent samples minus the number of parameters that must be estimated.

In Practice.

- In all cases thus far, it has been assumed that the N values drawn are all *independent* samples. Often, however, N samples of a geophysical variable are not independent, that is, they exhibit either spatial or temporal correlations. For example, geopotential height is highly auto-correlated so that each day's value is not independent from the previous day's. We will discuss how to deal with non-independence in Section 7.4 autocorrelation and degrees of freedom.

Worked Example 2.7.
The Southern Annular Mode (SAM) is the dominant mode of atmospheric variability in the Southern Hemisphere, and can be quantified by a monthly index which is approximately normally distributed with $\mu = 0$ and $\sigma = 1$.

(a) What is the probability that a particular month's SAM index is $\geqslant 0.5$?

(b) What is the probability that the average monthly SAM index over a 4 year period was $\geqslant 0.5$?

.....

(a) We are given that, $\mu = 0$ and $\sigma = 1$, and so we can calculate a z-score:

$$z = \frac{\overline{x} - \mu}{\sigma} = \frac{0.5 - 0}{1} = 0.5. \tag{2.51}$$

We want to know $\boldsymbol{Pr}(z > 0.5)$ (i.e. the area under the normal probability density curve that is to the right of 0.5) which can be computed using any software package, and the answer is 31%.

(b) Now, we want to test for the sample mean, with $N = 36$ months. In this case,

$$\overline{x} = 0.5, \sigma_{\overline{x}} = \frac{\sigma}{\sqrt{N}} = \frac{1}{\sqrt{36}} \tag{2.52}$$

$$z = \frac{\overline{x} - \mu_{\overline{X}}}{\sigma_{\overline{x}}} = \frac{0.50 - 0}{.1667} = 3.0 \tag{2.53}$$

The $\boldsymbol{Pr}(z \geqslant 3.0) = 0.1\%$. Such a low probability implies either a very rare event, or, that the dynamics of the SAM over those 4 years was different compared to the climatological SAM variability. Note that in this example we have assumed that each monthly sample is independent, so that the degrees of freedom of the data set equals the number of samples.

Unlike the z-distribution, the t-distribution depends on the size of the sample. The tails of the distribution are longer for smaller degrees of freedom (**Fig. 2.10**). For a large number of degrees of freedom the t-distribution approaches the z or normal distribution. Note that, although we sometimes speak of the t-distribution and contrast it with the normal distribution, the t-distribution is merely the probability density you expect to get when you take a small sample *from a normally distributed population.*

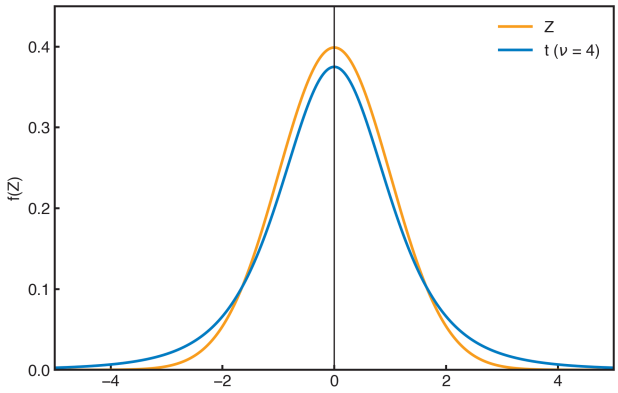

Figure 2.10 Probability density function of z— and t-distribution with $\nu = 4$ degrees of freedom.

In Practice.

- When using the t-statistic, you are making the strong assumption that the underlying distribution is normal. The Central Limit Theorem tells us that for a "large enough" sample size, the distribution of sample means is normal, no matter the distribution. For small sample sizes, the Central Limit Theorem does not apply. Thus, if the underlying population is not normally distributed, and you have a small sample size, you must use other methods.
- Smaller values of N lead to longer tails for the t-statistic, meaning you are more likely to get a sample mean far from the true value when N is smaller.
- Since the t-distribution approaches the normal distribution for large N, there is no theoretical reason to use the z-statistic in preference to the t-statistic, although it maybe be more convenient to do so.

The difference of means for the t-statistic is very similar to that for the z-statistic, but with slight modifications. Assume two samples of length N_1 and N_2 are drawn from an normally distributed population with true standard deviations $\sigma_1 = \sigma_2$, then,

$$t = \frac{(\overline{x_1} - \overline{x_2}) - \Delta_{1,2}}{\widehat{\sigma}\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}} \tag{2.54}$$

$$\widehat{\sigma} = \sqrt{\frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2 - 2}} \tag{2.55}$$

where $\nu = N_1 + N_2 - 2$ and $\Delta_{1,2}$ is the hypothesized difference. The pooled variance $\widehat{\sigma^2}$ is a weighted average of the sample variances.

2.5.2 Confidence intervals

Recall from our discussion of cumulative probability density function F, that

$$\boldsymbol{Pr}(a \leqslant x \leqslant b) = \int_a^b f(x)dx \tag{2.56}$$

$$\boldsymbol{Pr}(a \leqslant x \leqslant b) = F(b) - F(a). \tag{2.57}$$

For a standard normal random variable z, we determined that

$$\boldsymbol{Pr}(-1 \leqslant z \leqslant 1) = 68.27\% \tag{2.58}$$

$$\boldsymbol{Pr}(-2 \leqslant z \leqslant 2) = 95.45\% \tag{2.59}$$

These are confidence intervals for z. The first is the 68.27% confidence interval, and the second is the 95.45% interval.

One can instead first determine a confidence interval of interest, say 95%, and compute the lower-bound a and upper-bound b such that $\boldsymbol{Pr}(a \leqslant z \leqslant b) = 95\%$. Often, the confidence interval is discussed in terms of the parameter α, which is defined as 1 minus the confidence interval. For a 95% confidence interval, $\alpha = 0.05$.

To find the 95% confidence interval for z, or $\alpha = 0.05$, it helps to think in terms of the area under the standard normal probability distribution function (**Fig. 2.11**). That is, we want to find the critical z, denoted z_c, such that

$$\boldsymbol{Pr}\left(z_{c,\alpha/2} \leqslant z \leqslant z_{c,1-\alpha/2}\right) = 0.95. \tag{2.60}$$

Since the normal distribution is symmetric about zero, we can instead write

$$\boldsymbol{Pr}(z) \geqslant z_{c,1-\alpha/2} = 1 - \frac{\alpha}{2} = 0.975. \tag{2.61}$$

We look for 0.975 because we want the total area to add to 5% ($\alpha = 0.05$), and so 2.5% comes from the lower tail and 2.5% comes from the upper tail. Any statistical software can be used to find that for a 95% confidence interval of a normally distributed variable, $z_{c,0.975} = 1.96$.

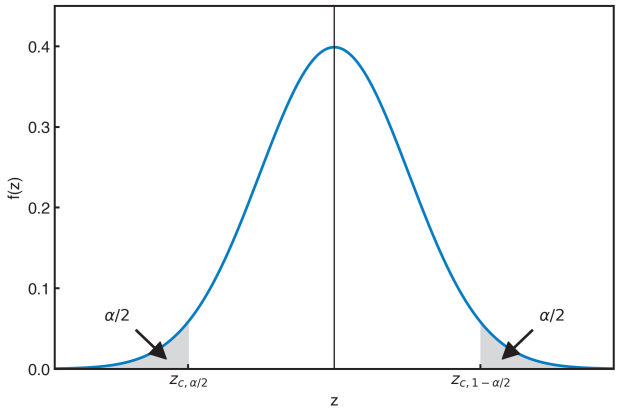

Figure 2.11 Illustration of the relation of the z-statistic probability density function to probability measure α.

The above examples with standardized data, are a relatively straight-forward application of the t-statistic and z-statistic because we are dealing with standardized data. However, what if your data are not standardized? You have two options: you can standardize your data and then do all of your analysis using standard normal variables (as above), or, you can use a modified equation for the confidence interval that takes into consideration the the data's non-zero mean and non-unity standard deviation as we will now demonstrate.

Plugging the definition of the z-statistic (2.33) into (2.60) leads to the 95% confidence interval for any sampled guassian variable x:

$$-z_{c,0.975} \leqslant \frac{x - \mu}{\sigma} \leqslant z_{c,0.975}. \tag{2.62}$$

Following similar steps for the sample mean,

$$-z_{c,0.975} \leqslant \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{N}}} \leqslant z_{c,0.975}. \tag{2.63}$$

From this we can deduce that the true mean μ falls within the following bounds 95% of the time:

$$\overline{x} - z_{c,0.975}\frac{\sigma}{\sqrt{N}} \leqslant \mu \leqslant \overline{x} + z_{c,0.975}\frac{\sigma}{\sqrt{N}}. \tag{2.64}$$

In general, confidence limits for population means of symmetric distributions can be represented by

$$\mu = \overline{x} \pm z_{c,1-\alpha/2}\frac{\sigma}{\sqrt{N}} \tag{2.65}$$

Confidence intervals for the sample mean t-statistic are defined similarly,

$$\mu = \overline{x} \pm t_{c,0.975}\frac{s}{\sqrt{N-1}}, \tag{2.66}$$

where t_c is the critical value for t and depends on the significance level desired and the sample size.

Worked Example 2.8.
You have 5 years of monthly-mean temperature data from the MSU4 satellite. The mean temperature around the 60°N latitude circle during January is −60° C and the standard deviation is 8° C. What are the 95% confidence limits on the true population mean? You can assume that monthly-mean temperatures are normally distributed.

.....

t-statistic
Since $N = 5$, we must use the t-statistic. The critical value is $t_{c,0.975} = \pm 2.78$ for $\nu = 5 - 1 = 4$. Thus, the population mean μ is expected to lie within

$$-60 \pm 2.78 \frac{8}{\sqrt{4}} \Rightarrow -67.0 \leqslant \mu \leqslant -53.0 \tag{2.67}$$

z-statistic
If we had erroneously used the z-statistic, the critical value is $z_{c,0.975} = \pm 1.96$ and the population mean μ would be expected to lie within

$$-60 \pm 1.96 \frac{8}{\sqrt{5}} \Rightarrow -71.1 \leqslant \mu \leqslant -48.9 \tag{2.68}$$

Using the t-statistic gives a wider confidence interval than the z-statistic, reflecting the additional uncertainty associated with small N. If we had erroneously used the z-statistic instead of the t-statistic we would have underestimated the 95% confidence bounds by 35%.

2.5.3 Chi-Squared Distribution: Tests of Variance

Sometimes we want to test if the sample variances are truly different. For this we cannot use t-statistic or z-statistic as these are for sample means, but we can use the Chi-Squared distribution. First, define a random variable χ^2:

$$\chi^2 = (N - 1) \frac{s^2}{\sigma^2} \tag{2.69}$$

This quantity can be used to test if a sample variance s^2 is different from a population variance σ^2. Note we are using a ratio, rather than a difference.

If the underlying distribution from which we draw N values to compute χ^2 is normally distributed with standard deviation σ, then the χ^2 values themselves will be distributed as follows:

$$f(\chi^2) = f_0(\nu)(\chi^2)^{(\frac{1}{2}\nu - 1)} \exp^{-\frac{1}{2}\chi^2}, \quad \nu = N - 1 \tag{2.70}$$

where f_0 is a normalization factor. This is the *Chi-Squared distribution* and can be used to estimate the significance of the ratio $\frac{s^2}{\sigma^2}$.

If you wish to determine confidence bounds on the true variance, you can move things around to obtain the confidence limits given your sample variance:

$$\frac{s^2(N-1)}{\chi^2_{c,0.975}} \leqslant \sigma^2 \leqslant \frac{s^2(N-1)}{\chi^2_{c,0.025}}, \quad \nu = N - 1. \tag{2.71}$$

Note that the Chi-squared distribution is not symmetric like the normal distribution, and so the lower and upper critical values $\chi^2_{c,0.025}$ and $\chi^2_{c,0.975}$ must both be computed and, like the t-distribution, are functions of the sample size N.

In Practice.

- For $\nu \gtrsim 30$, the Chi-Squared distribution approaches the Normal distribution.

2.6 The Binomial Distribution

2.6.1 Binomial Distribution

Suppose you have a set of N trials in which the outcome is either "success" or "failure". The probability of success in one trial is $p = \boldsymbol{Pr}$(success in one trial). If X is the total number of successes in N trials, then

$$\boldsymbol{Pr}(X = k) = \binom{N}{k} p^k (1-p)^{N-k} = \frac{N!}{(N-k)!\,k!}\, p^k (1-p)^{N-k}, \quad k = 1, 2, 3...N. \tag{2.72}$$

At first, the right-hand-side might look complicated, but note that it is just the probability of k successes times the probability of the rest being failures with an additional factor in front to account for the order of occurrence not mattering.

The binomial distribution is helpful in assessing "field significance", the significance of multiple tests succeeding when an array of variables are tested against the same hypothesis. An example would be correlating the sunspot index with a map of pressure at many points over the earth. How many individual "significant" events do you expect to get by chance in such cases?

As an example, **Fig. 2.12** shows for N tries of a test at the $p = 0.05$ significance level what the binomial distribution (2.72) says about how many you should get by chance alone.

Note that the probability of getting 5 successes or more in 30 tries is less than 0.05 and getting 10 successes or more in 100 tries is less than 0.05. That is 16.7% are successes for 30 tries and only 10% are successes for 100 tries at same probability level. For smaller samples, the fraction of total tries that can succeed by chance is greater. Even for 100 tries, 10% can succeed by chance, where the probability of each individual occurrence is p=0.05. The most likely outcome is shown by the peak of the blue line and is what you expect, about 5% of the chances will succeed. But the chances of getting significantly more than that are quite good, and 10-15% of the field points could succeed by chance at the 5% level (see also Wilks, 2011; Livezey and Chen, 1983; Wilks, 2016).

Worked Example 2.9.
Suppose 14 out of 20 different climate models project that Australia will become drier with increasing greenhouse gas concentrations. What is the probability of getting this result if one assumes that drying and wetting are actually both equally likely under this scenario? That is $\boldsymbol{Pr}(\text{drying}) = \boldsymbol{Pr}(\text{wetting}) = 0.5$?

.....

$$\boldsymbol{Pr}(X = 14) = \binom{20}{14} 0.5^{14}(1-0.5)^{20-14} = 0.037 \tag{2.73}$$

What is the probability that 14 *or more* models agree that Australia will become drier if we assume that drying and wetting are both equally likely?

$$\boldsymbol{Pr}(X \geqslant 14) = \sum_{k=14}^{20} \binom{20}{k} 0.5^k (1-0.5)^{20-k} = 0.058 \tag{2.74}$$

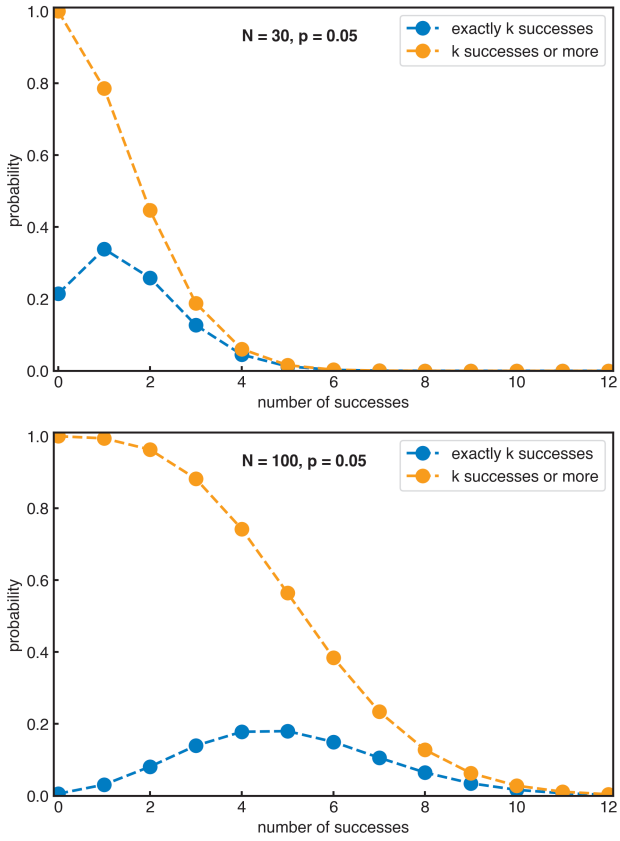

Figure 2.12 Probability of a given number of successes in N trials where the probability of a success is $p = 0.05$.

2.6.2 Normal Approximation to the Binomial

If you did the calculations above by hand you would find it tedious. This gets worse when the sample gets even larger. To assist in this, we can make use of theorem that allows us to use a Normal approximation when performing Binomial calculations.

From the central limit theorem, it can be shown that the distribution of sample means approaches the Normal Distribution, even if the population from which the means are derived is not normally distributed (see Section 2.4). This is also true for the Binomial distribution, for which values have a probability of being either zero or one, but nothing else. The distribution of sample means from a binomial population is nonetheless normally distributed about its mean value of 0.5.

Theorem 2.3 (DeMoivre-Laplace Theorem). *Let X denote a binomial variable defined on N independent trials, each having success probability p. Then, for any numbers a and b,*

$$\lim_{N\to\infty} \boldsymbol{Pr}\left(a < \frac{X - Np}{\sqrt{Np(1-p)}} < b\right) = \frac{1}{\sqrt{2\pi Np(1-p)}} \int_a^b e^{-x^2/2} dx \tag{2.75}$$

This theorem tells us that the statistic $z = \frac{X-Np}{\sqrt{Np(1-p)}}$ follows a normal distribution with $\mu = Np$ and $\sigma = \sqrt{Np(1-p)}$. An approximate two-tailed 95% confidence interval for the number of successes X is then given by

$$Np - 1.96 \cdot \sqrt{Np(1-p)} \leqslant X \leqslant Np_1.96 \cdot \sqrt{Np(1-p)} \tag{2.76}$$

We can use this to simplify the calculation of binomial problems, as illustrated in the examples below.

In Practice.

- When deciding whether a Normal approximation is appropriate to use for your Binomial random variable, some good rules-of-thumb are:
  - large N
  - $Np \geqslant 10$
  - $N(1-p) \geqslant 10$

Worked Example 2.10.
An earthquake forecaster has to forecast 200 earthquakes. How many times in 200 tries must she be successful so we can say with 95% confidence that she has non-zero skill?

.....

The null hypothesis is that she has no skill and the significance level is $\alpha = 0.05$. We then want

$$\boldsymbol{Pr}(s > s^*|H_0) = 0.025 = \sum_{s=s^*}^{200} \binom{200}{s} (0.5)^s (1-0.5)^{200-s} \tag{2.77}$$

Solving this equation for $s > s^*$, the number of occurrences necessary to leave only a 0.025 probability to the right, is extremely tedious to do by hand. Instead, we can use the Normal approximation to the Binomial to convert this to the following problem:

$$\boldsymbol{Pr}(s > s^*|H_0) = 0.025 = \boldsymbol{Pr}\left(\frac{s - Np}{\sqrt{Np(1-p)}} > \frac{s^* - Np}{\sqrt{Np(1-p)}} \right) \tag{2.78}$$

$$= \boldsymbol{Pr}\left(Z > \frac{s^* - Np}{\sqrt{Np(1-p}} \right) \tag{2.79}$$

where $\boldsymbol{Pr}(Z > 1.96) = 0.025$ from the standard normal distribution. So, we want

$$\frac{s^* - Np}{\sqrt{Np(1-p}} > 1.96, \quad \text{or} \quad s^* = 114 \tag{2.80}$$

So, to pass a no-skill test on a sample of this size, the forecaster must be right 57% of the time or more.

The 95% confidence interval for the number of successes expected if the forecaster has no skill (i.e. under the null hypothesis) is given by:

$$Np \pm 1.96 \cdot \sqrt{Np(1-0.5)} = \tag{2.81}$$

$$100 \pm 1.96 \cdot \sqrt{10 \cdot 0.5} = \tag{2.82}$$

$$100 \pm 13.86 \tag{2.83}$$

Worked Example 2.11.
Normal Approximation to Binomial: Out of 48 independent climate model simulations, how many must agree that global temperatures will increase by 2100 so that we can say with 95% certainty that the models do not agree purely by chance? What is the 95% confidence interval on the number of models with positive temperature trends under the null hypothesis?

Here, let a success be that the model says global temperatures will increase. Our null hypothesis is that the models randomly guess whether global temperatures will increase - thus, there is a 50% chance that any one model will predict a temperature increase ($p = 0.5$). We want to know k^* such that:

$$\boldsymbol{Pr}(X \geqslant k^* | H_0) \leqslant 0.05 \tag{2.84}$$

That is, k^* is the number of models that must show a temperature increase for us to believe it is more than chance (that the null hypothesis can be rejected).

$$\sum_{k=k^*}^{48} \binom{48}{k} (0.5)^k (1-0.5)^{48-k} \leqslant 0.05 \tag{2.85}$$

This would take a long time by hand, however, we can instead use the Normal approximation to the Binomial:

$$\boldsymbol{Pr}\left(Z > \frac{k^* - 48 \cdot 0.5}{\sqrt{48 \cdot 0.5 \cdot (1-0.5)}}\right) = 0.025 \tag{2.86}$$

$$\frac{k^* - 48 \cdot 0.5}{\sqrt{48 \cdot 0.5 \cdot (1-0.5)}} = 1.96 \tag{2.87}$$

$$k^* \geqslant 31. \tag{2.88}$$

So, at least 31 models must show increasing temperatures to reject the null hypothesis that the model agreement in a warming trend is due to random chance. As expected, more than half of the models must show an increase.

The 95% confidence interval under the null hypothesis is:

$$Np \pm 1.96 \cdot \sqrt{Np(1-p)} \tag{2.89}$$

$$24 \pm 1.96 \cdot \sqrt{24(1-.5)} = 24 \pm 7 \tag{2.90}$$

2.7 The Poisson Distribution

The Poisson distribution applies when you are counting the number of objects in a certain interval. The interval can be in space (volume, area or length) or time. You know the average number of counts per unit interval, and wish to know the chance of actually observing various numbers of objects or events. We denote the associated random variable N, since they are actual counts.

$$N \Rightarrow \text{Poisson}(\lambda) \tag{2.91}$$

There are three necessary and sufficient conditions for a Poisson Distribution.

1. Two or more events cannot occur simultaneously. This means that the events themselves occupy negligible space (e.g. volume, area, length, time).

2. Events occur at an average rate of λ (per unit e.g. volume, area, length, time). This means that λ cannot be a function of space or time.
3. Events occur independently (i.e. they do not know about each other)

The probability mass function of a Poission is defined by the probability that $N = n$ in a given interval of magnitude t according to:

$$\boldsymbol{Pr}(N = n) = \frac{(\lambda t)^n}{n!} e^{-\lambda t} \qquad \lambda > 0, t > 0, \text{ and } n = 0, 1, \ldots \tag{2.92}$$

The first and second moments (mean and variance) are given by:

$$\mu = \lambda t \qquad \sigma^2 = \lambda t \tag{2.93}$$

Estimating $\widehat{\lambda}$ from your data is quite straightforward. Let N be the number of observed events in time t, and assume that N is well-modelled as a Poisson Distribution with unknown rate parameter λ (where λ has units of "events per unit time"). The rather obvious formula for estimating the rate parameter is then simply the number of events divided by the time over which they were observed:

$$\widehat{\lambda} = \frac{N}{t} \tag{2.94}$$

The standard deviation (or standard error) of this estimator is

$$\sigma_{\widehat{\lambda}} = \sqrt{\frac{\lambda}{t}} \tag{2.95}$$

Putting these together, the approximate confidence interval for the true parameter λ (assuming the Central Limit Theorem applies, which it does for $N > 30$ or so) is given by

$$\boldsymbol{Pr}\left(\widehat{\lambda} - z_{\alpha/2}\sigma_{\widehat{\lambda}} \leqslant \lambda \leqslant \widehat{\lambda} + z_{\alpha/2}\sigma_{\widehat{\lambda}}\right) \tag{2.96}$$

Worked Example 2.12.
Poisson rate confidence interval: Let's say we count 137 events in 44 minutes. Our estimated rate parameter is

$$\widehat{\lambda} = \frac{N}{t} = \frac{137}{44} \approx 3.11 \text{ events per minute} \tag{2.97}$$

The approximate standard error for the estimated rate parameter is

$$\sigma_{\widehat{\lambda}} = \frac{\sqrt{N}}{t}\frac{\sqrt{137}}{44} \approx 0.27 \text{ events per minute} \tag{2.98}$$

and so the approximate 95% confidence interval for the true, but unknown, rate parameter is

$$\widehat{\lambda} - 1.96\sigma_{\widehat{\lambda}} \leqslant \lambda \leqslant \widehat{\lambda} + 1.96\sigma_{\widehat{\lambda}} \Rightarrow 2.58 \leqslant \lambda \leqslant 3.64 \tag{2.99}$$

It turns out that the Poisson Distribution has a close relationship with the Binomial Distribution. That is, for $n \to \infty$, $p \to 0$, with $np \to \lambda \neq 0$, the Binomial Distribution converges to the Poisson Distribution with parameter λ. In practice, the Binomial Distribution may be approximated by the Poisson when $p < 0.5$ and $n > 20$.

One might be interested in whether the Poisson rates in two samples are different or not. Suppose we have two rates λ_1 and λ_2 drawn from samples of size t_1 and t_2. Our null hypothesis is that $\lambda_1 - \lambda_2 = 0$. The pooled-rate test is based on a standard normal statistic defined as follows.

$$z = \frac{\lambda_1 - \lambda_2}{\sqrt{\widehat{\lambda}\left(\frac{1}{t_1} + \frac{1}{t_2}\right)}} \tag{2.100}$$

where

$$\widehat{\lambda} = \frac{t_1\lambda_1 + t_2\lambda_2}{t_1 + t_2} \tag{2.101}$$

Worked Example 2.13.
Poisson rate difference test: According to the ERA reanalysis data set, during the 28 years from 1952 to 1979 there were 14 major stratospheric warmings, and during the 42 years from 1980 to 2021 there were 25 warmings. The rates are thus 0.05 per year and 0.595 per year. Are these rates different at p=0.05?

$$\widehat{\lambda} = \frac{42 * 0.595 + 28 * 0.5}{42 + 28} = 0.557 \tag{2.102}$$

$$z = \frac{0.595 - 0.5}{\sqrt{0.557\left(\frac{1}{42} + \frac{1}{28}\right)}} = 0.52 \tag{2.103}$$

This is much less than the critical z value of 1.96 for a two-tailed test, so we cannot reject the null hypothesis that the rates of occurrence are the same for the two intervals.

2.8 Non-parametric Statistical Tests

The statistical tests applied above mostly assume that the samples come from populations for which the statistical distributions are known, or assumed, *a priori*. We very often assume that the statistics we are testing are Normally distributed, so we can use the shape of the Normal distribution in our tests. Tests have also been developed that do not require the assumption of a theoretical distribution. These are called *non-parametric* or *distribution-free* statistical tests.

2.8.1 Signs Test

Suppose we have paired data (x_i, y_i) and we want to know if the mean of x_i is different from the mean of y_i. By *paired data*, we mean that each x_i is uniquely associated with a y_i. If we have a suspicion that the data are not normally distributed, and we do not have enough data to invoke the Central Limit Theorem, we cannot use the t-test or the z-test. Instead, if we formulate our question in terms of the median ($\widetilde{\mu}$), rather than the mean, our null hypothesis is that the medians of x_i and y_i are the same, and the alternative is that they are not:

$$H_0 : \widetilde{\mu}_1 = \widetilde{\mu}_2 \quad H_1 : \widetilde{\mu}_1 \neq \widetilde{\mu}_2 \tag{2.104}$$

Let's reformulate this in terms of a probability that y_i is greater than x_i (noting that we could as easily formulate it as less than)

$$H_0 : \boldsymbol{Pr}(y_i > x_i) = 0.5 \quad H_1 : \boldsymbol{Pr}(y_i > x_i) \neq 0.5 \tag{2.105}$$

To test this null hypothesis, we can use the Signs Test. To perform this test, we simply replace each (x, y) pair with a signed integer equal to 1 according to the following rule:

$$y_i > x_i \rightarrow +1 \tag{2.106}$$
$$y_i < x_i \rightarrow -1 \tag{2.107}$$
$$\tag{2.108}$$

The null hypothesis would suggest that there will be a similar number of positive and negative ones (both are equally probable). With this setup we now have a set of Bernoulli trials with success (+1) and failure (−1). We know that the number of successes over N trials will be binomially distributed, and so we can use this distribution to determine whether our actual success rate is outside of what might be expected given our null hypothesis.

Worked Example 2.14.
Cloud Seeding Experiment: Ten pairs of very similar developing cumulus clouds were identified. One from each pair was seeded, and the other was not. Then the precipitation falling from the clouds later was measured with a radar. The data in the following table resulted:

Cloud Pair	x_i: Precip. (untreated)	y_i: Precip. (treated)	$y_i > x_i$?
1	10	12	+1
2	6	8	+1
3	48	10	−1
4	3	7	+1
5	5	6	+1
6	52	4	−1
7	12	14	+1
8	2	8	+1
9	17	29	+1
10	8	9	+1

Using the data above, we get 8 +1 and 2 −1. Are these results inconsistent with the null hypothesis that cloud seeding has no effect on precipitation? Can we confidently say that the median values of the two samples are different at 95% confidence? We can plug our values into the binomial distribution to determine the probability of getting 8 successes in 10 tried.

$$\boldsymbol{Pr}(k \geqslant 8) = \sum_{k=8}^{10} \binom{10}{k} 0.5^k (1-0.5)^{10-k} = 0.055 \tag{2.109}$$

If things are random (our null hypothesis is true), the chance of getting two or fewer successes is equally probable as getting 8 or more. Using a two-sided test we find that the probability our result is $p = 0.11$, which fails a 95% confidence test. We expect to toss 8 out of ten heads or tails about 11% of the time.

The Signs Test is one of the simplest non-parametric tests available, but it also has its limitations. For example, it will not tell you the magnitude of the difference of the medians. There are many other distribution-free tests that can be used, for example, the *Wilcoxon signed rank test* and the *Wilcoxon-Mann-Whitney test*.

2.8.2 Rank Sum Test

Another common and classical non-parametric test is the *Rank-Sum Test* (or Wilcoxon-Mann-Whitney Test). Suppose we have two samples S_1 and S_2 of sizes N_1 and N_2 and we wish to test the null hypothesis that

they both were sampled from the same distribution (whatever it is). The first step to the Rank-Sum Test is to combine them into a single sample $N = N_1 + N_2$ and rank them from smallest (rank $r = 1$) to largest rank $r = N$). Next, compute the sum of the ranks of each sample S_1 and S_2 and call these R_1 and R_2.

R_1/N_1 and R_2/N_2 should be similar if our null hypothesis is true and the two samples are from the same underlying distribution. Thinking a little harder, one can see that there are $N!/(N_1!N_2!)$ possible combinations of R_1 and R_2. Mann-Whitney showed that the U statistic could be used to determine the probability of a particular combination where

$$U_1 = R_1 - \frac{N_1}{2}(N_1 + 1) \tag{2.110}$$

$$U_2 = R_2 - \frac{N_2}{2}(N_2 + 1) \tag{2.111}$$

where

$$U_1 + U_2 = \frac{N_1 N_2}{2}. \tag{2.112}$$

The U statistic is approximately Normally distributed with mean and standard deviation

$$\mu = \frac{N_1 N_2}{2} \tag{2.113}$$

$$\sigma = \left(\frac{N_1 N_2 (N_1 + N_2 + 1)}{12} \right)^{1/2} \tag{2.114}$$

The statistical significance of U can then be tested with the standard z-score.

2.8.3 Runs Test (Wald-Wolfowitz Test)

The Runs Test is a non-parametric test to check whether a list of values is random or not. For example, imagine a time series of anomalies as shown below, where "+" denotes a positive anomaly and "-" denotes a negative anomaly:

$$\underbrace{++++}_{\text{Run 1}}\underbrace{---}_{\text{Run 2}}\underbrace{+++}_{\text{Run 3}}\underbrace{--}_{\text{Run 4}}\underbrace{+++++}_{\text{Run 5}}\underbrace{++}\underbrace{----}_{\text{Run 6}} \tag{2.115}$$

We now separate this series into *runs*; there are a total of $R = 6$ runs, three of which are runs of "+" and three of which are runs of "-". The Runs Test tests the null hypothesis that the data set is random. Under this null hypothesis, the number of runs (R) in a sequence of N elements is a random variable whose conditional distribution given the observation of N_+ positive values and N_- negative values ($N = N_+ + N_-$) has the following properties:

$$\mu = 1 + \frac{2N_+N_-}{N} \tag{2.116}$$

$$\sigma^2 = \frac{2N_+N_-(2N_+N_- - N)}{N^2(N-1)} = \frac{(\mu - 1)(\mu - 2)}{N - 1} \tag{2.117}$$

If N_+ and N_- are each sufficiently large (say, each greater than 30) then the number of runs R is well modeled by a Normal distribution with parameters μ and σ given above. One can then use a typical z-score to determine the probability of obtaining the number of observed runs R under the null hypothesis that the data set is random.

2.8.4 Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov Test (or *KS Test*) tests the equality of two continuous, one-dimensional probability distributions. The most standard version tests whether a particular sample distribution is the same as a specific reference distribution. Because it is a non-parametric test, you do not need to know what the true distribution of your data is, however, the test will not tell you what distribution your data follows either. It will only give you information about its similarity to another reference distribution. Finally, the KS Test is sensitive to both *location* and *shape* and thus cannot tell you why the distributions are different (e.g. is the sample distribution shifted compared to the reference or is the sample distribution wider than the reference?).

The KS Test works by comparing the cumulative density functions (CDFs) of a sample of length N and a reference distribution. Specifically, the difference between the two CDFs is computed, and the *maximum difference*, denoted as D, is used as the test statistic. The null hypothesis is rejected at the significance level α if

$$\sqrt{\mathsf{N}}\mathsf{D} > \mathsf{K}_\alpha \tag{2.118}$$

where K_α is defined as

$$\boldsymbol{Pr}(\mathsf{K} \leqslant \mathsf{K}_\alpha = 1 - \alpha) \tag{2.119}$$

and the probability density function of K is defined as

$$\boldsymbol{Pr}(\mathsf{K} \leqslant x) = 1 - 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2x^2} \tag{2.120}$$

If you are specifically interested in whether your sample distribution is normal, other tests may be better suited (e.g. the Shapiro-Wilks or the Anderson-Darling test). In addition, it is important that you do not estimate the parameters of the *reference distribution* from the data. The test is not valid if you do. Thus, if you are comparing to a normal distribution and don't know the true mean and standard deviation of your sample population, you should standardize your data first and compare the standardized sample to the standard normal. Finally, note that the above discussion only applies when you wish to compare a single sample to some reference distribution. What if instead you wish to compare two sample distributions? For that, you can use the *Two Sample KS-Test* which is similar to the standard KS Test and will not be discussed in detail here.

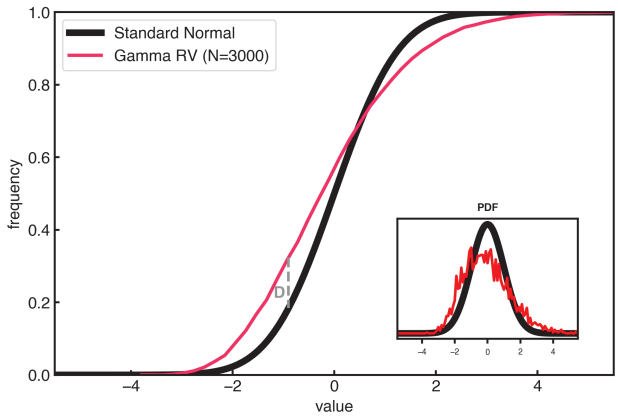

Figure 2.13 Comparison of the standard normal with a random variable drawn from a Gamma distribution. D is the maximum difference between the sample CDF and the reference (normal) CDF and is the test statistic used by the KS-test.

2.9 Hypothesis Testing

2.9.1 Terminology and symbology

- **significance level** [α]: the probability of a false positive (Type I error), often reported as $(1-\alpha)\%$ confidence level
- **critical value** [t_c, z_c]: the value that must be exceeded to reject the null hypothesis using a significance level of α
- **p-value**: the probability of observing an effect given that the null hypothesis is true (probability of the actual statistic you calculated from your data given the null hypothesis is true)

2.9.2 Setting-up the problem

Hypothesis testing involves stating a hypothesis (*null hypothesis*), and then computing statistics to quantify the extent to which your data set is (or is not) consistent with this hypothesis. The significance level (α) of a hypothesis test defines the probability of a false positive (i.e. Type I error), that is, stating that your data set is not consistent with the null hypothesis when in fact it is. This significance level is a choice that should be made by the scientist.

When performing a statistical hypothesis test there are five basic steps that should be followed in order:

1. State the significance level (α)
2. State the null hypothesis H_0 and the alternative H_1
3. State the statistic of interest
4. State the critical region
5. Evaluate the statistic and state the conclusion

Proper construction of the null hypothesis and its alternative is critical to the meaning of statistical significance testing. Careful logic must be employed to ensure that the null hypothesis is reasonable and that its rejection leads uniquely to its alternative. Usually the null hypothesis is a rigorous statement of the conventional wisdom or a "zero information conclusion", and its alternative is an interesting conclusion that follows directly and uniquely from the rejection of the null hypothesis. Usually the null hypothesis and its alternative are mutually exclusive.

Worked Example 2.15.
Examples of null hypotheses and their alternatives:

H_0: the means of two samples are equal

H_1: the means of two samples are not equal

H_0: the correlation coefficient is zero

H_1: the correlation coefficient is not zero

In Practice.

- Hypothesis testing tends to yield weak statements. All you can do is state whether or not the data are consistent with the null hypothesis. You cannot state whether the null hypothesis is true or whether the alternative hypothesis is true, or even whether either is false.

2.9.3 Type I and Type II errors in hypothesis testing

Even though you have applied a test and the test gives you a result, you can still be wrong. The following table illustrates the two different types of errors that can be made:

- **Type I:** reject the null hypothesis when it is actually true
- **Type II:** fail to reject the null hypothesis when it is actually false

	H_0 is true	H_0 is false
Fail to Reject H_0	No Error	Type II Error
Reject H_0	Type I Error	No Error

The way typical hypothesis tests are set up, a 95% confidence level means you have a 5% chance of making a *Type I Error*, that is, you reject the null hypothesis (e.g. think you found something interesting) when you should not have. It is much more difficult to asses the Type II Error - the probability you "play it safe and fail to reject H_0 when something interesting was there". For typical hypothesis testing, the probability of a Type II error can be very large.

In Practice.

- One often cares about the differences between probabilities of Type I and Type II errors. For example, if H_0 is that the bridge will hold-up if 10 semi-trucks cross at the same time, and H_1 is that the bridge will not hold-up, you might be happier with a Type I Error, which requires that you redesign the bridge, rather than a Type II Error, where you think the bridge will be fine, and it won't be.
- When performing a hypothesis test, it is good practice to determine α before performing any calculations. But which α should you choose? The choice of α depends on your risk tolerance, that is, the risk you are willing to take to have a Type I error - the smaller the α, the lower the risk. In atmospheric science, α is typically equal to 0.05, 0.01 or sometimes, 0.10, but it is up to the scientist to decide which value of α is best for the hypothesis being tested.

2.9.4 a priori vs. a posteriori

When performing hypothesis tests it is critical to make the distinction between *a priori* and *a posterior* information.

- ***a priori*:** you have reason to expect a particular relationship ahead of time
- ***a posteriori*:** you don't.

One place where such a distinction arises is whether to use a one-tailed or two-tailed hypothesis test. If you have an *a priori* expectation of the tail of interest, you can use a one-tailed test. Otherwise, you should use a two-tailed test. Since your *a priori* expectation might be regarded as subjective by another scientist, it is generally a better practice to use a two-tailed test.

Another common example is when the same hypothesis test is run many times for similar data sets, for example, testing the significance of anomalies at every grid point on the globe. If one does not take into consideration that the test was run hundreds, if not thousands, of times, they will likely be misled thinking there are more significant anomalies than there really are. In effect, you may be giving your hypothesis many chances to succeed. These concepts are perhaps best illustrated with examples (see below).

Worked Example 2.16.

***a priori vs a posteriori*:** You think that climate change has caused the frequency of severe weather to increase between the 1980's and today. You divide the globe into 20 regions and within each region analyze data for each of the 4 seasons. You test for changes in severe weather frequency using $\alpha = 0.05$ (95% confidence level). How many "significant" changes should you expect by chance alone? How might you apply *a posteriori* statistics?

.....

You have no *a priori* reason to expect a particular region or season should exhibit changes due to climate change, so you test them all. That is, you perform $N = 4 \times 20 = 80$ different hypothesis tests with the null hypothesis H_0 : the frequency of extreme weather has not changed. By chance alone, you expect on average 5% of these tests to reject the null hypothesis when it is in fact true, or, you expect 4 region/season combinations to produce "significant" changes, purely by chance.

$$\boldsymbol{Pr}(\text{correctly not reject } H_0 \text{ when it is true for one test}) = 0.95$$
$$\boldsymbol{Pr}(\text{correctly not reject } H_0 \text{ when it is true for all 80 test}) = 0.95^{80} \approx 1.7\%$$

Thus, your 95% confidence level is really a 1.7% confidence level! In other words, you have a 98.3% chance of finding at least one significant change, even if climate change has no impact.

Using *a posterior* statistics, we can instead calculate the significance level α for which $\alpha^{80} \approx 0.95$. In this case, $\texttt{alpha} \approx 0.9994$. Thus, if we require that severe weather frequency changes for each region/season combination pass at the 99.94% confidence level, the probability of correctly not rejecting the null hypothesis for all 80 chances will be 95%. We can also use the Binomial Distribution to assess the likelihood of getting some number of "significant" changes above the expected value of 4.

2.9.5 Field significance and False Discovery Rate

Much of geophysical research involves creating maps of a result, and often, scientists will assess the significance of each value on the map individually. As discussed above, one should expect a certain fraction of points to be "significant", even if the null hypothesis is true. Furthermore, many geophysical variables are spatially correlated, implying that significant points will likely appear clustered. An example of this is illustrated in **Fig. 2.14**. To create this figure, daily January 500 hPa geopotential heights at each latitude/longitude grid point is correlated with a time series X. Correlations different from zero at 95% confidence are stippled, and appear to show signals across the globe, with the largest signal in the tropical Pacific. The trick here is that X is a random Gaussian time series, with absolutely no physical meaning, and yet a large cluster of data points were found to be significantly correlated. Thus, in many applications, assessing the significance at each grid point is not enough - rather - one should assess the collective significance, or *field significance* over the entire map (Livezey and Chen, 1983).

Wilks (2016) outlined a straight-forward way to assess field significance by controlling the *false discovery rate* (FDR), or, the expected rate of rejected local null hypotheses where the actual null hypothesis is true. The general idea is that one sorts the list of p-values (across grid points), and then finds which p-value intersects the line defined by

$$y = \frac{i}{N}\alpha_{FDR} \tag{2.121}$$

where i is the position of the p-value in the sorted list, N is the total number of grid points, and α_{FDR} is a parameter that is chosen by the user. The p-value at the intersection is then the global p-value that each grid point must be smaller than to satisfy a particular false detection rate. An illustration of the calculation of this global p-value threshold is shown in the left panel of **Fig. 2.15** for the example plotted in the bottom panel of **Fig. 2.14**. There is no intersection of the actual p-values with the FDR criterion line given in **(2.121)**, and so, none of the stippled points in **Fig. 2.14** should be considered globally significant. For a case where we expect a physical relationship, that is, the correlation of daily January 500 hPa geopotential heights with the stratospheric zonal winds, an intersection occurs and the new global $p-value$ threshold is actually 0.066 rather than 0.05 (right panel of **Fig. 2.15**).

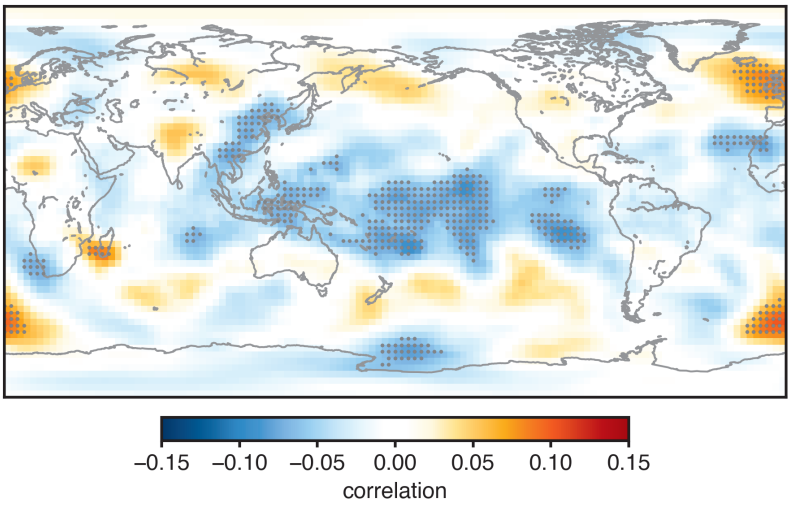

Figure 2.14 Correlation of daily January 500 hPa geopotential heights (1979-2011) with a random Gaussian time series. Statistically significant correlations at 95% confidence ($\alpha = 0.05$).

2.10 Extreme Value Theory

Extreme events are those that appear in the tails of the probability distribution (Coles, 2001). While rare, they can have very important impacts, and so understanding their frequency is very important. Design of physical and financial infrastructure must take into account the most extreme events that are likely to occur over some defined period of time. Therefore the study of extreme values is very important, particularly during this time of global change.

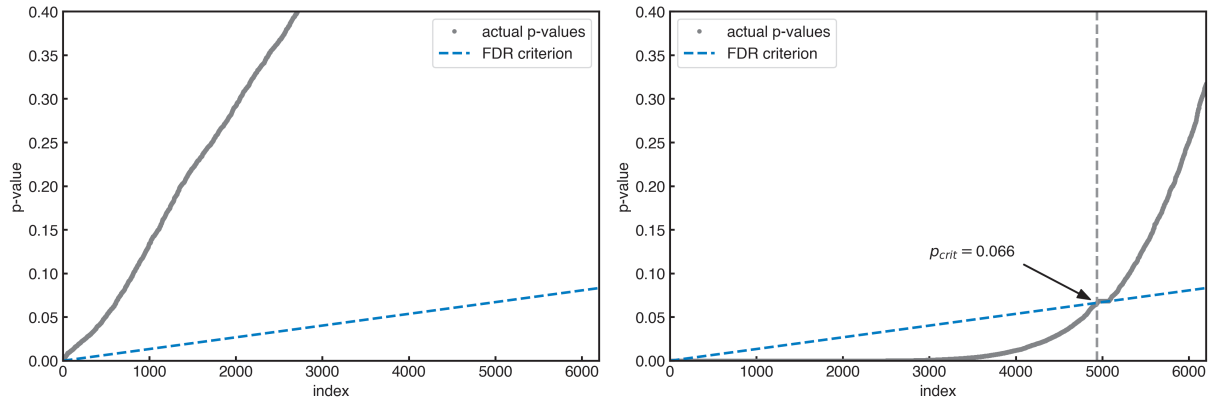

Figure 2.15 Illustration of the false detection rate criterion of Wilks (2016) for (left) correlations with a random time series as shown in **Fig. 2.14**, (right) correlations with daily January 100 hPa polar cap zonal winds averaged around the 65° latitude circle. In both panels $\alpha_{FDR} = 0.1$.

2.10.1 Fisher-Tippett Theorem and Generalized Extreme Value Distribution

Suppose we have a sample of n independent and identically-distributed random variables $[X_1, X_2, X_3, .., X_n]$, each of which has the same cumulative distribution function F. Suppose further that there exists two sequences of numbers $a_n > 0$ and $b_n \in \Re$ such that the following limits converge to a non-degenerate distribution function $G(x)$.

$$\lim_{n\to\infty} \boldsymbol{Pr}\left(\frac{\max\{X_1, \ldots, X_n\} - b_n}{a_n} \leqslant x\right) = G(x) \tag{2.122}$$

which is equivalent to

$$\lim_{n\to\infty} (F(a_n x + b_n))^n = G(x) \tag{2.123}$$

The Generalized Extreme Value (GEV) distribution has three parameters; location = μ, scale = σ and shape = ξ. Using the definition $s = (x - \mu)/\sigma$ the pdf of the GEV distribution is given by,

$$f(\, x \,|\, \mu, \sigma, \xi) = \frac{1}{\sigma}(1 + \xi\, s)^{-\frac{1}{\xi}-1}\, exp[-(1 + \xi\, s)] \tag{2.124}$$

For particular values of ξ the GEV divides into the Gumbel, Fréchet and Weibull families of distributions, corresponding to the cases $\xi = 0$, $\xi > 0$ and $\xi < 0$, respectively . Each of these distributions has a range of x in which they are supported, which depends on the shape and scale.

In addition, the Generalized Pareto Distribution is often used to describe extreme values. The pdf of the Pareto distribution is given by,

$$f(\, x \,|\, \sigma, \xi) = \frac{1}{\sigma}(1 + \frac{\xi\, x}{\sigma})^{-\frac{1}{\xi}-1} \quad x \in (0, \infty) \tag{2.125}$$

Dennis: *Need more detail and examples here. Finish project on Pareto fit to SeaTac Max temps*

2.11 Monte Carlo and Resampling

2.11.1 Monte Carlo Techniques

In the age of computers, sometimes it is easier to let the computer do the work by performing many intelligently designed calculations, and then infer a fact or statistical conclusion from the aggregate of these calculations. These techniques take advantage of the computational power at our fingertips and are incredibly powerful when data size is not an issue. The name *Monte Carlo* comes from the famous casino, not from the inventor of the method. It is a term that has no precise definition and covers a wide variety of techniques, which share in common the idea expressed in the first sentence.

One famous example is the calculation of π - the ratio of the circumference of a circle to its diameter. Rather than trying to derive it from basic principles, one can instead write a simple computer code to get a very accurate approximation. Specifically, π can be calculated by inscribing a circle within a square, dropping pebbles randomly on the square, and then counting the ratio of the pebbles in the square to those that fall within the circle. If the pebbles are dropped randomly, then this ratio should be the ratio of the areas of the circle to the square, which is $\pi/4$. If you do this many times you can get an arbitrarily good approximation to π.

2.11.2 Resampling via Bootstrapping

Bootstrap Resampling involves constructing a number of random *resamples* of a dataset of equal size to the true sample of interest. In this way, you do not need to assume anything about the underlying distribution of the data since it is already built into the original dataset. In essence, you ask, by random chance, what is the probability that a particular event (or sample statistic) occurred?

This method is also useful when you are determining statistics other than the mean (e.g. extrema, median, skewness) when we don't have simple statistics for these variables.

The advantage of this method is that you don't have to choose a model PDF and you can evaluate the number of successes in exceeding the criteria using the binomial distribution.

A question arises of whether one should perform the random sampling *with* or *without replacement*. Namely, should you be able to pick the same value twice for the same random sample? Most often, bootstrapping resampling is done *with replacement* as the data set used for sampling is meant to represent an entire population of possibilities. More practically, if your data set is large enough, with and without replacement should give nearly identical answers.

The technique is called "bootstrapping" because it almost appears you get something out without putting something in. This is a phrase often used in computer programming to refer to a small amount of simple software that can load more complex software that loads more complex software (etc., etc.,) almost as if the program is "pulling itself up by its bootstraps."

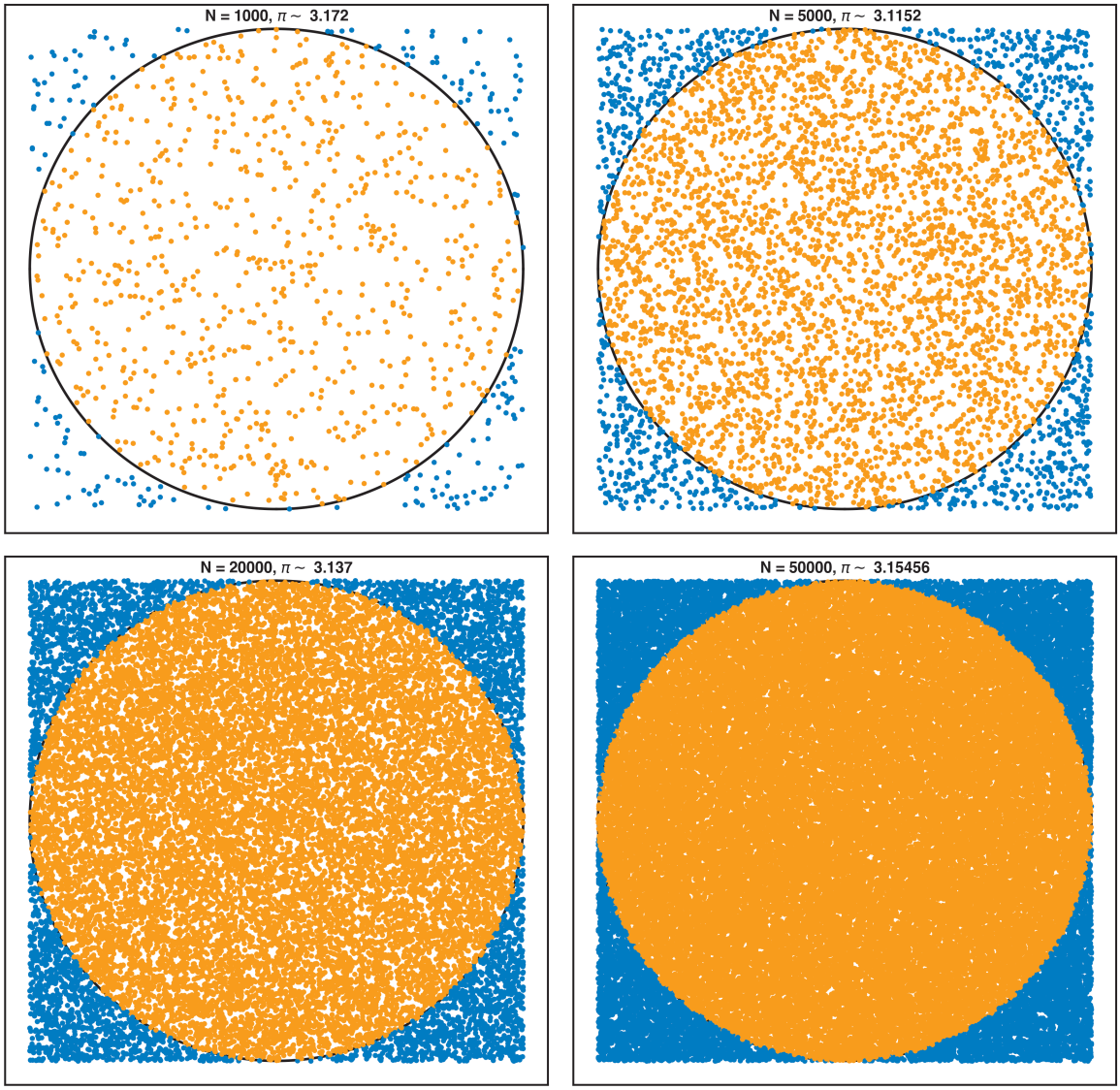

Figure 2.16 Estimate of π using a Monte Carlo approach.

Worked Example 2.17.
A frequently used application of the Bootstrapping method is to determine the significance of some field obtained through compositing (see Chapter 3). For example, say you in think that full moons have a strong impact on geopotential heights over Fort Collins, CO. You check this by calculating the average 500 hPa geopotential height on all full moon days (408 between 1979-2011) and obtain 5689.5m. The average over the entire record is 5696.9m - does this mean that full moons cause the geopotential heights to decrease?

The issue here is that you don't know that whether this difference of -7.4m is significant. That is, could it have just been due to random chance? In this case, our null hypothesis is that full moons have no effect on geopotential heights over Fort Collins, and so the anomaly of −7.4m is just due to random chance.

To test this using a bootstrap approach, we randomly draw a sample of geopotential heights of length 408 from all days between 1979-2011. We then calculate the mean across these 408 days and save it. We repeat this process 50,000 times. After we are done, we have 50,000 averages of 408 days - all under the null hypothesis. An example is shown in **Fig. 2.17**, where the gray line denotes the distribution of the 50,000 averages. Comparing our calculated geopotential heights under full moons, we see that they fall well within the bootrap samples. This tells us that we shouldn't reject the null hypothesis, since our actual results are well within the possibility of random chance. If our calculated value fell outside of the 95% bounds of the bootstrap samples, we would instead reject the null hypothesis and investigate further.

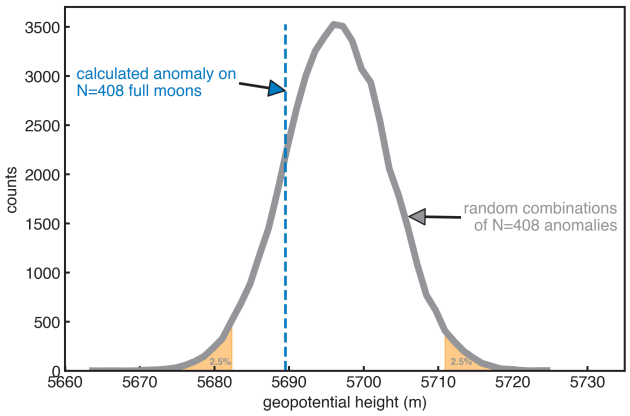

Figure 2.17 Distribution of 50,000 random averages of Fort Collins geopotential heights (N = 408).

2.11.3 Resampling via Jackknife

The jackknife method predates the bootstrap method and is a linear approximation of the bootstrap. It is a way of getting uncertainty estimates (or measuring the variance) of a particular statistic of your sample. The way it works is quite simple. You systematically remove one value from your sample (i.e. the $\mathtt{i}$th value), calculate the statistic of interest (call it $\mathtt{s}_\mathtt{i}$), then put the value back into the sample and remove the next value, calculate the statistic of interest...and on and on. In the end, you are left with many estimates of your statistic of interest, and from these, you can estimate its variance in the following way.

$$\mathbf{Var}(s) = \frac{N-1}{N} \sum_{i=1}^{N} \left(s_i - \overline{s}_{(.)}\right)^2 \tag{2.126}$$

where $\overline{s}_i$ is the estimate of statistic s leaving out the ith value and $\overline{s}_{(.)}$ is the average estimate of s over all leave-one-out estimates as given by

$$\overline{s}_{(.)} = \frac{1}{N} \sum_{i}^{N} s_i \tag{2.127}$$

Chapter 3
Compositing and Superposed Epoch Analysis

3.1 Introduction

Compositing, also sometimes called superposed epoch analysis when applied to time series, is one of the simplest analysis techniques, yet it can also be very powerful. It consists of sorting data into categories and comparing means for different categories. Although conceptually simple, compositing, like any other technique, must be applied logically, carefully and with proper concern for the possible appearance of spurious signals. Compositing is useful when you have many observations of some event and you are looking for responses to that event that are combined with noise from a lot of other influences. The idea is that if you average the data in some clever way in relation to the event, the event signal will remain and all other influences will tend to average out. Examples might include the climatic response to a volcanic eruption, the global weather response to el Niño, calculating the mean diurnal cycle of surface temperature in Dallas, Texas, or finding if precipitation responds to the phase of the moon. The last two of these relate to sorting out the true amplitude of cyclic responses. Often, compositing will reveal periodic phenomena with fixed phase that cannot be extracted from spectral analysis if the signal is small compared to the noise. Compositing makes no assumption of linearity, and it is good at separating small signals from noise, if your sample is large enough.

3.2 Steps in the Compositing Process

Setting up and executing a successful compositing study consists of several steps. If you think about these steps in the abstract, you will more easily see the ways to do it properly. It is easy to get excited with the specifics of a particular study and begin to lose your objectivity, especially if you hope to obtain a particular result. The steps are:

1. Select the basis for compositing and define the categories. The categories might be related to the phase of some cyclic phenomenon or forcing, or to time or distance from some event. Bases for compositing can range from the very commonplace, such as the hour of the day or the month of the year, to the relatively obscure, such as the passage of Earth through a reversal of the sun's magnetic field or the length of the solar cycle. It is highly desirable to have some believable hypothesis for why the event or cycle should affect the variables you are compositing. Otherwise you have greater risk finding a statistical coincidence with no physical meaning.
2. Compute the means and statistics for each category. One must be sure to be accurate and objective.
3. Organize and display the results. The results may be best shown in the form of tables, graphs or maps. It is important that whatever medium is used, a clear indication of the sense and significance of the results is achieved. This may mean adding confidence limits to the picture.

4. Validate the results. Validation of the results can be achieved in many ways. Statistical significance tests are only one of these. It is desirable to use as many of the following tests as possible, and any others that you can think of.
   - Use statistical significance tests. Using a model for the distribution of a variable about its mean, or nonparametric statistical tests, one can estimate the probability that the signal derived from the compositing exercise arose from chance. One should always do this type of testing, but it is not enough.
   - Subdivide the data set to show consistency. If you have enough data, you can divide the data set and see how well the derived relationships are maintained. This is especially useful if the statistical significance estimate is good, but the physical reason for the relationship is unclear. Monte Carlo techniques can be use here, *e.g.* divide the sample many ways randomly. How often does the result reproduce?
   - Show consistency in other ways.
     - Find some additional data and reproduce the results
     - Show that the results are consistent with space or time
     - Show that the results are consistent with a well-founded theory

3.3 Evaluating compositing studies

When evaluating a compositing study, ask yourself the following questions.

1. Do you have an *a priori* basis for expecting to find the relationships found in the study? *A priori* means beforehand: based in knowledge or hypothesis that was available to the investigator before the study was conducted. If there was not, then you might suspect that the relationship was found after trying a number of different things, in which case the probability that it occurred by chance is greatly increased, and *a priori* statistical tests should not be used. Every time you give your hypothesis another independent chance to succeed you need to multiply your certainty by the *a priori* probability that it is true (1.10).
2. What is the basis for compositing? Does it have a precise, unique, objective definition (*e.g.* time of day) or is it somewhat arbitrary? Does it have a distinct physical interpretation? Could another basis for compositing have been used just as well, or a better one defined?
3. How was the compositing performed? Can you easily visualize how the process might have been programmed for the computer? Could an opportunity for subjective judgment or subconscious bias have entered the procedure at some point?
4. How does the investigator argue that the results are statistically significant (*i.e.* that they would be reproducible in independent data sets)? List all the statistical arguments that are given and the physical or logical arguments as well. Can you think of alternative, perhaps simpler, explanations? Are there reasons to suspect that the method itself produced a signal that is not really in the data? Does the author have a justification for using *a priori* statistical tests?
5. Are you convinced of the validity of the results? Would you direct your research effort on the assumption that the results are correct? If not, what would it take to convince you?

3.4 Example: Daily Precipitation and Temperature

We download daily station data from the U.S. Climate Data Center. We choose the location of the Seattle SeaTac International Airport, which has records from January 1, 1948 to the present, about 69 years. We organize the data as a two-dimensional array by year from 1948 to 2016 and by day of the year from 1

(January 1) to 365 (December 31). We ignore the extra day in leap years. We compute the mean and standard deviation for each day of the year, using the sample of 69 years. We then plot the mean as a function of day of year.

Fig. 3.1 shows the raw mean of the maximum daily temperature and precipitation for the sample as a thin black line. Since the data are still noisy even with a sample of 69 years, we also show two different locally weighted scatterplot smoothings (LOWESS) of the data, with the red line showing a stronger smoothing. Also shown are dashed lines, which are the less smoothed data, plus or minus the smoothed standard deviation for each day. We see that the standard deviation of temperature is similar in winter and summer. Later we will show that it is actually a little bigger in summer.

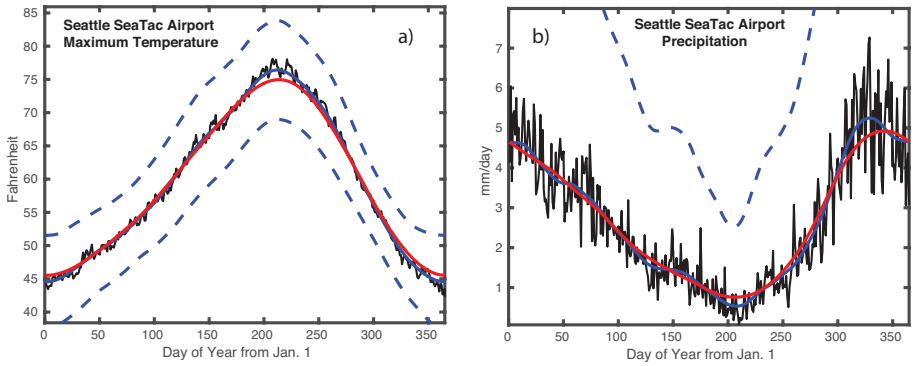

Figure 3.1 Composited a) daily maximum temperature and b) precipitation for 69 years of data from SeaTac Airport near Seattle, Washington. Blue and Red lines are two different smoothings of the daily composites. Dashed lines indicate plus or minus one smoothed standard deviation.

The annual cycle composite for daily precipitation shows that it rains more in the winter, and almost not at all for a brief period near July 31st (day 212). In this case we see that the standard deviation is larger than the mean, and that subtracting the standard deviation gives a non-physical negative value, which is not shown. This is because precipitation is not a normally distributed variable, since it has a minimum value of zero, which is also the most commonly occurring value. The skewness for the SeaTac maximum temperature data is 0.4 in July and -0.5 in January, both very close to the value of 0.0 for a normally distributed variable, while the kurtosis is 2.8 in July and 3.5 in January, also close the the normal distribution value 3.0. The precipitation data on the other hand have skewness of 5.2 in July and 3.0 in January, and kurtosis of 35.2 in July and 15.6 in January.

Fig. 3.2 shows the probability density functions for maximum daily temperature and daily precipitation for July and January obtained from the data by the kernel method. Temperature looks fairly Gaussian like a normally distributed variable, while the precipitation pdf peaks near zero and is better fit with a gamma distribution. The July maximum temperature pdf is wider and less peaked than the December one, indicating warmer, but more variable temperature in the summer. Because precipitation is less frequent in the summer, its pdf is even more strongly peaked at zero in July than January.

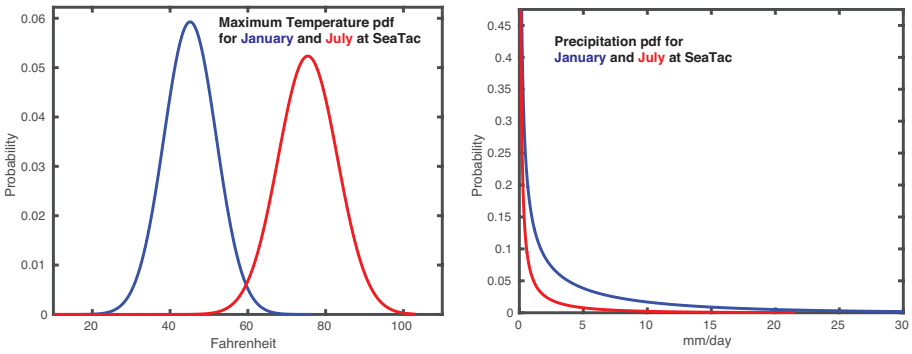

Figure 3.2 Smoothed Probability Density Functions for daily maximum temperature and precipitation for SeaTac Airport during July and January. Note that the precipitation pdf is unbounded at zero, but is cut off at 0.45 and at 30 mm/day so that the differences in the pdfs for precipitation can be better seen.

Chapter 4
Regression

In this chapter some aspects of linear statistical models or regression models will be reviewed. Topics covered will include linear least-squares fits of predictands to predictors, correlation coefficients, multiple regression, and statistical prediction. These are generally techniques for showing linear relationships between variables, or for modeling one variable (the predictand) in terms of others (the predictors). They are useful in exploring data and in fitting data. They are also a good introduction to more sophisticated methods of linear statistical modeling.

4.1 Ordinary linear least-squares regression

4.1.1 Independent variables are known

Suppose we have a collection of N paired data points (x_i, y_i) and that we wish to approximate the relationship between x and y with the expression:

$$\widehat{y} = a + b \cdot x + \epsilon \tag{4.1}$$

where a is called the *y-intercept* and b is the *slope of the line*. In what follows, we assume that x is known with precision, and that we wish to estimate y based on known values of x. The cases where both x and y contain uncertainties will be discussed next. The error, or residual, ϵ can be minimized in a least-squares sense by defining an error function Q in the following way:

$$Q = \frac{1}{N}\sum_{i=1}^{N} \epsilon^2 = \frac{1}{N}\sum_{i=1}^{N} (\widehat{y}_i - y_i)^2 = \frac{1}{N}\sum_{i=1}^{N} (bx_i + a - y_i)^2 \tag{4.2}$$

where the subscript i denotes i^{th} observation. Q is the sum of the squared differences between the data and our linear fit, and when it is minimized by choosing the parameters a and b we obtain the *least-squares linear fit* to the data.

The fact that the error is squared in the definition of Q has several important consequences.

- Q is positive definite.
- The minimization of Q (the derivative of Q) results in a linear problem to solve.
- Large errors are weighted more heavily than small errors.

The first two are very good consequences. The last can be good or bad depending on what you are trying to do. All the linear regression analysis techniques we will discuss in later chapters (EOF, SVD, PCA, etc.) share these same properties of linear least squares techniques.

We wish to select the constants a and b such that the error or risk functional Q is minimized. This is achieved in the usual way by finding the values of these constants that make the derivatives of Q with respect to them zero. Since the error is always positive and the error function has a parabolic shape, we know that

these zeros must correspond to minima of the error function

$$\frac{\partial Q}{\partial a} = 0 \text{ and } \frac{\partial Q}{\partial b} = 0 \quad \text{“The Normal Equations”} \tag{4.3}$$

It is straightforward to show that solutions to these equations results in the following

$$\frac{\partial Q}{\partial a} = 2aN + 2b\sum_{i=1}^{N} x_i - 2\sum_{i=1}^{N} y_i = 0 \tag{4.4}$$

$$\frac{\partial Q}{\partial b} = 2a\sum_{i=1}^{N} x_i + 2b\sum_{i=1}^{N} x_i^2 - 2\sum_{i=1}^{N} x_i y_i = 0 \tag{4.5}$$

Dividing both equations by N and moving the y terms to the left-hand-side results in

$$\overline{y} = b\overline{x} + a \tag{4.6}$$

$$\overline{xy} = b\overline{x^2} + a\overline{x} \tag{4.7}$$

where $\overline{(\cdot)}$ denotes the mean across all N observations and $(\cdot)'$ will denote departures from this mean. This system of equations can also be written in matrix form,

$$\begin{bmatrix} 1 & \overline{x} \\ \overline{x} & \overline{x^2} \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \overline{y} \\ \overline{xy} \end{bmatrix} \tag{4.8}$$

which is especially useful when one moves to multi-linear regression with more than one independent variable.

The solutions for the regression coefficients are:

$$a = \overline{y} - b\overline{x} \tag{4.9}$$

$$b = \frac{\overline{x'y'}}{\overline{x'^2}} \tag{4.10}$$

The term $\overline{x'y'}$ is given a special name, the *covariance* of x and y, and is defined as

$$\overline{x'y'} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y}) \tag{4.11}$$

Thus, we see that a_1 is the covariance of x and y normalized by the variance of the independent variable x. Also, note that a, also known as the *y-intercept* is zero if the variables x and y have mean zero.

One can show that the minimum value of the error functional obtained via ordinary least-squares regression is:

$$Q_{min} = \overline{y'^2} - \frac{\overline{x'y'}^2}{\overline{x'^2}} = \overline{y'^2} - b^2\overline{x'^2} \tag{4.12}$$

Thus, we see that the minimum error is the total variance minus the explained variance, which is related to the squared slope (b) and the variance of the predictor.

4.1.2 Independent and dependent variables are uncertain

Quite often the first attempt to quantify a relationship between two experimental variables is linear regression analysis. In many cases one of the variables is a precisely known independent variable, such as time or distance, and the regression minimizes the root mean square (rms) deviation of the dependent variable from the line, assuming that the measurements contain some random error. It often happens that both variables are subject to measurement error or noise, however. In this case, to perform simple linear regression analysis

one must choose which variables to define as dependent and independent. The two possible regression lines obtained by regressing y on x or x on y are the same only if the data are exactly collinear.

An alternative to simple regression is to minimize the perpendicular distance of the data points from the line in a two-dimensional space. This approach has a very long history scattered through the literature of many scientific disciplines (Adcock 1878; Pearson 1901; Kermack 1950; York 1966). The method can be elaborated to any degree desired, to take into account the different scales of the two variables in question, their uncertainty, or even the confidence one has in individual measurements (see Section 4.1.4.1.

One of the better, and more elegant, methods of doing linear fits between two variables is EOF/PC analysis, which is discussed in a later chapter of these notes. It turns out that, at least in two dimensions, doing EOF analysis minimizes the perpendicular distance from the regression line and is more elegant than the methods used by Kermack and Haldane (1950) and York (1966). EOF/PC analysis is also easily generalized to many dimensions. See Chapter 5.

4.1.3 Uncertainty estimates of ordinary least-squares regression

We want to fit a straight line to a time series of N observations y_i taken at time x_i. The linear fit is given by

$$y_i = a + bx_i + e_i, \quad i = 1, 2, ..., N \tag{4.13}$$

where e_i represents the residual error of the linear fit at each time x_i. From Chapter 4.1.1, we know that the ordinary least squares solution for parameters a and b are

$$\widehat{b} = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{N}(x_i - \overline{x})^2} = \frac{\overline{x'y'}}{\overline{x'^2}} \tag{4.14}$$

$$\widehat{a} = \overline{y} - \widehat{b}\overline{x} \tag{4.15}$$

and so, the errors of the fit, called the *residuals*, are

$$\widehat{e}_i = y_i - (\widehat{a} + \widehat{b}x_i) = y_i - \widehat{y}_i, \quad i = 1, 2, ..., N \tag{4.16}$$

Now, we would like to assign ranges, or confidence limits, on our estimates of a and b. We start with the unbiased estimate of the standard error variance of the residuals:

$$\widehat{\sigma}_e^2 = \frac{1}{N-2}\sum_{i=1}^{N}\widehat{e}_i^2 = \frac{N}{N-2}(1 - r_{xy}^2)\overline{y'^2} \tag{4.17}$$

where we divide by $N - 2$ to account for the fact that two degrees of freedom were used to estimate a and b. The expression that includes the correlation coefficient, r_{xy}, follows from the derivations in Chapter 4.2.

For the time being we will assume that all of these residuals are independent of one another, but if instead they are autocorrelated, we could use a model of red noise to estimate the true number of degrees of freedom N^*, and then replace N with N^* (see Chapter 7 for a discussion of degree of freedom estimates for autocorrelated data).

From the standard error variance of the residuals, $\widehat{\sigma}_e^2$, we can estimate the standard error variance of the of slope, $\widehat{\sigma}_b^2$ in the following way. First,

$$\widehat{\sigma}_b^2 = \frac{\widehat{\sigma}_e^2}{N\sigma_x^2}, \quad \sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2 \tag{4.18}$$

where we have assumed that the x_i's are precisely known. Putting the pieces together leads to

$$\widehat{\sigma}_b^2 = \frac{\frac{1}{N-2}\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{N}(x_i - \overline{x})^2} \tag{4.19}$$

Looking at this equation, you can see intuitively that it is somewhat like the error in our y estimate divided by the variance in our x values - sort of like a slope of errors or variances.

Since $\frac{\widehat{b}-b}{\widehat{\sigma}_b}$ is distributed like the t-statistic with $N-2$ degrees of freedom, we can put limits on the true slope b in the following way:

$$\widehat{b} - t_{\alpha/2}^{N/2} < b < \widehat{b} + t_{\alpha/2}^{N-2}\widehat{\sigma}_b \tag{4.20}$$

where t_{α}^{N-2} is the critical value of the t-statistic for confidence level α and degrees of freedom $N-2$.

We can apply these techniques to the record of annual mean land temperature from the Goddard Institute of Space Studies (GISS) for the period 1900-2016, as shown in **Fig. 4.1**. Note that the lower limits on the trends are all positive, so we can say that the trends on the intervals are positive at 95% confidence.

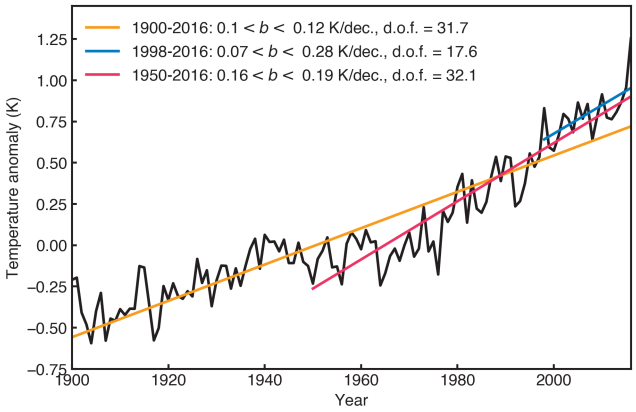

Figure 4.1 The GISS global surface temperature timeseries with linear trends (b) and $\pm$ 2.5 % uncertainties for various periods. Estimated degrees of freedom are also given. Units of the trends are Kelvin per decade.

Fig. 4.1 illustrates several aspects of linear fitting to time series. First, the result may depend sensitively on the end points of the analysis. Note that the procedure described in Section 4.1.3 assigns the shorter 1950-2017 period more degrees of freedom than the longer period from 1900-2017. This is because the longer period has an S-shape associated with the period of slow change from 1940-1980. As a result, the residuals from the linear fit for the longer period yield a large autocorrelation. The decades of the 1950's to 1970's are consistently below the line and the decades from 2000-2017 are consistently above the line. The period from 1950-2017 is better fit by a straight line and gives a larger number of degrees of freedom. Despite that the uncertainty is smaller for the longer period because the variance of the predictor is greater. The statistics support the notion that the recent trends are greater than the long-term trend at 95% significance. Starting the trend calculation at 1950 is not objective, since it was chosen by inspecting the time series, but it is true nonetheless that any starting point after 1950 or so yields the same conclusion that the recent warming is faster than the estimate for 1900-2017, unless the record is so short that the uncertainty is too great, as is the case for the 1998-2017 period.

4.1.4 Other least-squares fits

4.1.4.1 Orthogonal-least squares regression

Looking back at our derivations for ordinary least squares, it becomes apparent that the results are not symmetric for x and y. That is, it matters which variable you call x (the independent variable) and which you call y (the dependent variable). This can be seen for the example data provided in **Fig. 4.2**. If you cannot adequately justify which data should be x and which should be y, or it does not make sense to even try, *orthogonal least squares* may instead be what you want.

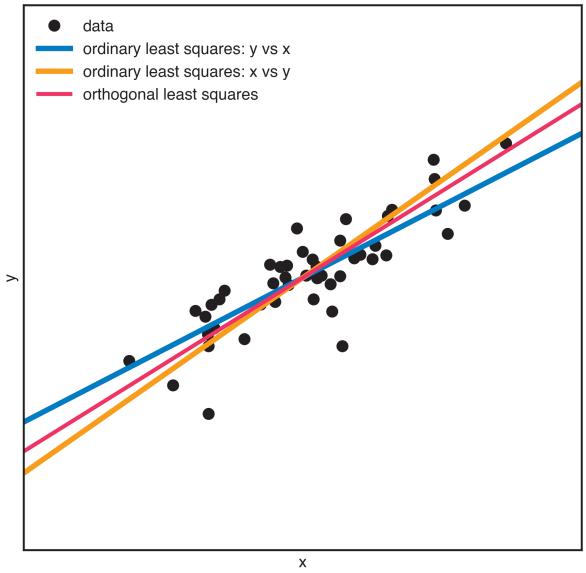

Figure 4.2 Three least-squares best fit lines calculated in different ways: (1) ordinary least squares where $\mathtt{x}$ is the independent variable and $\mathtt{y}$ is the dependent variable, (2) ordinary least squares where $\mathtt{y}$ is the independent variable and $\mathtt{x}$ is the dependent variable, and (3) using orthogonal least squares.

While ordinary linear least squares minimizes the vertical errors between the data and the best fit line (i.e. the error in $\mathtt{y}$), orthogonal linear least squares minimizes the orthogonal errors, as shown in **Fig. 4.3**. It so happens that EOF analysis (to be discussed in Chapter 5) in two-dimensions provides the orthogonal least squares fit.

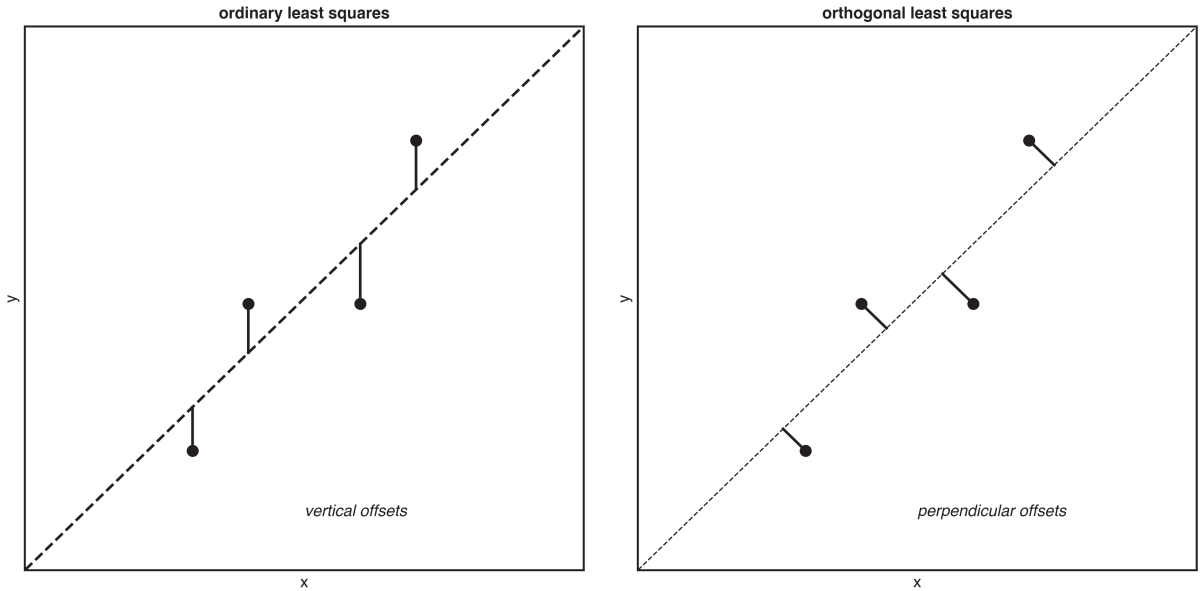

Figure 4.3 Depiction of (left) ordinary least squares regression which defines the errors based on vertical offsets and (right) orthogonal least squares regression which defines the error based on perpendicular offsets.

4.1.4.2 Power laws and polynomials

Many other curves besides a straight line can be fit to data using a similar procedure to that outlined for ordinary least-squares regression. Some common examples are power laws and polynomials such as

$$y = ax^b \Rightarrow \ln y = \ln a + b \ln x \tag{4.21}$$
$$y = ae^{bx} \Rightarrow \ln y = \ln a + bx \tag{4.22}$$
$$y = a_0 + a_1 x + a_x^2 + a_3 x^3 + ... + a_n x^n \tag{4.23}$$

In some cases, like that of power laws (e.g. **(4.21)**, **(4.22)**), one can use logarithms to turn the problem into a linear one, in which case, standard linear least squares methods can be used to estimate the parameters.

4.2 Correlation

4.2.1 How good is the linear fit?

How much we believe the computed regression coefficient ($\widehat{b}$) depends on the spread of the dots about the best fit line. If the dots are closely packed about the regression line, then the fit is good. The spread of the dots is given by the *correlation coefficient* r.

Here is one way to derive the correlation coefficient. By definition, the total variance of $y(t)$ is

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y})^2 \tag{4.24}$$

and by definition, the total variance of the fit of $x(t)$ to $y(t)$ (i.e. the variance of $\widehat{y}$) is

$$\frac{1}{N}\sum_{i=1}^{N}(\widehat{y}_i - \widehat{\overline{y}})^2 = \frac{1}{N}\sum_{i=1}^{N}(\widehat{y}_i - \overline{y})^2 \tag{4.25}$$

where we have used the fact that

$$\widehat{\overline{y}} = a + b\overline{x} = \overline{y} \tag{4.26}$$

The percent of the total variance in y explained by the fit $\widehat{y}$ is thus given by the ratio

$$r^2 = \frac{\text{explained variance}}{\text{total variance}} \tag{4.27}$$

$$= \frac{\sum_{i=1}^{N}(\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{N}(y_i - \overline{y})^2} \tag{4.28}$$

$$= \frac{\sum_{i=1}^{N}(bx_i + a - \overline{y})^2}{\sum_{i=1}^{N}(y_i'^2)} \tag{4.29}$$

$$= \frac{\sum_{i=1}^{N}(bx_i + \overline{y} - b\overline{x} - \overline{y})^2}{\sum_{i=1}^{N}(y_i'^2)} \tag{4.30}$$

$$= \frac{\sum_{i=1}^{N}(bx_i')^2}{\sum_{i=1}^{N}(y_i'^2)} \tag{4.31}$$

$$= \frac{(\frac{\overline{x'y'}}{\overline{x'^2}})^2 \sum_{i=1}^{N}(x_i')^2}{\sum_{i=1}^{N}(y_i'^2)} \tag{4.32}$$

$$= \frac{(\overline{x'y'})^2 \sum_{i=1}^{N}(x_i')^2}{(\overline{x'^2})^2 \sum_{i=1}^{N}(y_i'^2)} \tag{4.33}$$

$$= \frac{(\overline{x'y'})^2 \frac{1}{N}\sum_{i=1}^{N}(x_i')^2}{(\overline{x'^2})^2 \frac{1}{N}\sum_{i=1}^{N}(y_i'^2)} \tag{4.34}$$

$$= \frac{(\overline{x'y'})^2 \cdot \overline{x'^2}}{(\overline{x'^2})^2 \cdot \overline{y'^2}} \tag{4.35}$$

$$= \frac{(\overline{x'y'})^2}{\overline{x'^2} \cdot \overline{y'^2}} \tag{4.36}$$

Hence,

$$r = \frac{\overline{x'y'}}{\widehat{\sigma_x}\widehat{\sigma_y}} \tag{4.37}$$

Some important points about the correlation coefficient r:

- r^2 is the fraction of variance explained by the linear least-squares fit between the two variables
- r varies between -1 and 1 and r^2 varies between 0 and 1

Note that if $\sigma_x = \sigma_y = 1$ and $\overline{x} = \overline{y}$ (that is, both x and y are standardized), then the correlation r is equal to the regression coefficient $\widehat{b}$. More generally, there is a strong relationship between the regression line and the correlation coefficient:

$$\widehat{b} = r\frac{\sigma_y}{\sigma_x} \tag{4.38}$$

Thus, the regression coefficient can be thought of as the correlation coefficient multiplied by the ratio of the standard deviations of y and x.

Worked Example 4.1.
Suppose that the correlation coefficient between sunspots and five-year mean global temperature is 0.5 ($r = 0.5$). Then the fraction of the variance of 5-year mean global temperature that is linearly explained by sunspots is $r^2 = 0.25$. That is, the fraction of unexplained variance is still 75%. The *root-mean-square error* (RMS error), normalized by the total variance is thus:

$$\left(\frac{\text{MS Error}}{\text{Total Variance}}\right)^{1/2} = \sqrt{1 - r^2} - \sqrt{0.75} = 0.87 \tag{4.39}$$

Thus, only a 13% reduction in RMS error results from a correlation coefficient of 0.5. The implications of this are further illustrated in the following table:

r	r^2	**RMS error**
0.98	.960	20.0%
0.9	.81	43.6%
0.8	.64	60.0%
0.5	.25	86.6%
0.3	.09	95.4%
0.1	.01	99.5%

In Practice.

- As **Worked Example 4.1** illustrates, statistically significant correlations are not necessarily useful for forecasting. If you have enough data you may be able to show that a measured $r = 0.3$ correlation coefficient reflects that the true correlation coefficient is different from zero at the 99% confidence level, but such a correlation, however real, is often useless for forecasting. The RMS error would be 96% of the variance. The exception to this statement about the uselessness of small correlations comes where you have a very large number of trials or chances. If you have a large volume of business (billions of dollars) spread over a large number of transactions and you shade your trades properly using the 0.3 correlation prediction, then you can actually make a lot of money...sometimes.

In Practice.

- The correlation will only show the linear relationships clearly. Nonlinear relationships may exist for which the correlation coefficient will be zero. For example, if the true relationship is parabolic, and the data are evenly sampled, the correlation coefficient would be close to zero, even though an exact parabolic relationship may exist between the two data sets.
- The correlation cannot reveal quadrature relationships (although lagged correlations often will). For example, meridional wind and geopotential are approximately uncorrelated along latitudes even though the winds are approximately geostrophic and easily approximated from the geopotential. They are in quadrature (90 degrees out of phase).
- The statistical tests (to be described next) apply to independent data. Often the sample data are not independent. The actual number of degrees of freedom may be much smaller than the sample size.
- Watch out for nonsense correlations that may occur even though the two variables have no direct relation to each other. The correlations may occur by chance or because the two variables are each related to some third variable. For example, over the past 50 years the number of books published and professional baseball games played have both increased, so that they are positively correlated. Does this mean that, if there is a players' strike, book publishing will take a nose dive?
- **Fig. 4.4** illustrates some of the problems that can arise when using linear regression and correlation coefficients to describe relationships between two data sets. This set of four examples is famously known as *Anscombe's Quartet*, as each panel has exactly the same correlation coefficient of $r = 0.82$.

4.2.2 Sampling Theory of Correlation (Pearson's correlation)

4.2.2.1 Statistical significance of correlations

The correlation, r, between two time series, $x(t)$ and $y(t)$, gives a measure of how well the two time series vary linearly with one another (or do not). But how can you tell whether the correlation you calculate is significantly different from zero? In this section we will review the techniques for testing the statistical significance of correlation coefficients.

Suppose we have N pairs of values (x_i, y_i) from which we have calculated a sample correlation coefficient r. The theoretical true value is denoted by ρ. For now, we will assume that we are sampling x and y from Normal distributions.

When the true correlation coefficient is zero, that is, when $\rho = 0$, the distribution of r is symmetric about zero and we are able to make use of the z- and t-statistic. Namely, the random variable t

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \tag{4.40}$$

will follow the t distribution with degrees of freedom $\nu = N - 2$.

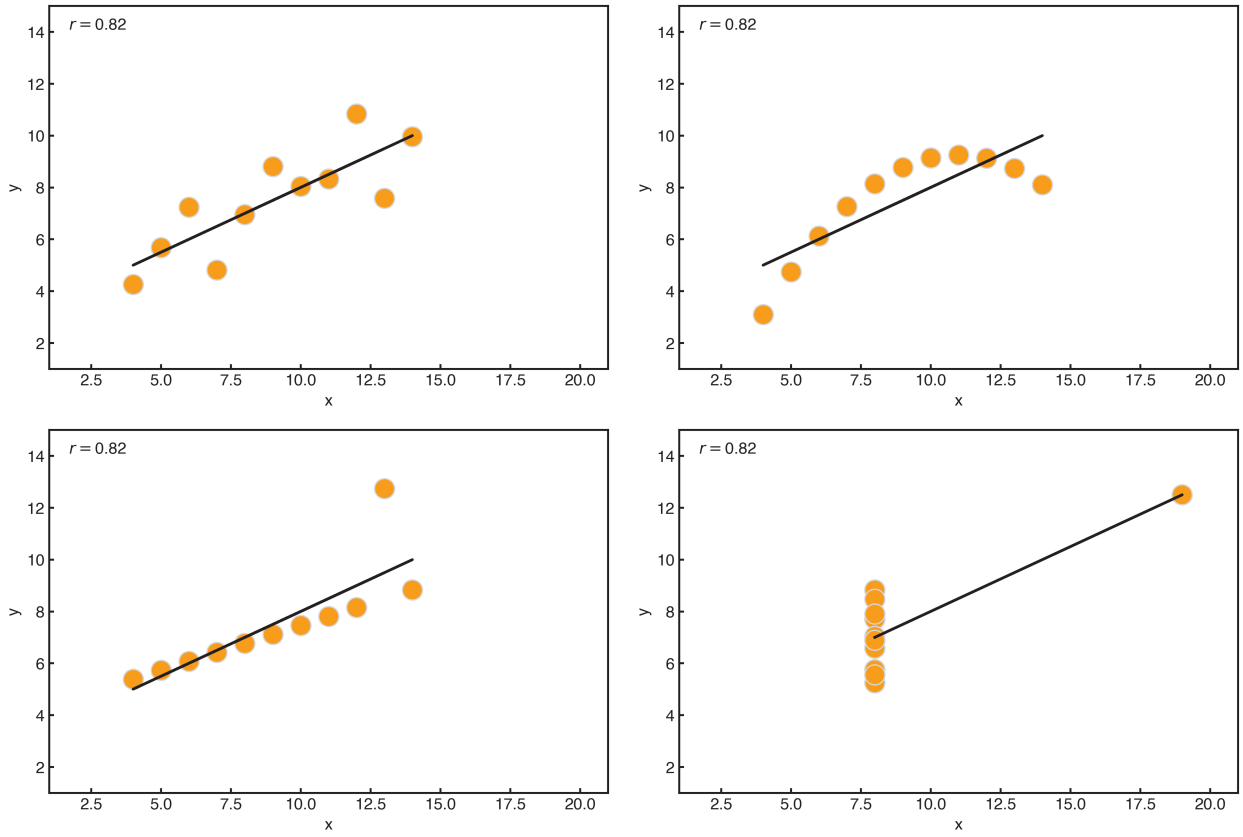

Figure 4.4 Four sets of data, known as *Anscombe's Quartet* where the correlations are all $r = 0.82$.

Worked Example 4.2.
We have two time series, each of length $N = 20$, and they are correlated at $r = 0.6$. Does this correlation exceed the 95% confidence interval under the null hypothesis that $\rho = 0$? You can assume both time series are sampled from underlying normal distributions and that the 20 observations in each data set represent 20 degrees of freedom.

We had no prior knowledge (before getting the samples) that the correlation would be positive or negative, so we will use a two-tailed t-test.

$t_c = 2.1$ for $\nu = N - 2 = 18$, so we want to know if the sample statistic $t > 2.1$.

$$t = \frac{0.6\sqrt{20-2}}{\sqrt{1-.6^2}} = 3.18. \tag{4.41}$$

Since $t = 3.18 > 2.1$, we can reject the null hypothesis that the true correlation is zero at 95% confidence.

In Practice.

- It turns out that the t-statistic is only applicable for $\rho = 0$ if the underlying distributions of the data are both normal, or if N is big enough that the central limit theorem applies. For well behaved distributions, a good rule of thumb is that an $N > 20$ should be sufficient for the central limit theorem to apply and the t-statistic to be appropriate for testing the null hypothesis that $\rho = 0$. Examples of the t values obtained from a range of distributions is given in **Fig. 4.5**.

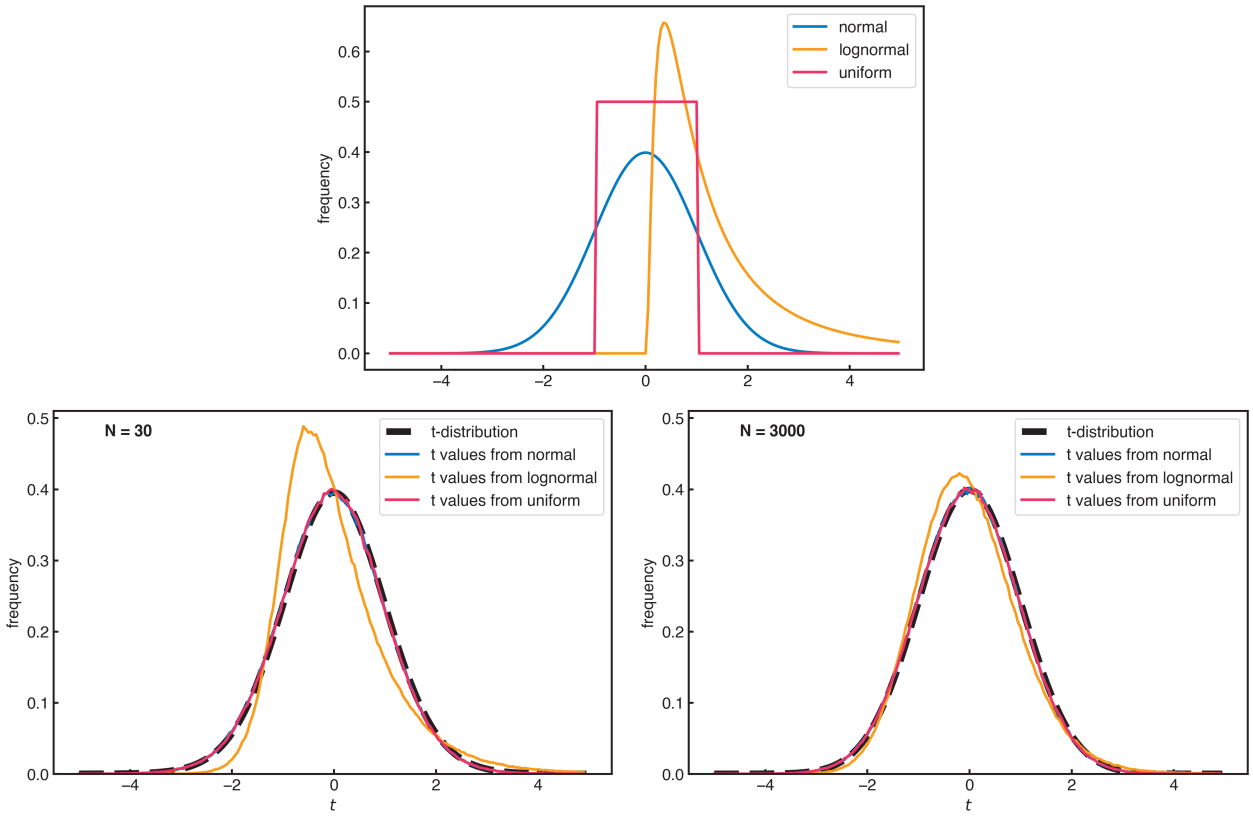

Figure 4.5 Three underlying sampling distributions and the resulting distribution of t values **(4.40)** computed from correlations obtained using $N = 30$ and $N = 3000$. The theoretical t-distribution is denoted by the black dashed line.

When the true correlation coefficient is not expected to be zero (i.e. $\rho \neq 0$), we cannot assume that the sampled correlations r will come from a symmetric, normal distribution. Instead, the distribution will be skewed due to the fact that correlations cannot exceed -1 or 1. In this instance, we must use the *Fisher-Z Transformation* to convert the distribution of r into something that is normally distributed (Z).

$$Z = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right) \tag{4.42}$$

The Fisher-Z statistic is then normally distributed with the following mean and standard deviation:

$$\mu_Z = \frac{1}{2} \ln \left(\frac{1+\rho}{1-\rho} \right) \quad \sigma_Z = \frac{1}{\sqrt{N-3}} \tag{4.43}$$

Thus, the confidence bounds for Z become

$$Z - t_c \sigma_Z \leqslant \mu_Z \leqslant Z + t_c \sigma_Z \tag{4.44}$$

If you have μ_Z and want the corresponding actual correlation ρ, you can use the following handy transformation

$$\rho = \frac{e^{2\mu_Z} - 1}{e^{2\mu_Z} + 1} = \frac{e^{\mu_Z} - e^{-\mu_Z}}{e^{\mu_Z} + e^{-\mu_Z}} = \tanh(\mu_Z) \tag{4.45}$$

Worked Example 4.3.
What are the 95% confidence limits on the true correlation ρ if you draw 21 samples from a normal distribution and obtain and $r = 0.8$?

Since we want the confidence bounds, we need to employ the Fisher-Z transformation in (4.42)

$$Z = \frac{1}{2} \ln\left(\frac{1 + 0.8}{1 - 0.8}\right) = 1.0986 \tag{4.46}$$

$$\sigma_Z = \frac{1}{\sqrt{21 - 3}} = .235 \tag{4.47}$$

Calculating $t_{0.025} = 2.1$ (using $\nu = 21 - 3$) leads to:

$$Z - 2.1\sigma_Z \leqslant \mu_Z \leqslant Z + 2.1\sigma_Z \tag{4.48}$$

$$0.61 \leqslant \mu_Z \leqslant 1.59 \tag{4.49}$$

While interesting, knowing μ_Z is not very helpful unless we convert it back to a correlation. So, plugging the bounds into (4.45) leads to

$$0.54 \leqslant \rho \leqslant 0.92 \tag{4.50}$$

Tests for the significance of the difference between two non-zero correlation coefficients are made by applying the Z statistic using the fact that it is normally distributed. For example, suppose we have two samples, one of size N_1 and one of size N_2, and each produce a correlation coefficient of r_1 and r_2, respectively. We test for a significant difference between these correlations by first calculating the Fisher-Z transformations for each:

$$Z_1 = \frac{1}{2} \ln\left(\frac{1 + r_1}{1 - r_1}\right); \quad Z_2 = \frac{1}{2} \ln\left(\frac{1 + r_2}{1 - r_2}\right) \tag{4.51}$$

From these we can calculate the typical z-score from

$$z = \frac{Z_1 - Z_2 - \Delta_{1,2}}{\sigma_{1,2}} \tag{4.52}$$

where

$$\Delta_{1,2} = \mu_1 - \mu_2 \tag{4.53}$$

is the transformed difference you expect (your null hypothesis). If you null hypothesis is that the true correlations of the two samples are equal (i.e. $\rho_1 = \rho_2$), then $\Delta_{1,2} = 0$. The denominator in (4.52) is given by

$$\sigma_{1,2} = \sqrt{\sigma_1^2 + \sigma_2^2} = \sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}} \tag{4.54}$$

4.2.2.2 Spearman's rank correlation

Spearman's rank correlation is a nonparametric test that determines whether a set of paired data monotonically vary together, but it is not concerned with the actual amplitude of the variations, just the ranks of the values. Since this is a nonparametric test, no assumption about normality needs to be made.

The idea is very simple, the original paired data x_i and y_i get converted into ranks (position in a sorted list) X_i and Y_i and Spearman's rank correlation ρ_R is given by

$$\rho_R = \frac{\sum_i (X_i - \overline{X_i})(Y_i - \overline{Y_i})}{\sqrt{\sum_i (X_i - \overline{X_i})^2 (Y_i - \overline{Y_i})^2}} \tag{4.55}$$

When computing the ranks, if there are duplicate values the ranks are equal to the average rank/position.

The standard error of Spearman's rank correlation ρ is given by

$$\sigma_\rho = \frac{0.6325}{(N-1)^{1/2}} \tag{4.56}$$

For significance testing on ρ_R, one can use the Fisher-Z test or the t-test (when the null hypothesis is that $\rho_R = 0$) as for the standard Pearson correlation.

There are many other nonparametric methods for calculating correlations, for example, Kendall's Tau Rank Correlation. We will not delve into these here.

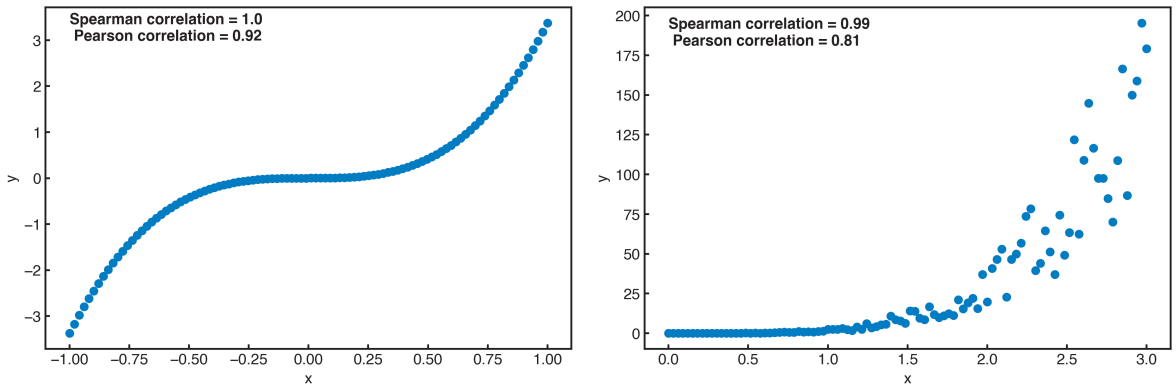

Figure 4.6 Paired data and their Spearman and Pearson correlations.

Worked Example 4.4.

Given paired data x and y, one can calculate Spearman's rank correlation using (4.55). For the paired data in the table below:

Spearman rank correlation: 0.99
Pearson correlation: 0.81

x	y	rank X	rank Y
1.04	1.39	2	1
1.46	6.78	6	5
1.03	2.21	1	2
1.66	13.46	7	8
1.29	6.3	4	4
1.70	11.31	8.5	6
1.27	4.37	3	3
1.70	20.42	8.5	9
1.97	22.22	10	10
1.43	11.81	5	7

4.3 Multiple Linear Regression

4.3.1 Generalized Normal Equations

Multiple regression is the regression of more than two variables. The basic idea is that you wish to use multiple predictors x_i to predict your predictand y. That is, you wish to find the regression coefficients a_i that provide the best guess $\widehat{y}$ for your predictand y

$$\widehat{y} = a_0 + a_1x_1 + a_2x_2 + ... + a_nx_n \tag{4.57}$$

For a single predictor x, we wanted to minimize the cost function Q, defined as:

$$Q = \sum_i^N (\widehat{y}_i - y_i)^2 = \sum_i^N (a_1x_i + a_0 - y_i)^2 \tag{4.58}$$

For the multiple predictor case (predictors $x_1, x_2, x_3..., x_n$), we want to minimize

$$Q = \sum_i^N (\widehat{y}_i - y_i)^2 = \sum_i^N (a_0 + a_1x_{1,i} + a_2x_{2,i} + a_3x_{3,i} + ... + a_nx_{n,i} - y_i)^2 \tag{4.59}$$

where n is the number of predictors and N is the number of time steps. Thus, $x_{2,i}$ denotes the predictor x_2 at time step i.

For n predictors, we have $n+1$ equations derived by setting

$$\frac{\partial Q}{\partial a_i} = 0 \tag{4.60}$$

where i goes from 0 to n.

$$\overline{y} = a_0 + a_1\overline{x_1} + a_2\overline{x_2} + ... + a_n\overline{x_n} \tag{4.61}$$

$$\overline{x_1 y} = a_0\overline{x_1} + a_1\overline{x_1^2} + a_2\overline{x_1 x_2} + ... + a_n\overline{x_1 x_n} \tag{4.62}$$

$$\overline{x_2 y} = a_0\overline{x_2} + a_1\overline{x_2 x_1} + a_2\overline{x_2^2} + ... + a_n\overline{x_2 x_n} \tag{4.63}$$

$$... \tag{4.64}$$

$$\overline{x_n y} = a_0\overline{x_n} + a_1\overline{x_n x_1} + a_2\overline{x_n x_2} + ... + a_n\overline{x_n^2} \tag{4.65}$$

If we assume the mean has been removed from every variable, these simplify to n equations and n unknowns (since we now know that $a_0 = 0$ and so **(4.61)** is no longer useful).

For the jth equation:

$$\overline{x_j y} = \sum_{i=1}^{n} a_i \overline{x_j x_i} \tag{4.66}$$

One can write this in matrix form as:

$$\begin{bmatrix} \overline{x_1^2} & \overline{x_1 x_2} & \overline{x_1 x_3} & ... \\ \overline{x_2 x_1} & \overline{x_2^2} & \overline{x_2 x_3} & ... \\ \overline{x_3 x_1} & \overline{x_3 x_2} & \overline{x_3^2} & ... \\ ... & ... & ... & ... \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ ... \end{bmatrix} = \begin{bmatrix} \overline{x_1 y} \\ \overline{x_2 y} \\ \overline{x_3 y} \\ ... \end{bmatrix} \tag{4.67}$$

Since we have removed the means of all variables, the overbarred quantities are actually covariances. These covariances are closely related to the variance calculations from Chapter 2. If x and y are scalars, then the covariance C_{xy} is

$$C_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y}), \tag{4.68}$$

so if $\overline{x} = \overline{y} = 0$ then $\overline{xy} = C_{xy}$. The correlation between x and y is computed by dividing the covariance by the standard deviations of both variables:

$$r_{xy} = \frac{C_{xy}}{\sigma_x \sigma_y} \tag{4.69}$$

All of these manipulations can be done much more neatly in vector/matrix notation, and the extension to the case where y is a vector is straightforward in that context (see next section).

Using our knowledge of the covariance, we now see that if the means of our variables have been subtracted, the left-hand-side of **(4.67)** is a matrix of the covariances (termed the *covariance matrix* across all of the x_i's and the right-hand-side is a vector of the covariances between x_i and y. In this case, each horizontal line in **(4.67)** can be written in matrix notation as

$$\mathbf{C_{x_i x_j}} a_j = \mathbf{C_{x_i y}} \tag{4.70}$$

Since the ultimate goal is to determine the a_j coefficients, one can solve for this vector by inverting the real, symmetric matrix on the left, and multiplying the inverse times the vector on the right, at least in theory.

$$\mathbf{C_{x_i x_j}}^{-1}\mathbf{C_{x_i x_j}} a_j = \mathbf{C_{x_i x_j}}^{-1} C_{x_i y} \tag{4.71}$$

$$a_j = \mathbf{C_{x_i x_j}}^{-1} C_{x_i y} \tag{4.72}$$

However, many of the methods for computing the inverse of the covariance matrix require that $\mathbf{C_{x_i x_j}}$ j be invertible and not singular. In the following chapters we will discuss how singular value decomposition can be used to derive a very robust solution for the a_j's that is optimal even when the problem is over-determined and $\mathbf{C_{x_i x_j}}$ is singular.

In Practice.

- if each variables has been standardized (mean of 0 and standard deviation of 1), the left-hand-side of (4.67) is the *correlation matrix* of the x_j's, and the right-hand-side is the *correlation vector* between the x_j's and y.
- if the x_j's are time series at different locations in a data set, the covariance matrix yields information about the structures of the dominate data, and tells you something about the spatial variability of the different points
- if the predictors are linearly independent, the off diagonal elements are all 0 and the a_j's can be found algebraically

4.3.2 Derivation of the Normal Equations using Matrix Notation

Matrix notation is very powerful and compact for doing complex minimization problems and we will need to use it a lot to do more powerful methods later. As an example, then, let's derive (4.67) using matrix algebra. First some definitions.

Let's think of $\mathbf{y}$ and $\mathbf{a}$ as row vectors of length N and n, respectively, and the data matrix $\mathbf{X}$ as an $N \times n$ matrix, where N is the sample size and n is the number of predictors, x_j, as before.

$$\mathbf{y} = \begin{bmatrix} y_1 & y_2 & y_3 & \dots & y_N \end{bmatrix} \tag{4.73}$$

$$\mathbf{a} = \begin{bmatrix} a_1 & a_2 & a_3 & \dots & a_n \end{bmatrix} \tag{4.74}$$

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{21} & x_{31} & \dots & x_{N1} \\ x_{12} & x_{22} & x_{32} & \dots & x_{N2} \\ x_{13} & x_{23} & x_{33} & \dots & x_{N3} \\ \dots & \dots & \dots & \dots & \dots \\ x_{1n} & x_{2n} & x_{3n} & \dots & x_{Nn} \end{bmatrix} \tag{4.75}$$

Now we can express our desired regression equation in a very compact form

$$\widehat{\mathbf{y}} = \mathbf{aX} \tag{4.76}$$

where we get the vector of predicted values of y, $\widehat{\mathbf{y}}$, by multiplying the vector of coefficients $\mathbf{a}$ by the data matrix $\mathbf{X}$.

Our goal is to determine values for the vector of coefficients $\mathbf{a}$, and we do this by minimizing the squared error of our fit (i.e. the cost function Q). In matrix notation, we compute Q by taking the inner product of the error vector with itself

$$Q = (\mathbf{y} - \mathbf{aX})\,(\mathbf{y} - \mathbf{aX})^T \tag{4.77}$$

Here, $(\cdot)^T$ indicates the transpose of a matrix, and we will utilize the fact that $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$. Expanding the right-hand-side of (4.77) leads to

$$Q = \mathbf{yy}^T - \mathbf{yX}^T\mathbf{a}^T - \mathbf{aXy}^T + \mathbf{aXX}^T\mathbf{a}^T \tag{4.78}$$

The next step is to differentiate Q with respect to the coefficients a_j to obtain an equation for the values of $\mathbf{a}$ that minimize the error. Doing this leads to

$$\frac{\partial Q}{\partial \mathbf{a}} = \mathbf{0} - \mathbf{yX}^T - \mathbf{Xy}^T + \mathbf{XX}^T\mathbf{a}^T + \mathbf{aXX}^T \tag{4.79}$$

$$= \left(\mathbf{aXX}^T - \mathbf{yX}^T\right) + \left(\mathbf{aXX}^T - \mathbf{yX}^T\right)^T \tag{4.80}$$

Note that the right hand side of (4.80) can be organized into two terms that are the transposes of each other. If a quantity is zero, then its transpose is also zero. Therefore, we can use either of the two forms above to express the minimization. We will carry along both forms in the next couple of equations, although they mean the same thing.

We obtain the optimal solution for the a_j's that minimizes the error, Q, by setting the right hand side of (4.80) equal to zero, or,

$$\mathbf{aXX^T} = \mathbf{yX^T} \text{ or } \mathbf{XX^Ta^T} = \mathbf{Xy^T} \tag{4.81}$$

from which,

$$\mathbf{a} = \mathbf{yX^T}\left(\mathbf{XX^T}\right)^{-1} \text{ or } \mathbf{a^T} = \left(\mathbf{XX^T}\right)^{-1}\mathbf{Xy^T} \tag{4.82}$$

Looking back at (4.72), we see that it is equivalent to $\mathbf{a^T} = \left(\mathbf{XX^T}\right)^{-1}\mathbf{Xy^T}$ since

$$\mathbf{XX^T} = N\mathbf{C_{x_ix_j}} \text{ and } \mathbf{Xy^T} = N\mathbf{C_{x_iy}} \tag{4.83}$$

Worked Example 4.5.
You may consider Fourier harmonic analysis to be a special case of a multiple linear least-squares regression model. In this case, the predictors are sines and cosines in sampling dimension z of length N. For example:

$$x_1 = \sin\frac{2\pi z}{L}; x_2 = \cos\frac{2\pi z}{L}; x_3 = \sin\frac{4\pi z}{L}; x_4 = \cos\frac{4\pi z}{L}; \ldots \tag{4.84}$$

If you are unfamiliar with Fourier analysis, you may want to come back to this section after studying the description of Fourier analysis in Chapter 7.

If we take a multiple linear regression approach, this technique will work for unevenly spaced z_i, whereas standard Fourier Transform techniques will not. For evenly spaced data (evenly spaced z_i) and orthogonal predictors, as is the case for these sines and cosines,

$$a_j = \frac{\overline{x_j y}}{\overline{x_j^2}}; \text{ but } \overline{x_j^2} = \frac{1}{2} \text{ for all } N > 0 \tag{4.85}$$

so that

$$a_j = \frac{2}{N}\sum_{i=1}^{N} y_i \cdot x_j(z_i) \tag{4.86}$$

for example (4.87)

$$a_1 = \frac{2}{N}\sum_{i=1}^{N} y_i \cdot \sin\left(\frac{2\pi z_i}{L}\right) \tag{4.88}$$

and these coefficients are equivalent to what is used in Fourier decomposition, demonstrating that Fourier analysis is optimal in a least-squares sense.

4.3.3 Multiple Regression - How many predictors should I use?

Multiple regression allows for one to use nearly an infinite number of predictors to predict y. However, the question is then *"how many predictors should I use?"*. To make things a bit easier, in this section we will consider standardized variables, although one should keep in mind that all equations can be rewritten without this assumption.

In the case of standardized variables, the normal equations for multiple linear-least-squares regression can be written in the following way

$$r_{x_i x_j} a_i = r_{x_j y} \tag{4.89}$$

where once again r represents the correlation. We start with the simplest case of only two predictors

$$\widehat{y} = a_1 x_1 + a_2 x_2 \tag{4.90}$$

Then the normal equations can be expanded as

$$r_{x_1 x_1} a_1 + r_{x_1 x_2} a_2 = r_{x_1 y} \tag{4.91}$$

$$r_{x_2 x_1} a_1 + r_{x_2 x_2} a_2 = r_{x_2 y} \tag{4.92}$$

or in matrix notation,

$$\begin{bmatrix} r_{x_1 x_1} & r_{x_1 x_2} \\ r_{x_2 x_1} & r_{x_2 x_2} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} r_{x_1 y} \\ r_{x_2 y} \end{bmatrix} \tag{4.93}$$

But, since $r_{x_1 x_1} = r_{x_2 x_2} = 1$ and $r_{x_1 x_2} = r_{x_2 x_1}$, this can be rewritten as

$$\begin{bmatrix} 1 & r_{x_1 x_2} \\ r_{x_1 x_2} & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} r_{x_1 y} \\ r_{x_2 y} \end{bmatrix} \tag{4.94}$$

We solve for a_1 and a_2 and find that

$$a_1 = \frac{r_{x_1,y} - r_{x_1,x_2} r_{x_2,y}}{1 - r_{x_1,2}^2} \tag{4.95}$$

$$a_2 = \frac{r_{x_2,y} - r_{x_1,x_2} r_{x_1,y}}{1 - r_{x_1,2}^2} \tag{4.96}$$

If $\widehat{y}$ is the best-fit, then we can write the explained and unexplained variance as

$$\overline{y^2} = \overline{(y_i - \widehat{y})^2} + \overline{(\widehat{y} - \overline{y})^2} \tag{4.97}$$

$$\text{Total Variance} = \text{Unexplained Variance} + \text{Explained Variance}$$

Using the fact that $\widehat{y} = a_1 x_1 + a_2 x_2$ it can be shown that

$$1 = \frac{\overline{(y_i - \widehat{y})^2}}{\overline{y^2}} + R^2 \tag{4.98}$$

where the fraction of explained variance R^2 is given by

$$R^2 = \frac{r_{x_1,y}^2 + r_{x_2,y}^2 - 2 r_{x_1,y} r_{x_2,y} r_{x_1,x_2}}{1 - r_{x_1,x_2}^2} \tag{4.99}$$

In analogy with the case of simple regression, R can be defined as the multiple correlation coefficient, since its square is the fraction of explained variance.

It turns out that in multiple regression, if too many predictors are used, then the predictions associated with the regression will perform badly on independent data—worse than if fewer predictors were used in the first place. This is because using too many predictors can result in large coefficients for variables that are not actually highly correlated with the predictand. These coefficients help to fit the dependent data, but make the application to independent data unstable and potentially wildly in error. That is because you start to fit the noise, and when the noise changes the prediction is really bad. Also, sometimes these variables are better correlated with each other than they are with the predictand, which will also produce unstable predictions when used with independent data. In this case the covariance matrix you formally invert (i.e. **(4.72)**) is nearly singular.

Adding x_2 as a predictor does not always increase the explained variance. No benefit is derived from additional predictors, unless their correlation coefficient with the predictand exceeds the *minimum useful correlation* - the critical correlation required for a beneficial effect increases with the number of predictors used. Unless predictors can be found that are well correlated with the predictand and relatively uncorrelated with the other predictors, the optimum number of predictors will usually be small. The minimal useful correlation is defined as the minimum correlation between x_2 with y that will allow the addition of x_2 to improve the regression R^2. In math, this is,

$$|r_{x_2 y}|_{\text{min useful}} > |r_{x_1 y} r_{x_1 x_2}| \tag{4.100}$$

We can show this by substituting $r_{x_2 y} = r_{x_2 y_{\text{min useful}}} = r_{x_1 y} r_{x_1 x_2}$ into the expression for the explained fraction of the variance in the two-predictor case:

$$R^2 = \frac{r^2_{x_1,y} + r^2_{x_2,y} - 2 r_{x_1,y} r_{x_2,y} r_{x_1,x_2}}{1 - r^2_{x_1,x_2}} \tag{4.101}$$

$$= \frac{r^2_{x_1,y} + r^2_{x_2,y} - 2 r^2_{x_1,y} r^2_{x_1,x_2}}{1 - r^2_{x_1,x_2}} \tag{4.102}$$

$$= r_{x_1 y} \tag{4.103}$$

Thus, we have shown that when $r_{x_2 y}$ equals the minimum useful correlation, including the second predictor has no influence on the explained variance. What is not obvious at this point is that including such a useless predictor can actually have a detrimental effect on the performance of the prediction equation when applied to independent data, data that were not used in the original regressions. Note that the lower the value of $r_{x_1 x_2}$, that is, the more independent the predictors, the better chance that both predictors will be useful, assuming that they are both correlated with the predictand. Ideally we would like completely independent predictors, i.e. $r_{x_1 x_2} = 0$. Completely dependent predictors, $r_{x_1 x_2} = 1$, are useless since only one of them is enough (although you can usually reduce the noise by adding them together with some judicious weighting). The desire for independent predictors is part of the motivation for empirical orthogonal functions (EOFs), which will be described in Chapter 5.

Similar, but more complicated considerations apply when deciding to use a third predictor. In general, the more predictors used, the fewer degrees of freedom are inherent in the coefficients a_j, the lower the statistical significance of the "fit" to the data points, and the less likely that the regression equations will work equally well on independent data. If predictors are added indiscriminately, you come to a point where adding predictors makes the regression work less well on independent data, even though you are accounting for more of the variance of the dependent data set. This is because you can *over fit* the data, in essence, you will use the predictors to fit the noise rather than the signal. It is a good idea to use as few predictors as possible, while still getting most of the skill you can. Later we will describe how to pick the optimal set of predictors.

Worked Example 4.6.
Say you have two predictors, x_1 and x_2 and both are correlated with the predictand y at 0.5, and are correlated with each other at 0.5, that is

$$r_{1,y} = r_{2,y} = r_{1,2} = 0.5 \tag{4.104}$$

Does adding x_2 increase your R^2 compared to a regression with x_1 alone?

For the first predictor only, the variance explained is

$$R_1^2 = r_{1,y}^2 = 0.25 \tag{4.105}$$

Adding a second predictor, x_2, leads to

$$R_{1,2}^2 = \frac{0.5^2 + 0.5^2 - 2 \times 0.5 \times 0.5 \times 0.5}{1 - 0.5^2} = 0.33 \tag{4.106}$$

Thus, adding a second predictor helps explain more of the variance of y.

Now let us assume that $r_{2,y} = 0.25$ and everything else remains the same. Adding the second predictor leads to

$$R_{1,2}^2 = \frac{0.5^2 + 0.25^2 - 2 \times 0.5 \times 0.25 \times 0.5}{1 - 0.5^2} = 0.25 \tag{4.107}$$

In this case, adding the second predictor does not increase the explained variance!

4.3.3.1 Adjusted R^2

In most situations, increasing the number of predictors will always increase the explained variance (R^2) because at some point the predictors will start fitting the noise rather than the signal. This will become evident when the regression model is applied to independent data - as the model will be worse than a model with fewer predictors.

How do you know when you are starting to fit the noise and thus should stop adding predictors? One tool for determining this is the *adjusted* R^2, which attempts to quantify when the additional variance explained by a new predictor is not enough to warrant its addition in the full model. The adjusted R^2 is defined as

$$\overline{R}^2 = 1 - (1 - R^2)\frac{n - 1}{n - p - 1} \tag{4.108}$$

where p is the number of predictors (not including a constant term) and n is the sample size.

Unlike R^2, $\overline{R}^2$ only increases with the addition of a new predictor when the increase in R^2 is more than what would be expected by chance. Thus, one can use the adjusted R^2 to determine the number of predictors by plotting $\overline{R}^2$ for each additional predictor and determining when it reaches a maximum. Beyond this maximum, additional predictors will likely only degrade the fit when independent data is analyzed.

Chapter 5
Seeking Structure in Data

5.1 Introduction

In this chapter we discuss the use of matrix methods from linear algebra, primarily as a means of searching for structure in data sets.

Empirical Orthogonal Function (EOF) analysis seeks structures that explain the maximum amount of variance in a two-dimensional data set. One dimension in the data set represents the dimension in which we are seeking to find structure, and the other dimension represents the dimension in which realizations of this structure are sampled. In seeking characteristic spatial structures that vary with time, for example, we would use space as the structure dimension and time as the sampling dimension. The analysis produces a set of structures in the first dimension, which we call the EOF's, and which we can think of as being the structures in the spatial dimension. The complementary set of structures in the sampling dimension (e.g. time) we can call the Principal Components (PC's), and they are related one-to-one to the EOF's. Both sets of structures are orthogonal in their own dimension. Sometimes it is helpful to sacrifice one or both of these orthogonalities to produce more compact or physically appealing structures, a process called rotation of EOF's.

Singular Value Decomposition (SVD) is a general decomposition of a matrix. It can be used on data matrices to find both the EOF's and PC's simultaneously. In SVD analysis we often speak of the left singular vectors and the right singular vectors, which are analogous in most ways to the empirical orthogonal functions and the corresponding principal components.

If SVD is applied to the covariance matrix between two data sets, then it picks out structures in each data set that are best correlated with structures in the other data set. They are structures that 'explain' the maximum amount of covariance between two data sets in a similar way that EOF's and PC's are the structures that explain the most variance in a data set. It is reasonable to call this Maximum Covariance Analysis (MCA).

Canonical Correlation Analysis (CCA) is a combination of EOF and MCA analysis. The two input fields are first expressed in terms of EOF's, the time series of PC's of these structures are then normalized, a subset of the EOF/PC pairs that explain the most variance is selected, and then the covariance (or correlation) matrix of the PC's is subjected to SVD analysis. So CCA is MCA of a covariance matrix of a truncated set of PC's. The idea here is that the noise is first reduced by doing the EOF analysis and so including only the coherent structures in two or more data sets. Then the time series of the amplitudes of these EOFs are normalized to unit variance, so that all count the same, regardless of amplitude explained or the units in which they are expressed. These time series of normalized PCs are then subjected to MCA analysis to see which fields are related.

5.2 Data Sets as Two-Dimensional Matrices

Imagine that you have a data set that is two-dimensional. The easiest example to imagine is a data set that consists of observations of several variables at one instant of time, but includes many realizations of these variable values taken at different times. The variables might be temperature and salinity at one point in the ocean taken every day for a year. Then you would have a data matrix that is 2 by 365; 2 variables measured 365 times. So one dimension is the variable and the other dimension is time. Another example might be measurements of the concentrations of 12 chemical species at 10 locations in the atmosphere. Then you would have a data matrix that is 12x10 (or 10x12). One can imagine several possible generic types of data matrices.

- A space-time array: Measurements of a single variable at M locations taken at N different times, where M and N are integers.
- A parameter-time array: Measurements of M variables (e.g. temperature, pressure, relative humidity, rainfall, . . .) taken at one location at N times.
- A parameter-space array: Measurements of M variables taken at N different locations at a single time.

You might imagine still other possibilities. If your data set is inherently three dimensional, then you can string two variables along one axis and reduce the data set to two dimensions. For example: if you have observations at L longitudes and K latitudes and N times, you can make the spatial structure into a big vector LxK=M long, and then analyze the resulting (LxK)xN=MxN data matrix. (A vector is a matrix where one dimension is of length 1, e.g. an 1xN matrix is a vector).

So we can visualize a two-dimensional data matrix X as follows:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & . . & x_{1N} \\ x_{21} & x_{22} & . . & x_{2N} \\ x_{31} & x_{32} & . . & x_{3N} \\ . & . & . . & . \\ x_{M1} & x_{M2} & . . & x_{MN} \end{bmatrix} = X_{ij}, \quad i = 1, M; \quad j = 1, N \tag{5.1}$$

So the state space, which might represent spatial grid points or a set of biological or chemical measurements has a dimension of M, which is the first index. This state vector is observed N times, which is the second index. So a column of this matrix represents the state of the system, and the N different columns represent different samples or times of measurement. So M is the state space and N is the sampling space.

Here we have included the symbolic bold $\mathbf{X}$ to indicate a matrix and the subscript notation to indicate the same matrix.

We define the transpose of the matrix by reversing the order of the indices to make it an NxM matrix.

$$\mathbf{X}^T = N \overbrace{\begin{bmatrix} . . \\ . . \\ . . \end{bmatrix}}^{M} = X_{ji}, \quad j = 1, N; \quad i = 1, M \tag{5.2}$$

In multiplying a matrix times itself we generally need to transpose it once to form an inner product, which results in two possible "dispersion" matrices.

$$\mathbf{X}\mathbf{X}^T = M \overbrace{\begin{bmatrix} . . . \\ . . . \end{bmatrix}}^{N} \overbrace{\begin{bmatrix} . . \\ . . \\ . . \end{bmatrix}}^{M} N = D_{ij}, \quad i = 1, M; \quad j = 1, M \tag{5.3}$$

Of course, in this multiplication, each element of the first row of $\mathbf{X}$ is multiplied times the corresponding element of the first column of $\mathbf{X}^T$, and the sum of these products becomes the first (first row, first column) element of $\mathbf{X}\mathbf{X}^T$. And so it goes on down the line for the other elements. This explains matrix multiplication for those who may be rusty on this. So the dimension that you sum over, in this case N, disappears and we get an M x M product matrix. In this projection of a matrix onto itself, one of the dimensions gets removed

and we are left with a measure of the dispersion of the structure with itself across the removed dimension (or the sampling dimension). If the sampling dimension is time, and we have removed the time mean of the data, then the resulting dispersion matrix is the matrix of the covariance of the spatial locations with each other, as determined by their variations in time. One can also compute the other dispersion matrix in which the roles of the structure and sampling variables are reversed.

$$\mathbf{X}^{\mathsf{T}}\mathbf{X} = N \overbrace{\begin{bmatrix} . & . \\ . & . \\ . & . \end{bmatrix}}^{M} \overbrace{\begin{bmatrix} . & . & . \\ . & . & . \end{bmatrix}}^{N} M, \quad i = 1, N; \quad j = 1, N \tag{5.4}$$

Both of the dispersion matrices obtained by taking inner products of a data matrix with itself are symmetric matrices. They become covariance matrices, if we divide by the sample size.

$$\mathbf{X}\mathbf{X}^{\mathsf{T}}/N = \mathbf{C} = C_{ij}, \quad i = 1, M; \quad j = 1, M \tag{5.5}$$

In the second case the covariance at different times is obtained by projecting on the sample of different spatial points. Either of these dispersion matrices may be used to study the temporal/spatial structure of data sets.

5.3 Empirical Orthogonal Functions

Suppose we wish to define a vectors $\mathbf{e}$ that has maximum similarity to a data set $\mathbf{X}$. To measure the similarity we can project the vector $\mathbf{e}$ onto the data set $\mathbf{X}$ as follows,

$$\mathbf{e}^{\mathsf{T}}\mathbf{X} = [e_1, e_2, ., ., e_M] \begin{bmatrix} x_{11} & x_{12} & . & . & x_{1N} \\ x_{21} & x_{22} & . & . & x_{2N} \\ x_{31} & x_{32} & . & . & x_{3N} \\ . & . & . & . & . \\ x_{M1} & x_{M2} & . & . & x_{MN} \end{bmatrix} = [q_1, q_2, ., ., q_N] = \mathbf{q} \tag{5.6}$$

We can make a scalar measure of the similarity by taking the inner product of this projection time series $\mathbf{q}$ with itself and dividing by the sample size, N, to make the measure of similarity independent of the sample size.

$$\mathbf{q}\mathbf{q}^{\mathsf{T}}/N = \mathbf{e}^{\mathsf{T}}\mathbf{X}\mathbf{X}^{\mathsf{T}}\mathbf{e}/N = \mathbf{e}^{\mathsf{T}}\mathbf{C}\mathbf{e} = \lambda \tag{5.7}$$

where λ is the similarity measure that we want to maximize, which has units of the square of $\mathbf{X}$. One final constraint we need to apply is to limit the length of $\mathbf{e}$, which we can do by requiring that it have a length of one, $\mathbf{e}\mathbf{e}^{\mathsf{T}} = 1$. The problem

$$\mathbf{e}^{\mathsf{T}}\mathbf{C}\mathbf{e} = \lambda, \quad \text{subject to,} \quad \mathbf{e}\mathbf{e}^{\mathsf{T}} = 1 \tag{5.8}$$

is the standard eigenanalysis problem, which can be applied to any symmetric matrix. Up to M such eigenvectors and eigenvalues can be found.

$$\mathbf{e_i}^{\mathsf{T}}\mathbf{C}\mathbf{e_i} = \lambda_i, \quad = 1, M, \tag{5.9}$$

The problem can be written in matrix form as follows.

$$\mathbf{E}^{\mathsf{T}}\mathbf{C}\mathbf{E} = \tag{5.10}$$

Where $\mathbf{E}$ is the matrix with the eigenvectors $\mathbf{e}_i$ as its columns, and Λ is the matrix with the eigenvalues λ_i, along its diagonal and zeros elsewhere.

$$\mathbf{E} = \begin{bmatrix} e_{11} & e_{12} & . & . & e_{1M} \\ e_{21} & e_{22} & . & . & e_{2M} \\ e_{31} & e_{32} & . & . & e_{3M} \\ . & . & . & . & . \\ e_{M1} & e_{M2} & . & . & e_{MM} \end{bmatrix} = E_{ij}, \quad i = 1, M; \quad j = 1, M \tag{5.11}$$

Each column in 5.11 is a distinct eigenvector, with the order chosen so that the first eigenvalue is the largest, so that the first eigenvector explains the most variance.

$$= \begin{bmatrix} \lambda_1 & 0 & . & . & 0 \\ 0 & \lambda_2 & . & . & 0 \\ 0 & 0 & \lambda_3 & . & 0 \\ . & . & . & . & . \\ 0 & 0 & . & . & \lambda_M \end{bmatrix} = \Lambda_{ij}, \quad i = 1, M; \quad j = 1, M \tag{5.12}$$

The set of eigenvectors, e_{ij}, and associated eigenvalues, λ_j, represent a coordinate transformation into a coordinate space where the covariance matrix $\mathbf{C}$ becomes diagonal. Because the covariance matrix is diagonal in this new coordinate space, the variations in these new directions are uncorrelated with each other, at least for the sample that has been used to construct the original covariance matrix. The eigenvectors define directions in the initial coordinate space along which the maximum possible variance can be explained, and in which variance in one direction is orthogonal to the variance explained by other directions defined by the other eigenvectors. The eigenvalues indicate how much variance is explained by each eigenvector. If you arrange the eigenvector/eigenvalue pairs with the biggest eigenvalues first, then you may be able to explain a large amount of the variance in the original data set with relatively few coordinate directions, which correspond to characteristic structures in the original structure space.

The sum of all the eigenvalues is the variance, so that the eigenvalue/divided by the sum of the eigenvalues gives the fraction of variance explained by each EOF. Since the EOFs are ordered from largest eigenvalue to smallest, plotting the eigenvalues as a function of their index provides a measure of how effective the eigenanalysis has been in explaining the variance with a small number of structures.

5.3.1 Two-Dimensional Example

It is simplest to visualize EOFs in two-dimensions as a coordinate rotation that maximizes the efficiency with which variance is explained. Consider the following scatter plot of paired data (x_1, x_2). The eigenvectors are shown as lines in this plot. The first one points down the axis of the most variability, and the second is orthogonal to it.

5.3.2 EOF/Principal Component Analysis - Introduction

In this section we will talk about what is called Empirical Orthogonal Function (EOF), Principle Component Analysis (PCA), or Factor Analysis, depending on the tradition in the discipline of interest. EOF analysis follows naturally from the preceding discussion of regression analysis and linear modeling, where we found that correlations between the predictors causes them to be redundant with each other and causes the regression equations involving them to perform poorly on independent data. EOF analysis allows a set of predictors to be rearranged into a new set of predictors that are orthogonal with each other and that maximizes the amount of variance in the dependent sample that can be explained with the smallest number of EOF predictors. It was in this context that Lorenz (1956) introduced EOF's into the meteorological literature. The same mathematical tools are used in many other disciplines, under a variety of different names. In addition to providing better predictors for statistical forecasting, EOF analysis can be used to explore the structure of the variability within a data set in an objective way, and to analyze relationships within a set of variables.

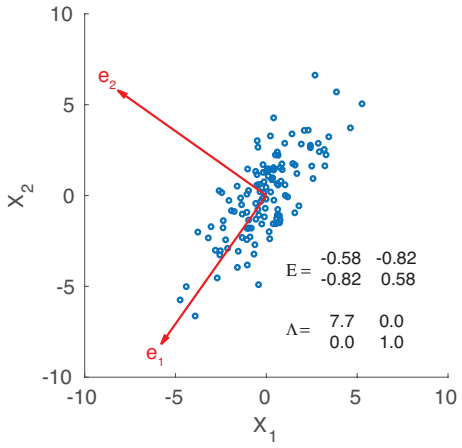

Figure 5.1 Scatter plot of two variables x_2 vs x_1. Eigenvectors are shown as red arrows and are elongated by a factor of ten for better viewing. Eigenvector matrix $\mathbf{E}$ and eigenvalue matrix Λ are shown as arrays of numbers.

Examples include searching for characteristic spatial structures of disturbances and for characteristic relations between parameters. The relationships between parameters may be of scientific interest in themselves, quite apart from their effect on statistical forecasting. The physical interpretation of EOFs is tricky, however. They are constructed from mathematical constraints, and may not have any particular physical significance. No clear-cut rules are available for determining when EOFs correspond to physical entities, and their interpretation always requires judgment based on physical facts or intuition.

One area where EOF analysis is useful is in fitting a line to data. In ordinary least squares the line that is obtained depends on which variable is chosen to be the independent variable and which is chosen to be the dependent variable, so long as the data are not perfectly colinear. EOF analysis minimizes the perpendicular distance from the line and is not dependent on which variable is viewed as independent. Generally, both variables are noisy so that EOF line fitting is more robust.

5.4 Principal Components and EOFs

The EOFs form an orthogonal coordinate system that is a rotation of the original coordinate system. It is convenient to order the eigenvalues and eigenvectors in order of decreasing magnitude of the eigenvalue. The first eigenvector thus has the largest and explains the largest amount of variance in the data set used to construct the covariance matrix. In this new coordinate system, each data point is defined by a set of distances in the new coordinate system. To get these distances we project the eigenvectors onto the original data. We will call these new distances the Principal Components (PCs). To obtain the PCs we project the eigenvectors onto the original data, since the eigenvectors are directions defined in the original coordinate space. If we define the PC matrix as $\mathbf{Z}$, then the projection to obtain $\mathbf{Z}$ is a simple matrix multiplication.

$$\mathbf{Z} = \mathbf{E}^{\mathsf{T}}\mathbf{X} \tag{5.13}$$

Because the eigenvectors are orthogonal,

$$\mathbf{E}^{\mathsf{T}}\mathbf{E} = \mathbf{I} \tag{5.14}$$

where $\mathbf{I}$ is the identity matrix, with ones down the diagonal and zeros off the diagonal. Because of the orthogonality of the EOFs we can multiply (5.13) on the left by $\mathbf{E}$ and obtain an equation for the original data in terms of the PC matrix $\mathbf{Z}$.

$$\mathbf{X} = \mathbf{E}\mathbf{Z} \tag{5.15}$$

So it is easy (EZ) to get the original data back from the PCs since they are both expressions for the same data set, but in different reference frames or coordinate systems. The PC matrix $\mathbf{Z}$ has the same shape and information as the original data matrix $\mathbf{X}$ (5.1), but the information is expressed in a different coordinate system, which is defined by the EOF directions.

$$\mathbf{Z} = \begin{bmatrix} z_{11} & z_{12} & . . & z_{1N} \\ z_{21} & z_{22} & . . & z_{2N} \\ z_{31} & z_{32} & . . & z_{3N} \\ . & . & . . & . \\ z_{M1} & z_{M2} & . . & z_{MN} \end{bmatrix} = Z_{ij}, \quad i = 1, M; \quad j = 1, N \tag{5.16}$$

Now the columns represent the state of the system at different times, but in the new, more efficient, coordinate system.

5.4.1 Orthogonality of the Principle Components

Now we can easily show that the principle component time series, the time series of the amplitudes of each eigenvector, are uncorrelated in the sample space (*e.g.* time). The covariance matrix of the PCs is written as,

$$\mathbf{C}_Z = \mathbf{Z}\mathbf{Z}^T/N \tag{5.17}$$

Substituting in the expression for $\mathbf{Z}$ from (5.13), and using (5.10) we get,

$$\mathbf{C}_Z = \mathbf{Z}\mathbf{Z}^T/N = \mathbf{E}^T\mathbf{X}\mathbf{X}^T\mathbf{E}/N = \mathbf{E}^T\mathbf{C}\mathbf{E} = \tag{5.18}$$

From which we see that the covariance matrix of the PCs is also diagonal and equal to the eigenvalue matrix of the eigenanalysis.

The PCs are useful for prediction, since the PC time series are uncorrelated with each other, thus they share no variance between them. Truncating the EOF and PC representation of the data by removing EOFs that explain a small amount of variance may allow one to construct a set of predictors for which noise is reduced and the predictors are uncorrelated in the sample space. Noise is reduced if one assumes that the signal is correlated in space, so that eliminating EOFs that explain a small amount of variance is removing noise and not signal.

5.4.2 EOF Analysis via Singular Vector Decomposition

EOF/PC analysis can also be done by direct singular value decomposition of the data matrix $\mathbf{X}$ instead of doing an eigenanalysis of the covariance matrix. If we take the two-dimensional data matrix of structure (e.g. space) versus sampling (e.g. time) dimension, and do direct singular value decomposition of this matrix, we recover the EOFs, eigenvalues, and normalized PC's directly in one step. If the data set is relatively small, this may be easier than computing the dispersion matrices and doing the eigenanalysis of them. If the sample size is large, it may be computationally more efficient to use the eigenvalue method. Remember first our definition of SVD of a matrix:

Singular Value Decomposition: Any m by n matrix $\mathbf{X}$ can be factored into

$$\mathbf{X} = \mathbf{U}\ \mathbf{V}^T \tag{5.19}$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal and is diagonal. The columns of $\mathbf{U}$ (m by m) are the eigenvectors of $\mathbf{X}\mathbf{X}^T$, and the columns of $\mathbf{V}$ (n by n) are the eigenvectors of $\mathbf{X}^T\mathbf{X}$. The r singular values on the diagonal of (m by n) are the square roots of the nonzero eigenvalues of both $\mathbf{X}\mathbf{X}^T$ and $\mathbf{X}^T\mathbf{X}$.

So we suppose that the data matrix $\mathbf{X}$ is MxN, where M is the space or structure dimension and N is the time or sampling dimension. More generally, we could think of the dimensions as the structure dimension M and the sampling dimension N, but for concreteness and brevity let's call them space and time. Now $\mathbf{X}\mathbf{X}^\mathsf{T}$ is the dispersion matrix obtained by taking an inner product over time, leaving the covariance between spatial points. Thus the eigenvectors of $\mathbf{X}\mathbf{X}^\mathsf{T}$ are the spatial eigenvectors, and appear as the columns of $\mathbf{U}$ in the SVD. Conversely, $\mathbf{X}^\mathsf{T}\mathbf{X}$ is the dispersion matrix where the inner product is taken over space and it represents the covariance in time obtained by using space as the sampling dimension. So the columns of $\mathbf{V}$ are the normalized principal components that are associated uniquely with each EOF. The columns of U and V are linked by the singular values, which are down the diagonal of . These eigenvalues represent the amplitude explained, however, and not the variance explained, and so are the proportional to the square roots of the eigenvalues that would be obtained by eigenanalysis of the dispersion matrices. The eigenvectors and PC's will have the same structure, regardless of which method is used, however, so long as both are normalized to unit length.

To illustrate the relationship between the singular values of SVD of the data matrix and the eigenvalues of the covariance matrix, consider the following manipulations. Let's assume that we have modified the data matrix $\mathbf{X}$ to remove the sample mean from every element of the state vector, so that $\mathbf{X} = \mathbf{X} - \overline{\mathbf{X}}$. The covariance matrix is given by

$$\mathbf{C} = \mathbf{U}\,\mathbf{V}^\mathsf{T} \tag{5.20}$$

and the eigenvectors and eigenvectors are defined by the diagonalization of $\mathbf{C}$.

$$\mathbf{C} = \mathbf{E}\,\mathbf{E}^\mathsf{T} \tag{5.21}$$

Now if we take the SVD of the data matrix, , and use it to compute the covariance matrix, we get,

$$\mathbf{C} = \mathbf{U}\,\mathbf{V}^\mathsf{T}(\mathbf{U}\,\mathbf{V}^\mathsf{T})^\mathsf{T})/\mathrm{N} = \mathbf{U}\,\mathbf{V}^\mathsf{T}\mathbf{V}\,^\mathsf{T}\mathbf{U}^\mathsf{T}/\mathrm{N} = \mathbf{U}\,\,^\mathsf{T}\mathbf{U}^\mathsf{T}/\mathrm{N} \tag{5.22}$$

Comparing 5.21 and 5.22 one can infer that $\mathbf{U} = \mathbf{E}$ and $= \,^\mathsf{T}/\mathrm{N}$ or $\lambda_\mathrm{i} = \sigma_\mathrm{i}^2/\mathrm{n}$. The singular values represent amplitudes across $\mathbf{X}$, and the eigenvalues represent variance.

We can also see that $\mathbf{V}$ represents normalized PC time series in the following way,

$$\mathbf{Z} = \mathbf{E}^\mathsf{T}\mathbf{X} = \mathbf{E}^\mathsf{T}\mathbf{U}\,\mathbf{V}^\mathsf{T} = \mathbf{E}^\mathsf{T}\mathbf{E}\,\mathbf{V}^\mathsf{T} = \,\mathbf{V}^\mathsf{T} \tag{5.23}$$

Here we have used 5.13, 5.14 and 5.19.

Notice that as far as the mathematics is concerned, both dimensions of the data set are equivalent. You must choose which dimension of the data matrix contains interesting structure, and which contains sampling variability. In practice, sometimes only one dimension has meaningful structure, and the other is noise. At other times both can have meaningful structure, as with wavelike phenomena, and sometimes there is no meaningful structure in either dimension.

Note that in the eigenanalysis,

$$\mathbf{Z}\mathbf{Z}^\mathsf{T} = \mathbf{E}^\mathsf{T}\mathbf{X}\mathbf{X}^\mathsf{T}\mathbf{E} = \mathbf{E}^\mathsf{T}\mathbf{C}\mathbf{E}\mathrm{N} = \,\mathrm{N} \tag{5.24}$$

while in the SVD computation,

$$\mathbf{Z}\mathbf{Z}^\mathsf{T} = \,\mathbf{V}^\mathsf{T}\mathbf{V}\,^\mathsf{T} = \,\,^\mathsf{T} \tag{5.25}$$

so we must have, as show before, that,

$$^\mathsf{T} = \,\mathrm{N} \quad \text{or} \quad \sigma_\mathrm{i}^2 = \lambda_\mathrm{i}\mathrm{N} \tag{5.26}$$

From **(5.24)** and **(5.25)** we see that the covariance matrix of the principle components is diagnonal, and the principle components are orthogonal (uncorrelated) in the structure dimension.

5.4.3 A very simple example

Consider the following simple 4x2 data set. Imagine that the structure dimension is 2 and the sampling dimension is 4.

$$\mathbf{X} = \begin{bmatrix} 2 & 4 & -6 & 8 \\ 1 & 2 & -3 & 4 \end{bmatrix}$$

Do SVD of that data matrix to find its component parts.

$$\mathbf{U}\,\mathbf{V}^{\mathsf{T}} = \mathbf{X} \tag{5.27}$$

The singular value matrix contains only one non-zero value. This means the data matrix is singular and one structure function and one temporal function can explain all of the data, so only the first column of the spatial eigenvector matrix is significant. The data points in mathbfX all fall on the same line. The singular value contains all of the amplitude information. The spatial and temporal singular vectors are both of unit length.

$$= \begin{bmatrix} 12.25 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Next U, which contains the spatial singular vector in the first column.

$$\mathbf{U} = \begin{bmatrix} 0.894 & -0.447 \\ 0.447 & 0.894 \end{bmatrix}$$

Finally, the temporal structure matrix **V**. Only the first column is meaningful in this context and it gives the normalized temporal variation of the amplitude of the first spatial structure function.

$$\mathbf{V} = \begin{bmatrix} 0.183 & -0.119 & -0.976 & 0.000 \\ 0.365 & -0.239 & 0.098 & -0.894 \\ -0.548 & -0.837 & 0.000 & 0.000 \\ 0.730 & -0.478 & 0.195 & 0.447 \end{bmatrix}$$

We can reconstruct the data matrix by first multiplying the singular value matrix times the transpose of the temporal variation matrix to form a traditional PC matrix.

$$\mathbf{Z} = \mathbf{V}^{\mathsf{T}} \tag{5.28}$$

$$\mathbf{Z} = \begin{bmatrix} 2.236 & 4.472 & -6.7082 & 8.944 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Only the first row of this matrix has nonzero values, because the amplitude of the second structure function is zero. The second spatial structure is the left null space of the data matrix. If you multiply it on the left of the data matrix, it returns a row of zeros. The first row of **Z** is the principal component vector for the first EOF, including the dimensional amplitude. Finally we can recover the data matrix by multiplying the spatial eigenvector matrix times the previous product of the singular value and the temporal structure matrix. This is equivalent to multiplying the eigenvector matrix times the PC matrix, and gives us the original data back.

$$\mathbf{X} = \mathbf{U}\mathbf{Z} \tag{5.29}$$

$$\mathbf{X} = \begin{bmatrix} 2.00 & 4.00 & -6.00 & 8.00 \\ 1.00 & 2.00 & -3.00 & 4.00 \end{bmatrix}$$

5.5 Presenting the Results of EOF and PC Analysis

After completing EOF analysis of a data set, we have a set of eigenvectors, or structure functions, which are ordered according to the amount of variance of the original data set that they explain. In addition, we have the principal components, which are the amplitudes of these structure functions at each sampling time. Normally, we only concern ourselves with the first few EOFs, since they are the ones that explain the most variance and are most likely to be scientifically meaningful. The manner in which these are displayed depends on the application at hand. If the EOFs represent spatial structure, then it is logical to map them in the spatial domain as line plots or contour plots, possibly in a map projection that shows their relation to geographical features.

One can plot the EOFs directly in their normalized form, but it is often desirable to present them in a way that indicates how much real amplitude they represent. One way to represent their amplitude is to take the time series of principal components for the spatial structure (EOF) of interest, standardize this time series to unit variance, and then regress it against the original data set. This produces a map with the sign and dimensional amplitude of the field of interest that is explained by the EOF in question. The map has the shape of the EOF, but the amplitude actually corresponds to the amplitude in the real data with which this structure is associated. Thus we get structure and amplitude information in a single plot. If we have other variables, we can regress them all on the PC of one EOF and show the structure of several variables with the correct amplitude relationship. For example, SST and surface vector wind fields can both be regressed on PCs of SST.

5.5.1 How to scale and plot EOFs and PCs

Let's suppose we have done EOF/PC analysis using either the SVD of the data, or the eigenanalysis of the covariance matrix. We next want to plot the EOF's to show the structure in the state space of the data. The EOFs are normalized to unit length, but we would like to combine this structure with some amplitude information in a single plot. One way to do this is to scale the eigenvectors according to the amplitude in the data set that they represent. A simple way to do this is to multiply the eigenvectors by the square root of the eigenvalue. We learned in the previous section that,

$$\mathbf{E} = \mathbf{U} \quad \text{and} \quad = \Sigma^2/\mathsf{N} \tag{5.30}$$

We define the EOF with amplitude as $\mathbf{D}$.

$$\mathbf{D}^{\mathsf{EOF}} = \mathbf{E}\,^{1/2} = \mathbf{D}^{\mathsf{SVD}} = \mathsf{N}^{-1/2}\mathbf{U} \tag{5.31}$$

In each case you can show that $\mathbf{D}\mathbf{D}^{\mathsf{T}} = \mathbf{C}$, so if you put in the amplitudes and take the inner product you get back the covariance matrix of the input data. Note also that $\mathbf{D}^{\mathsf{T}}\mathbf{D} =$. The columns of the matrix $\mathbf{D}$ are the eigenvectors, scaled by the amplitude that they represent in the original data. It might be more interesting to plot $\mathbf{D}$ than $\mathbf{E}$, because then you can see how much amplitude in an RMS sense is associated with each EOF, and you can plot the patterns in hectoPascals, °C, kg, or whatever units in which the data are given. These methods work if the data analyzed is dimensional, if the data have first been standardized, then the principle components can be regressed onto the un-standardized data to get structures with dimensional amplitudes.

In most cases the sample mean of each state variable would be removed before doing EOF or PC analysis. Sometimes it is also advisable to take the amplitudes out of the data before conducting the EOF analysis by dividing each observation by its standard deviation over the sample. Subtracting the mean and dividing by the standard deviation can be called standardizing the data, and we can denote the standardized data set as $\widetilde{\mathbf{X}}$. Reasons for standardizing might be: 1) the state vector is a combination of things with different units or 2) the variance of the state vector varies from point to point so much that this distorts the patterns in the data. If you are looking for persistent connections between the data, rather than just an efficient expression of variance, you may want to look at correlation rather than covariance. In such cases the data

can be standardized such that the variance of the time series of each element of the state vector is 1. The covariance matrix of this standardized data set is a correlation matrix.

If the data have been standarized, we can still get the amplitude into the structure by regressing the principle components of the EOF analysis onto the original dimensional data. Regression can also be used to see how the EOF of the state variable is related to other variables not in the state vector used to do the EOF analysis.

To do the regression analysis, we first want to get the PC time series normalized so that they have unit variance in time. Regardless of whether the original data were standarized or not, we can obtain a standardized principle component time series by dividing by the square root of the eigenvalue.

$$\widetilde{\mathbf{Z}} = {}^{-1/2}\mathbf{Z} \tag{5.32}$$

As an exercise, show that

$$\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^{\mathsf{T}}/\mathsf{N} = \mathbf{I} \tag{5.33}$$

The regression to determine $\mathbf{D}$ is then,

$$\mathbf{D} = \mathbf{X}\widetilde{\mathbf{Z}}^{\mathsf{T}}/\mathsf{N} \tag{5.34}$$

The columns of the matrix $\mathbf{D}$ are the EOFs, except with amplitudes that are equal to the amplitude in the original X field that is associated with a one standard deviation variation of the PC time series. This is a reasonable number to look at, since it is the amplitude that you might typically see associated with this EOF.

5.6 Significance of EOF Analysis

EOF/PC analysis makes sense when the data contain a lot of correlation between the elements of the state space, so that the variance in the data can be explained with a number of EOFs that is smaller than the number of elements in the state vector $\mathbf{X}$. The fraction of variance explained is measured by the eigenvalue λ_i divided by the sum of the eigenvalues, since the sum of the M eigenvalues is the total variance.

$$\mathsf{FoV}_i = \frac{\lambda_i}{\sum_{m=1}^{M} \lambda_i} \tag{5.35}$$

For a random state vector sampled a finite number of times, however, it is likely that EOF analysis will by chance find some EOFs that explain more variance than others, even if the state vectors are uncorrelated with each other. This happens due to random chance, but the EOF analysis orders the eigenvalues from largest to smallest, and so it appears that the data is structured, when if fact it is not. When the analysis is performed on another sample from the same uncorrelated data set, a different set of EOFs are found that appear, by chance, to explain more variance than expected for an uncorrelated data set.

This can be illustrated with a simple example. Suppose we consider a state vector of Gaussian white noise that has no correlation in the state space (M=20) or the sample space N. We consider samples of size N=20, 50, 200, 5,000. Since each eigenvalue spectrum will be a little different due to sampling variations, we do this 1000 times and average the eigenvalue spectra to give the most likely spectrum (**Fig. 5.2**). The true eigenvalue spectrum is uniform with a each eigenvalue equal to one. A large sample of N=5,000, nearly produces a uniform eigenvalue spectrum, but for smaller samples the eigenvalue spectrum is strongly sloped, indicating that the first eigenvector can explain much more variance than the last one. This happens because for any small sample some structure will explain a lot of the variance by chance. This structure will be assigned the first position. Each time this is done the EOF that explains the most variance is different, however, so the large explained variance by the first EOF is not robust or meaningful.

5.6.1 The North Test

North et al. (1982) suggested a "rule-of-thumb" for assessing the statistical significance of the eigenvalue spectrum. For a Gaussian distribution the eigenvalues should fall within a range given by a standard error that depends on the eigenvalue, λ and the number of independent samples, N.

$$\Delta\lambda = \lambda\left(\frac{2}{N}\right)^{1/2} \tag{5.36}$$

They argued that 68 percent of sample eigenvalues should fall within these limits, and that if the uncertainties of two adjacent eigenvalues overlapped, then eigenvalues are not really distinct and the EOFs associated with those eigenvalues will show large inter-sample variability, would not be robust, and should not be taken seriously. **Fig. 5.2** includes the the North et al. (1982) eigenvalue uncertainty bars. In each case the uncertainties of adjacent eigenvalues overlap, so we conclude that none of the eigenvalues or eigenvectors are meaningful. We could illustrate the inter-sample variation in the eigenvectors by plotting the eigenvector that explains the most variance for each of the 1000 realizations associated with the cases in **Fig. 5.2** and this would show that they are random in appearance. The rule of thumb expressed by **(5.36)** correctly suggests that the eigenvalues are not distinct for all the cases shown in **Fig. 5.2**.

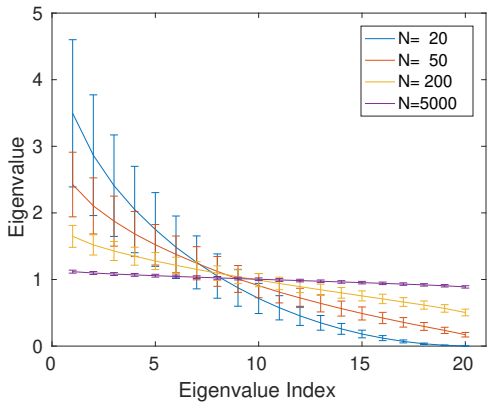

Figure 5.2 Eigenvalue spectra for white noise in a state space of M=20 with different sample sizes N=20, 50, 200, and 5,000. The North et al. (1982) uncertainty estimates are shown for each spectrum.

When structure is present in the data, the first few eigenvalues will set themselves apart from the rest, and the error bars suggested by North et al. (1982) will not overlap with the others as shown in **Fig. 5.3**.

5.6.2 Assessing Physical Significance

If the North criterion is satisfied for a particular analysis, one must still be careful in interpreting the EOFs. In interpreting EOFs one must remember exactly what they are. They are mathematical constructs that are chosen to represent the variance over the domain and sample of interest as efficiently as possible, and also be orthogonal with each other. Sometimes these mathematical constraints will select scientifically interesting structures in a data set, but not always. EOF analysis will always pick out some structures that represent more of the variance than the others will, and they will often tend to look wavelike, because they are constrained to be orthogonal. If you put autocorrelated noise through EOF analysis, it will produce structures that resemble the Fourier modes for the domain of interest, even when the data set is pure noise. If the data are autocorrelated in time or space, for which we can use the model of red noise, then the eigenvalue spectrum will be peaked, with some EOFs explaining much more of the variance than others. These will be smoothly varying large-scale structure, and it is tempting to interpret them physically, although they are

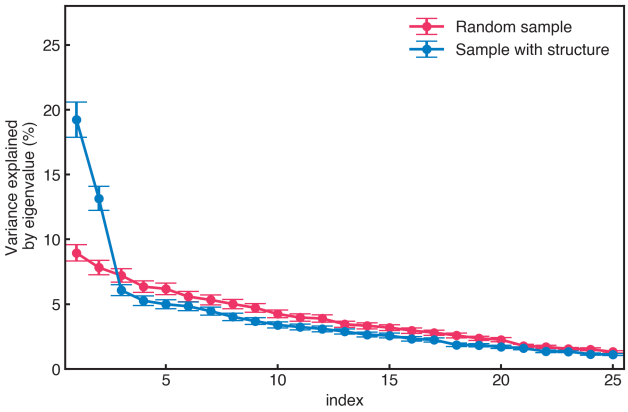

Figure 5.3 Eigenvalue spectrum of white noise (purple) and data with structure (blue) for a state space of $M = 25$ and sample of size $N = 100$.

telling you nothing but that the adjacent data points are correlated with each other. The particular structures obtained will depend on the particular spatial and temporal slice of data that was used to compute them. Just because EOF analysis produces large-scale wavelike structures does not mean the data contain coherent wave structures shaped like the sample EOFs. Sometimes, in an effort to explain the variance of a finite sample, EOF analysis will combine distinct physical modes into a single EOF, which may be called mode mixing. Below are some suggestions for steps to follow in attempting to ascertain whether EOFs produced in an analysis are a reflection of scientifically meaningful structures in the data.

1. Is the variance explained by the EOF more than you would expect if the data had no structure? What is your null hypothesis for the data field? Do you expect adjacent data points to be correlated? Can the EOFs of interest support a rejection of the uninteresting null hypothesis that adjacent points are correlated?
2. Do you have any *a priori* reason for expecting the structures that you find? Are the structures explainable in terms of some theory? Do the spatial and temporal structures of the modes behave consistently with theory and *a priori* expectation?
3. How robust are the structures to the choice of structure domain? If you change the domain of the analysis, do the structures change significantly? If the structure is defined in geographical space, and you change the size of the region, do the structures change significantly? If the structures are defined in a parameter space, and you add or remove a parameter, do the results change in a sensible manner, or randomly?
4. How robust are the structures to the sample used? If you divide the sample into randomly chosen halves and do the analysis on each half, do you consistently get the same structures?

5.7 Applications of EOF/PC Analysis

Rearrangement of data into Empirical Orthogonal Functions (EOFs) and their Principal Components (PCs) is useful in a variety of contexts.

5.7.1 Data Compression

EOF/PC analysis is a kind of functional representation, where a set of spatial/structure functions is derived that explains the largest amount of variance with the smallest number of functions. The functions are

arranged according to their rank in explaining variance. EOFs represent the correlated structures in the data. Often a large amount of the variance of a data set can be represented with a relatively small number of EOFs, so that when the data are stored as the PCs, the volume required is small, if the PCs of EOFs that explain a small amount of variance are discarded.

For example, the human fingerprint can be represented in great detail with about 10 EOFs. It can be shown that Fourier analysis is optimal in a least squares sense, but EOF analysis will often beat Fourier analysis in terms of efficiency of representation, when the data contain structures that are represented well by a small range of Fourier wavelengths. Fingerprints are better represented by EOFs than by Fourier series because the fingerprint patterns are simpler than they appear, being composed of a set of whorls that occupy a rather small range of wavenumbers in the x-y space of fingerprint area. Fourier analysis has to carry along all of the wavenumbers needed to span a print of a certain size, whereas EOF analysis can concentrate on the scales and shapes that contain the real information, and thus require far fewer stored numbers to reproduce a given individual's print. In general EOF analysis performs well when most of the variability is contained in a small number of structures. This is indicated by the eigenvalue spectrum. If the first few eigenvalues are large and most of the rest are small, then the number of degrees of freedom of the data set is much less than the number of data points, and use of EOFs to compress the data can provide benefits.

5.7.2 Determining Degrees of Freedom

If a spatiotemporal data set has large correlations between the state-space variables, then the data set may have fewer independent degrees of freedom than the number of state-space variables (*e.g.* spatial grid points or parameters). For many purposes it is important to evaluate how many independent degrees of freedom a data set has. This can be assessed using EOF analysis, as has been reviewed by (Bretherton et al., 1999).

Consider a spatial data set of dimension m that is stationary on the time interval for which it is sampled. Define a quadratic functional of some vector variable $\mathbf{X}(t)$, where the vector is of length m.

$$E(t) = \sum_{i=1}^{m} \mathbf{X}_i^2(t) \tag{5.37}$$

The number of spatial degrees of freedom m^* is defined to be the number of uncorrelated random normal variables a_k, each having zero mean and the same population variance $\langle a^2 \rangle$, for which the χ^2 distribution for the specified functional most closely matches the PDF of $E(t)$. In order to approximate this one can require that the χ^2 distribution match the observed distributions ensemble mean value $\langle E \rangle$ and the temporal variance about this mean,

$$var(E) = \langle E'^2 \rangle = \left\langle \left(E - \langle E \rangle \right)^2 \right\rangle \tag{5.38}$$

For the χ^2 distribution $\langle E \rangle = m^* \langle a^2 \rangle$ and $var(E) = 2m^* \langle a^2 \rangle^2$. We can then solve for the spatial degrees of freedom that matches the first two moments of the normal distribution of variance. This is a "moment matching" estimate of the effective number of degrees of freedom.

$$m^*_{mm} = \frac{2\langle E \rangle^2}{var(E)} \qquad \langle a^2 \rangle^2_{mm} = \frac{var(E)}{\langle E \rangle} \tag{5.39}$$

These estimates can be obtained from the mxm covariance matrix of $\mathbf{X}$, $\mathbf{C}_{xx} = \mathbf{C}$, if $\mathbf{X}(t)$ is normally distributed and we know $\mathbf{C}$ well enough. Suppose we have the eigenvalues λ_k and the standardized principle components $z_k(t)$ of $\mathbf{C}$. We can now calculate m^* from the eigenvalues in the following way.

$$E(t) = \sum_{k=1}^{m} \lambda_k z_k^2(t) \qquad \langle E \rangle = \sum_{k=1}^{m} \lambda_k \tag{5.40}$$

and

$$\begin{aligned} var(E) = \sum_{k=1}^{m} \lambda_k^2 var(z_k^2(t)) &= \sum_{k=1}^{m} \lambda_k^2 \left\langle var\left(z_k^2 - \langle z_k^2 \rangle\right)^2 \right\rangle \\ &= \sum_{k=1}^{m} \lambda_k^2 \left\langle z_k^4 - \langle z_k^2 \rangle^2 \right\rangle \end{aligned} \tag{5.41}$$

Since we are assuming that the PCs are standardized Gaussian normal variables their variance is one and their kurtosis is 3, and we have that,

$$var(E) = \sum_{k=1}^{m} \lambda_k^2 \left\langle z_k^4 - \langle z_k^2 \rangle^2 \right\rangle = \sum_{k=1}^{m} \lambda_k^2 (3-1) = 2 \sum_{k=1}^{m} \lambda_k^2 \tag{5.42}$$

We can now write down an eigenvalue based estimate for the effective number of spatial degrees of freedom by substituting 5.40 and 5.42 into 5.39.

$$m_{eff}^* = \frac{\left(\sum_{k=1}^{m} \lambda_k\right)^2}{\sum_{k=1}^{m} \lambda_k^2} = \frac{\left(m\bar{\lambda}\right)^2}{m\overline{\lambda^2}} \tag{5.43}$$

This formula can also be written in terms of the covariance matrix from which the eigenvalues were derived.

$$m_{eff}^* = \frac{\left(\sum_{i=1}^{m} C_{ii}\right)^2}{\sum_{i=1}^{m} \sum_{j=1}^{m} C_{ij}^2} = \frac{\left(\operatorname{tr} \mathbf{C}\right)^2}{\operatorname{tr}\left(\mathbf{C}^2\right)} \tag{5.44}$$

Here tr indicates the trace of the matrix. Note that the denominator in **(5.43)** equals the square of the Frobenius norm of **C**.

5.7.3 Prefiltering

It might be reasonable to assume that EOFs with large eigenvalues are the structure in a data set and EOFs with small eigenvalues are noise. One could argue then that doing EOF analysis and removing the variance associated with small eigenvalues is a noise reduction procedure. For example, when reconstituting the original data from the EOF expansion, the PC time series associated with the smaller eigenvalues could be set to zero. This would produce a data set with the effects of those less correlated structures removed.

5.7.4 Statistical Prediction

EOFs are orthogonal in the structure dimension and in the sampling dimension. In statistical prediction, correlation between predictors is undesirable [reference earlier section]. PC series are uncorrelated in the sampling dimension, and the covariance of the PC matrices is diagnonal with the eigenvalues down the diagonal.

$$\mathbf{Z}\mathbf{Z}^T/N = \mathbf{C}_{ZZ} = \tag{5.45}$$

It is thus advantageous to first do EOF analysis and use the principal components of the EOFs as predictors rather than any other combination of the predictor variables. This makes the predictors uncorrelated and the inversion to obtain the regresson coefficients much easier.

5.7.5 Exploratory Data Analysis

Suppose we have a state vector of some system for which we have a large sample. EOF/PC analysis can be used as a tool to search for characteristics structures in the spatial, parameter, or time dimensions of the data set. The spatial structure and associated temporal structure of a data field may be helpful in identifying mechanisms that produce variability in the data set. If the data set shows that much of the variance can be explained by a few EOFs, then the analysis can be focused by limiting the analysis to these few structures and their variations within the sample.

Wavelike phenomena are easily picked up by EOF analysis. For example, suppose that the data set consists of a standing wave with a spatial pattern that oscillates in time,

$$w(x,t) = \cos(2\pi x/L) \times \cos(2\pi t/T) \tag{5.46}$$

The EOF analysis will show one nonzero eigenvalue corresponding to a wave with wavelength L in space and period T in time. A traveling wave has a formula as **(5.47)**.

$$\begin{aligned} w(x,t) &= \cos(2\pi x/L - 2\pi t/T) \\ &= \cos(2\pi x/L) \times \cos(2\pi t/T) + \sin(2\pi x/L) \times \sin(2\pi t/T) \end{aligned} \tag{5.47}$$

Representing a traveling wave requires two EOFs and corresponding PCs, each 90-degrees out of phase. These would have the same eigenvalue, since their amplitudes are equal.

5.8 Rotation of Empirical Orthogonal Functions

EOF analysis enforces orthogonality on the eigenvectors and their corresponding principal components. Sometimes the orthogonality constraint will cause structures to have significant amplitude all over the domain (e.g. spatial domain) of the analysis, when physical reasoning suggests that the structures should be much more localized. To reduce the effect of the orthogonality constraint and allow more localized structures to emerge, we can consider rotation of the eigenvectors (Horel 1984, Richman 1986).

The procedure begins by selecting a subset of the eigenvectors, say those that explain 70% of the variance, and discarding the rest. The constraint of orthogonality is relaxed and replaced with another constraint that is designed to make the EOFs as "simple" as possible. Relaxing one orthogonality is an orthogonal rotation, and relaxing both orthogonalities is an oblique rotation. Simplicity of structure is defined to occur when most of the elements of the eigenvector are either of order one (absolute value) or zero, but not in between. The selected eigenvectors are rotated until the criterion is maximized.

The Quartimax Criterion seeks an orthogonal rotation of the eigenvector matrix $e_{ij} = \mathbf{E}$ into a new factor matrix $b_{ij} = \mathbf{B}$ for which the variance of squared elements of the eigenvector is a maximum. The quantity to be maximized is,

$$Q = \frac{1}{Mm} \sum_{i=1}^{M} \sum_{j=1}^{m} \left(b_{ij}^2 - \overline{b^2} \right) \tag{5.48}$$

where

$$\overline{b^2} = \frac{1}{Mm} \sum_{i=1}^{M} \sum_{j=1}^{m} b_{ij}^2 \tag{5.49}$$

The quantity **(5.48)** to be maximized can be simplified to,

$$Q = \frac{1}{Mm} \sum_{i=1}^{M} \sum_{j=1}^{m} b_{ij}^2 - \overline{b^2} \tag{5.50}$$

Since the mean-squared loading remains constant under orthogonal rotations, the criterion is simply equivalent to maximizing the sum of the fourth power of the loadings, hence the name Quartimax.

One often used criterion for defining simplicity is the Varimax Method. The simplicity of an individual rotated eigenvector is defined as the variance of its squared loadings, b_{ij}, where i is the loading index for an eigenvector and j denotes the eigenvector.

$$V_j = \frac{1}{M}\sum_{i=1}^{M}\left(b_{ij}^2\right)^2 - \frac{1}{M^2}\left(\sum_{j=1}^{M} b_{ij}^2\right)^2 \qquad j = 1, 2, ..., m \tag{5.51}$$

When the variance V_j is at a maximum the jth rotated eigenvector has its greatest simplicity in the sense that its loading tends toward unity or zero. The criterion of simplicity of the complete rotated eigenvector matrix is defined as the maximization of the sum of the simplicities of the individual eigenvectors,

$$V = \sum_{j=1}^{m} V_j \tag{5.52}$$

Equation (5.52) is called the raw Varimax criterion. It is sometimes useful to weight the individual eigenvectors with a weight h_j, for example by the variance explained. The final normalized Varimax criterion is then,

$$V = M\sum_{j=1}^{m}\sum_{i=1}^{M}\left\{\frac{b_{ij}}{h_j}\right\}^4 - \sum_{j=1}^{m}\left\{\sum_{i=1}^{M}\frac{b_{ij}^2}{h_j}\right\}^2 \tag{5.53}$$

The Varimax method is often preferred over the Quartimax method because the sensitivity to changes in the number (or choice) of variables is less. The difference between the results obtained with the two methods in practice is usually small. Many other criteria closely related to these can be found in available software packages.

5.8.1 The Eight Physical Variables Example

An interesting example of the use of EOF rotation in factor analysis is the 'Eight Physical Variables Example' that looks at the correlations between eight measures of human anatomy. The eight physical variables are listed in the table below.

Eight Physical Variables	
1. Height	5. Weight
2. Arm Span	6. Bitrochantric Diameter
3. Forearm Length	7. Chest Girth
4. Lower Leg Length	8. Chest Width

These eight variables are highly redundant, since length of forearm, arm span, and height are all highly correlated. In cases such as this factor analysis can describe the anatomy with fewer variables. The correlation matrix for the eight physical variables is shown below.

$$\mathbf{C} = \begin{bmatrix} 1.00 & .846 & .805 & .859 & .473 & .398 & .301 & .382 \\ .846 & 1.00 & .881 & .826 & .376 & .326 & .277 & .415 \\ .805 & .881 & 1.00 & .801 & .380 & .319 & .237 & .345 \\ .859 & .826 & .801 & 1.00 & .436 & .329 & .327 & .365 \\ .473 & .376 & .380 & .436 & 1.00 & .762 & .731 & .629 \\ .398 & .326 & .319 & .329 & .762 & 1.00 & .583 & .577 \\ .301 & .277 & .237 & .327 & .731 & .583 & 1.00 & .539 \\ .382 & .425 & .345 & .365 & .629 & .577 & .539 & 1.00 \end{bmatrix}$$

The correlation matrix shows that all the physical variables are positively correlated. The eigenvalue spectrum derived from eigenanalysis of the correlation matrix of the eight physical variables shows that the

first two eigenvectors explain 85% and 12% of the correlation, respectively, so that only these two vectors are considered and included in the rotation.

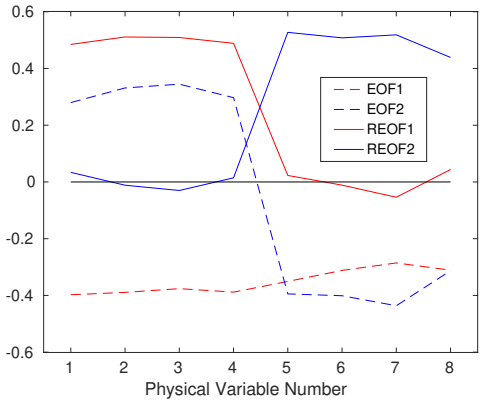

Figure 5.4 Eigenvectors and orthogonally rotated eigenvectors for the Eight Physical Variables data set. Varimax rotation was used.

Fig. 5.4 shows the eigenvectors derived from the correlation matrix of the eight physical variables. The rotated EOFs obtained from an orthogonal varimax rotation are also shown. The first eigenvector indicates that all the physical variables tend to go up and down together, so long people are also broad and heavy. The second eigenvector contains the implication that the people who are long, also tend to be thin. These indications are somewhat contrary to the basic correlation matrix that shows the correlation between the first 4 and the last 4 physical variables are positive, but weak. The structure of the eigenvectors is heavily constrained by the requirement of orgthogonality. Rotation of the eigenvectors gives an alternative interpretation. The rotated eigenvectors lead to an interpretaion of the data in which one factor is length, or the "bone factor", and a second factor is related to width and weight, or the "flesh factor". This interpretation of the data recognizes the strong positive correlation among the length variables and among the weight variables, but does not impose an artificial negative correlation between the length and weight variables. The second interpretation seems more acceptable, and explains the data just as well. In this case the bone and flesh factors are still orthogonal in the parameter domain, since an orthogonal rotation was used.

5.8.2 EOF Analysis of Red Noise

Another example of the use of rotated EOFs is red noise that is autocorrelated in space and time, but has no other structure. Suppose we take a sample of such a space-time series that has 50 spatial grid points and 400 sample times. The autocorrelation in space is 0.9 and the autocorrelation in time is 0.2, so that all the samples are essentially independent. We perform EOF analysis on the standardized data, so that each point has unit variance in time. Before doing the EOF analysis, though, let's consider in **Fig. 5.5**a a one point correlation map between grid point 25 and all the other grid points. This plot indicates that grid point 25 is well correlated with adjacent data points, but not correlated at all with distant points. Now let's do EOF analysis of this data set.

The spectrum of eigenvalues of our red noise data set is shown in **Fig. 5.5**b. Most of the variance is explained by the first few eigenvectors, and the North Test suggests that the largest eigenvalues are distinct and the associated eigenvectors should be robust and reproducible. This is true because the data is highly autocorrelated in space, so that slowly varying functions should explain a lot of the variance. Beyond telling us that the data are correlated in space, however, these eigenvectors are not physically meaningful, since we

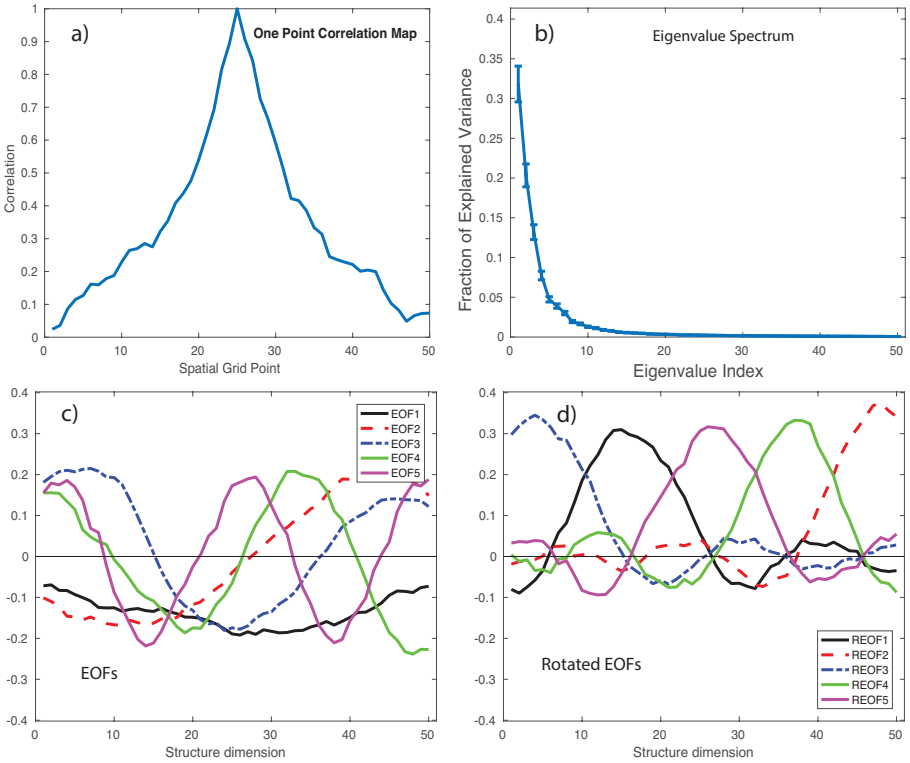

Figure 5.5 a) Autocorrelation of grid point 25 with all the other grid points of a sample of red noise with a spatial autocorrelation of 0.9. b) Eigenvalue spectrum for red noise data set with North uncertainty bars. c) First five eigenvectors. d) First five rotated eigenvectors. An orthogonal varimax rotation was used.

know the data are just red noise in space. An assessment of the effective number of degrees of freedom in this data set using **(5.43)** indicates that the data set has fewer than 6 spatial degrees of freedom despite having 50 grid points.

The raw eigenvectors are shown in **Fig. 5.5**c. The first eigenvector is of one sign everywhere, like a constant first term in a Fourier series. It suggests that all the points go up and down together, but we know from **Fig. 5.5**a that the data at distant points are in fact not correlated. The second eigenvector looks like a sine wave with a wavelength equal to the size of the model domain. Here EOF analysis is just constructing the functions of a Fourier series that is orthogonal within the spatial domain.

The retained eigenvectors that explain 70% of the total variance were rotated using an orthogonal Varimax criterion. The rotated eigenvectors shown in **Fig. 5.5**d are localized in space and span a region that is similar in scale to the autocorrelation shown in **Fig. 5.5**a. In this case then, the rotated EOFs are much more representative of the data than are the raw EOFS. They show a sequence of blobs that are localized in space, but near zero elsewhere, which is a good representation of red noise. These rotated EOFs are also orthogonal in the space dimension, so there is no shared variance between them. They each explain about an equal fraction of the total variance. The negative values could be brought much closer to zero by using an oblique rotation.

5.8.3 Wintertime 500hPa Height Example

As a more realistic example of the use of rotation of EOFs, consider the wintertime anomalies of 500hPa height with time scales longer than 30 days. To perform this analysis we first remove the climatological annual cycle, or it would dominate the analysis. We want to explore the structure of the slowly varying anomalies

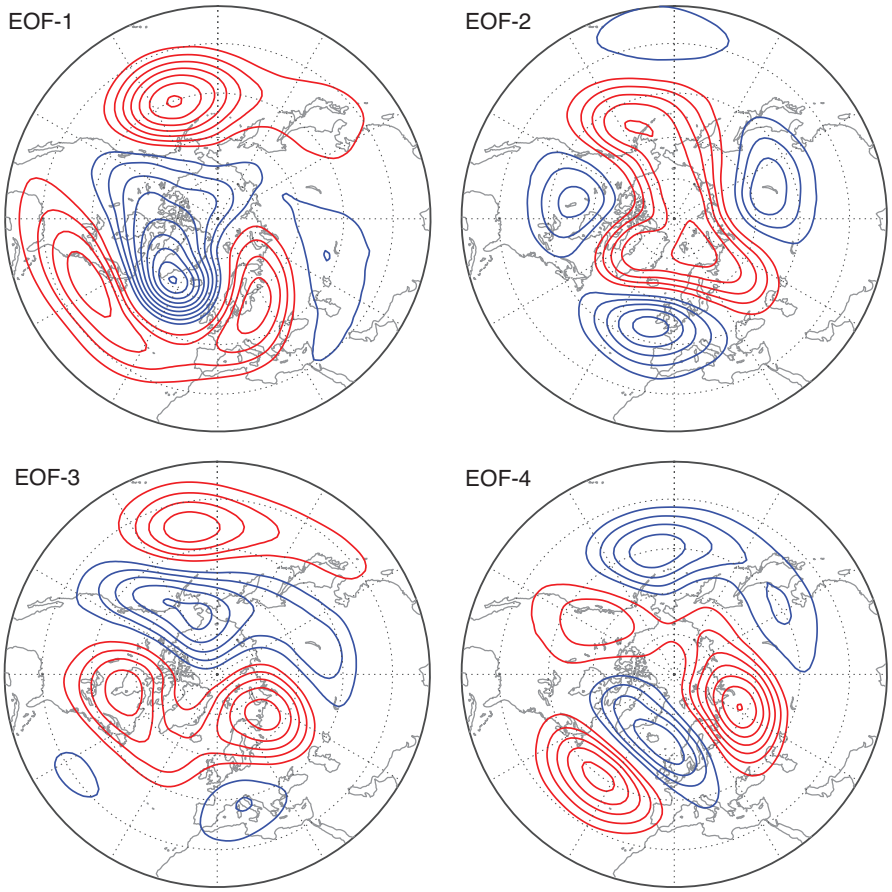

Figure 5.6 First four EOFs of the wintertime 500hPa height anomalies north of 20N.

of the 500hPa height field, which characterizes the structure of the mid-troposphere. To focus on the winter season we consider only the period from October 1 to March 31. To focus on the Northern hemisphere, we consider only the region poleward of 20N. Because the data are in latitude-longitude coordinates, we weight the variance by cosine of latitude, so so that the more closely spaced data in high latitudes do not bias the analysis toward high latitudes. The raw anomalies are used, unstandardized, so that regions of large variance have a bigger impact on the structures obtained.

The eigenvalue spectrum is smoothly decreasing with eigenvalue number, and 70% of the variance is explained by the first 11 eigenvectors, indicating that the low-frequency variability of the 500hPa height is strongly autocorrelated and of large spatial scale compared to the grid spacing of the basic data. The first 11 eigenvectors are rotated, but for economy of presentation we show only the first four in **Fig. 5.6**. These generally have some amplitude everywhere in the Northern Hemisphere, and tend to show strong connections between the Atlantic and the Pacific Ocean basins.

When the eigenvectors are rotated, they become more local and tend to look more meteorological (**Fig. 5.7**). The first rotated EOF corresponds to the North Atlantic Oscillation, which was discovered in local correlation maps before good global analyses were available. The second and third rotated EOFs correspond to known patterns that reflect the propagation of Rossby waves along great circle routes. The fourth rotated EOF appears to be a Rossby wave train propagating across Eurasia. In this case the rotation to produce more localized structures leads to a much more sensible physical interpretation than the raw EOFs, which try very hard to maximize the variance explained over the whole domain with a set of orthogonal structures. Rotation produces structures more similar to what you get from one-point correlation maps.

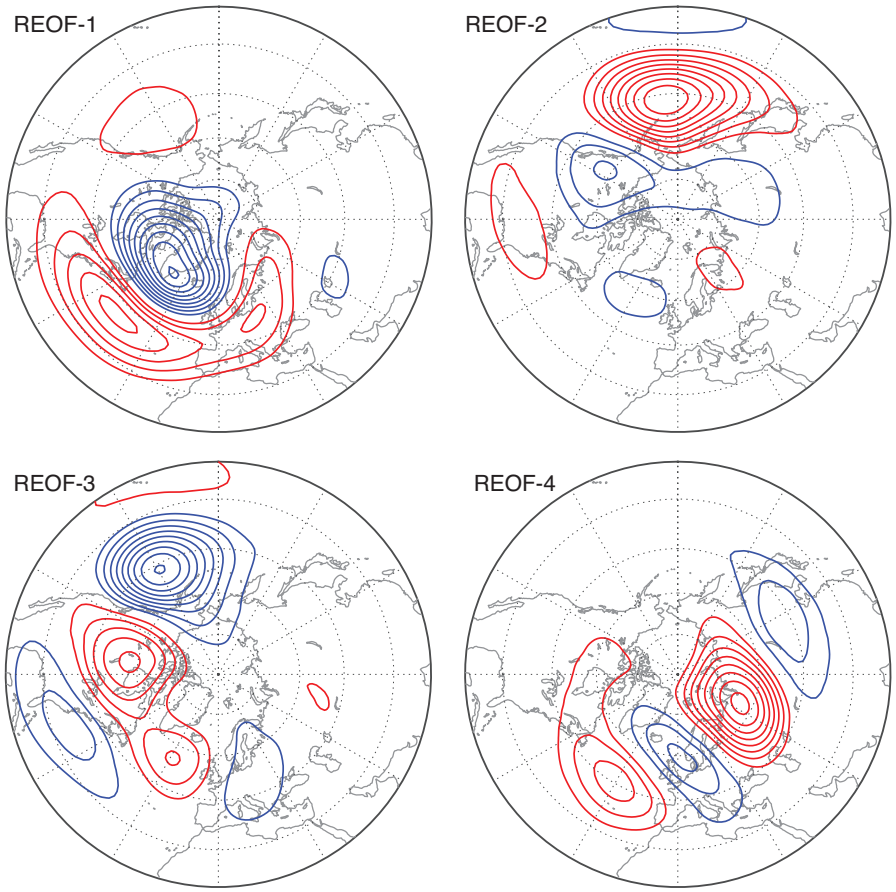

Figure 5.7 First four rotated EOFs of the wintertime 500hPa height anomalies north of 20N. Orthogonal varimax rotation was used.

5.9 Maximum Covariance Analysis

EOF analysis searches for structures that explain the maximum amount of variance in some data matrix. The data matrix is presumed to have a structure dimension, for example spatial location, and a sampling dimension, for example time. In Maximum Covariance Analysis (MCA) two data matrices with different structures, or state spaces, are considered, but they share a common sampling dimension. For example, one could consider the fields of sea surface temperature and surface chlorophyll content, measured at the same set of times. Or the sampling dimension could be a collection of hospital patients, and the two state vectors could be their "Eight Physical Variables" and their cholesterol data (say, 3 numbers). Suppose the sample of patients is N. One could address the relationship between body dimensions and cholesterol by making an augmented state vector consisting of the 11 numbers that include their 8 physical variables and their 3 cholesterol values. EOF analysis of the 11 x N data matrix might determine if particular combinations of variables vary together and in that way explain a lot of the combined variance. One might expect that if the physical variables and the cholesterol variables vary together, then they should show up in structures that efficiently explain the variance of the augmented or combined state vector. This is a form of augmented EOF analysis that tries to explain the maximum amount of variance over a combined data set. Where the variables have different units, one would normally standardize the data to unit variance before the conducting the EOF analysis.

In MCA one looks for structures in the data set that are well correlated with structures in the other data set. To analyze the relationship between the eight physical variables and the cholesterol data using MCA, one first computes the covariance (or correlation) matrix between the 8 x N and 3 x N data sets to make an

8x3 covariance matrix. Then SVD analysis of this covariance matrix yields structures that are well correlated between body dimensions and cholesterol levels, if such exist. The resulting singular vectors and singular values would tell you about structures in one data set that are correlated with structures in the other data set as you sample across the population of your hospital patients. For example, do the flesh variables correlate strongly with high levels of "bad" cholesterol? The singular values tell you the amount of covariance that is explained by each pair of structures.

Prohaska Prohaska (1976) first perhaps used MCA in the meteorological literature, although it has long been used in the social sciences. Bretherton et al. Bretherton et al. (1992) and Wallace et al. Wallace et al. (1992) popularized it for meteorological and oceanographic use.

5.9.1 MCA Mathematics

Let us suppose we have two data matrices $\mathbf{X}$ and $\mathbf{Y}$ of size MxN and LxN, where M and L are the structure dimensions and N is the shared sampling dimension. We begin by taking the inner product of these two matrices to obtain an MxL covariance matrix.

$$\mathbf{X}\mathbf{Y}^{\mathsf{T}}/N = \mathbf{C}_{XY} \tag{5.54}$$

Normally, we would remove the time mean (average over the sample N) from X and Y, so that $\mathbf{C}_{XY}$ is indeed a covariance matrix in the usual sense. If the two fields have different units, then the correlation matrix should be used.

Having formed the covariance matrix between the two data sets by projecting over their sampling dimension, SVD analysis can be done to decompose the covariance matrix into its column space and row space and associated singular values. The column space will be structures in the dimension M that are orthogonal and have a partner in the row space of dimension L. Together these pairs of vectors efficiently and orthogonally represent the structure of the covariance matrix. The hypothesis is that these pairs of functions represent scientifically meaningful structures that explain the covariance between the two data sets. Let's set the deeper issues aside for a moment and just look at some of the features of the mathematics. We consider the SVD of the MxL covariance matrix.

$$\mathbf{C}_{XY} = \mathbf{U}\,\mathbf{V}^{\mathsf{T}} \tag{5.55}$$

The columns of $\mathbf{U}$ (MxM)are the column space of $\mathbf{C}_{XY}$ and represent the structures in $\mathbf{X}$ that are highly correlated with $\mathbf{Y}$. The columns of $\mathbf{V}$ are the row space of $\mathbf{C}_{XY}$ and are those structures in the Y space that best explain the covariance. The singular values are down the diagonal of the matrix and give the amount of covariance explained by each pair of left and right singular vectors. The sum of the squares of the singular values σ_k is equal to the sum of the squared covariances between the original elements of $\mathbf{X}$ and $\mathbf{Y}$. The number of non-zero singular values will be less than or equal to the smaller of M or L, $k \leqslant K \leqslant M \cap L$.

$$\|\mathbf{C}_{XY}\|^2 = \sum_{i=1}^{M}\sum_{j=1}^{L}\left(\overline{x_i y_j}\right)^2 = \sum_{k=1}^{K}\sigma_k^2 \tag{5.56}$$

Since the input matrix is a covariance matrix, the singular values have units of covariance, or correlation if the original matrix is a correlation matrix.

As in EOF analysis we can project the left an right singular vectors onto the data to express the initial data in the coordinate space of the optimal directions for expressing covariance.

$$\mathbf{X}^* = \mathbf{U}^{\mathsf{T}}\mathbf{X} \quad ; \quad \mathbf{Y}^* = \mathbf{V}^{\mathsf{T}}\mathbf{Y} \tag{5.57}$$

Using (5.54), (5.54) and the orthonormality of the singular vectors, it is easy to show that the covariance matrix of the amplitude loadings of the left and right singular vectors computed in (5.56) is diagonal and equal to the singular value matrix of the original covariance matrix.

$$\mathbf{C}_{X^*Y^*} = \mathbf{X}^*\mathbf{Y}^{*T}/N = \tag{5.58}$$

The sum of the squares of the singular values is equal to the square of the Frobenius Norm (the sum of the squares of the elements) of the covariance matrix, which is the total squared covariance. One can ask whether one mode stands out over the others by asking whether it explains a large fraction of the covariance, although it is also necessary that the total covariance between the two data sets be large, or the results are not meaningful.

5.9.2 Normalized Root Mean Squared Covariance

The total squared covariance, sum of the squares of all the elements of the covariance matrix is a useful measure of the strength of the simultaneous linear relationship between the fields. We can normalize this with the product of the variance of the left and right fields and call it the normalized root mean squared covariance. If this statistic is very small, then the covariance between the two data sets is small, and it may not make sense to search for structure in the covariance using MCA.

$$RMSC = \left(\frac{\sum_{i=1}^{M} \sum_{j=1}^{L} \overline{x_i y_j}^2}{\left(\sum_{i=1}^{M} \overline{x_i^2} \right) \left(\sum_{j=1}^{L} \overline{y_j^2} \right)} \right)^{1/2} \tag{5.59}$$

The normalized root mean square covariance should be on the order of 0.1 or greater to indicate well-correlated fields so that MCA is justified.

5.9.3 Heterogeneous and Homogeneous Regression Maps

The singular vectors are normalized and non-dimensional, whereas the expansion coefficients have the dimensions of the original data. Like EOFs, singular vectors can be scaled and displayed in a number of ways. The sign is arbitrary, but if you change the sign of one component, you must change the sign of everything, including either left or right singular vectors and their corresponding expansion coefficients. One must remember that the singular vectors, as defined here, are constructed to efficiently represent covariance, and they may not, in general, be very good at representing the variance structure.

As in EOF/PC analysis, the dimensional amplitude of the left and right singular vector patterns can be obtained by regressing the original data onto the normalized loading vectors (5.56). MCA analysis is a little different than EOF analysis, since to get the structure of the left field, you can project the left field data onto the expansion coefficient of the right singular vector, and vice versa. These are heterogeneous regression maps.

$$\mathbf{D}_{XY} = \mathbf{X}\widetilde{\mathbf{Y}}^{*T}/N/quad \quad \mathbf{D}_{YX} = \mathbf{Y}\widetilde{\mathbf{X}}^{*T}/N \tag{5.60}$$

Where here the tilde indicates that the time series of the loading vectors has been normalized to unit variance. One can also construct homogeneous regression maps by regressing the left variable onto the normalized left loading vectors.

$$\mathbf{D}_{XX} = \mathbf{X}\widetilde{\mathbf{X}}^{*T}/N \qquad \mathbf{D}_{YY} = \mathbf{Y}\widetilde{\mathbf{Y}}^{*T}/N \tag{5.61}$$

Heteorgeneous regressions show the amplitude and structure of the patterns that best explain the covariance between two data sets. The homogenous regressions show how the singular vectors do in explaining the

variance of their own data set. If the patterns that explain covariance between two data sets are similar to the patterns that explain the variance in each data set, then the homogenous and the heterogeneous patterns should be similar. Another way to check this is to compare the singular vectors with the EOFs of each data set. If they are similar then the structures that explain the variance well also explain the covariance.

In contrast to the principal component time series of EOF analysis, the expansion coefficient time series of MCA are not mutually orthogonal. The correlation coefficient between the expansion coefficients for corresponding left and right singular vectors is a measure of the strength of the coupling between the two patterns in the two fields.

5.9.4 Statistical significance of MCA

MCA is subject to the usual sampling fluctuations. Sampling errors can be significant if the number of degrees of freedom in the original data set is modest compared to the degrees of freedom in the structure (e.g. spatial). Some effort should be made to evaluate how many degrees of freedom the data set really has. The usual method of dividing the data set should be used, if possible, to test sensitivity of the results to the specfic sample chosen. One should also try to evaluate how much variance and covariance are explained with a pair of patterns that may be of interest. Comparison against Monte Carlo experiments may also give insight into how probable it is that a given correlation pattern could have arisen by chance from essentially random data.

Many caveats and criticisms have been offered for MCA analysis. Newman and Sardeshmukh(1995) showed that MCA would reveal a linear operator $\mathsf{y} = \mathsf{L}\mathsf{x}$ only under the very restrictive condition that the operator was orthogonal, $\mathsf{L}^{\mathsf{T}} = \mathsf{L}^{-1}$.

Cherry(1996) recommended extreme caution in applying MCA, since it tends to produce spurious spatial patterns. Cherry(1997) showed that singular vectors could be thought of as orthogonally rotated PC patterns, rotated so as to produce maximum correlation between pairs of rotated PCs. He recommends first carrying out separate PC analysis on the two data sets. It is less likely that patterns picked out from two data sets for the ability to explain variance in their own domain will be correlated with patterns in another domain, purely by chance. Hu (1997) pointed out some lack of uniqueness problems with MCA analysis.

5.10 Canonical Correlation Analysis

Principal component analysis and MCA analysis can be performed in sequence, and we can call the result Canonical Correlation Analysis (CCA) (e.g. Barnett and Preisendorfer; 1987), First performing EOF analysis, and then truncating the EOF expansion will reduce noise and make the search for correlation less subject to noise. The raw state variables can be first subjected to EOF/PC analysis, and new state variables formed from the subset the EOF/PCs that explains most of the variance. The reduction of the dimension of the data set to the strongest PCs reduces the possibility that correlated patterns will emerge by chance from essentially random data. If desired, one can then normalize the time series of PCs so that they have unit variance. One then calculates the SVD from the correlation matrix of these normalized PC time series. This means that the correlation between patterns of EOFs is maximized by the SVD analysis, rather than the covariance. It is argued that CCA is more discriminating than MCA analysis, in that it is not overly influenced by patterns with high variance, but weak correlation, but it is also susceptible to sampling variability.

The first step in CCA is to perform EOF analysis on the original data for both the left and right fields and construct the time series of the PCs, which are the amplitudes of the EOFs at each sampling time for each data set. Of course this step presumes that the original data are highly correlated, so that EOF analysis makes sense. So the first step is an orthogonal rotation of the coordinate systems so that the first direction explains most of the variance, and so forth.

The data are next truncated by retaining only those PCs that explain a lot of variance, thus reducing the number of degrees of freedom in the input data sets from the original structure dimensions of the input fields x and y to some smaller dimension. Since the PCs are efficient in explaining the variance, a small number

can explain a large fraction of the variance. In choosing the number of modes to be retained, one faces a trade-off between statistical significance and explaining as much variance as possible. To have statistical significance argues for as few modes as possible so that the number of samples will be large compared to the number of degrees of freedom in the structure dimension. To include as much variance as possible, one would include more PCs in the analysis. In any case, the number of degrees of freedom in the structure that are retained should be much less than the number of independent samples, or the results will have neither stability nor statistical significance. This is especially important when you have many more spatial grid points than independent samples, as is often the case in investigating interannual variability of global data fields. If the coupling between the fields is large and the sample size is sufficiently large, the spatial patterns should be insensitive to the number of modes retained over a range of truncations.

The retained PC time series are next normalized to make the variance over the sampling dimension unity for each PC. If the sampling dimension is time, this is just dividing each PC by its standard deviation in time. Hence all the PC time series are weighted equally, regardless of the amount of variance in the original data that they explain. After these modifications, the modified data matrices no longer contain the information necessary to reconstruct the original data sets.

The remainder of the analysis is very similar to MCA analysis. First construct the inner product across sampling dimension to form the covariance matrix between the two truncated and normalized PC data sets. Since these data set time series have been normalized, the covariance matrix is a correlation matrix between the retained PCs. The Frobenius norm of the correlation matrix may be interpreted as the total fraction of the variance of the left modified data set that is explained by the right modified data set, and vice versa.

Because the SVD is done on a correlation matrix, the singular values may be interpreted as correlation coefficients or "canonical correlations". SVD rearranges the PCs into combinations so that the first set in each modified input data series explains as much as possible of the correlation with the other modified data set. The structures in each field associated with these canonical correlations can be called the canonical correlation vectors, if you like.

**********end Dennis **********

5.11 Cluster Analysis

Cluster analysis is a popular technique in data mining and its goal is to group a set of observations into a number of distinct clusters. The end result is that each observation belongs to only one cluster, and this clustering is determined based on a particular cost function. Determining the optimal set of clusters is often an iterative process that begins with an initial guess.

Recall that the principal components in EOF analysis tell you how much a given observation looks like a particular EOF pattern. While, an observation may look predominately like one particular EOF (thus, the principal component for this EOF is large), it likely still has non-zero principal component values for the other EOFs. In cluster analysis, this is not the case, as each observation is tied to only one cluster.

5.11.1 k-*means Clustering*

k-means clustering is one of the more popular clustering techniques in Earth science due to its relative simplicity. The aim is to group N observations in k clusters in which each observation belongs to the cluster with the nearest center. This results in the data being partitioned into Voronoi cells. The optimal distribution of the cluster centers is the distribution that minimizes the within cluster sum of squares (i.e. the sum of the distance functions of each point to the cluster center). Mathematically, this is written as:

$$\underset{\mathbf{S}}{\operatorname{argmin}} = \sum_{i=1}^{k} \sum_{\mathbf{x} \in \mathbf{S}_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 \tag{5.62}$$

where $\mathbf{S} = S_1, S_2, ..., S_k$ and denotes the set of k clusters, $\mathbf{x}$ denotes the set of observations, $\boldsymbol{\mu}_i$ denotes the center of cluster S_i defined as the mean of all points in S_i, and argmin specifies that we are looking for the minimum over varying options of $\mathbf{S}$.

While the idea behind k-means is straightforward, actually computing the most optimal solution is very difficult and time intensive, especially for large data sets. Thus, most applications of k-means clustering employ heuristic algorithms that converge quickly to a local optimal solution that may or may not be the global optimal solution. Thus, in practice, one typically runs the k-means clustering algorithm many times (with each instance starting with a unique initial guess) and takes the iteration that produced the smallest minimum within cluster sum of squares (i.e. **(5.62)**).

The standard algorithm for k-means clustering is Lloyd's algorithm, which begins with an initial guess of the cluster centers $\boldsymbol{\mu}$ and then iteratively refines this guess until the it converges. **Fig. 5.8** shows an example of this iterative process.

To apply k-means clustering, two decisions that must be made:

1. **How many clusters do you want?** This is a decision that must be made by the user, and is not determined by the clustering algorithm. One will obtain different results based on the choice of k as seen in **Fig. 5.9**. There are methods such as the gap statistic that can be used to help determine which k to use (Tibshirani et al. 2001).
2. **How should you seed your initial guess?** There are many options available for how to best make your initial guess at the cluster centers. In addition, running k-means clustering many times, and choosing the optimal solution out of these is a way to avoid your final solution being too dependent on the initial guess.

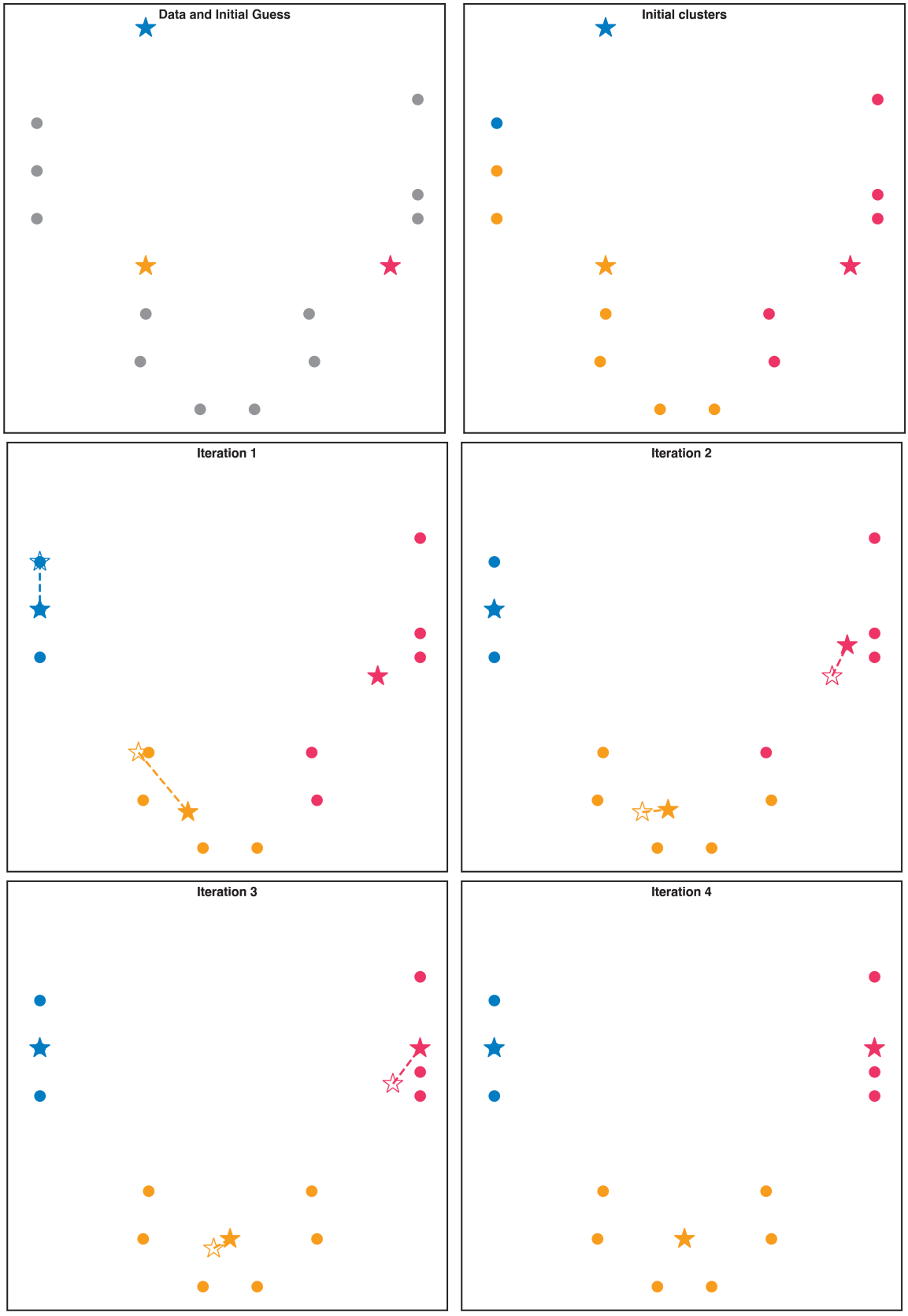

Figure 5.8 An example of the iterative refinement of k-means clustering using Lloyd's algorithm. Cluster centers are shows as stars, and empty stars denote the previous cluster center. The algorithm has converged after the third iteration.

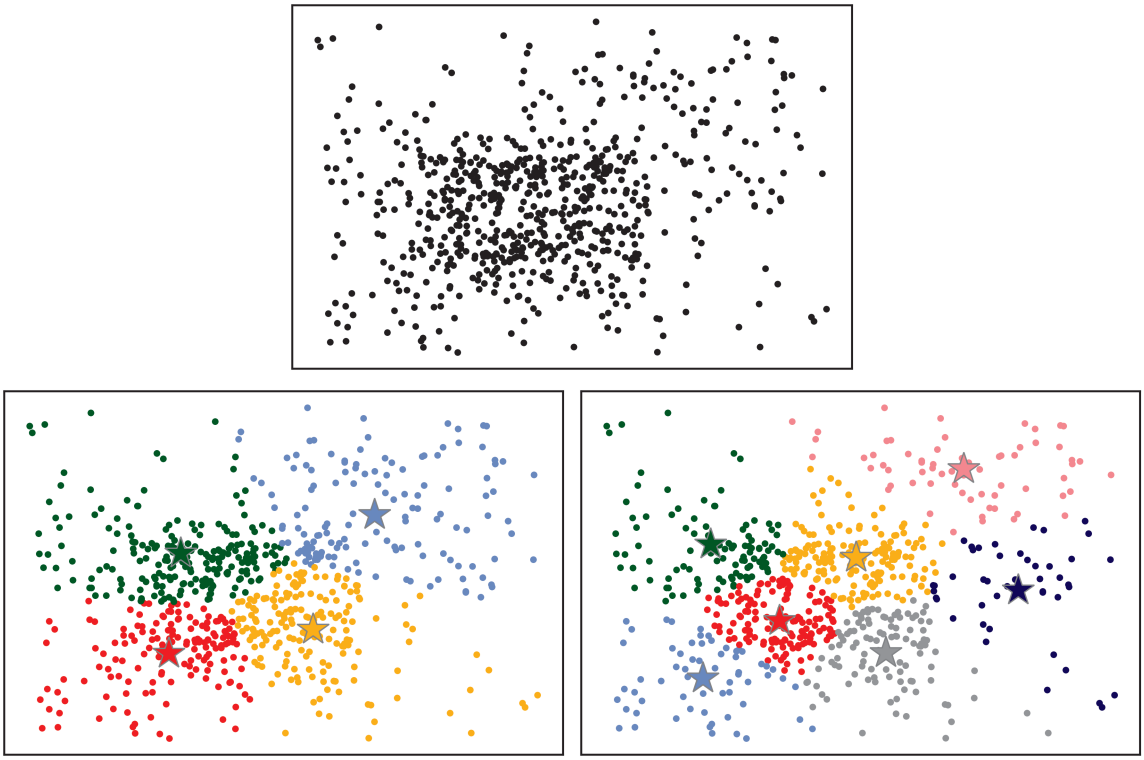

Figure 5.9 Using k-means clustering to identify 4 (bottom left) and 7 (bottom right) clusters in the data. The top panel shows the raw data. The stars denote the centroids of each of the seven clusters. The k-means algorithm was run 20 different times and the iteration with the optimal cost function was retained.

5.11.2 Self-Organizing Maps (SOMs)

A self-organizing map, also known as a Kohonen map, is yet another form of cluster analysis that uses unsupervised machine learning to train an artificial neural network. Under certain conditions it is identical to the k-means clustering method, however, the distinguishing feature of SOMs is that when one cluster center is updated, the neighboring cluster centers can also be updated in the a similar way. This results in a set of cluster centers (or *nodes*) that are organized such that nodes that are most similar are located near one another and nodes that are least similar are further apart when the are organized into an $m \times p$ grid.

Like most artificial neural networks, the SOMs algorithm can be summarized as a two step process: training and mapping.

1. **Training:** The training phase is when the SOM algorithm does the heavy thinking. During training, the SOM nodes are updated by comparing them to input examples (i.e. observations) that are ingested one at a time. Often, the same training data is iterated through multiple times. The result is a set of trained SOM nodes.
2. **Mapping:** After training is complete, the mapping phase involves automatically classifying a new observation as belonging to one of the nodes. In some cases, the training data is distinct from the data you are interested in mapping, while in other cases they may be one and the same.

As with k-means clustering, the user must first determine how many SOM nodes they wish to create to represent their data. Unlike k-means, however, this information is provided as an $m \times p$ grid, and thus, the user must determine both m and p. In the example below we will choose a SOM size of 20×20 giving a total of 400 nodes, however, it is important to note that the resuting SOM nodes are highly dependent on the dimensions chosen.

SOM training begins by first initializing the $m \cdot p$ nodes. This can be done either with random data, or using the principal components of the observations. Then, the algorithm proceeds by ingesting the training data either one observation at a time (*online training*) or as a single group (*batch training*). Then the observations are compared to each of the $m \cdot p$ nodes and the *best match unit* (BMU) is identified as the node with the smallest Euclidean distance to the observation in question. We will denote this BMU as n_i to signify the i^{th} node, where i could take a value from 1 to $m \cdot p$. Finally, the BMU and neighboring nodes are updated in the following manner:

$$n_j(t+1) = n_j(t+1) + \alpha(t) h_{ci}(t)[x(t) - n_i(t)] \tag{5.63}$$

where $1 \leqslant j \leqslant m \cdot p$, t is the training time, α is the learning rate parameter and h_{ci} is the neighborhood function. Specific descriptions of these inputs are given below.

- **training time (t):** how many iterations have been performed
- **learning rate parameter (α):** how strongly to update the nodes given the observation; typically decreases with training time
- **neighborhood function (h):** a function describing how many surrounding nodes to update besides the BMU; typically decreases with training time

The learning rate parameter and the neighborhood function must all be chosen by the user. Often, the learning rate parameter starts large, and then decreases linearly with training time. There are many neighborhood functions to choose from, but some common ones are a 2D Guassian or the Epanechikov function. It is typically considered good practice to initially set your neighborhood function parameters to include the entire SOM-space to ensure that all nodes are updated. This can be modified at later training times.

Typically, the SOM algorithm is implemented in multiple stages, with varying combinations of the above parameters.

Since SOM analysis has a large number of free parameters chosen by the user, there are many possible SOMs that can be computed from a single set of observations. One question is, how do I choose which one to use? Two key metrics should be considered:

- **quantization error:** the average distance between each observation and its BMU
- **topographic error:** the fraction of all observations who's first and second BMUs are *not* adjacent units

One wants to minimize both the quantization error and the topographic error, and typically, the topographic error is kept near zero while the quantization error is minimized. Thus, it is good practice to perform SOM analysis multiple times on the same data set, using different paramater values, and then choose the best based on the quantization and topographic errors.

To demonstrate SOM analysis and how the results can be visualized, we use hourly observations from Christman Field in Fort Collins, Colorado from January 1 - December 31, 2016. The data input into the algorithm is a 2D matrix X with dimensions (8784, 6). That is, 8784 hourly observation periods and 6 variables: temperature (degrees F), relative humidity (%), wind speed (mph), pressure (mb), solar radiation (W/m^2) and precipitation (mm). SOM training was performed in two phases: rough and finetune, each of which had different training lengths and initial and final radii of influence (for use in the neighborhood function). The SOM nodes were initialized with random data and a 20× 20 grid was chosen. The resulting SOM nodes are plotted in **Fig. 5.10**, where each panel denotes a different variable and each grid denotes a different SOM. For example, we can see that times of very strong wind are often associated with low relative humidity and precipitation. **Fig. 5.11** displays the frequency of occurrence of each of the SOM nodes (i.e. how many observations had that particular node as their BMU). We can see that the nodes on the left hand side of our grid tend to occur most frequently, and from **Fig. 5.10**, these periods are defined by low wind, high relative humidity, generally lo temperatures on average. [1]

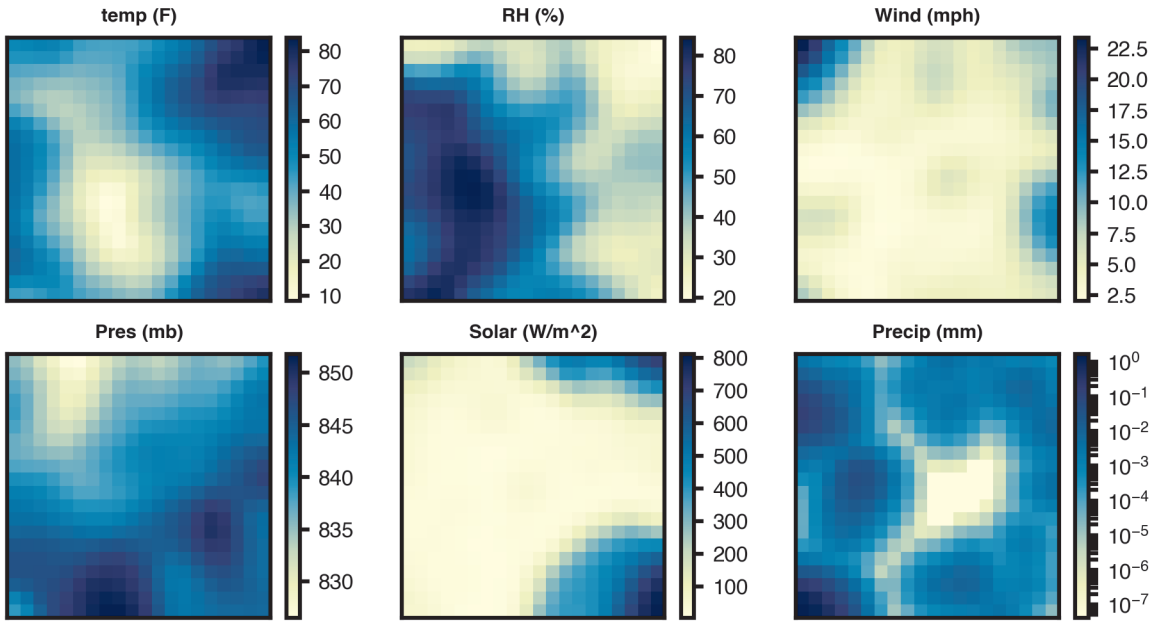

Figure 5.10 SOM weights for each variable displayed in a 20 × 20 grid.

[1] This example as created using the SOMPY code available here: https://github.com/sevamoo/SOMPY/tree/master/sompy

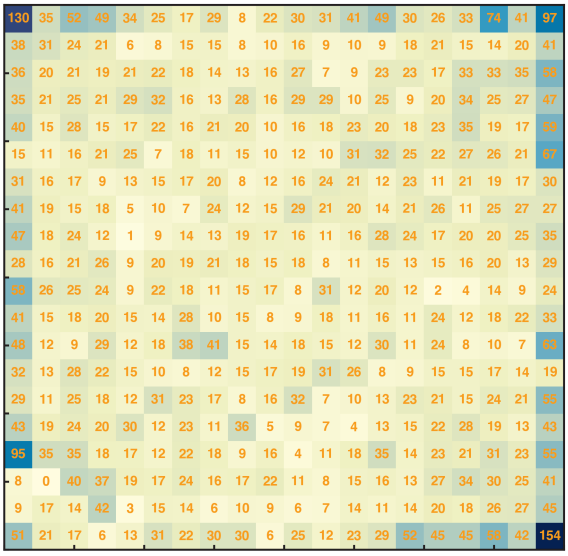

Figure 5.11 The frequency of occurrence of each of the 400 SOM patterns. Darker colors denote more frequent, and the number of occurrences out of the 8784 observations is written in orange.

Chapter 6
Mapping Data to a Grid and Data Assimilation

6.1 Placing data on a regular grid

In dynamical meteorology, oceanography, and numerical prediction one is often presented with the following problem. Data are available at a number of observation points (usually located near cities or at field stations, along ship cruise tracks, at moorings, or perhaps located by the observation points of an orbiting satellite) that are unevenly distributed over the domain of interest (the globe, for example). In order to compute derivatives of the field variables, as would be required in diagnostic studies or in the initialization of a numerical model, or simply to perform a sensible averaging process, one often requires values of the variables at points on a regular grid. Assigning the best values at the grid points, given data at arbitrarily located stations and perhaps a first guess at regular grid points, is what has traditionally been called objective analysis when done on a computer rather than graphically by hand.

We will use the example of making weather maps from rawinsonde data as the particular example of the mapping problem here. In fact the methods described are applicable to any problem where the data you are given do not fill the domain of interest fully, and/or where the data must be interpolated to a regular grid. The regridding can be in space, in time, or both. You may also find yourself in the position of wanting to plot a continuous function of an observation in two parameter dimensions, and have samples at only a few points. We will proceed through some of the methods in the order that they arose in the history of numerical weather forecasting. In this way we show the weaknesses of some of the most obvious methods such as function fitting, to the correction method, and ultimately to statistically optimized correction methods such as optimum interpolation. Current assimilation schemes in numerical forecast models us a combination of optimum interpolation and use of the governing equations of the model, which we can call *Kalman filtering*, which is discussed in elementary terms in Chapter **??**.

6.1.1 Interpolation with polynomial fits

Let's say we want to estimate the temperature at a point. However, we don't have any observations at that exact location. How might we use our observations to still get an estimate of the temperature at our point? The answer could be to perform some sort of interpolation. Probably one of the most intuitive methods for interpolating is to fit some polynomial to all of our station values, and then use that curve to get the temperature at a location between the observations. For example,

$$\Phi(x, y) = a_0 + a_1 x + a_2 x^2 + b_2 y^2 + 2c_2 xy + \dots \tag{6.1}$$

It turns out this isn't a very good method when you have sparse data due to the unstable nature of the polynomial fit. Removing just one point can wildly change the curve/interpolation in the vicinity of this point and will impact the values at many other points too. The problem gets worse as the order of the polynomial is increased. An example of this is depicted in **Fig. 6.1**. Note how wildly the two curves depart from each other in the vicinity of the missing point. Such problems can be avoided by stepping away from

polynomial fits and rather, utilizing a reasonable "first guess", and then only modifying it when and where data are available. Also, if the new data departs too wildly from the first guess, one suspects that the data are faulty.

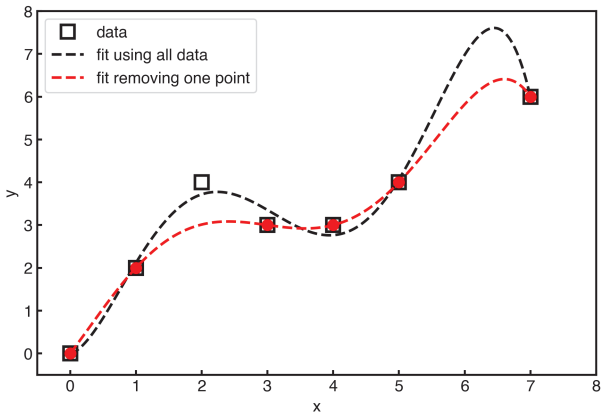

Figure 6.1 Illustration of the unstable nature of polynomial fits when one data point is removed using a 5th order polynomial.

A polynomial fit that actually got adopted by the US National Meteorological Center for its routine operational products was proposed by Flattery (1971). In this scheme, *Hough functions* were used as the interpolating polynomials. These functions are an orthogonal set that are the solutions of the linearized equations for a resting atmosphere (the tidal equations). The idea was that if you expressed the data in terms of actual solutions of the dynamical equations, then your fit between the data points would have some dynamical consistency. The Hough functions are global functions and so all of the observations were used simultaneously to define the global Hough function coefficients and produce a global map. Only the Hough functions describing slowly varying rotational modes were used. The gravity wave modes were zeroed out to produce a well-initialized field. This method replaced Cressman's correction method (Cressman, 1959) for global analyses in about 1972 and was replaced by Optimum Interpolation (see Chapter 6.1.2) in 1978.

This method has some dynamical and mathematical appeal, but is in truth just a glorified polynomial fit and has all of the problems of polynomial fits. First of all, the atmosphere is highly nonlinear and strongly forced by heating, especially in the tropics. The Hough modes chosen were primarily the free, non-divergent Rossby modes, which constitute a large, but not dominant, fraction of the variance. Therefore this aspect of the Flattery method did not buy much. In the tropics, where highly divergent motions forced by heating are important, the analyses constructed with the Flattery method are very much in error, especially in their estimates of divergence, which they set to essentially zero. In addition the Hough function fits are wildly unstable in regions of sparse data, like any polynomial fit. The NMC tropical analyses produced before 1978 are almost totally useless because they were made with the Flattery analysis system. Normal mode fits are still used in numerical initialization schemes to remove fast gravity waves, but this does not really affect the slowly changing meteorological flow. Modern reanalysis data products are based on data assimilation methods that take into account both the data and the model forecast and the uncertainty in both.

6.1.2 Optimum Interpolation

> *"The interpolation which is linear relative to the initial data and whose root-mean- square error is minimum is called the optimum interpolation." - Wiener, 1949*

The difference between optimum interpolation and linear regression is that the coefficients are not determined anew each time. Suppose we consider deviations from some "normal" state. This could be climatology or a

first guess, depending upon the application.

$$\phi^{'} = \phi - \phi_{norm} \quad \phi_{norm} = \overline{\phi} \text{ or a first guess} \tag{6.2}$$

Then we try to approximate the value of ϕ at a grid point, ϕ_g, in terms of a linear combination of the values of ϕ at neighboring station points, ϕ_s.

$$\phi_g^{'} = \sum_{i=1}^{N} p_i \phi_i^{'} \tag{6.3}$$

The coefficients p_i are to be determined by minimizing the mean squared error

$$E = \overline{\left(\phi_g^{'} - \sum_{i=1}^{N} p_i \phi_i^{'}\right)^2} \tag{6.4}$$

We can write the normalized error as

$$\epsilon \equiv \frac{E}{\overline{\phi_g^{'2}}} = 1 - 2\sum_{i=1}^{N} p_i r_{gi} + \sum_{i=1}^{N}\sum_{j=1}^{N} p_i p_j r_{ij} \tag{6.5}$$

$$\text{where } r_{gi} = \frac{\overline{\phi_g^{'}\phi_i^{'}}}{\overline{\phi_g^{'2}}} \quad r_{ij} = \frac{\overline{\phi_i^{'}\phi_j^{'}}}{\overline{\phi_g^{'2}}} \tag{6.6}$$

Differentation with respect to the coefficients leads to the condition of minimization used to determine them.

$$\frac{\partial \epsilon}{\partial p_i} = -2r_{gi} + 2\sum_{j=1}^{N} p_j r_{ij} = 0 \quad i = 1, 2, \ldots, N \tag{6.7}$$

(6.7) constitutes a system of N linear equations for the N p's. By substituting the conditions **(6.7)** into the expression for the error **(6.5)**, it can be shown that the error obtained after fitting the coefficients is

$$\epsilon = 1 - \sum_{i=1}^{2} r_{gi} p_i \tag{6.8}$$

Note that in this simple example, if one of the observation points, k, coincides with a grid point, then $r_{gk} = 1$, and we expect the regression procedure to return $p_k = 1$ and all the other weights zero. In this case the error is zero, $\epsilon = 0$, since we have assumed the data are perfect. If the station points are uncorrelated with the grid point in question, then $p_i = 0$ and $\epsilon = 1$, the climatic norm. That is, the error will equal the standard deviation, but no worse.

6.1.2.1 Adding measurement error

In what we have done so far the observations have been assumed to be perfect. Let us now consider what happens if we explicitly take account of the fact that our observations will always contain some error, δ_i.

$$\phi_i^{'} = \phi_{ia}^{'} + \delta_i \tag{6.9}$$

Let's assume, as is usually reasonable, that the error is unbiased (zero mean) and ucorrelated with the true value, that is,

$$\overline{\phi_{ia}^{'}\delta_i} = 0 \tag{6.10}$$

and that the errors at the various stations where we have data are also uncorrelated

$$\overline{\delta_i \delta_j} = \begin{cases} \overline{\delta^2} \text{ for } i = j \\ 0 \text{ for } i \neq j \end{cases} \tag{6.11}$$

In this case, rather than **(6.5)**, we obtain

$$\epsilon = 1 - 2\sum_{i=1}^{N} p_i r_{gi} + \sum_{i=1}^{N}\sum_{j=1}^{N} p_i p_j r_{ij} + \eta \sum_{i=1}^{N} p_i^2 \tag{6.12}$$

where r_{ij} is the correlation between the two points and where η is the ratio of the error variance to the measurement variance - in other words, the *signal-to-noise ratio.*

$$\eta = \frac{\overline{\delta^2}}{\overline{\phi_g'^2}} \tag{6.13}$$

Minimization of the error leads to the condition

$$\sum_{j=1}^{N} r_{ij} p_j + \eta p_i = r_{gi} \quad \text{for } i = 1, 2, 3, \ldots, N \tag{6.14}$$

In this case the normalized error is

$$\epsilon = 1 \sum_{i=1}^{N}\sum_{j=1}^{N} p_i p_j r_{ij} + \eta \sum_{i=1}^{N} p_i^2 \tag{6.15}$$

6.1.2.2 What is the effect of including noise in the measurements?

In order to see how optimum interpolation treats the *a priori* information that the measurements include some error, it is instructive to compare the results **(6.14)** and **(6.15)** with the results **(6.5)** and **(6.7)** obtained assuming perfect data. In the case of perfect data, **(6.7)** gives

$$r_{ij} p_j = r_{gi} \quad \text{or } p_j = r_{ij}^{-1} r_{gi} \tag{6.16}$$

When noise is included we get, rather, the result **(6.14)**, which can be written

$$\{r_i j + \eta \mathbf{I}_{ij}\} p_j = r_{gi} \quad \text{or } p_j = \{r_i j + \eta \mathbf{I}_{ij}\}^{-1} r_{gi} \tag{6.17}$$

where $\mathbf{I}_{ij}$ is the unit matrix. Looking at the right-hand member of the pair of equations in **(6.17)**, it is easy to see that the coefficients p_j will be smaller when the error is large. This is most obvious if we assume that $r_i j$ is diagonal. Thus we see that the inclusion of error makes the coefficients in **(6.3)** smaller and that therefore, by **(6.2)**, the estimate we make will be closer to climatology. If we include error, then Optimum Interpolation will draw more closely to climatology or the first guess and tend to weight new observations less heavily. This is desirable. By putting different values of η_j along the diagonal, one can put information on the confidence one has in individual stations into the analysis scheme and weight more heavily those stations in which one has more confidence.

6.1.2.3 What do we need to make Optimum Interpolation work?

In order to make the above schemes work, we need the correlation matrices r_{ij} and r_{gi}. The first of these is easily calculable from observations, but the second is not since it involves correlations between the station points and the grid points. We do not have data at the grid points, or we would not need an analysis scheme. In practice, not even the r_{ij} are calculated in full generality. It is possible to assume that correlations between

points depend only on the distance between them and not on location or direction (although it would be possible to include directionally dependent (anisotropic) correlations). In this case the single isotropic correlation function can be estimated from station data. This is a crude approximation since correlations between stations depend on the location of the stations and whether longitude or latitude separates them. An example illustrating the anisotropy of correlation functions in 500 hPa geopotential heights is shown in **Fig. 6.2**.

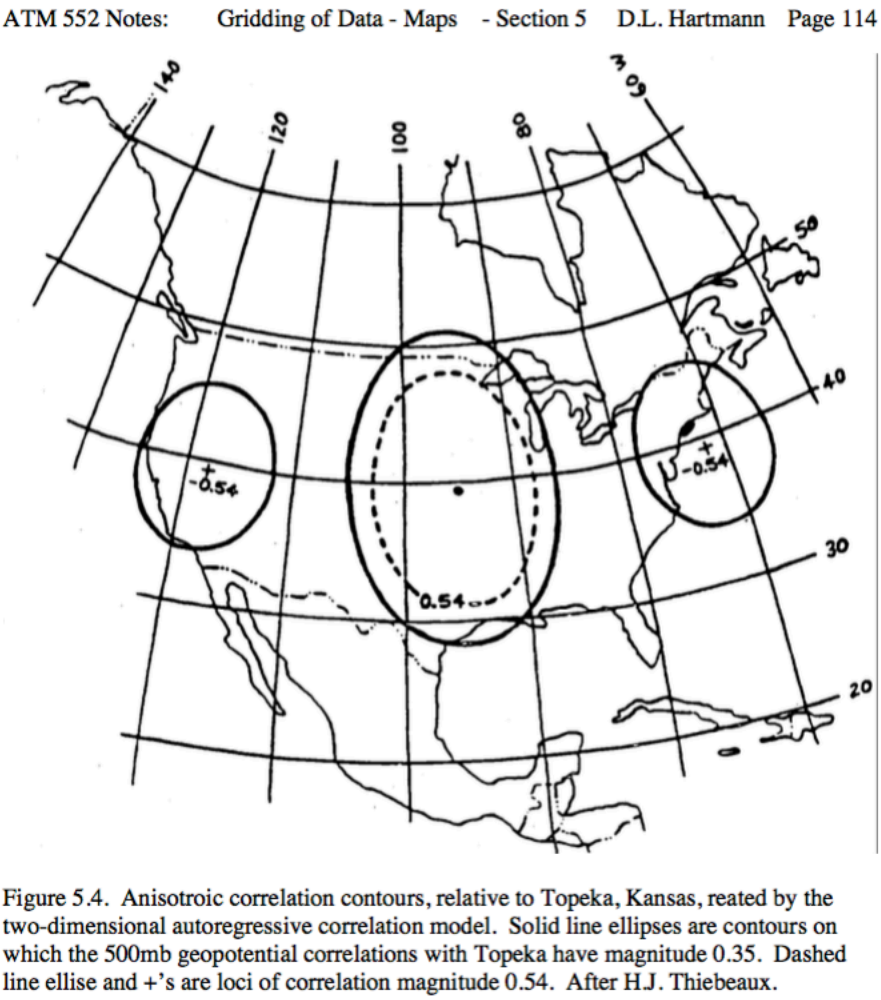

Figure 5.4. Anisotroic correlation contours, relative to Topeka, Kansas, reated by the two-dimensional autoregressive correlation model. Solid line ellipses are contours on which the 500mb geopotential correlations with Topeka have magnitude 0.35. Dashed line ellise and +'s are loci of correlation magnitude 0.54. After HJ. Thiebeaux.

Figure 6.2 Anisotropic correlation contours, relative to Topeka, Kansas, created by a two-dimensional autoregressive correlation model. Solid line ellipses are contours on which the 500 hPa geopotential correlations with Topeka have magnitude 0.35. Dashed lines denote the correlation of 0.54 and +'s are loci of correlation magnitude 0.54. After H.J. Thiebeaux.

Libby: *stopped at Dennis' Chapter 5.4*

Chapter 7
Time Series Analysis

7.1 Introduction

In this chapter we will consider some common aspects of time series analysis including autocorrelation, statistical prediction, harmonic analysis, power spectrum analysis, and cross-spectrum analysis. We will also consider space-time cross spectral analysis, a combination of time-Fourier and space-Fourier analysis, which is often used in meteorology. The techniques of time series analysis described here are frequently encountered in all of geoscience and in many other fields. We will spend most of our time on classical Fourier spectral analysis, but will mention briefly other approaches such as Maximum Entropy (MEM), Singular Spectrum Analysis (SSA) and the Multi-Taper Method (MTM). Although we include a discussion of the historical Lag-correlation spectral analysis method, we will focus primarily on the Fast Fourier Transform (FFT) approach.

7.2 Autocorrelation and Red Noise

7.2.1 The Autocorrelation Function

Given a continuous function $x(t)$, defined in the interval $t_1 < t < t_2$, the autocovariance function is

$$\Phi(\tau) = \frac{1}{t_2 - t_1 - \tau} \int_{t_1}^{t_2-\tau} x'(t)x'(t+\tau)dt \tag{7.1}$$

where primes indicate deviations from the mean value, and we have assumed that $\tau > 0$. In the discrete case where x is defined at N points spaced at an interval of Δt , $k = 1, 2, .., N$, we can calculate the autocovariance at lag L.

$$\Phi(L) = \frac{1}{N - 2L} \sum_{k=L}^{N-L} x'_k x'_{k+L} = \overline{x'(t)x'(t + L\Delta t)} \tag{7.2}$$

The autocovariance is the covariance of a variable with itself (Greek autos = self) at some other time, measured by a time lag (or lead) τ. Note that $\Phi(0) = \overline{x'^2}$, so that the autocovariance at lag zero is just the variance of the variable.

The Autocorrelation function is the normalized autocovariance function $r(\tau) = \Phi(\tau)/\Phi(0)$ and $-1 < r(\tau) < 1$. If x is not periodic then $r(\tau) \to 0$, as $\tau \to \infty$ It is normally assumed that data sets subjected to time series analysis are stationary . The term stationary time series normally implies that the true mean of the variable and its higher-order statistical moments are independent of the particular time in question and so do not vary within the sample. Therefore it is usually necessary to remove any trends in the time

series before analysis. This also implies that the autocorrelation function can be assumed to be symmetric, $r(\tau) = r(-\tau)$.

7.2.2 White Noise

In the special case $r(\Delta t) = a = 0$, our time series is a series of random numbers, uncorrelated in time so that $r(\tau) = \delta(0)$ a delta function. For such a "white noise" time series, the present value is of no help in projecting into the future. The probability density function of white noise can vary, but we will generally use Gaussian Normal white noise whose probability distribution is Gaussian around a mean value of zero. Figure 7.1 shows a sample of Gaussian and uniformly distributed white noise, along with their corresponding sample probability density functions. Gaussian noise is both more likely to be near zero and to depart far from zero than uniformly distributed noise. A Gaussian distribution fits many naturally occurring time series. In both cases the autocorrelation is zero for any nonzero lag.

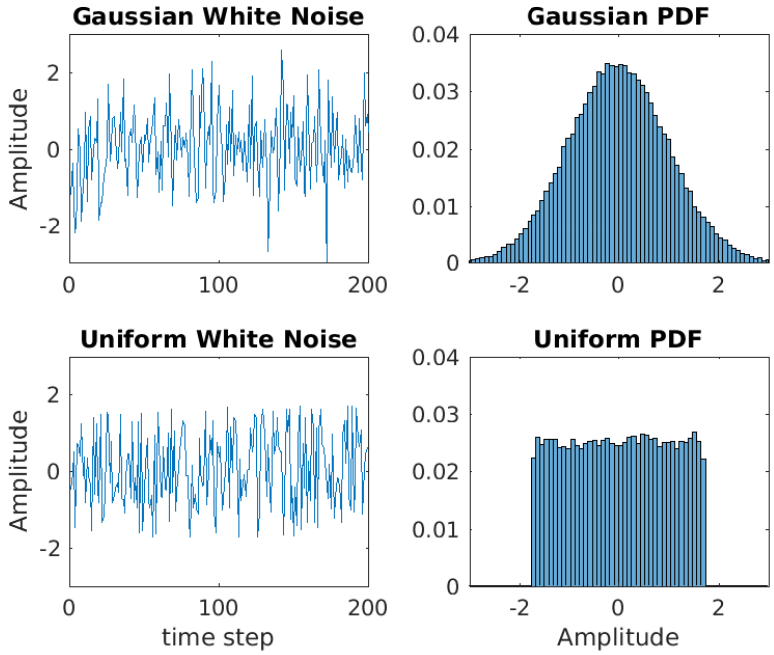

Figure 7.1 Samples of Gaussian and Uniform White Noise and their sample probability distributions. Both samples have a standard deviation of 1.0. The probability distributions are based on a sample of 50,000 time steps.

7.2.3 Red Noise

We define a "red noise" time series as being of the form:

$$x(t) = ax(t - \Delta t) + \left(1 - a^2\right)^{1/2} \epsilon(t) \tag{7.3}$$

where x is a standardized variable $\overline{x} = 0$ and $\overline{x'^2} = 1$. The constant a is on the interval between zero and one $(0 < a < 1)$ and measures the degree to which memory of previous states is retained. The function ϵ represents a series of random numbers drawn from a standardized normal distribution, and Δt is the time

interval between data points. This is also called a Markov Process or an Auto-Regressive, or AR-1 Process, since it remembers only the previous value.

Multiply 7.3 by $x(t-\Delta t)$ and average over time to show that a is the one-lag autocorrelation, or the autocorrelation at one time step Δt.

$$\overline{x(t)x(t-\Delta t)} = a\overline{x(t-\Delta t)x(t-\Delta t)} + \left(1-a^2\right)^{1/2}\overline{\epsilon(t)x(t-\Delta t)} \tag{7.4}$$

Since the time series has unit variance, and the noise is uncorrelated with the time series, 7.4 becomes,

$$r(\Delta t) = a \tag{7.5}$$

so that a is the autocorrelation at one time step.

Using 7.4 multiple times we can show how the autocorrelation depends on the time interval.

$$\begin{aligned} x(t+\Delta t) &= ax(t) + \left(1-a^2\right)^{1/2}\epsilon(t) \\ x(t+\Delta t) &= a(ax(t-\Delta t) + \left(1-a^2\right)^{1/2}\epsilon(t)) + \left(1-a^2\right)^{1/2}\epsilon(t) \\ \overline{x(t+\Delta t)x(t-\Delta t)} &= a^2\overline{x(t-\Delta t)x(t-\Delta t)} + 0 \\ r(2\Delta t) &= a^2 \end{aligned} \tag{7.6}$$

From 7.6 we determine by induction that,

$$r(n\Delta t) = a^n \tag{7.7}$$

So for a red noise time series, the autocorrelation at a lag of n time steps is equal to the autocorrelation at one lag, raised to the power n. A function that has this property is the exponential function, $e^{nx} = \left(e^x\right)^n$, so we may hypothesize that the autocorrelation function for red noise has an exponential shape.

$$r(n\Delta t) = \exp\left(-n\Delta t/T\right) \tag{7.8}$$

where $T = -\Delta t/\ln a$ is the e-folding time of the autocorrelation, and if $\tau = n\Delta t$, then

$$r(\tau) = \exp\left(-|\tau|/T\right) \tag{7.9}$$

The autocorrelation function 7.9 is shown in Fig. 7.2.

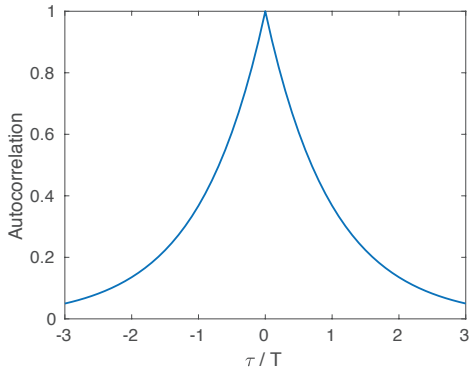

Figure 7.2 Autocorrelation function for red noise 7.9 plotted as a function of lag time τ divided by e=folding time T.

In summary, if we are given an auto-regressive (AR-1) process 7.3, then its autocorrelation is given by 7.9, where the e-folding time is $T = -\Delta t/\ln a$.

7.3 Statistical Prediction and Red Noise

Consider a prediction equation of the form

$$\widehat{x}(t+\Delta t) = a_1 x(t) + a_2 x(t-\Delta t) \tag{7.10}$$

where a_1 and a_2 are chosen to minimize the root-mean-square error on dependent data. Recall from our discussion of multiple regression that for two predictors x_1 and x_2 used to predict y, the second predictor is only useful if,

$$\left|r(x_2, y)\right| \geqslant \left|r(x_1, y) r(x_1, x_2)\right| \tag{7.11}$$

In the case where the equality holds, $r(x_2, y)$ is equal to the "minimum useful correlation" discussed in 4.3.3 and will not improve the forecasting skill beyond the level possible by using x_1 alone. In the case of trying to predict future values from prior times, $r(x_2, y) = r(2\Delta t)$, and $r(x_1, y) = r(x_1, x_2) = r(\Delta t)$ so that we must have,

$$\left|r(2\Delta t)\right| \geqslant r(\Delta t)^2 \tag{7.12}$$

in order to justify using a second predictor at two time steps in the past. Note that for red noise

$$r(2\Delta t) = r(\Delta t)^2 \tag{7.13}$$

so that the value at two lags previous to now always contributes exactly the minimum useful, and nearly automatic, correlation, and not more. For red noise, then, nothing is gained by using a second predictor. All we can use productively is the present value and the autocorrelation function. Using one predictor one time step before is called the persistence forecast, where we assume tomorrow will be like today.

7.4 Degrees of Freedom with Gaussian Red Noise

For a red noise process, adjacent values are correlated and so not independent. One cannot gain more information about a red noise process by sampling it at finer and finer temporal resolution. Information, or degrees of freedom, increase as a function of sample length. Leith (1973) used a Gaussian red noise model to assess the number of degrees of freedom for assessing the uncertainty of sample means. He proposed that the number of independent samples $N*$ contained in a time series of N time steps separated by Δt with an e-folding time of $T = -\Delta t / \ln(r\Delta t)$ is given by,

$$N^* = \frac{N\Delta t}{2T} = \frac{\text{Total time series length}}{\text{Two times the autocorrelation e-folding time}} \tag{7.14}$$

In other words, the number of degrees of freedom is the total length of the time sample divided by twice the e-folding time of the autocorrelation. Separation of two e-folding times between independent measurements is required since the intervening points can be mostly predicted by two points separated by a smaller time interval than $2T$. Leith's formula can also be written as,

$$\frac{N^*}{N} = \frac{1}{2}\ln(r(\Delta t)) \tag{7.15}$$

Leith's formula is consistent with Taylor (1921), who said that

$$\frac{N^*}{N} = \frac{1}{2L} \tag{7.16}$$

where L is given by

$$L = \int_0^\infty r(\tau')d\tau' \tag{7.17}$$

If we substitute the formula for the autocorrelation of red noise 7.8 into 7.17, and take into account that Taylor was using non-dimensional time $t' = t/\Delta t$, then we can show that $L = T$, and Leith's formula is the same as Taylor's.

The factor of two comes into the bottom of the above expressions for N* so that the intervening point is not easily predictable from the ones immediately before and after. If you divide the time series into units of e-folding time of the auto-correlation, T, One can show that, for a red noise process, the value at a midpoint, which is separated from its two adjacent points by the time period T, can be predicted from the two adjoining values with combined correlation coefficient of about $2e^{-1}$, or about 0.52, so about 25% of the variance can be explained at that point, and at all other intervening points more can be explained.

Bretherton et al. (1999) provide a nice review of efforts to assess spatial and temporal degrees of freedom. They use an approach in which the spatial or temporal statistics are fitted to a Chi-Squared distirbution. An alternative formula for effective degrees of freedom to be used in assessing the statistical significance of temporal means is given as,

$$\frac{N^*}{N} = \frac{(1 - r_1(\Delta t))}{(1 + r_1(\Delta t))} \tag{7.18}$$

If one is looking at a first order process, such as the calculation of a mean value, or the computation of a trend where the exact value of the time is know, then the formula 7.18 should be used.

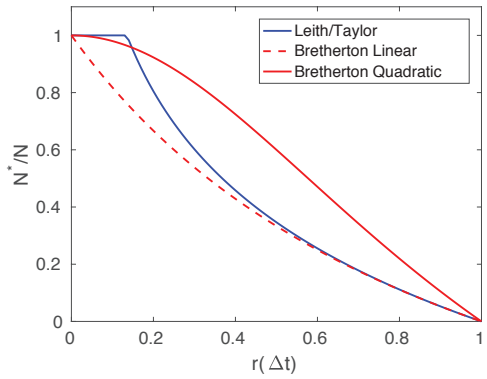

Figure 7.3 Ratio of degrees of freedom N^* to sample size N as a function of autocorrelation at one time step $r(\Delta t)$ for use in assessing the statistical significance of means(linear) and correlations (quadratic).

For quadratic statistics, such as variance and covariance analysis between two variables x_1 and x_2, a good approximation to use is:

$$\frac{N^*}{N} = \frac{(1 - r_1(\Delta t)r_2(\Delta t)}{(1 + r_1(\Delta t)r_2(\Delta t)} \tag{7.19}$$

where, of course, if we are covarying a variable with itself, then $r_1(\Delta t)r_2(\Delta t) = r(\Delta t)^2$. This formulation goes back as far as Bartlett (1935). Of course, if the time or space series is not Gaussian red noise, then the formula is not accurate. But it is still good practice to use it, and many geophysical time series contain a good measure of red noise.

Figure 7.3 shows $N*/N$ as a function of $r(\Delta t)$ for the formulas introduced in section 7.4. Because in quadratic statistics such as covariance, the noise is multiplied by itself, the reduction of degrees of freedom with increasing autocorrelation is slower than for linear statistics such as the mean value.

7.5 Harmonic Analysis and the Fourier Transform

Harmonic analysis is the interpretation of a time or space series as a summation of contributions from harmonic functions, each with a characteristic time or space scale. Consider that a variable $y(t)$ is defined on the interval $0 < t < T$. Then we can write any time series $y(t)$ as

$$y(t) = \sum_{k=0}^{\infty} \left[A_k \cos\left(2\pi k \frac{t}{T}\right) + B_k \sin\left(2\pi k \frac{t}{T}\right) \right] \tag{7.20}$$

where T is here the length of the period of record and k is a positive integer. $y(t)$ is a continuous function of t. The coefficients of this expansion in sines and cosines can be obtained by multiplying by a test function on the left, say $\cos\left(2\pi n \frac{t}{T}\right)$, where n is any positive integer.

Because the sines and cosines are orthogonal to each other on the interval $0 < t < T$, after we integrate over time we obtain simple algebraic equations for the coefficients.

$$A_0 = \overline{y} \qquad A_k = \frac{2}{T}\int_0^T y(t)\cos\left(2\pi k \frac{t}{T}\right)dt \qquad B_k = \frac{2}{T}\int_0^T y(t)\sin\left(2\pi k \frac{t}{T}\right)dt \qquad k > 0 \tag{7.21}$$

A time series of real data is usually presented at discrete values of time separated by a time step Δt. In that case the integrals presented in 7.21 are approximated with a summation over the time series consisting of a sample of N equally-spaced values.

$$A_0 = \overline{y} \qquad A_k = \frac{2}{N}\sum_{i=1}^{N} y(t_i)\cos\left(2\pi k \frac{i\Delta t}{T}\right) \qquad B_k = \frac{2}{N}\sum_{i=1}^{N} y(t_i)\sin\left(2\pi k \frac{i\Delta t}{T}\right) \qquad k > 0 \tag{7.22}$$

Here $T = N\Delta t$. Where the data are not equally spaced, A_k and B_k can be estimated by regression. In the case of equally spaced data, one can estimate N coefficients from N data points, but the $k = N/2$ is the maximum wavenumber that can be computed. $A_{k=0} = \overline{y}$ is the mean of the time series and $B_{k=0} = B_{k=N/2} = 0$. The highest wavenumber that can be computed, k=N/2, is for a wavelength of $2\Delta t$ for which the amplitude, but not the phase can be computed. The highest resolvable frequency $1/2\Delta t$, is half the sampling frequency and is called the "Nyquist" frequency after Harry Nyquist, a Swedish born American electronic engineer. The highest frequency resolved thus depends on the sampling interval. The lowest frequency resolved is $1/T$, and is thus determined by the total length of the sample. The separation between frequencies is also $1/T$, which can be called the "bandwidth" of the analysis. The frequencies resolved by a sample of N values separated by Δt are thus,

$$f_i = \frac{i}{T}; \qquad i = 0, 1, 2, ...N/2 \tag{7.23}$$

Since the sines and cosines are orthogonal on the interval $0 < t < T$, the time series can be reconstructed exactly from the Fourier coefficients.

$$y(t_i) = \overline{y} + \sum_{k=1}^{N/2} A_k \cos\left(2\pi k \frac{i\Delta t}{T}\right) + B_k \sin\left(2\pi k \frac{i\Delta t}{T}\right) \tag{7.24}$$

This can be rearranged slightly to be,

$$y(t) = \overline{y} + \sum_{k=1}^{N/2} C_k \cos\left(\frac{2\pi k}{T}(t - t_k)\right) \tag{7.25}$$

where,

$$C_k^2 = A_k^2 + B_k^2 \quad \text{and} \quad t_k = \frac{T}{2\pi k}\arctan\left(\frac{B_k}{A_k}\right) \tag{7.26}$$

In 7.26 the time series is represented by a summation of cosine waves, each with an amplitude C_k and a phase delay t_k.

7.6 The Power Spectrum

In many cases of interest, it is useful to know how the variance of a time series is distributed across the frequency domain. Having performed the Fourier analysis of a discrete time series and expressed it in terms of a set of cosine waves with different frequencies, we can easily express the variance of y in the following way.

$$\overline{y'^2} = \frac{1}{2}\sum_{k=1}^{N/2} C_k^2 \tag{7.27}$$

So the power as a function of frequency f_k can be written as,

$$\Phi(f_k) = \frac{1}{2}C_k^2 \qquad \text{where} \qquad f_k = \frac{k}{T} \tag{7.28}$$

From 7.27 and 7.28 we infer that if we plot the power spectrum as a function of frequency, then the area under the curve will be proportional to the variance, a useful thing. If the frequency range is very large, one can plot $f_k\Phi(f_k)$ versus $\log f_k$ and preserve the area - variance relationship. If the power range and frequency range are both very large, one can plot the log of power versus the log of frequency, but in that case the area under the curve is no longer proportional to variance, and one must be careful when interpreting the contribution of different frequencies to the total variance. Since the power spectrum is based on N data points, and the resulting power spectrum has only $N/2$ values, each realization of a power spectrum has about two degrees of freedom. We don't need to worry about autocorrelation, since spectral analysis takes this into account explicitly.

7.7 Methods of Computing Power Spectra

7.7.1 Direct Fourier Transform

The power spectrum can be obtained by direct Fourier transform, as described in section 7.5. This is feasible for large data sets because of the Fast Fourier Transform (FFT), which greatly speeds up the computation of Fourier transforms if the length of the record is a power of two, $N = 2^m$, where m is any integer. Mixed radix FFTs are also available for $N = 2^m 3^n 5^p$. As previously mentioned, however, each spectral estimate has only about 2 degrees of freedom. To give the estimate of the power spectrum more statistical robustness, one can average adjacent frequencies together, or divide the record up into shorter segments, and average the spectral estimates from these shorter segments into an average spectrum with more degrees of freedom. In either case, one is forced to consider a trade off between spectral resolution, which depends on the length of the chunk of data given to the spectral analysis, and the number of degrees of freedom in the final spectrum. If L spectral estimates result from the analysis, and N total data are used to make that estimate, then the number of degrees of freedom is slightly more than N/L. Generally, the more degrees of freedom in the final spectral estimate, the better the quality of the analysis. One should also insist on adequate quality.

7.7.2 Lag Correlation Method

Norbert Wiener showed that the autocovariance and the power spectrum are Fourier transforms of each other. So we can obtain the power spectrum by performing harmonic analysis on the autocovariance function. This

was the method preferred before fast computers, because the number of computations required is much less than a direct Fourier transform. The number of lags can be chosen to achieve the desired frequency resolution, and the number of degrees of freedom of the resulting power spectrum increases with the length of the available record. Suppose we consider time lags, τ on the interval $T_L < \tau < T_L$. The Fourier transform pair of the continuous spectrum and the continuous lag correlation are then,

$$\Phi(k) = \int_{-T_L}^{T_L} r(\tau) e^{-ik\tau} d\tau \tag{7.29}$$

$$r(\tau) = \int_{-k^*}^{k^*} \Phi(k) e^{ik\tau} dk \tag{7.30}$$

The maximum lag, T_L, which is $L\Delta t$ in discrete mathematics, determines the bandwidth of the spectrum and the number of degrees of freedom associated with each spectral estimate. The bandwidth is $1/2T_L$, and frequencies resolved are $k/(2L\delta t$, where $k = 0, 1, 2, ...L$. There are L spectral estimates produced, and if N data are used to compute them, then each spectral estimate has about N/L degrees of freedom. So if $N = 1000$ data points are used to compute spectral estimates at $L = 50$ frequencies, each estimate has 20 degrees of freedom.

The lag correlation method is rarely used nowadays, because Fast Fourier Transform algorithms are more efficient and widespread. The lag correlation method is important for intellectual and historical reasons, and because it comes up again in higher order spectral analysis.

7.8 The Complex Fourier Transform and Spectral Analysis

In previous sections we presented the Fourier Transform in real arithmetic using sine and cosine functions. It is much more compact and efficient to write the Fourier Transform and its associated manipulations in complex arithmetic. In a domain of continuous time and frequency, we can write the Fourier Transform Pair as integrals:

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} F(\omega) e^{i\omega t} d\omega \tag{7.31}$$

$$F(\omega) = \int_{-\infty}^{+\infty} f(t) e^{-i\omega t} dt \tag{7.32}$$

Here $f(t)$ is some real time series in the independent variable t, and $F(\omega)$ is the Fourier Transform of $f(t)$, and is generally a complex number with a real and imaginary part. ω is the frequency in radians per unit time. If the period of the oscillation is T, then the radian frequency is $\omega = 2\pi/T$, and the frequency in cycles per unit time is $f = 1/T$.

7.8.1 Parseval's Theorem

The following theorem by Parseval is important in spectral analysis and filtering theory. It is derived by considering two functions $f_1(t)$ and $f_2(t)$ with Fourier Transforms $F_1(\omega)$ and $F_2(\omega)$. We consider the following manipulation.

$$\begin{aligned}\int_{-\infty}^{+\infty} f_1(t)f_2(t)dt &= \int_{-\infty}^{+\infty} f_1(t)\left[\frac{1}{2\pi}\int_{-\infty}^{+\infty} F_2(\omega)e^{i\omega t}d\omega\right]dt \\ &= \frac{1}{2\pi}\int_{-\infty}^{+\infty} F_2(\omega)\int_{-\infty}^{+\infty} f_1(t)e^{i\omega t}dtd\omega \\ &= \frac{1}{2\pi}\int_{-\infty}^{+\infty} F_2(\omega)F_1(-\omega)d\omega \\ &= \frac{1}{2\pi}\int_{-\infty}^{+\infty} F_1(\omega)F_2^*(\omega)d\omega\end{aligned} \tag{7.33}$$

Here the asterisk indicates a complex conjugate. Since the original time series are real, we must have that $F(-\omega) = F^*(\omega)$. In the special case where $f_1(t) = f_2(t) = f(t)$ we obtain,

$$\int_{-\infty}^{+\infty} f(t)^2 dt = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \left|F(\omega)\right|^2 d\omega \tag{7.34}$$

Thus the square of the time series integrated over time is equal to the square (inner product) of the Fourier transform integrated over frequency. This shows that the integrated variance in time is equal to the power spectrum integrated over frequency.

7.8.2 The Time Shifting Theorem

The time shifting theorem indicates that introducing a time shift in complex Fourier space is a simple multiplication by a complex number. To see this consider the Fourier transform of a time series shifted by a time increment τ.

$$f(t \pm \tau) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} F(\omega)e^{i\omega(t\pm\tau)}d\omega = \frac{1}{2\pi}\int_{-\infty}^{+\infty} F(\omega)e^{\pm i\omega\tau}e^{i\omega t}d\omega \tag{7.35}$$

From 7.35 we see that the Fourier transform of $f(t \pm \tau)$ is the Fourier transform of $f(t)$ multiplied by the factor $e^{\pm i\omega\tau}$.

7.8.3 Lagged Covariance and the Power Spectrum

The continuous form of the definition of the lag covariance function for a time series $f(t)$is,

$$r(\tau) = \int_{-\infty}^{+\infty} f(t)f(t+\tau)dt \tag{7.36}$$

We can use Parseval's theorem and the time shifting theorem to write,

$$\begin{aligned} r(\tau) &= \int_{-\infty}^{+\infty} f(t) f(t+\tau) dt = \frac{1}{2\pi} \int_{-\infty}^{+\infty} F(\omega) F^*(\omega) e^{i\omega\tau} d\omega \\ &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} \Phi(\omega) e^{i\omega\tau} d\omega \end{aligned} \tag{7.37}$$

where $\Phi(\omega)$ is the power spectrum. The power spectrum is thus the Fourier Transform of the autocovariance function, so that they form a Fourier transform pair.

$$\begin{aligned} r(\tau) &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} \Phi(\omega) e^{i\omega\tau} d\omega \\ \Phi(\omega) &= \int_{-\infty}^{+\infty} r(\tau) e^{-i\omega\tau} d\tau \end{aligned} \tag{7.38}$$

This methodology can also be applied to the covariance between two different time series $f_1(t)$ and $f_1(t)$. A similar relationship occurs between the covariance between two time series and the cross power $\Phi_{12}(\omega)$. The cross power has a real part, or cospectrum, and an imaginary part, or quadrature spectrum, which together specify the phase between the two time series.

$$\begin{aligned} r_{12}(\tau) &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} \Phi_{12}(\omega) e^{i\omega\tau} d\omega \\ \Phi_{12}(\omega) &= \int_{-\infty}^{+\infty} r_{12}(\tau) e^{-i\omega\tau} d\tau \end{aligned} \tag{7.39}$$

7.8.4 Example: The Power Spectrum of Red Noise

The autocorrelation function for red noise is,

$$r(\tau) = e^{-\tau/T} \tag{7.40}$$

where T is the e-folding time of the autocorrelation, *i.e.* autocorrelation time. Using 7.38 and inserting 7.40 we obtain,

$$\Phi(\omega) = \int_{-\infty}^{+\infty} e^{-\tau/T} e^{-i\omega\tau} d\tau = \frac{2T}{1+\omega^2 T^2} \tag{7.41}$$

Figure 7.4 shows that the power spectrum of red noise peaks strongly at low frequencies as the autocorrelation time, T, increases. Note that the range of frequency in this theory is infinite, so as $T \to 0$ $\Phi(\omega) \to 0$, but takes on a uniform value for $0 < \omega < \infty$. The total variance is unchanged. In Figure 7.4 we show only the range $0 < \omega < \pi$, which would be the Nyquist range for $\Delta t = 1$ using discrete data. Later we will consider how this theoretical spectrum for red noise is modified for a finite sample of discrete data.

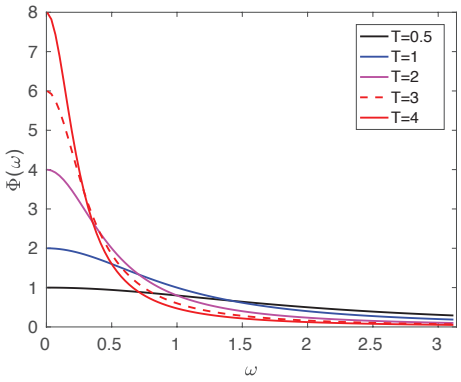

Figure 7.4 Power spectrum as a function of frequency in radians for the theoretical spectrum in 7.41 for autocorrelation e-folding times of 0.5, 1, 2, 3, and 4 time units.

7.9 Data Windows and Window Carpentry

In the analytic case we presume an infinite domain so that the true spectrum can be calculated exactly, provided the analytic function satisfies certain conditions.

$$F(\omega) = \int_{-\infty}^{+\infty} f(t) e^{-i\omega t} dt \tag{7.42}$$

In real cases, however, where we cannot observe $f(t)$ on the interval $-\infty < t < \infty$, we are looking at the function through a "window." The window can be represented by a function $w(t)$. The window function for the ideal analytic case where we know the function for all time is $w(t) = 1$ on the interval $-\infty < t < \infty$, but in the more realistic case where we know the function only on some finite interval, say $-T/2 < t < T/2$, then $w(t) = 1$ on that interval and $w(t) = 0$ everywhere else. In the case of a finite window, we do not see the true time series $f(t)$, but that time series multiplied by a window function $f(t)w(t)$. In order to understand the effect on the Fourier transform of multiplying the time series by a window function, we can use a form of the convolution theorem that states that the Fourier transform of the product of two functions is the convolution of their individual Fourier transforms 7.43.

$$\int_{-\infty}^{+\infty} f(t) w(t) e^{-i\omega t} dt = \frac{1}{2\pi} \int_{-\infty}^{+\infty} F(\overline{\omega}) W(\omega - \overline{\omega}) d\overline{\omega} \tag{7.43}$$

It is useful then to know the Fourier transform of the window function to see how it modifies the true Fourier transform, which is related closely to the power spectrum by 7.34. Consider the following Boxcar function window.

$$w(t) = \begin{cases} 1/T & \text{for} \quad T/2 < t < T/2 \\ 0 & \text{otherwise} \end{cases} \tag{7.44}$$

The Fourier transform of the boxcar function 7.44 is easily obtained.

$$\begin{aligned} W(\omega) &= \int_{-\infty}^{+\infty} w(t) e^{-i\omega t} dt = \frac{1}{T} \int_{-T/2}^{T/2} e^{-i\omega t} dt \\ &= \frac{1}{i\omega T} \left[e^{i\omega T/2} - e^{-i\omega T/2} \right] = \frac{\sin\left(\frac{\omega T}{2}\right)}{\frac{\omega T}{2}} = \text{sinc}\left(\frac{\omega T}{2}\right) \end{aligned} \tag{7.45}$$

The Fourier transform of the boxcar function is a sinc function, $\mathrm{sinc}(x) = \sin(x)/x$, which has the unfortunate characteristics of relatively large negative side lobes that decay slowly, so that it spreads variance around in frequency space in a spurious way. Figure 7.5 shows this behavior. The first zero crossing of the response function $W(\omega)$ occurs at $\omega = 2\pi/T$, which for a discrete Fourier analysis is the lowest frequency resolved and also the separation between resolved frequencies (*i.e.* the bandwidth). The first negative side lobe always has an amplitude of about -0.22, so a spurious signal of about 22% appears a bandwidth or so away from the actual frequency.

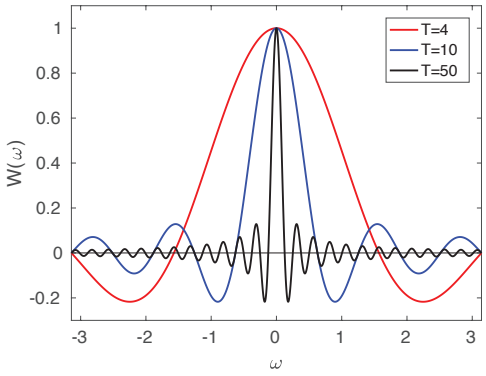

Figure 7.5 Fourier transform of the rectangular window function for various values of the length of record T in time units. The plot is terminated at $\omega = \pm\pi$, which would be the Nyquist interval in radian frequency if the time step was one time unit.

A finite window acts as a smoothing on the true spectrum. In addition to the smoothing effect, the side lobes of the frequency window lead to spectral leakage to other frequencies. The degree of smoothing and the range of the spread in frequency depend on the length of the data window T. The shorter that T is, the stronger the smoothing/spread will be. Therefore we can see that we must carefully design the window through which we view the data prior to spectral analysis, if we want to obtain the best results. The response function for the rectangular window is shown below. Note that while the response does peak strongly at the central frequency, significant negative side lobes are present. This means that our spectral analysis will introduce spurious oscillations at higher and lower frequencies that are out of phase with the actual oscillation.

To improve the quality of the resulting spectral analysis, it is desirable to modify the data window from the naive rectangular window. The ideal window response function, $W(\omega)$, would have a narrow central lobe and insignificant side lobes. We can improve on the naive rectangular window function and the rather unsatisfactory sinc function window response through modifications of the window function $w(t)$. Since much of the poor behavior of the rectangular window comes from the Gibbs effect of its sharp edges, rounding the edges of the window is an intuitive approach. Many data windows have been proposed for different purposes, most of which taper the window to near zero at the ends and maximize in the middle of the data window Harris (1978). In addition to the rectangular window already discussed, we will introduce just two that are suitable for general use; the Hann or Hanning window, and the Hamming window.

7.9.1 The Hanning Window

The Hanning window is named after Julius von Hann and is also called the elevated cosine or cosine bell window.

$$w(t) = \frac{1}{2}\left(1 - \cos\left(\frac{2\pi t}{T}\right)\right) \quad T/2 < t < T/2 \tag{7.46}$$

$$W(\omega) = \mathrm{sinc}\left(\frac{\omega T}{2}\right) + \frac{1}{2}\left(\mathrm{sinc}\left(\frac{\omega T}{2} + \pi\right) + \mathrm{sinc}\left(\frac{\omega T}{2} - \pi\right)\right) \tag{7.47}$$

We can see that the first part of the response function for the cosine bell window is a sinc function exactly like that of the rectangular window. In addition, however, we have two additional sinc functions of reduced amplitude that maximize in the center of the negative lobes on either side of the central lobe. The effect of these is to nearly cancel the negative side lobes of the rectangular response function while significantly broadening the central lobe (Fig. 7.6). We like the fact that the side lobes are now smaller, but the broadened central lobe means that the spectrum will be slightly smoothed compared to a rectangular window. This smoothing is not a disaster in most applications, since we often end up doing some smoothing anyway, and the smoothing effect of the cosine bell window allows us to claim a small increase in the number of degrees of freedom per spectral estimate. If greater frequency resolution is required, then a longer chunk of data must be used (*i.e.* increase T)

7.9.2 The Hamming Window

If a slightly better cancellation of the side lobes is desired, then the Hamming window (Richard W. Hamming) can be used. In addition to slightly better cancelling of the side lobes, a narrower central lobe is achieved compared to the Hanning window. Figure 7.6 shows the Fourier transforms of the Rectangular, Hanning and Hamming windows for a window length of $T = 100$. The side lobes are greatly reduced by tapering the window. The differences between the Hanning and Hamming windows are modest.

$$w(t) = \frac{1}{2} - 0.426\cos\left(\frac{2\pi t}{T}\right) \quad T/2 < t < T/2 \tag{7.48}$$

$$W(\omega) = \mathrm{sinc}\left(\frac{\omega T}{2}\right) + 0.426\left(\mathrm{sinc}\left(\frac{\omega T}{2} + \pi\right) + \mathrm{sinc}\left(\frac{\omega T}{2} - \pi\right)\right) \tag{7.49}$$

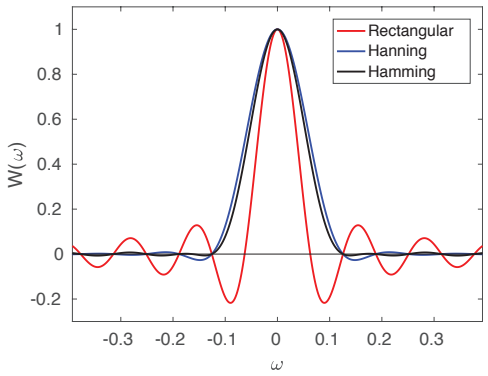

Figure 7.6 Response functions $W(\omega)$ for the rectangular, Hanning and Hamming windows for a window length of $T = 100$.

7.9.3 Welches Overlapping Segment Analysis: WOSA

The most common methods of spectral analysis used employ the Fast Fourier Transform method. In this method a direct Fourier Transform is made of the data using an efficient algorithm that makes use of the fact that the length of the time series has been chosen to be an integer power of two $M = 2^n$. Mixed-

radix FFT's are also available for which $M = 2^n 3^m 5^j$. In applying these methods the total time series of length $N\Delta t$ can be broken up into a series of smaller chunks of length M. Since a tapered window like the Hanning window will normally be applied, it is better to overlap the segments so that the data near the break points are not "wasted" by receiving a small weight. Overlapping the data by 50% will ensure that all the data are counted equally in the average spectrum that will be accumulated by averaging the results from each individual segment. This averaged spectrum will have approximately 2N/M degrees of freedom, since each power spectrum will have only M/2 estimates. The number of degrees of freedom will actually be slightly larger than this, depending on how much smoothing the data window provides, making the number of independent spectral estimates for each realization of the spectrum smaller than M/2. The spectra and cross-spectra for these smaller chunks can be averaged into a grand spectrum that has some degree of statistical reliability if N >> M. This is called Welch's Overlapping Segment Analysis, or WOSA.

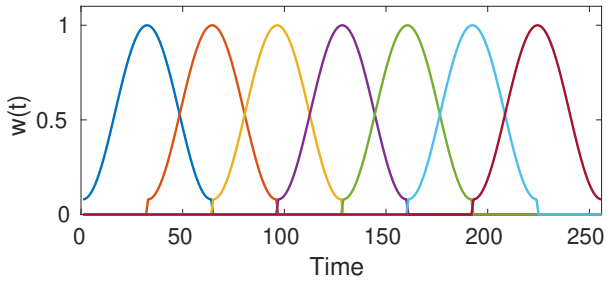

Figure 7.7 Illustration of WOSA analysis in which Hamming windows of length M = 64 are overlapped by 50% on a data set of length N = 256.

Figure 7.7 illustrates how a set of Hamming windows of length M = 64 overlapped by 50% can be used to analyze a data set of length N = 256 . All the data are equally weighted in the composite spectrum that results, except for the data on either end of the data set that are under weighted. Since the resulting spectrum will contain only M/2 spectral estimates, and N data are used, the number of degrees of freedom per spectral estimate is approximately 2N/M, times some factor to take into account that the spectrum is smoothed by the window. In the case of the Hamming window this factor is about 1.2.

Figure 7.8 shows example results for computing a spectrum using WOSA analysis and three different windows, the rectangular window, the Hanning taper and the Hamming taper. The input time series is a cosine wave with a period of 4.2 time units. The total record is of length N = 2048, a chunk length of M = 64 was used with an overlap of 50% = 32 time units. The power scale is logarithmic to show the sensitivity where the estimated power is very small. As expected from 7.6 one can see that the tapered windows greatly reduce the amplitude away from the line center at $\omega = 2|\mathrm{pi}/2.4$, especiallly for the first side lobe, but they also widen, smooth out, the central peak to be twice qs wide as for the rectangular window. The Hamming window does a more effective job than Hanning of removing the first side lobes, which are the biggest, but do allow more variance to pass far away from the line center. These amplitudes are weaker by a factor of 10^{-4} from the peak power and would not be a concern in typical geophysical settings with lots of noise.

7.10 Designing a Power Spectral Analysis

When considering spectral analysis of a time or space series, one typically has a hypothesis that a peak in the spectra my occur in a particular frequency range, which would indicate a larger than expected amount of variance with the corresponding period. It is essential to establish an *a priori* argument for where that spectral peak should be. Once this is known the nature of the required data set and an effective approach to spectral analysis can be designed.

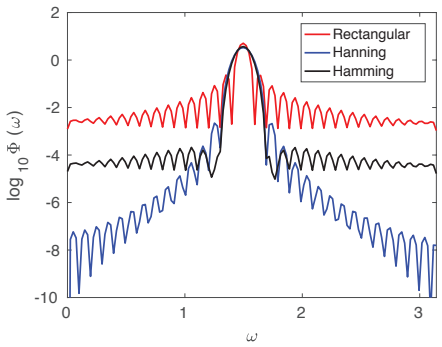

Figure 7.8 Power spectrum of a cosine wave with a period of 4.2 time units estimated with WOSA analysis in which Hamming windows of length $M = 64$ are overlapped by 50% on a data set of length $N = 2048$. The results with a rectangular window, a Hanning taper and a Hamming taper are shown.

7.10.1 Bandwith and Chunk Length

The first thing that to consider is how much resolution of frequency is required to isolate the peak of interest. Is it a very sharp peak, at a particular frequency, or is it expected to be a broad peak? If another peak is expected to be nearby, how much bandwidth do you need to separate the peak you seek from others you know to be around. The frequency separation between a period of P and a period of P' is $\Delta f = 1/P' - 1/P$. Supposing that you want at least a couple frequencies in between to show the separation between these peaks, and you are using a finite window that will smooth the spectrum a bit, you probably want a bandwith of at most one fourth of the frequency separation between the periods of interest. The bandwith of a spectral analysis is $\Delta f = 1/M\delta t$, so you want a chunk length of at least four times the longer of the two periods P and P'.

7.10.2 Time Step

Suppose we know that we expect a peak in the variance at a period of P time units, or a frequency of $f = 1/P$ cycles per unit time. A time step of $\Delta t = P/4$ will put that spectral peak right in the center of the Nyquist interval $0 < f < 1/2\delta t = 2/P$. So half the resolved frequencies will be higher than the frequency of interest and half will be lower. There is little point in having a smaller time step than this, since it just adds more high frequencies and does not help at all with the frequency of interest. Being in the middle of the Nyquist interval is more than enough to reduce any problems with aliasing from frequencies higher than those resolvable by the time step chosen.

7.10.3 Robustness and Degrees of Freedom

In designing a spectral analysis procedure, one must take into account the tradeoff between spectral resolution and degrees of freedom, if the length of the available time series is limited. We stated before that the number of degrees of freedom in a spectral estimate is approximately the number of data points divided by the number of independent spectral estimates. The number of degrees of freedom required depends on the relative strength of the peak we are looking for. We will come back to this in the section on testing the statistical significance of spectral peaks, but for now let's assume a general rule that we don't take a spectrum seriously unless we have about 20 degrees of freedom, so this means we need a data set that is 10 times the length of the chunk, $N = 10xM$. With that let's consider an example to fix ideas.

7.10.4 Example of 5-Day Wave

Suppose we have an *a priori* expectation that a peak in variance will occur at 5 days. What is the time step Δt and chunk length M that will resolve this well, and how much data do we need to get a robust result? Our frequency of interest is $f = 0.2$ cycles per day (cpd), and a time step of one day will give a Nyquist frequency of $f_{Nyquist} = 0.5$ cpd. Since it should be easy to get daily data, let's choose that as our sampling interval.

Next we need to decide what bandwith to use. Suppose for the sake of argument that we want to be able to distinguish a 5-day wave from a six day wave. So, $\Delta f = 1/5 - 1/6 = 1/30$. If we want three frequencies in between, then we need a bandwidth of about $\Delta f == 1/30x4 = 1/120$. Since we'll be using an FFT and WOSA analysis, let's choose a chunk length of $M = 128$, which is the nearest power of 2 greater than 120. To get about 20 degrees of freedom, we'll need 1280 days of data, or about 3.5 years.

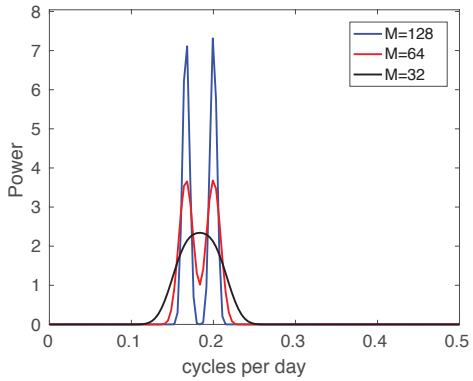

Figure 7.9 Power spectrum of a data set consisting of a 5-day and 6-day harmonic, using a Hamming taper with chunk length of M =128, 64 and 32. In this plot the total variance is the area under the curve, which is the same in each case.

Figure 7.9 shows the spectra computed from an input time series consisting of cosine waves of 5 and 6 days using our design chunk length of M =128, as well as shorter chunks of 64 and 32. As we expected, the 128-day chunk length allows the 5 and 6-day waves to be separated, with zero variance showing at 2 intervening frequencies. The separation is still observable with a chunk length of 64, but at a chunk length of 32 the two peaks merge into one broad peak.

7.11 Statistical Significance of Spectral Peaks

The statistical significance of a peak in a power spectrum is assessed as in any case by stating the significance level desired, and then stating the null hypothesis and its alternative. The null hypothesis is usually that the time series is not periodic in the region of interest, but simply noise. Since most geophysical time series contain a good measure of noise, we can usefully compare the amplitude of a spectral peak to a background value determined by a red noise fit to the spectrum. We illustrate here one simple method of determining if a spectral peak is statistically significant.

One way to test significance of a spectral peak is to compute the ratio of the observed power, Φ to the power expected from your null hypothesis Φ_0 and compare this value with a "Chi-Squared" test with the corresponding number of degrees of freedom,

$$\chi^2 = (n-1)\frac{s^2}{\sigma^2} \quad \nu = (n-1) \tag{7.50}$$

where the Number of degrees of freedom ν is one less than the number of independent samples, n. The number of independent samples is estimated from,

$$n = \frac{2N}{M} f_w \tag{7.51}$$

where,

$$\begin{aligned} N &= \text{total sample size} \\ M &= \text{chunk length of FFT} \\ f_w &= \text{a factor to account for the smoothing by the window} \end{aligned} \tag{7.52}$$

The factor of two arises because the spectrum has only $M/2$ estimates, since the phase information is not included in the power spectrum. The factor f_w is generally between 1 and 1.5, depending on how much smoothing of the spectrum that the particular window does. For the Hanning window, $f_w = 1.2$, To be on the conservative side, we can assume $f_w = 1.0$, but in marginal cases this factor can be used to get a more accurate significance estimate.

A more convenient way to apply this test is to use the F-test, based on the F distribution.
Theorem: If σ_1^2 and σ_2^2 are the variances of independent random samples of size n_1, and n_2, respectively, taken from two normal populations having the same variance, then

$$F_{\nu_1}^{\nu_2} = \frac{\sigma_2^2}{\sigma_1^2} = \frac{\Phi}{\Phi_0} \tag{7.53}$$

is a value of a random variable having the F distribution with the parameters $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$, giving the number of degrees of freedom for the upper and lower variance, respectively.

The F distribution can be used to determine whether two sample variances are different in a statistical sense at a chosen probability level. Tables of the F distribution are included in **??**. You should not confuse this F with $F(z)$, the cumulative distribution of the Normal distribution. In assigning a confidence level and interpreting statistical tests of significance one must be concerned very much with the distinction between *a priori* and *a posteriori* statistics.

a priori: If we have stated in advance that we expect a peak at a particular frequency (and given a good reason beforehand), then we can simply test the significance of the spectral peak above the background using the normal confidence limits set forth for the chi-squared or F statistics.

a posteriori: If we have not stated at which frequency we expect the peak, then we must determine the probability that one frequency out of the $M/2$ we have computed should show a significant peak. The usual way to do this would be to take the probability of a type II error (accepting a false hypothesis as true) and multiply this by the number of chances we have given the spectrum to exceed the required level.

7.11.1 Example: a priori versus a posteriori Spectral Peaks

Suppose we have a spectral peak which exceeds the background significantly at the 95% probability level.

1. If we had predicted this frequency before computing the spectrum, then we can use the 95% probability level and infer that only a 5% possibility exists that this spectral peak could have occurred by chance. We have an *a priori* reason for expecting this peak and we can therefore use *a priori* statistics.

2. If we had not predicted the frequency of the peak, then we must test the probability that one frequency out of our sample of $M/2f_w$ independent estimates should show a significant peak. $M/2$ is the number of frequencies retained in our spectrum and f_w is the factor indicating the degree of smoothing by the window. If the number of independent frequencies in our spectrum is 50, then our chance of getting a spectrum with no significant peaks is $(0.95)^5 0 = 0.08$ and it is likely that at least one peak will exceed the 95% confidence level by chance. If we had started with a 99% confidence limit the significance would be $(0.99)50 = 0.6$, and there is still a 40% chance that one peak will exceed the 99% confidence level. Therefore, in practice we need to have an *a priori* reason for expecting a peak at the frequency where one occurs, or our confidence

level is very low. This *a priori* reason could be a theory, or the observation of a peak at the frequency of interests in an independent data set. The probability of getting spectral peaks with 95% significance at the same frequency by chance in two independent data sets is small.

7.11.2 Example: Statistical Design

Example: Suppose we wish to examine climatic fluctuations in the 10-1000 year range. Assume that a physically meaningful peak should have twice the variance of the background spectrum. What should be the spacing of the observations and how long a time series is required to get meaningful statistics?

A time step of two years would resolve the shortest period of interest very well, since the Nyquist period would be 4 years and a period of 8 years would be in the middle of the Nyquist interval ($0 < f < 1/(2\Delta t)$. The next question would be how much bandwidth is required, which determines the chunk length, M. Suppose for the sake of argument that we'd like to be able to distinguish a 500 year period from a 1000 year period. If we choose a chunk length of 4096 years, then the fourth frequency resoloved would be 1/1024 cycles per year, and the eighth frequency would be 1/512 cycles per year. This is fine because it leaves three frequencies lower than the lowest frequency of interest, and it leaves three frequencies between the two frequencies of interest 1/1000 and 1/500 cycles per year. Since the time step is two years, we need a chunk length of $M = 2048$. To see how many degrees of freedom we need, we first need to decide what our significance level is. Let's say we want 99% significance if our spectral peak exceeds the background by a factor of two. We then ask how many degrees of freedom are required so that an F-statistic of 2.0 is 99% significant. We need to know the number of degrees of freedom for our null hypothesis spectrum. This is usually a two-parameter fit, so let us assume we have at least 100 degrees of freedom for it. Then from the F-statistic table, we determine that the sample spectrum must have at least 23 degrees of freedom. This means we need about 12 independent realizations of our spectrum or about 12x2x2048 = 49,152 years of data. So if we find a spectral peak with a variance ratio of 2 using a chunk length of 4096 years, then this peak will be significant at 99%, if we've predicted the frequency of the peak *a priori*. Two-year averages or data taken every other year are fine for this analysis. Using yearly data would just double the Nyquist frequency and add many high frequencies of no interest to our spectrum.

7.11.3 The Red Noise Null Hypothesis

A useful null hypothesis for many geophysical time series is that the time series consists of autocorrelated Gaussian noise. The degree of autocorrelation can be a very important physical characteristic and the reasons why geophysical time series are autocorrelated are interesting, but here we focus on the exceptions where the time series contains some periodic phenomena immersed in autocorrelated noise. The theoretical spectrum for autocorrelated "red" noise was presented in 7.2.3 , but here we consider how the red noise spectrum is modified by being viewed at discrete times separated by Δt. We begin with the equation for an autocorrelated random walk 7.3.

$$x(t) = \alpha x(t - \Delta t) + \left(1 - \alpha^2\right)^{1/2} \epsilon(t) \tag{7.54}$$

where $\alpha = r(\Delta t)$ Using the "time shifting theorem" 7.35, we can write the Fourier transform of 7.54 as,

$$\begin{aligned} X(\omega) &= \alpha X(\omega) e^{-i\omega\Delta t} + \left(1 - \alpha^2\right)^{1/2} E(\omega) \\ &= \frac{C E(\omega)}{1 - \alpha e^{-i\omega\Delta t}} \end{aligned} \tag{7.55}$$

where $X(\omega)$ is the Fourier transform of $x(t)$, $E(\omega)$ is the Fourier transform of Gaussian white noise, and $C = \left(1 - \alpha^2\right)^{1/2}$. We know from Parseval's Theorem that the power spectrum of $x(t)$ is

$$\Phi_{xx}(\omega) = X(\omega)X^*(\omega) = \frac{C^2 E(\omega)E^*(\omega)}{\left(1-\alpha e^{-i\omega\Delta t}\right)\left(1-\alpha e^{i\omega\Delta t}\right)} = \frac{1-\alpha^2}{1-2\alpha\cos(\omega\Delta t)+\alpha^2} \quad \text{for} \quad 0<\omega<\frac{\pi}{\Delta t} \tag{7.56}$$

where we have used the identity $\cos(x) = \frac{\left(e^{ix}+e^{-ix}\right)}{2}$. The formula 7.56 for the power spectrum of red noise was discussed by Gilman et al. (1963).

To fit the shape function 7.56 to a real spectrum we need the one-lag autocorrelation, α from the input time series. A slightly more robust estimate of the parameter α from the original time series is the average of the one-lag autocorrelation and the square root of the two-lag autocorrelation. It is presumed here that the time step is chosen appropriately for the variability present in the time series. If the time step is too small or too large for the characteristic variability in the time series, then the results will be poor. We then multiply the shape 7.56 for that value of α by a factor that will make the variance equal to that of the time series in question. One simple way to do this is to match the total variance. The total variance is the sum of the power over all non-zero frequencies. So sum the power in the observed spectrum and the power in the idealized red noise spectrum. Then multiply the idealized red noise spectrum by this ratio, so that the red noise spectrum has the same total variance as the observed spectrum. The null-hypothesis spectrum is thus a two-parameter fit that has the same total variance and same one-lag autocorrelation as the observed time series.

7.11.4 Continuous and Discrete Red Noise Spectra

It may be helpful to compare the discrete red noise spectrum 7.56 with the continuous red noise spectrum in section 7.8.4 . For the continuous spectrum 7.4 the integration is performed for $0<\omega<\infty$, whereas for the discrete spectrum 7.56 the integration is performed for $0<\omega<\pi/\Delta t$. The discrete spectrum has an integral of π, so that if $\alpha = 0$ the spectrum has a uniform value of 1. For the continuous spectrum, $\alpha = 0$ gives a vanishingly small value that stretches to infinity, although the total variance is conserved as α increases and the variance peaks near $\omega = 0$. To compare the shapes of the two formulas in the Nyquist interval, we can integrate the continuous spectrum over the Nyquist interval to establish its norm, and use that to rescale the continuous spectrum so that it has the same variance as the discrete spectrum on the Nyquist interval.

$$\int_0^{\pi/\Delta t} \frac{2T}{(1+\omega^2 T^2)}\,d\omega = -2\tan^{-1}\left(\frac{\pi}{\ln\alpha}\right) \tag{7.57}$$

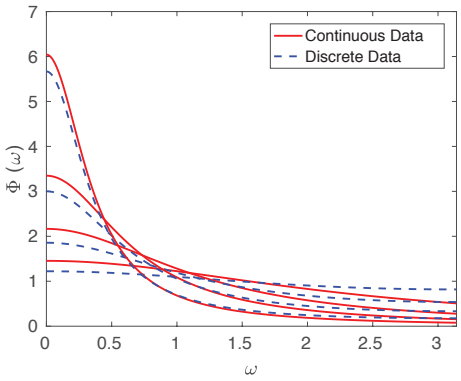

Figure 7.10 Power spectra of red noise computed using the continuous 7.4 and discrete 7.56 formulations. Curves are shown for $\alpha = 0.1$, 0.3 0.5 and 0.7. The continuous spectrum is normalized for the Nyquist interval using 7.57.

Comparison of the continuous and discrete approximations of the power spectrum of red noise are shown in Figure 7.10. It can be seen that the discrete approximation smooths the spectrum a little, with lower values at low frequencies nd higher values at high frequencies. The discrete approximation will be used to test experimental spectra computed with discrete data.

7.11.5 Example: Sampling Red Noise

In this section we consider samples of red noise and use the red noise null hypothesis and F-statistic to test for the significance of the peaks that appear. This example shows that one sample of red noise can show a spectrum that appears to have peaks, but proper use of statistics shows these peaks to be insignificant. As an example, we generate a time series of daily observations with an autocorrelation of $\alpha = 0.5$ at a one-day lag. We choose a chunk length of 256 days, which yields a spectrum with 128 estimates. To assess statistical significance we use a red noise spectrum with the shape of 7.56 and the same variance as the observed spectrum. We assess the number of degrees of freedom in the spectrum and then multiply the null-hypothesis spectrum by the appropriate F-statistic.

$$\Phi_{99} = \Phi_{\mathrm{Null}} \times \mathrm{F}_{\nu_0}^{\nu 0.01} \tag{7.58}$$

Here the F-statistic has three parameters, the number of degrees of freedom in the null hypothesis ν_0, the number of degrees of freedom in the sample ν, and the significance level, $p = 0.01$ for 99% confidence. To increase the degrees of freedom and improve the robustness of the spectrum we can average multiple realizations of the spectrum by using independent chunks of data.

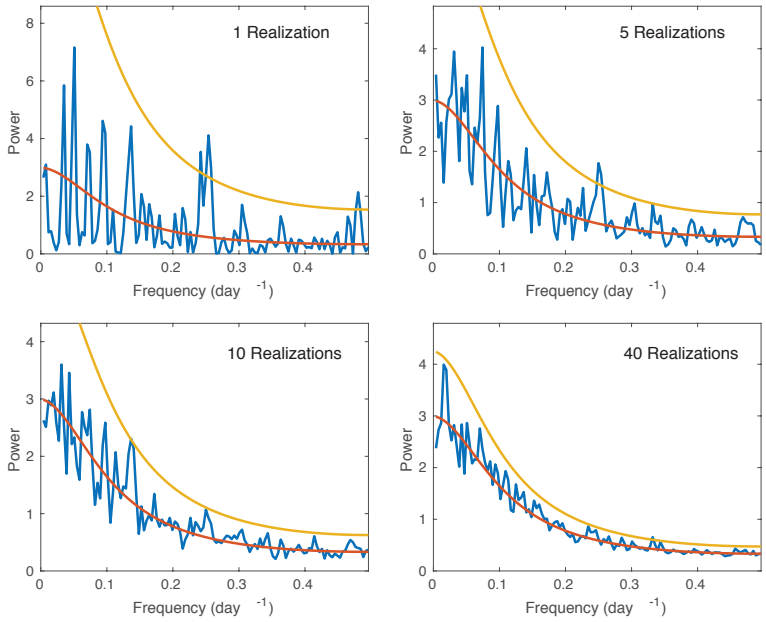

Figure 7.11 Power spectra of red noise with an autocorrelation of $\alpha = 0.5$ at one day, using a Hamming window and a chunk length of M=256 days. Examples are show for averages over 1, 5, 10 and 40 realizations. The red line is best fit using 7.56. The orange line is the 99% confidence limit, calculated as explained in the text.

Figure 7.11 illustrates the impact of averaging multiple realizations of a spectrum. A single realization shows a peak around f =0.25 cycles per day that exceeds the 99% confidence level indicated by the orange line. Since the spectrum has 128 chances to exceed the 99% confidence level, it is not surprising that one or two frequencies exceed the 99% threshold. Since we did not predict a peak at f =0.25 and we know that the observed spectrum has only about 2 degrees of freedom, we would not take this peak seriously. It is highly

likely that a sample of red noise will produce a spikey spectrum by chance. Since it got off to such a good start, by chance, with the first realization, the spectral peak around $f = 0.25$ still passes the 99% threshold of significance when 4 more realizations are added. It is still not significant, since we have no *a priori* reason to expect it in the spectrum with 5 realizations. When we add 5 more realizations for 10 total, the $f = 0.25$ peak finally is below the significance level, but a new peak at $f = 0.13$ appears to touch the 99% threshold, and again we have not predicted that peak in advance, so we would not take it seriously without further analysis. If we include a total of 40 realizations the two peaks previously considered fall below the line, and another one jumps out around $f = 0.34$. Again, we expect 1 or 2 frequencies out of 128 to pass the 99% threshold by chance. An extremely large sample would be needed to get all of the kinks and wiggles out of the experimental spectrum. The lesson here is to insist on sufficient quality and don't become excited by spurious peaks that appear by chance.

7.12 Prewhitening

The previous section discussed one method for determining the significance of spectral peaks emersed in red noise. It is also possible to remove the red noise from a data set prior to spectral analysis, a process called "prewhitening" the time series. The approach is very simple, one just subtracts the red noise approximation to the time series. A prewhitened time series $x_w(t)$ is computed from $x(t)$ in the following way.

$$x_w(t) = x(t) - \alpha x(t - \Delta t) \tag{7.59}$$

Here $\alpha = r(\Delta t)$. To illustrate how prewhitening can work we consider red noise with an autocorrelation of $\alpha = 0.5$ for daily sampling, and we add pure periodicities with periods of 10 and 4 days. Figure 7.12 considers red noise as in Figure 7.11, but adds periodicities at 4 and 10 days. To illustrate the effect of chunk length, we include spectra computed with $M = 256$ and $M = 128$. To be fair, we give the shorter chunk length twice as many realizations, so it has the same amount of data to work with. In this case with only red noise and periodicities in the time series, prewhitening works beautifully and produces a white noise spectrum with periodicities. The significance assessment is the same, however, and both methods and both chunk lengths properly identify the periodicities as significant, assuming we had an *a priori* reason for expecting peaks at 4 and 10 days. Note that the noise used in each of the cases in Figure 7.12 was the same realization, but different from the noise realization used in Figure 7.11.

7.12.1 Rossby-Gravity Waves in Reanalysis

Spectral analysis is an effective tool for identifying phenomena with a well-defined period. In the days before global weather analyses, spectral analysis was used to identify waves in time series from a few stations where balloon observations of wind and temperature were available. A classic example is the definition of tropical waves from rawinsonde stations in the tropical Pacific. Examples include Kelvin waves (Wallace and Kousky, 1968), Mixed Rossby-Gravity waves (Yanai et al., 1968) and the Madden-Julian Oscillation (Madden and Julian, 1971). Nowadays reanalysis products reconstruct global maps of wind, temperature and other meteorological variables, even where few observations are present, by simulating the dynamics and thermodynamics of the atmosphere as part of the interpolation process. Here we will look for the Mixed Rossby-Gravity wave discovered theoretically by Matsuno (1966) and later observed by Yanai et al. (1968). We expect the MRG wave to have a periodicity around 4-5 days in the meridional (north-south) wind at the equator. We will look for it in the central equatorial Pacific between 190E and 210E. We use data every 5-degrees of longitude and average the spectra together. Data for calendar years 2000-2015 are used and the annual cycle is removed.

Figure 7.13 shows power spectra of the meridional wind at 850hPa at the equator using three different methodologies, plus a contour plot of the fraction of variance in the 3 to 6-day period band. Our expectation is that a spectral peak will appear near 4 to 5 days associated with the theoretical prediction of the Mixed

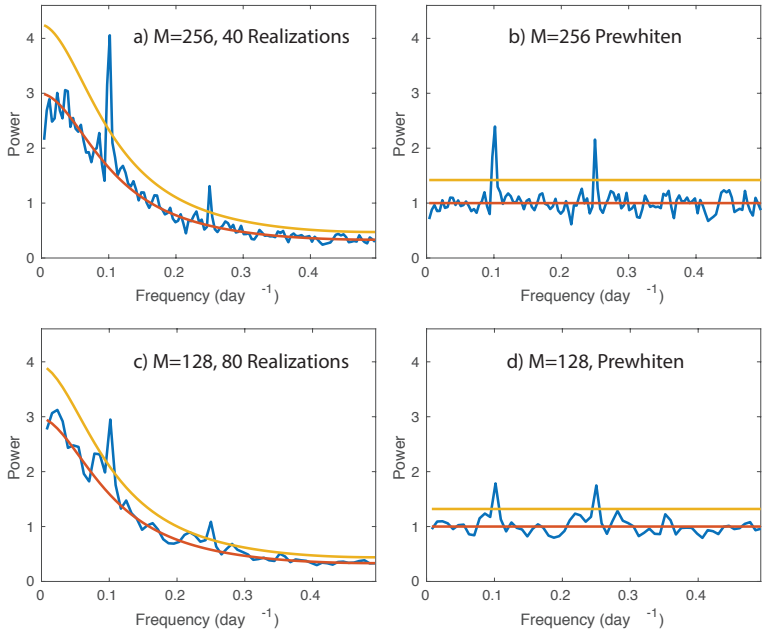

Figure 7.12 Power spectra of red noise with an autocorrelation of $\alpha = 0.5$ at one day plus harmonics with periods of 4 and 10 days. Spectra are computed using a Hamming window and chunk lengths of M=256 and M=128 days. Spectra with and without prewhitening are shown. To be fair, the same number of data are presented to each analysis, so the shorter chunk length of 128 days is given 80 realizations compared to 40 for the longer chunk length. The orange line is the 99% confidence limit, calculated as explained in the text.

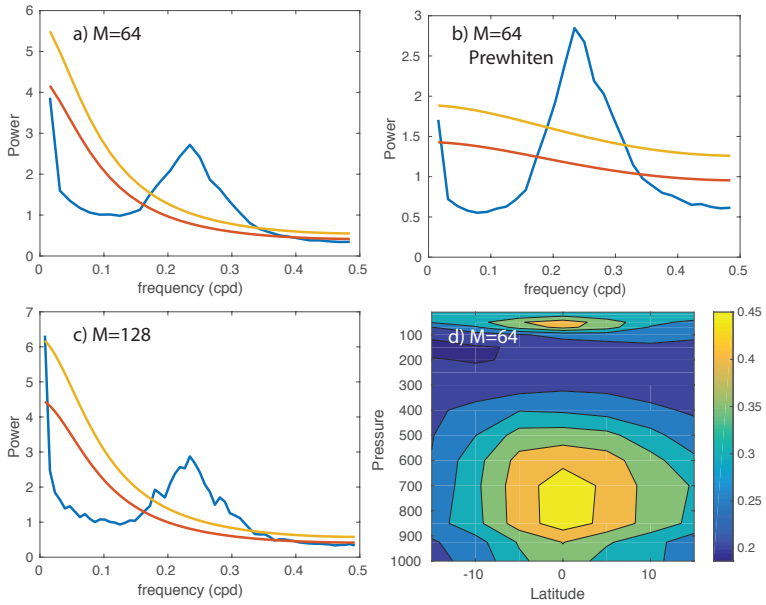

Figure 7.13 a) Power spectrum of the meridional component of wind at the equator and 850hPa averaged over the longitudes of 190-210E using a chunklength of a) $M = 64$, b) M=64 with prewhitening, c) $M = 128$. d) contour plot of the fraction of the total power that is contained in the 3-6 day period band as a function of latitude and pressure, computed with $M = 64$ without prewhitening. Red and orange curves indicate the red noise null hypothesis and 99% confidence level.

Rossby-Gravity wave. To distinguish 4 from 5 days we need a bandwith of $\Delta f = \left(\frac{1}{4} - \frac{1}{5}\right)/3 = \frac{1}{60}$, which will be nicely accomplished with a chunk length of $M = 64$. For comparison, we also show a chunk length of $M = 128$, which shows some additional detail that does not appear to be significant. Because the timeseries has such a strong peak in variance near 4.5 days, a red noise fit is not a good fit to the spectrum. If we remove our best fit to red noise from the time series by prewhitening, then we get a slightly more pleasing appearance (Fig. 7.13b), but no real difference in interpretation. Some autocorrelation remains after the red noise is removed because the periodicity itself will contribute autocorrelation . Plotting the ratio of the variance in the 3 to 6-day period range to the total variance (Fig. 7.13d) shows that the periodicity is confined near the equator and to the lower troposphere, although it does appear again in the lower stratosphere.

7.13 Multi-Taper Method of Spectral Analysis

In the Welch's Overlapped Segment Averaging (WOSA) method described in section 7.9.3, a single window function is applied to different segments of the time series and then averaged to produce a final spectral estimate with robustness and good spectral response properties. The choice of chunk length and window shape is based on the spectral resolution and degrees of freedom desired. Different chunk lengths and window shapes give optimal performance for different ranges of frequency. As the name implies, the multi-taper method (Thomson, 1982) uses a set of different data tapers to try to provide an optimal estimate of the spectrum of the time series. A nice description is given in Percival and Walden (1993). The idea is to choose a set of orthogonal tapers that provide optimal resolution and minimum leakage. A set of such tapers is devised that are harmonics on the data interval, and then the average of spectra computed using a number of these tapers is formed. Such a technique can give better results in some cases.

7.14 Maximum Entropy Spectral Analysis

Maximum entropy spectral analysis is a technique that can be used when you have a short period of record, but you need more spectral resolution that you can get by doing traditional Fourier Spectral analysis on the available data (Marple, 1987). It provides this extra spectral resolution by extending the autocovariance matrix in a way that adds the least information to the covariance matrix (maximum entropy). It will tend to strongly localize spectral peaks, so you can determine their location very precisely. The problem is that it has adjustable parameters that can be used to get a spectrum of arbitrary sharpness, and it may split peaks if the order of approximation is too high. The tools for assigning statistical significance to the results of such an analysis are uncertain. It is best used in conjunction with traditional Fourier spectral analysis, after you have established the significance of periodicities in the time series of interest.

7.15 Cross Spectrum Analysis

Cross spectral analysis allows one to determine the relationship between two time series as a function of frequency. Normally, cross-spectral analysis makes sense when statistically significant peaks at the same frequency have been shown in two time series. In that case we wish to know if these periodicities are related with each other and, if so, what the phase relationship is between them. One may extend this concept a bit by considering whether it may make sense to do cross-spectral analysis even in the absence of peaks in the power spectrum. Suppose we have two time series whose power spectra both are indistinguishable from red noise? Under these circumstances what might cross-spectral analysis still be able to reveal? It might be that within this red noise spectrum there are in fact coherent modes at particular frequencies. We can test for this by looking at the coherency spectrum.

7.15.1 Complex Fourier Transform of Cross-Spectrum

The formulation for cross-spectral is based on Fourier Analysis. It is most compact to express this in complex form, where variables have a real and an imaginary component. Any time series on the interval $0 < t < T$ can be expressed as a complex Fourier transform.

$$x(t) = \sum_{k=-N/2}^{N/2} F_x^k e^{i\frac{2\pi k t}{T}} \qquad y(t) = \sum_{k=-N/2}^{N/2} F_y^k e^{i\frac{2\pi k t}{T}} \tag{7.60}$$

The cross spectrum of x and y is performed by taking the product of the Fourier coefficients.

$$C_{xy}^k = F_x^k F_y^{k*} \tag{7.61}$$

where the asterisk indicates a complex conjugate. The complex Fourier coefficients can be written in polar form.

$$F_x^k = F_x^{kRe} + iF_x^{kIm} = R_x^k e^{i\theta_{xk}} \tag{7.62}$$

where

$$R_x^k = ((F_x^{kRe})^2 + (F_x^{kIm})^2)^{1/2} \tag{7.63}$$

and R_x^k is real.

With these definitions, and integrating over time, we find that the cross spectrum of x and y is,

$$C_{xy}^k = R_x^k R_y^k e^{i(\theta_x^k - \theta_y^k)} = R_x^k R_y^k e^{i\Delta\theta^k} = R_x^k R_y^k (\cos\Delta\Theta^k + i\sin\Delta\Theta^k) = Co^k + iQ^k \tag{7.64}$$

We thus see that the cross spectrum has a real part, the cospectrum, Co, and an imaginary part the quadrature spectrum, Q, whose ratio determines the phase difference between the two time series at each frequency indicated by the index k.

$$\Delta\theta^k = \arctan Q^k / Co^k \tag{7.65}$$

The cospectrum and quadrature spectrum can be averaged over many realizations and also over frequency to compute robust phase differences between time series. The coherence uses the averaged spectra and cross spectra to measure the degree to which the phase and amplitude differences remain constant across the realizations and frequencies that are averaged over. The coherence-squared is expressed in terms of the averaged Power spectra of the two time series, Φ_x^k and Φ_y^k and the averaged co-spectra and quadrature spectra squared.

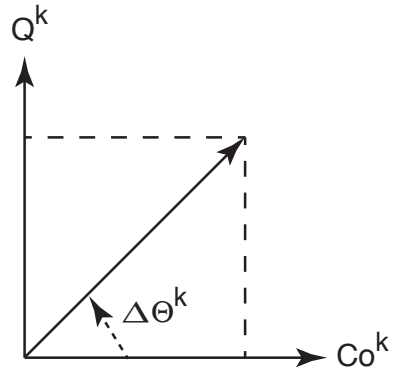

Figure 7.14 Diagram showing the relationship between the co-spectrum Co^k, quadrature spectrum Q^k and the phase angle between two time series $\Delta\Theta^k$ as a function of frequency index, k.

$$\mathrm{Coh}^2(k) = \frac{\mathrm{Co}(k)^2 + \mathrm{Q}(k)^2}{\Phi(k)\Phi(k)} \tag{7.66}$$

The coherence is analogous to the correlation coefficient, except it is a function of frequency. For a single realization of a single frequency the coherency is one. As different realizations are averaged together, the coherency will decline if the phase difference or amplitude ratio of the two time series varies from realization to realization. Coherence measures the consistency of the linear relationship between the two time series as the number of realizations is increased. Tables for the statistical significance of the coherence have been prepared that can test the null hypothesis that the coherence is zero (Amos and Koopmans, 1963) 10.1. The uncertainty in the phase difference can also be related to the coherence (Goodman, 1957). The uncertainty in the phase difference estimation increases as the coherence is reduced (Hartmann, 1974).

7.15.2 Example: Rossby-Gravity Wave Cross-Spectral Analysis

In Figure 7.13d it is shown that the ratio of meridional wind variance in the 3 to 6 day period band in the central equatorial Pacific peaks in the lower troposphere on the equator. This is the structure we expect for the Mixed Rossby-Gravity wave. In this section we perform cross-spectral analysis between the meridional wind at the point on the equator at 210E and 850hPa and the meridional wind at other latitudes, longitudes and pressures. Figure 7.15 shows coherence and phase as functions of both longitude and latitude and pressure and latitude. The spectra are averaged over many realizations and all frequencies corresponding to periods between 3.5 and 5 days. This analysis includes 91 realizations of its 64-day chunk length, and 7 frequencies are averaged together, so the cross-spectral analysis has about 600 degrees of freedom. The coherence has statistically significant and reasonably large values over most of the domain shown. The phase decreases toward the west and upward, indicating waves with westward and upward phase movement. The waves move westward with time and tilt eastward with height in the troposphere. In the stratosphere the waves tilt the opposite direction, consistent with upward propagation there (Holton and Hakim, 2012). They change phase by about 90-degrees in 20-degrees of longitude and so have an effective wavelength of about 8,000km, and move westward at about $10 \mathrm{m s}^{-1}$.

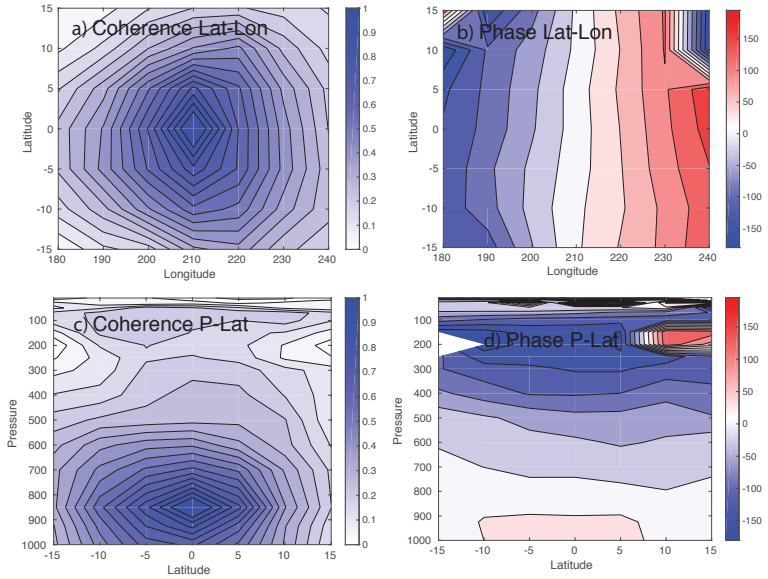

Figure 7.15 Cross-spectral analysis of the meridional wind at the equator, 850hPa and 210E, for periods in the range of 3.5 to 5 days. a) Coherence with the meridional wind at other latitudes and longitudes at 850hPa. b) Phase difference with other latitudes and longitudes at 850hPa. c) Coherence with the meridional wind at other pressures and latitudes. d) Phase difference with the meridional wind at other pressures and latitudes. A chunk length of $M = 64$ was used. Contour interval for coherence is 0.05 and for phase is 30 degrees.

7.16 Space-Time Spectrum Analysis

Mixed space-time spectral analysis is a straightforward extension of harmonic analysis to two dimensions. It is most convenient if the spatial dimension is cyclically continuous, such as in the case of latitude circles, or at least that the spatial dimension has fixed boundaries, like an ocean basin. In such cases we can look for modes of variability in which spatial scales have particular temporal scales. If the behavior is indeed harmonic (wavelike), then we expect mixed space-time spectral analysis to isolate any such modes that are present. For example, if one did mixed space-time spectral analysis of a stringed instrument, one would definitely expect to find a definite relationship between the length scales and the time scales of the oscillations.

Suppose we have a function of longitude, λ , and time, t. We can write:

$$x(\lambda, t) = \sum_{k} \sum_{\pm\omega} X_{k,\pm\omega} \cos(k\lambda \pm \omega t + \Theta_{k,\pm\omega}) \tag{7.67}$$

where $+\omega$ and $-\omega$ correspond to eastward- and westward-moving waves, respectively (Hayashi, 1971, 1979). The zonal wavenumber, k, is the number of zero crossings of the cosine wave around a latitude circle.

If we have such an expansion then we can write the power spectrum as function of both wavenumber and frequency as,

$$P_{k,\pm\omega}(x) = \frac{1}{2} X^2_{k,\pm\omega} \tag{7.68}$$

If we have two time series $x(\lambda, t)$ and $y(\lambda, t)$ we can write the cospectrum between x and y as,

$$Co_{k,\pm\omega}(x, y) = \frac{1}{2} X_{k,\pm\omega} Y_{k,\pm\omega} \cos\big(\Theta_{k,\pm\omega}(y) - \Theta_{k,\pm\omega}(x)\big) \tag{7.69}$$

and the quadrature spectrum is,

$$Q_{k,\pm\omega}(x, y) = \frac{1}{2} X_{k,\pm\omega} Y_{k,\pm\omega} \sin\big(\Theta_{k,\pm\omega}(y) - \Theta_{k,\pm\omega}(x)\big) \tag{7.70}$$

And so the coherence-squared is written,

$$Coh^2_{k,\pm\omega}(x, y) = \frac{Co^2_{k,\pm\omega}(x, y) + Q^2_{k,\pm\omega}(x, y)}{P_{k,\pm\omega}(x) \cdot P_{k,\pm\omega}(y)} \tag{7.71}$$

To obtain the expansion 7.67 we proceed by first performing a Fourier analysis in the longitude coordinate.

$$x(\lambda, t) = \sum_{k} C_k(t) \cos(k\lambda) + S_k(t) \sin(k\lambda) \tag{7.72}$$

We then perform a Fourier analysis in time of these cosine and sine coefficent time series.

$$\begin{aligned} C_k(t) &= \sum_{\omega} A_{k,\omega} \cos(\omega t) + B_{k,\omega} \sin(\omega t) \\ S_k(t) &= \sum_{\omega} a_{k,\omega} \cos(\omega t) + b_{k,\omega} \sin(\omega t) \end{aligned} \tag{7.73}$$

Hayashi (Hayashi, 1971) shows through a straightforward manipulation that A,B,a and b can be related to $X_{k,\pm\omega}$ as follows.

$$4X^2_{k,\pm\omega} = (A \mp b)^2 + (\pm B - a)^2 \tag{7.74}$$

where $X^2_{k,\pm\omega}$ is the space-time power spectrum we desire and the phase is given by,

$$\phi_{k,\pm\omega} = \tan^{-1}\Big(\frac{\mp B - a}{A \pm b}\Big) \tag{7.75}$$

7.16.1 Standing Waves

In space-time spectral analysis standing waves appear as equal contributions from eastward and westward propagating waves. For example, consider the following simple analytic wave.

$$\begin{aligned} x(\lambda, t) &= \cos(k\lambda)\cos(\omega t) \\ &= \frac{1}{2}\big(\cos(k\lambda - \omega t) + \cos(k\lambda + \omega t)\big) \end{aligned} \tag{7.76}$$

We thus see that a stationary wave whose amplitude oscillates in time is equivalent to equal amplitude eastward and westward propagating waves with the same wavenumber and frequency. If we do mixed space-time spectral analysis of such a wave it is hard to tell if it is two independent waves traveling eastward and westward with the same wavenumber and frequency, or a stationary wave whose amplitude oscillates in time. Figure 7.16 shows the space-time spectrum of an analytic zonal wavenumber 5 with a period of 5 days computed with a window of 32 daily observations and based on a record of 264 days. Other details of the procedure are discussed in section 7.16.2.

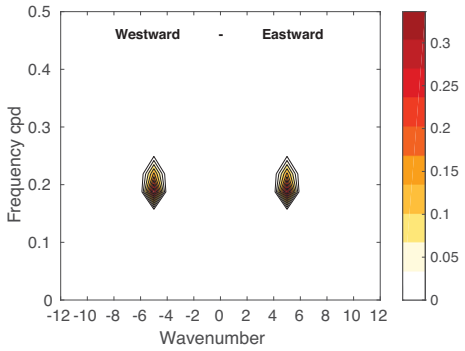

Figure 7.16 Space-time spectral analysis of a standing wave of wavenumber 5 and period 5 days. The time spectral analysis was done with a 32-day chunk and based on 264 days of record. Negative wavenumbers indicate westward traveling waves. Frequency is given in cycles per day.

Such a spectrum presents two possibilities; A standing wave or two independent waves traveling in opposite directions. How does one distinguish these two possibilities? One way to approach this problem is to ask if the eastward and westward waves are related, are they coherent with each other, do they bear a constant phase relationship to each other (standing wave), or are the eastward and westward waves linearly independent? Two somewhat different approaches to this question have been presented. One way to judge this is to formulate a coherence-squared between the eastward and westward waves (Pratt (1976); Hayashi (1977, 1979)). Another method is to look at the coherence in time between the sine and cosine coefficients of a particular wavenumber. Schäfer (1979) uses the coherence in time of the sine and cosine coefficients to ask whether what is seen are "waves" or "noise".

An alternative method of separating standing and traveling oscillations has been proposed by Watt-Meyer and Kushner (2015). This method assumes that the standing wave is the minimum amplitude of the eastward and westward components, and defines the traveling component as the difference between the actual value and the minimum value.

$$X^{Standing}_{k,\pm\omega} = \min\Big(X_{k,+\omega}, X_{k,-\omega}\Big) \tag{7.77}$$

$$X^{Traveling}_{k,\pm\omega} = X_{k,\pm\omega} - X^{Standing}_{k,\pm\omega} \tag{7.78}$$

Here $X_{k,\pm\omega}$ is defined in 7.67. Using the definitions in 7.77 and 7.78 has the advantages that the standing and traveling components can be reconstructed for display, and the possibly important covariance between the standing and traveling components can be explicitly computed.

7.16.2 Example of Space-Time Spectral Analysis

In this section we will discuss an example of space-time spectral analysis of tropical winds and outgoing longwave radiation (OLR). We obtain wind data from the ERA-Interim reanalysis project (Dee et al., 2011) and OLR data from the NOAA daily record (Liebmann and Smith, 1996). Daily instantaneous 0Z ERA Interim data were used beginning in 1979 and extending for 16 years. The daily OLR data used extend from 1979 through 2013.

Prior to spectral analysis the data were processed in the following way. The annual cycle was first removed by fitting the first 4 harmonics of the annual period (365.25 days) to the entire data set. Next The data were averaged in latitude from 15S to 15N to measure the equatorially symmetric part of the variability, and a difference was taken of the average from the equator to 15N minus the average of the equator to 15S to measure the equatorially anti-symmetric part of the variability. At this point we have time series at each longitude. Next we prewhiten each time series by removing the red noise using the autocorrelation for each time series following the procedure outlines in section 7.12. This is more objective than using the smoothing and ratioing procedure employed by Wheeler and Kiladis (1999), and it retains the correct ratio of variance at different frequencies. The data are now ready for analysis. The first step is to do a Fourier transform in longitude to divide the variance into different zonal wavenumbers as in 7.72. The data are spaced at 2.5 degrees, so that there are 72 grid points around a latitude circle. Since this is not a power of two, this Fourier Transform is done by regressing sine and cosine functions onto the data in longitude at each time.

Next we need to perform the time Fourier analysis as in 7.73. For this we wish to use a fast Fourier Transform, and therefore need to divide the time series into chunks whose lengths are powers of 2. If we choose a chunk length of 128 days we have 45 realizations in a time series of 16 years. We therefore divide the time series in to chunks of 128 days. Because we are going to use a Hamming filter, we overlap these chunks by 50%. Once we have selected a chunk, we remove the linear trend and multiply it by the Hamming window function. The FFT program we use returns the complex coefficients of a complex Fourier transform, which we wish to translate into the real coefficients of a sine and cosine expansion.

$$\begin{aligned} x(t) &= \sum_{-\omega}^{\omega} F_\omega e^{i\omega t} \\ x(t) &= \sum_{\omega} A_\omega \cos(\omega t) + B_\omega \sin(\omega t) \end{aligned} \tag{7.79}$$

If the original time series $x(t)$ is real then the complex Fourier coefficients corresponding to positive and negative ω must be complex congugates of each other. We can then easily show that,

$$A_\omega = 2\mathrm{Real}(F_\omega) \qquad B_\omega = -2\mathrm{Imag}(F_\omega) \tag{7.80}$$

With these identities we can rationalize the real formulation of (Hayashi, 1971) with a standard complex FFT. We cannot average the Fourier transforms, but must calculate the power spectrum $X^2_{k,\pm\omega}$ for each chunk and then average those together to get our averaged spectrum.

We plot the averaged space-time spectra with zonal wavenumber as the abscissa and frequency as the ordinate, and denote westward moving waves with negative wavenumbers. This has the advantage of connecting the Mixed Rossby-Gravity wave across zero wavenumber in a way that would be less clear if the axes were reversed. Figure 7.17 shows a few plots of the space-time spectra of OLR and wind that illustrate the power of this technique.

Figure 7.17a shows the space-time power spectrum of equatorially symmetric OLR. Westward propagating Rossby waves are seen, but more prominent are the eastward-propagating Madden-Julian Oscillation (MJO) at low frequencies and low zonal wavenumbers and eastward propagating convectively-coupled Kelvin waves spanning wavenumbers 2 to 8 and periods from 10 to 3 days. The anti-symmetric OLR (Fig. 7.17b)shows some low-frequency variability across a larger range of wavenumbers, the westward propagating Rossby waves and a new feature centered at zonal wavenumber zero and a period of 4 days that represents a broad range of mixed Rossby-gravity waves corresponding to those investigated in section 7.12.1.

Space-time power spectra for symmetric 850hPa winds are shown in Figures 7.17c,d. The MJO is very apparent in the symmetric zonal wind, as are the Kelvin waves. The symmetric zonal wind also shows a

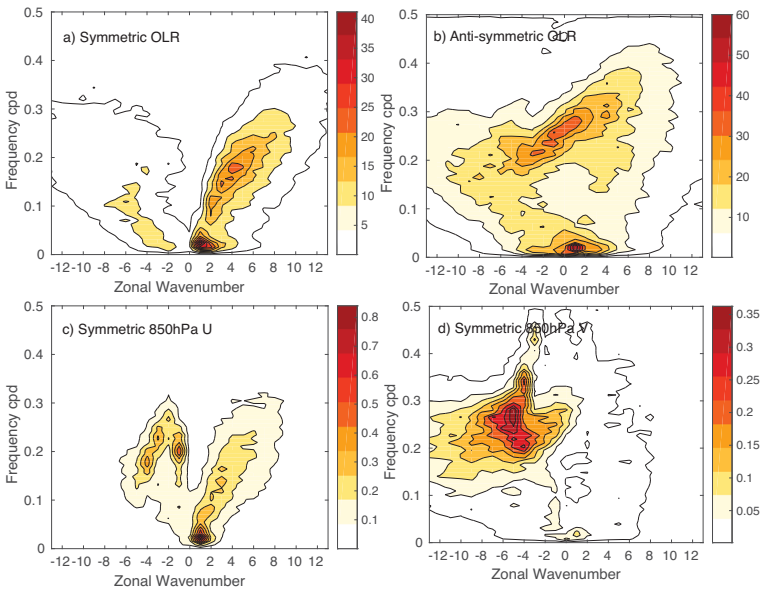

Figure 7.17 Space-time spectral analysis of equatorial waves. a) OLR averaged over 15S-15N, b) OLR averaged over 0-15N minus OLR averaged over 0-15S, c) Zonal wind at 850hPa averaged over 15S-15N, d) Meridional wind averaged over 15S-15N. Negative wavenumbers indicate westward traveling waves. Frequency is given in cycles per day.

nice peak in variance associated with the 5-day wave of zonal wavenumber 1 (Geisler and Dickinson, 1976; Hendon and Wheeler, 2008). The mixed Rossby-gravity waves are seen in the symmetric meridional wind. The westward propagating Rossby waves seen in the OLR show up better in the space-time power spectra of the asymmetric components of velocity, which are not shown here.

Chapter 8
Filtering

8.1 Introduction

In this chapter we will consider the problem of filtering time or space series so that certain frequencies or wavenumbers are removed and some are retained. Filtering is an often used and sometimes abused method of accentuating certain frequencies and removing others. The technique can be used to isolate frequencies that are of physical interest from those that are not. It can be used to remove high frequency noise or low frequency trends from time series and leave unaltered the frequencies of interest. These applications are called low-pass and high-pass filtering, respectively. A band-pass filter will remove both high frequencies and low frequencies and leave only frequencies in a band in the middle. Band-pass filters tend to make even noise look periodic, or at least quasi-periodic, so one should verify that the frequency range of interest has some physically meaningful content before selecting it with a band-pass filter. We will begin by noting a few important theorems that constitute the fundamental tools of filtering.

8.1.1 The Convolution Theorem

If two functions $f_1(t)$ and $f_2(t)$ have Fourier Transforms $F_1(\omega)$ and $F_2(\omega)$, then the Fourier transform of the product of $f_1(t)$ and $f_2(t)$ is

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} F_1(\lambda)\, F_2(\omega-\lambda)\, d\lambda \tag{8.1}$$

and the Fourier transform of

$$\int_{-\infty}^{\infty} f_1(\tau)\, f_2(t-\tau)\, d\tau \tag{8.2}$$

is $F_1(\omega) \times F_2(\omega)$. This latter result is the most useful in filtering, since it says that the Fourier transform of the convolution of two functions in time is just the product of the Fourier transforms of the two individual functions.

8.1.2 Parseval's Theorem

Parseval's theorem states that

$$\int_{-\infty}^{\infty} f_1(t)\, f_2(t)\, dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} F_2(\omega)\, F_1(\omega)^*\, d\omega \tag{8.3}$$

Here $F_1(\omega)^*$ indicates the complex conjugate of $F_1(\omega)$. For the case where $f_1(t) = f_2(t) = f(t)$, Parseval's theorem yields,

$$\int_{-\infty}^{\infty} f(t)^2 \, dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} |F(\omega)|^2 \, d\omega \tag{8.4}$$

Equation 8.4 equates the integral over time of the variance to the integral over frequency of the power. Thus the area under the power spectrum plotted versus frequency is equal to the variance of the time series.

8.2 Filtering

Suppose we wish to modify oscillations of certain frequencies in a time series while keeping other frequencies the same. For example, remove high frequency oscillations (low-pass filter), remove low frequencies (high-pass filter), or both (band-pass filter). The ozone layer of Earth s atmosphere is a low-pass filter for sunlight in the sense that it absorbs all energy with wavelengths shorter than 300 nanometers before it reaches the surface. Several different approaches to filtering can be taken.

8.2.1 Fourier Method

One possible method is to Fourier transform the timeseries of interest, multiply the Fourier coefficients by a suitable set of weights to remove or amplify the frequencies of interest, then reconstitute the time series by inverting the Fourier transform of the modified Fourier coefficients to produce the filtered timeseries. Here we show the proposed mathematical operation for a timeseries represented by a cosine series.

$$f(t) = \sum_{i=0}^{M} C_{\omega_i} \cos(\omega_i t - \phi_i) \tag{8.5}$$

$$f_{filtered}(t) = \sum_{i=0}^{M} R(\omega_i) \times C_{\omega_i} \cos(\omega_i t - \phi_i) \tag{8.6}$$

Here $R(\omega)$ is the amplitude response function for our filter, which would typically vary from zero (removal of the frequency ω) to one (passing the frequency through the filter unchanged).

$$R(\omega) = \frac{C_{\omega - filtered}}{C_{\omega - original}} \tag{8.7}$$

The problem with this method is that the reconstructed time series may not resemble the original one, particularly near the ends. This is the same general characteristic of functional fits discussed in an earlier chapter. Also, you need the whole record of data before you can produce a single filtered data point, and the most recently acquired values are at the end of the data stream, where the problems with the Fourier method are worst.

For realistic applications we most often use a local system of weights, so that the filtered timeseries always resembles the original timeseries at each point. These filter weight methods can be recursive, in which the already filtered data points are used in the filter, or non-recursive, in which only original non-filtered data are used to construct the filtered time series.

8.2.2 Centered, Non-recursive Filtering Method

A good place to start is with centered, non-recursive weights, which will introduce us to filtering with a simple but practical method. In this method, the original time series is subjected to a weighted running average, so that the filtered point is a weighted sum of surrounding points.

$$f_{filtered}(t) = \sum_{k=-J}^{J} w_k \, f(t + k\Delta t) \tag{8.8}$$

Some data points will be lost from each end of the time series since we do not have the values to compute the smoothed series for $i < J$ and $i > N - J$. It seems obvious that such an operation can produce only smoothed time series if the weights are positive and hence constitutes a low-pass filter. However, a high-pass filter can be constructed quite simply by subtracting the low-pass filtered time series from the original time series. The new high-pass response function will then be,

$$R_H(\omega) = 1 - R_L(\omega) \tag{8.9}$$

Where the subscripts H and L refer to high- and low-pass filters. One can then design a high-pass filter by first designing a low-pass filter that removes just those frequencies one wishes to retain. You can also make a band-pass filter by applying a low pass filter to a time series that has already been high-passed (or vice versa), in which case the response function is the product of the two response functions (center case in Figure 8.1). Or you can subtract a low pass filtered version of the data set from another one with a cutoff at a higher frequency, as illustrated on the right of Figure 8.1.

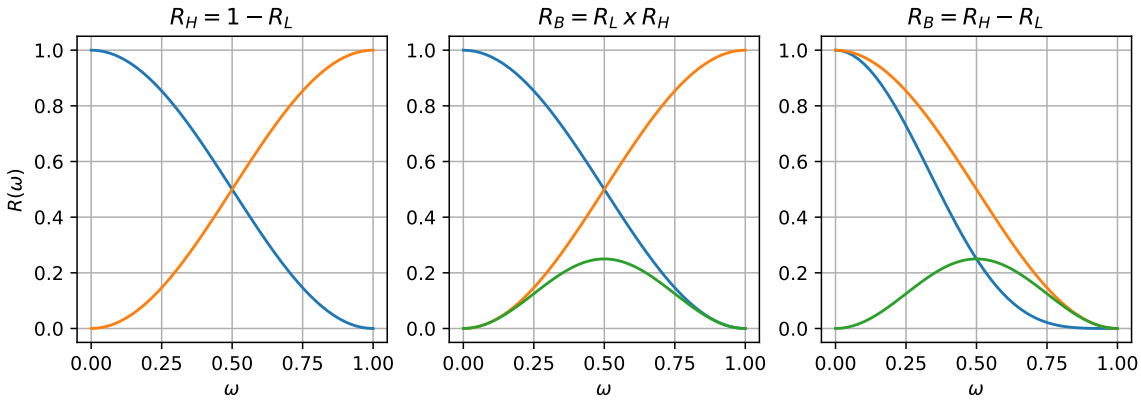

Figure 8.1 Examples of response functions for low(blue), high(orange) and band-pass(green) filters. High-pass can be obtained by subtracting the low-pass from the original data. Band-pass can be obtained by applying both a high-pass and low pass filter, or subtracting low-pass filtered data from low-pass filtered data where the cutoff is at a higher frequency.

8.2.3 Obtaining the Response Function

The response function is the spectrum of amplitude modifications made to all frequencies by the filtering function. It is the ratio of the filtered output amplitude to the unfiltered input amplitude.

$$R(\omega) = \frac{\text{filtered time series at frequency } \omega}{\text{original time series at frequency } \omega} \tag{8.10}$$

Filtering can alter both the amplitude and the phase, so that $R(\omega)$ may be real or imaginary, but we may be interested only in the change of amplitude or power of the response function.

$$|R(\omega)| = \frac{\text{amplitude of filtered time series at frequency } \omega}{\text{amplitude original time series at frequency } \omega} \tag{8.11}$$

$$|R(\omega)|^2 = \frac{\text{power of filtered time series at frequency } \omega}{\text{power of original time series at frequency } \omega} \tag{8.12}$$

Some filtering is done by nature and instruments and these may introduce phase errors. Filters can be designed with a real response function, unless we desire to introduce a phase shift. Phase shifting filters are not too commonly used in meteorological or oceanographic data analysis or modeling, and so we will not discuss them except in the context of recursive filtering, where a phase shift is often introduced with a single pass of a recursive filter. Centered, symmetric, non-recursive filters have the feature of giving a real response function, and so do not introduce phase changes into the filtered time series.

How do we design a centered, non-recursive set of weights with the desired frequency response? Our smoothing operation can be written,

$$g(t) = \sum_{k=-J}^{J} f(t + k\Delta t)\, w(k\Delta t) \tag{8.13}$$

where $g(t)$ is the smoothed time series, $f(t)$ is the original time series and $w(k\Delta t)$ are the discrete weights applied at 2J+1 time points. In the continuous case we can write this as

$$g(t) = \int_{-\infty}^{\infty} f(\tau)\, w(t-\tau)\, d\tau \tag{8.14}$$

The filtered output $g(t)$ is just the convolution of the unfiltered input series $f(t)$ and the filter weighting function $w(t)w(t)$. From the convolution theorem the Fourier transform of

$$\int_{-\infty}^{\infty} f(\tau)\, w(t-\tau)\, d\tau \quad \text{is} \quad F(\omega)\, W(\omega) \tag{8.15}$$

so that the Fourier transform of the filtered time series is

$$G(\omega) = F(\omega)\, W(\omega) \tag{8.16}$$

So to obtain the Fourier coefficients of the filtered time series we multiply the Fourier transform of the input time series by the Fourier transform of the weighting function. The power spectrum of the filtered time series is thus,

$$P_g(\omega) = G(\omega)\, G(\omega)^* = F(\omega)W(\omega)\, (F(\omega)W(\omega))^* = |F(\omega)|^2\, |W(\omega)|^2 \tag{8.17}$$

From 8.17 we can infer that the response function for the power spectrum is just the power spectrum of the weighting function.

8.2.4 Simple Example of Cosine Wave

Suppose our input time series consists of a single cosine wave with amplitude 1.0. Assuming that the weights are symmetric, the filtered signal is then

$$g(t) = \sum_{k=0}^{J} w(k\Delta t)\, \cos(\omega(t + k\Delta t)) \tag{8.18}$$

The Fourier Transform of the filtered time series is

$$G(\omega) = 2\int_0^\infty g(t)\,\cos\omega t\,dt \tag{8.19}$$

Substituting in our expressiom for $g(t)$ we get,

$$\begin{aligned} G(\omega) &= 2\int \sum_k w_k \cos(\omega t + \omega k\Delta t)\cos\omega t\,dt \\ &= \sum_k w_k\,\cos\omega k\Delta t \\ &= \sum_k w_k\,\cos\omega t_k \end{aligned} \tag{8.20}$$

For a single cosine wave at frequency ω, the Fourier transform is $F(\omega) = 1$, so that

$$R(\omega) = \frac{G(\omega)}{F(\omega)} = \sum_k w_k\,\cos\omega t_k = W(\omega) \tag{8.21}$$

The response function $R(\omega)$ is just the Fourier transform of the filter weights. If we assume that the weighting function and the response function are symmetric and real.

$$R(\omega) = W(\omega) = 2\int_0^\infty w(t)\,\cos\omega\tau\,d\tau \qquad \tau = k\Delta t \tag{8.22}$$

and conversely, the weighting function can be obtained from a specified response function

$$w(\tau) = 2\int_0^\infty R(\omega)\,\cos\omega\tau\,d\omega \tag{8.23}$$

and $w(\tau)$ and $R(\omega)$ constitute a Fourier transform pair.

8.2.5 The Running Mean Smoother

A popular but very sub-optimal smoother is the running mean smoother, for which the weights are a boxcar function.

$$w(\tau) = \frac{1}{T} \text{ on the interval } 0 < \tau < T \tag{8.24}$$

and zero elsewhere. As we recall from chapter 7, the Fourier transform of a boxcar function is a sinc function.

$$R(\omega) = \frac{\sin\frac{\omega T}{2}}{\frac{\omega T}{2}} \tag{8.25}$$

This response function approaches one at zero frequency, hence it has no effect on frequencies that are very long compared to the averaging interval T. As one can see from Figure 8.2, the response function is zero at every harmonic of the averaging interval T, when $\omega T = 2\pi n, n = 1, 2, 3, ...$ The running mean thus removes exactly all harmonics of the fundamental period T. The running mean has several major weaknesses as a filter, however. It cuts off very slowly and then passes through zero and has a wiggle at higher frequencies that maxes out at about -0.2 near $\omega = 3\pi/2$, and decays very slowly. The sharp cutoff of the weighting

function gives rise to Gibbs phenomena in the response function. We might do better to select weighting functions that were more tapered.

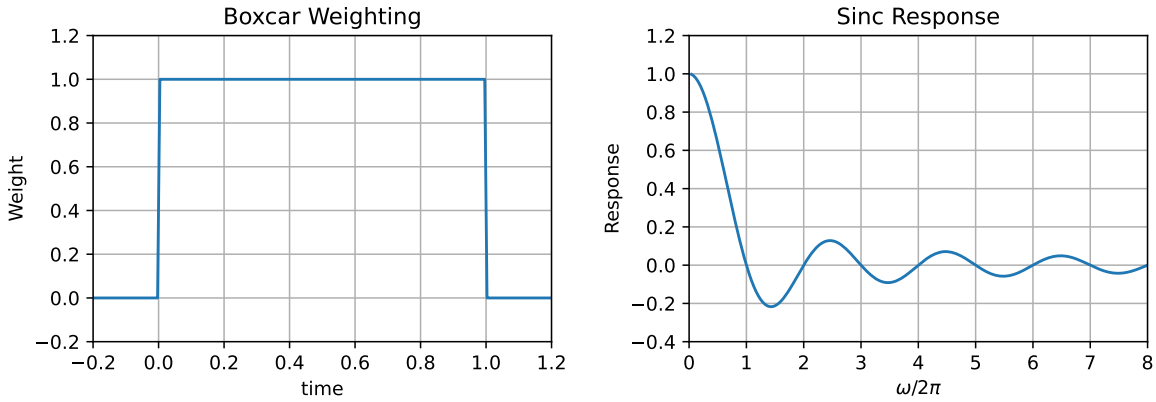

Figure 8.2 Boxcar (running mean) weights and the associated sinc frequency response function.

8.2.6 Construction of Symmetric Non-recursive Filters

In this section we will describe methods for the construction of simple non-recursive filters. Suppose we consider a simple symmetric non-recursive filter.

$$y_n = \sum_{k=-N}^{N} C_k \, x_{n-k} \qquad where \; C_{-k} = C_k \tag{8.26}$$

To perform a Fourier transform of (8.26) it is useful to introduce the *Time Shifting Theorem*. Suppose we wish to calculate the Fourier transform of a time series $f(t)$, which has been shifted by a time interval $\Delta t = a$. Begin by replacing t with $t \pm a$ in the Fourier integral representation of $f(t)$,

$$f(t \pm a) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega) \, e^{i\omega(t \pm a)} d\omega \tag{8.27}$$

which can be slightly arranged to,

$$f(t \pm a) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega) \, e^{\pm i\omega a} \, e^{i\omega t} d\omega \tag{8.28}$$

from which we infer that the Fourier transform of $f(t \pm a)$ is $F(\omega) \, e^{i\omega a}$. Thus the Fourier transform of the time-shifted time series is the Fourier transform of the original time series multiplied by the factor $z = e^{i\omega \Delta t}$.

We can then Fourier transform (8.26) and use the Time Shifting Theorem to obtain,

$$Y(\omega) = \left[\sum_{k=-N}^{N} C_k \, e^{i\omega k \delta t} \right] X(\omega) \tag{8.29}$$

Here $Y(\omega)$ and $X(\omega)$ are the Fourier transforms of $y(t)$ and $x(t)$, *respectively*. Because the weights are symmetric, $C_{-k} = C_k$, and we know the identity,

$$\cos x = \frac{e^{ix} + e^{-ix}}{2} \tag{8.30}$$

we can write (8.29) as,

$$R(\omega) = \frac{Y(\omega)}{X(\omega)} = C_0 + 2\sum_{i=1}^{N} C_k \cos(\omega k \Delta t) \tag{8.31}$$

8.2.7 Frequency Response of Simple Filters

Armed with the simple formula 8.31 we can investigate the frequency response functions of several simple centered, non-recursive filters.

8.2.7.1 The Running Mean Smoother

The running mean smoother replaces the central value on an interval with the average of the values surrounding that point. The running mean can be taken over an arbitrary number of points, e.g. 2, 3, 5, 7. Referring to (8.31) again, a running mean smoother has $C_k = 1/(2N+1$, where $-N < k < N$. The length of the running mean smoother is $2N + 1$. We write below the response functions for running-mean smoothers of length 3,5, and 7. These response functions apply to the Nyquist interval $0 < \omega < \pi/2$

$$\begin{aligned}
2N+1 = 3: &\qquad R(\omega) = \frac{1}{3} + \frac{2}{3}\cos(\omega\Delta t) \\
2N+1 = 5: &\qquad R(\omega) = \frac{1}{5} + \frac{2}{5}\cos(\omega\Delta t) + \frac{2}{5}\cos(2\omega\Delta t) \\
2N+1 = 7: &\qquad R(\omega) = \frac{1}{7} + \frac{2}{7}\cos(\omega\Delta t) + \frac{2}{7}\cos(2\omega\Delta t) + \frac{2}{7}\cos(3\omega\Delta t)
\end{aligned} \tag{8.32}$$

These square weighting functions give damped sine wave response functions (sinc functions), which are generally undesirable, as previously noted. A slightly tapered weighting function, such as the 1-2-1 filter gives a much nicer response function.

$$1-2-1 \text{ Filter}: \qquad R(\omega) = \frac{1}{2} + \frac{1}{2}\cos(\omega\Delta t) \tag{8.33}$$

We have to alter 98.31) a bit to compute the response function for a 1-1 Filter, a running mean that just averages adjacent values. The result is:

$$1-1 \text{ Filter}: \qquad R(\omega) = \frac{1}{2}\cos(\frac{1}{2}\omega\Delta t) \tag{8.34}$$

All these response functions are plotted in Figure 8.3. Note how the 1-2-1 filter cuts off more sharply than the 1-1 filter (running mean 2), but does not have the ugly negative side lobes of the running mean filters and exactly removes the highest resolved frequency. The 1-2-1 is a good simple filter and can be applied multiple times if a stronger low-pass filter is desired.

8.3 General Symmetric Non-recursive Filter Weights

We can find the weights that would give a desired response function as follows. Multiply both sides of (8.31) by $\cos(j\omega\Delta t)$ $j = 0, 1, 2, \ldots N$ and then integrate frequency ω over the Nyquist interval $0 - \pi/\Delta t$

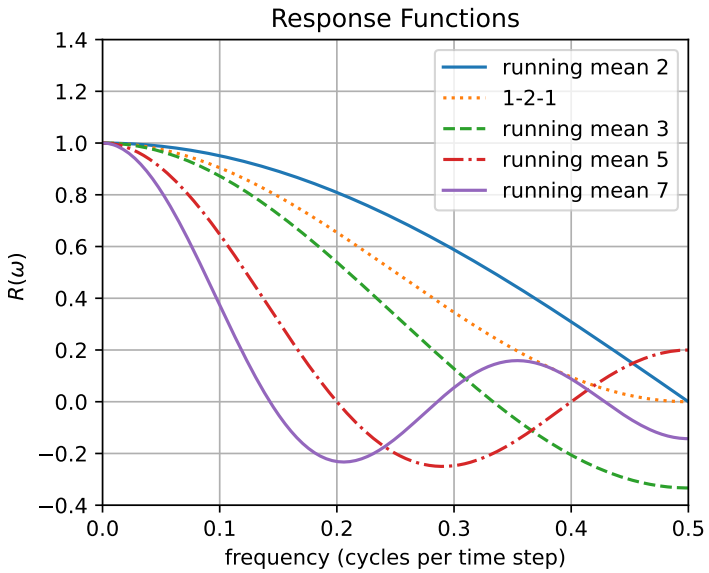

Figure 8.3 Response Functions on the Nyquist interval for various centered, non-recursive filters.

$$\int_0^{\pi/\Delta t} \cos(j\omega\delta t) R(\omega) d\omega = 2C_j \int_0^{\pi/\Delta t} \cos(j\omega\Delta t)\cos(k\omega\Delta t) d\omega \tag{8.35}$$

This yields,

$$C_k = \frac{1}{\pi}\int_0^{\pi} \cos(k\omega')\, R(\omega') d\omega' \tag{8.36}$$

Here $\omega' = \omega\Delta t$, so that $0 < \omega' < \pi$ is the Nyquist interval. From (8.36) we can derive the appropriate weighting coefficient for any arbitrary desired response function $R(\omega)$.

The ideal low-pass filter response function might be one up to some chosen frequency and then cut off abruptly to zero. Let's suppose we want the cutoff to appear at a frequency that is some fraction α $(0 < \alpha < 1)$ of the Nyquist interval, as follows.

$$R(\omega) = \begin{cases} 1 & \omega < \alpha\pi \\ 0 & \omega > \alpha\pi \end{cases} \tag{8.37}$$

Using (8.37) we can perform the integral in (8.36) and obtain,

$$\begin{aligned} C_k &= \frac{1}{\pi}\int_0^{\alpha\pi} \cos(k\omega)\, d\omega \\ C_k &= \frac{1}{k\pi}\sin(\alpha k\omega') \end{aligned} \tag{8.38}$$

Note that the amplitude of the coefficients drops off as k^{-1}, which is rather slow. The coefficients, or weights, C_k, are a $\mathtt{sinc}$ function in k, as shown previously in Section 8.2.5. To get a really sharp cutoff of the response function we need to use a large number of weights. Usually we want to keep the number of points to a minimum, because we lose $N-1$ data off each end of the time series and because the computations take time. The computation time problem can be alleviated with the use of recursive filters.

If we truncate (8.37) at some arbitrary value of N, then the response function will be less sharp than we would like and will have wiggles associated with Gibb's phenomenon. This is shown in Figure 8.4, which

shows the response functions (8.31) for the weights (8.38) for truncation of N=6, N=12, and N=24. The value of $\alpha = 0.5$ was chosen to cut the Nyquist interval in the center. One can see that more filter weights give a sharper cutoff, but that the Gibb's Phenomena wiggles are undesirable.

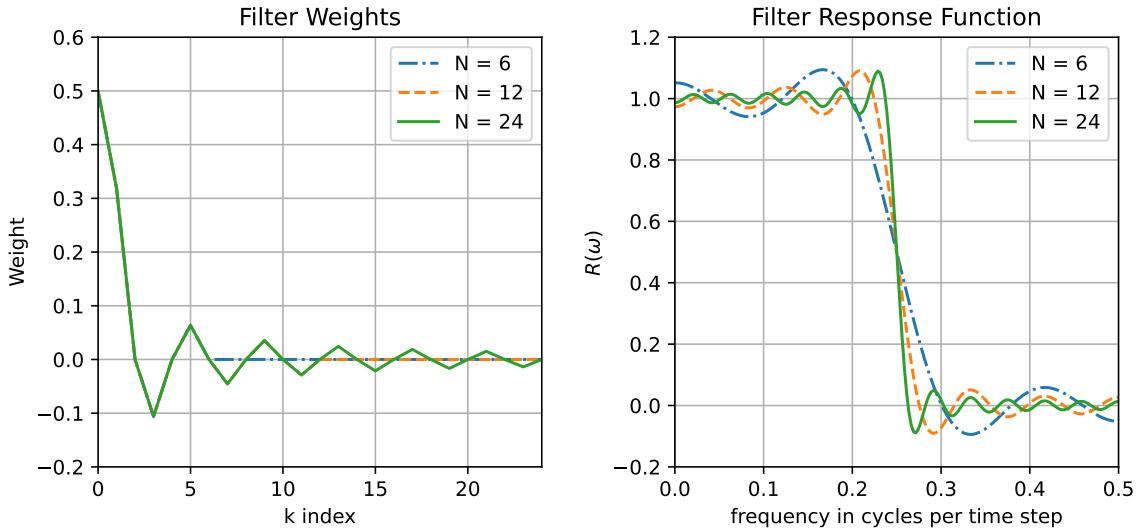

Figure 8.4 Filter weights and response functions for centered non-recursive filters with $\alpha = 0.5$ and N = 6, 12 and 24. Note that only the non-negative k indices are shown.

8.3.1 Lanczos Smoothing of Filter Weights

The wiggles in the response functions of Figure 8.4 have a wavelength of approximately the last included or the first excluded harmonic of (8.38). We can remove this harmonic by smoothing the response function. The running mean smoother exactly removes oscillations with a period equal to that of the length of the running mean smoother. The wavelength of the last harmonic included in (8.38) is $2\pi/N\Delta t$, which suggests that we smooth the response function in the following way.

$$\widetilde{R}(\omega) = \frac{N\Delta t}{2\pi} \int_{-\pi/N\Delta t}^{\pi/N\Delta t} R(\omega) d\omega \tag{8.39}$$

Substituting our equation for the response function (8.4) into (8.39) we get,

$$\begin{aligned} \widetilde{R}(\omega) &= C_0 + \frac{N\Delta t}{2\pi} \int_{-\pi/N\Delta t}^{\pi/N\Delta t} 2 \sum_{k=1}^{N} C_k \cos(k\Delta t\omega) d\omega \\ &= C_0 + \frac{2N}{\pi} \sum_{k=1}^{N} \frac{C_k}{k} \cos(k\Delta t\omega) \sin(\frac{k\pi}{N}) \end{aligned} \tag{8.40}$$

Rearranging this a little, we obtain,

$$\widetilde{R}(\omega) = C_0 + 2\sum_{k=1}^{N}\left\{\frac{\sin(\frac{k\pi}{N})}{\frac{\pi k}{N}}\right\} C_k \cos(k\Delta t\omega) \tag{8.41}$$

Here we can clearly see that the running mean smoother of the response function has given us a new set of filter weights, where we multiply the original weights by the factor $\mathrm{sinc}(\frac{\pi k}{N})$.

$$\widetilde{C}_k = \mathrm{sinc}\left\{\frac{\pi k}{N}\right\} C_k \tag{8.42}$$

These factors are sometimes called the sigma factors. Note that the last coefficient, C_N, disappears entirely because the sigma factor is zero ($\sin\pi = 0$).

The effect of applying the Lanczos smoothing to the filter weights and response functions are shown in Figure 8.5. The Gibbs phenomenon is greatly reduced and the cutoff is considerably more gradual than for the unsmoothed filter weights. It is clear that one must use a relatively large number of weights to get a good result, since the resulting response function for $N = 6$ is not very functional, although the $N = 12$ $N = 24$ cases look OK.

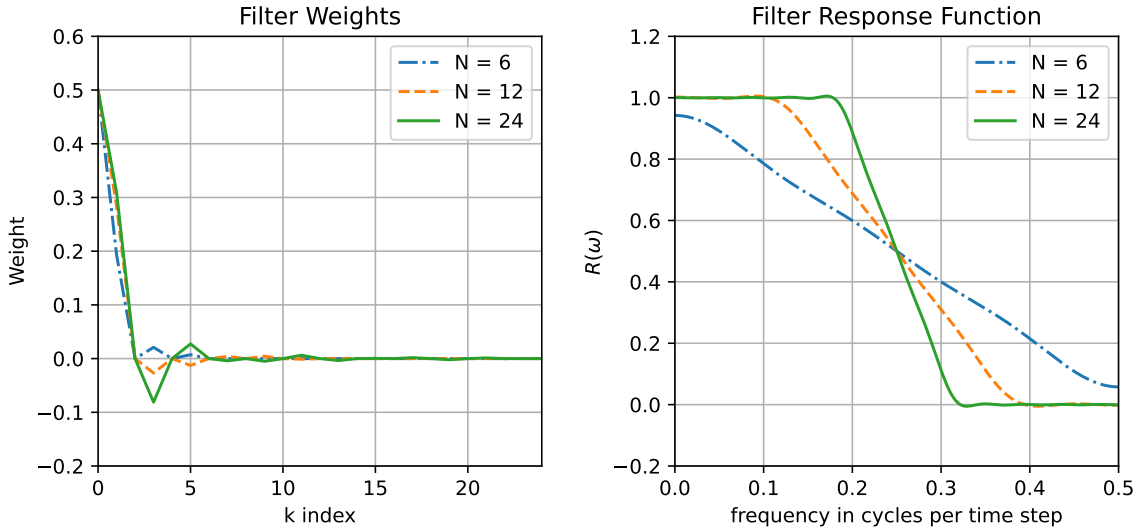

Figure 8.5 Filter weights and response functions for centered non-recursive filters with $\alpha = 0.5$ and N = 6, 12 and 24 after applying Lanczos smoothing.

8.4 Recursive Filters

The filters we have discussed so far are obtained by convolving the input series $x(n\Delta t) = x_n$ with a weighting function w_k, in the following way.

$$y_n = \sum_{k=-K}^{K} w_k\, x_{n+k} \tag{8.43}$$

Such filtering schemes will always be stable, but it can require a large number of weights to achieve a desired response function. If greater efficiency of computation is desired, then it may be attractive to consider a recursive filter of the general form,

$$y_n = \sum_{k=-0}^{K} b_k\, x_{n-k} + \sum_{j=-0}^{J} a_j\, y_{n-j} \tag{8.44}$$

In this case the filtered value depends not only on the unfiltered input series, but also on previous values of the filtered time series. In general, sharper response functions can be obtained with fewer weights and thereby fewer computations than with non-recursive filters. The method of constructing the weights for a recursive filter from a desired response function is not as easy as with convolution filters, and the filtering process is not necessarily stable.

8.4.1 Response Functions for General Linear Filters

To find the response function for the general linear filter (8.44) let's first rearrange it in the following way,

$$y_n - \sum_{j=-0}^{J} a_j\, y_{n-j} = \sum_{k=-0}^{K} b_k\, x_{n-k} + \sum_{j=-0}^{J} a_j\, y_{n-j} \tag{8.45}$$

We next take the Fourier transform of (8.45) and use the time shifting theorem.

$$Y(\omega)\left\{1 - \sum_{j=1}^{J} a_j\, z^{-j}\right\} = X(\omega)\left\{\sum_{k=1}^{K} b_k\, z^{-k}\right\} \tag{8.46}$$

Here $z = e^{i\omega\Delta t}$. From this we can obtain the response function.

$$R(\omega) = \frac{Y(\omega)}{X(\omega)} = \frac{\left\{\sum_{k=1}^{K} b_k\, z^{-k}\right\}}{\left\{\sum_{j=1}^{J} a_j\, z^{-j}\right\}} \tag{8.47}$$

Here $R(\omega)$ is the system function of the general recursive filter and measures the ratio of the Fourier transform of the output function to the input function. In general $R(\omega)$ will be complex for recursive filters, which means that recursive filters will introduce a phase shift in the frequencies that they modify. This is because the filters are not symmetric, in general. Physically realizable filters, as might be employed to real-time data or in electric circuits, cannot be symmetric, since the future data are not known at the time the filtration of the present value must be produced. In data applications where we want to remove the phase shift, we generally run the filter forward and backward across the data set. The real squared amplitude response function can be obtained from,

$$|R(\omega)|^2 = R(\omega)R(\omega)^* \tag{8.48}$$

where the asterisk indicates the complex conjugate.

8.4.2 A Simple Recursive Filter

We can illustrate some important facts about recursive filters by considering the simple example of a recursive filter given by,

$$y_n = x_n + 0.95\, y_{n-1} \tag{8.49}$$

The response function for this filter can be gotten from the general formula (8.47).

$$R(\omega) = \frac{1.0}{1.0 - 0.95z^{-1}} \tag{8.50}$$

We can find the equivalent non-recursive filter by dividing out the rational factor in (8.50) to obtain a polynomial in z. The result is,

$$R(\omega) = \frac{1.0}{1.0 - 0.95z^{-1}} = 1.0 - 0.95z^{-1} + 0.9025z^{-2} - 0.8574z^{-3} + 0.8145z^{-2} + \ldots \tag{8.51}$$

or

$$R(\omega) = \frac{1.0}{1.0 - 0.95z^{-1}} = 1.0 + \sum_{n=1}^{\infty} 0.95^n \, z^{-n} \tag{8.52}$$

Notice how slowly the coefficients of the polynomial decay. These coefficients are also the weights of the equivalent non-recursive filter. Thus many, many points are necessary to replicate the effect of the recursive filter with a non-recursive filter.

8.4.3 Impulse Response of a Recursive Filter

It is important to know how many data points a recursive filter must pass over before its response begins to settle out. This will indicate how many points must be disregarded off the end of a time series that has been filtered recursively. We can address this question by asking how the filter responds to a unit impulse time series of the form.

$$x_n = \begin{cases} 1.0 & n = 1 \\ 0.0 & n \neq 1 \end{cases} \tag{8.53}$$

The time series that results from filtering the time series (8.53) can be called the *impulse response* of the filter. The filter (8.49) acting on the input time series (8.53) will produce the following filtered time series.

$$y_0 = 1.0,\ y_1 = 0.95,\ y_2 = 0.9025,\ y_3 = 0.8574,\ y_4 = 0.8145,\ \ldots \tag{8.54}$$

The impulse response(8.54) of the recursive filter (8.49) is identical to the equivalent non-recursive filter and decays very slowly. So we conclude that we lose about the same number of endpoints with both types of filter. The only apparent advantage of the recursive filter is that it requires far fewer computations to achieve the same effect. The phase errors introduced by recursive filters can be reduced or eliminated by passing over the time series twice, once in the forward direction and once backward in time. The resulting amplitude response function is the square of the response function for a single application. Most high-level programming languages (e.g. Python, Matlab) have a filter function that passes twice over the data in this way.

8.4.4 Construction of Recursive Filters

The construction of appropriate weights from a system function, or response function of desired shape is not quite as straightforward for recursive filters as for non-recursive filters, and requires different mathematics. From what we have done so far, it might be obvious that what we need to do is construct the appropriate polynomial in $z = e^{i\omega\Delta t}$ that produces the desired system function as described in (8.47) Recall that z maps into the unit circle in the complex plane, as the frequency varies from zero to the Nyquist frequency. For a recursive filter to be stable, all zeros of the polynomial in the system function must be within the unit circle. It is also useful to realize that the z transform is linear, so that the system function of the sum of two filters is the sum of the system functions for each filter. Also, if two filters are applied successively, the system function of the result is the product of the system function for the two filters.

The construction of recursive filters, that is finding the appropriate weights, can be simplified by transforming the z variable to a w variable defined in the following way.

$$z = e^{i\omega\Delta t} = \frac{1+iw}{1-iw} \tag{8.55}$$

or

$$w = i\left(\frac{1-z}{1+z}\right) \tag{8.56}$$

8.4.5 Butterworth Filters

As a useful and common example we can consider the Butterworth family of filters with response functions defined as follows.

$$R(\omega)R(\omega)^* = \frac{1}{1+\left(\frac{w}{w_c}\right)^{2N}} \tag{8.57}$$

The filter has the desirable property of smoothness and high tangency at the origin and infinity. It contains two design parameters, w_c and N, which can be used to design a filter with a cutoff at the desired frequency and the appropriate amount of sharpness to the cutoff. We will not delve further into the specifics of how to derive the filter weights, as most people will simply use an off-the-shelf routine to do this. Butterworth filters are smooth and monotonic, which are generally good characteristics for data work. If a sharper cutoff is required and negative side lobes are tolerable, there are other filtering schemes with these characteristics.

In applying a Butterworth filter, one is given the choice of a cutoff frequency, our parameter α, and an order N. Let us explore the response functions and impulse response functions for several choices of N. Figure 8.6 illustrates how higher orders give sharper response functions, but also very elongated impulse response functions. These features are emphasized as the cutoff is moved to lower frequencies. For $N = 4$ the impulse response function becomes very small after about 20 time steps, while $N = 9$ requires many more time steps for its full effect to be felt.

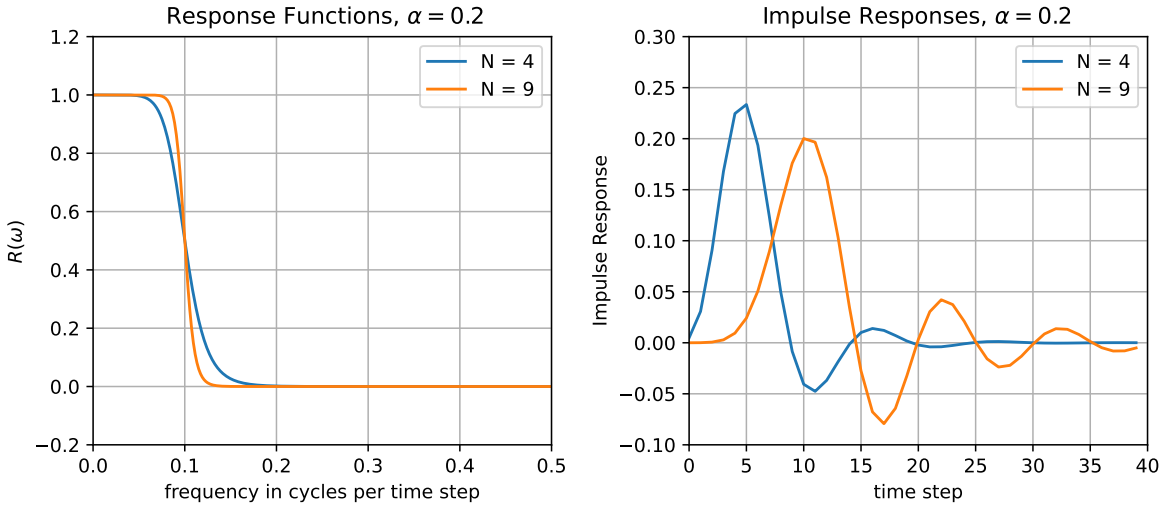

Figure 8.6 Response functions and impulse response functions for Butterworth filters with orders $N = 4$ and $N = 9$, each with $\alpha = 0.2$.

8.4.6 Example using Butterworth Filter

Suppose we have a time series that consists of low-frequency red noise $\mathrm{RN}(t)$ plus a cosine wave with a period of 5 days whose amplitude is inversely proportional to the departure of the red noise from zero. This might be caused by a wave instability that only grows when the mean value of the variable is near zero.

$$f(t) = \mathrm{RN}(t) + \frac{1}{\mathrm{RN}(t) + \epsilon} \cos\left(\frac{2\pi t}{5}\right) \tag{8.58}$$

Here ϵ is a suitably small number. Power spectral analysis of this time series shows red noise with a peak in variance at 5 days, but since the phase information is lost in spectral analysis, we would not learn about the sporadic nature of the 5-day oscillation and its relation to the low-frequency variability (Figure 8.7).

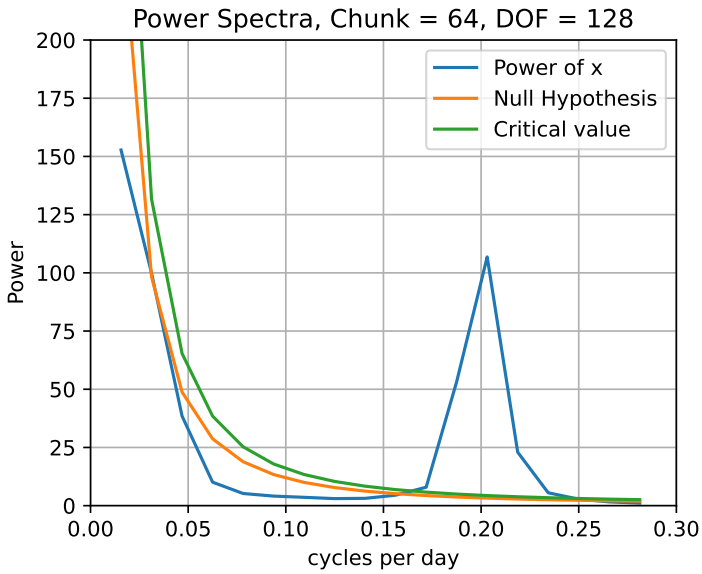

Figure 8.7 Power spectrum of a timeseries with red noise, plus a large and significnt 5-day oscillation.

The power spectrum in Figure 8.7 suggests using a Butterworth filter to separate highpass and lowpass variance at a frequency of about 0.15 cycles per day or $\alpha = 0.3$. We then plot a part of the data showing the lowpass and highpass data in Figure 8.8. One clearly gets the notion that the 5-day oscillation is episodic and has the most amplitude when the low-pass time series is near zero.

To show this another way, in Figure 8.9 we make a scatter plot of the absolute value of the high-passed timeseries versus the low-passed data. This clearly shows a peak in the amplitude of the 5-day oscillation when the low-passed time series is near zero.

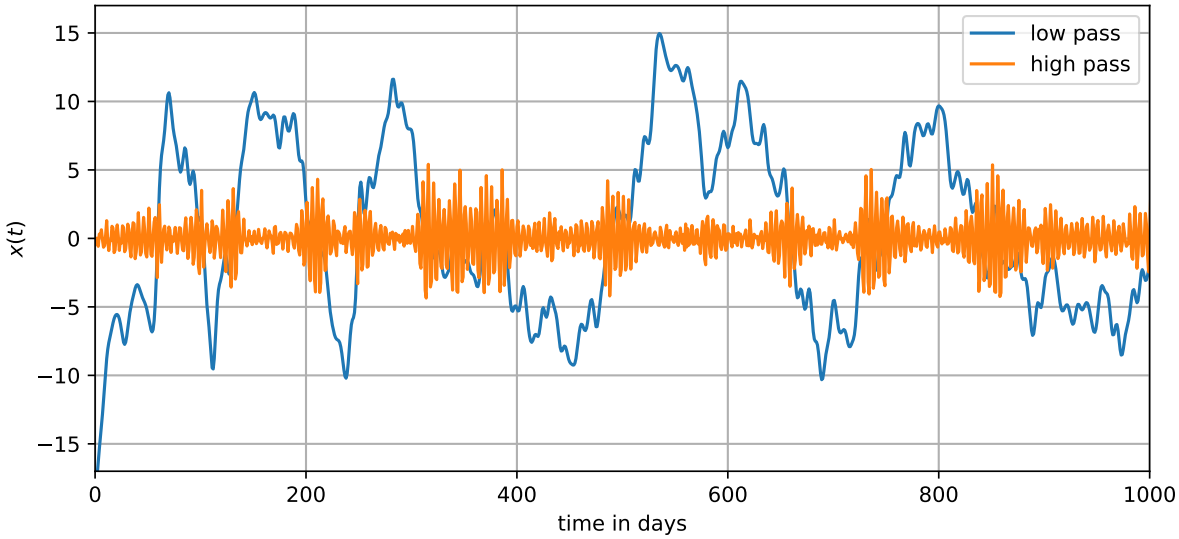

Figure 8.8 The first 1000 days of the high-passed and low-passed timeseries consisting of red noise, plus a 5-day oscillation.

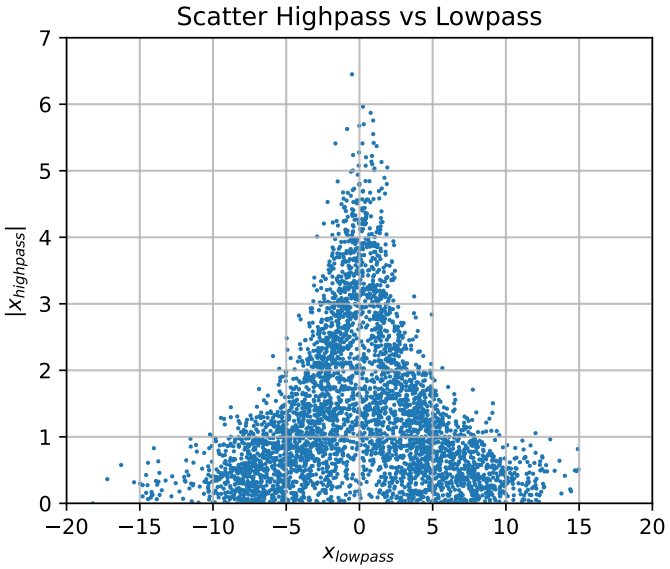

Figure 8.9 Scatter diagram of the absolute value of the high-passed time series versus the value of the low-passed time series.

Chapter 9
Wavelets

9.1 Introduction

A wavelet is a wave-like oscillation that is localized in the sense that it grows from zero, reaches a maximum amplitude, and then decreases back to zero amplitude again. It thus has a location where it maximizes, a characteristic oscillation period, and also a scale over which it amplifies and declines. Wavelet analysis developed in the largely mathematical literature in the 1980's and began to be used commonly in geophysics in the 1990's. Wavelets can be used in signal analysis, image processing and data compression. They are useful for sorting out scale information, while still maintaining some degree of time or space locality. Wavelets have been used to compress and store fingerprint information. Because the wavelet and scaling functions are obtained by scaling and translating one or two "mother functions", time-scale wavelets are particularly appropriate for analyzing fields that are fractal. Wavelets can be appropriate for analyzing non-stationary time series, whereas Fourier analysis generally is not. They can be applied to time series as a sort of fusion (or compromise) between filtering and Fourier analysis. Wavelets can be used to compress the information in two-dimensional images from satellites or ground based remote sensing techniques such as radars. Wavelets are useful because as you remove the highest frequencies, local information is retained and the image looks like a low resolution version of the full pictures. With Fourier analysis, or other global functional fits, the image may lose all resemblance to the picture, after a few harmonics are removed. This is because wavelets are a hierarchy of local fits, and retain some time localization information, and Fourier or polynomial fits are global fits, usually.

In general, you can think of wavelets as a compromise between looking at digital data at the sampled times, in which case you maximize the information about how things are located in time, and looking at data through a Fourier analysis in frequency space, in which you maximize your information about how things are localized in frequency and give up all information about how things are located in time. In wavelet analysis we retain some frequency localization and some time localization, so it is a compromise.

9.2 Wavelet Types

According to Meyer(1993), two fundamental types of wavelets can be considered, the Grossmann-Morlet time-scale wavelets and the Gabor-Malvar time-frequency wavelets. The more commonly used type in geophysics is probably the time-scale wavelet. These wavelets form bases in which a signal can be decomposed into a wide range of scales, in what is called a "multiresolution analysis". From this comes the obvious application in image compression, as one can call up additional detail as required until the exact image at the original resolution is reconstructed. The intervening coarse resolution images will look like the full resolution one, just fuzzier. This is not true in general of Fourier analysis, where throwing out the last few harmonics can cause the picture to change dramatically.

Time-scale wavelets are defined in reference to a "mother function" $\psi(t)$ of some real variable t. The mother function is required to have several characteristics: it must oscillate, and it must be localized in the

sense that it decreases rapidly to zero as $|t|$ tends to infinity. It is also very helpfult to require that the mother function have a certain number of zero moments, according to,

$$\int_{-\infty}^{\infty} t^{m-1}\,\psi(t)\,dt = 0 \quad m = 1, 2, 3, \ldots \tag{9.1}$$

Here m is the *approximation condition order* of the wavelet. If the order is one, the mean of the wavelet is zero; if the order is two, the trend of the wavelet is zero, and so forth.

The mother function can be used to generate a whole family of wavelets by translating and scaling the mother wavelet.

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\,\psi\Big(\frac{t-b}{a}\Big), \quad a > 0,\ b \in \Re \tag{9.2}$$

Here b is the translation parameter and a is the scaling parameter. Provided that $\psi(t)$ is real-valued, this collection of wavelets can be used as an orthonormal basis to describe any function $f(t)$. The coefficients of this expansion can be obtained through the usual projection.

$$\Psi_{a,b} = \int_{-\infty}^{\infty} \psi_{a,b}(t)\,f(t)\,dt \tag{9.3}$$

These coefficients measure the variations of the field $f(t)$ about the point b, with the scale given by a. A set of parameters a_k and b_j , representing different scales and locations, can be chosen to form an orthonormal basis set. In that case we can reconstruct the original data from the wavelets and their coefficients.

$$f(t) = \sum_j \sum_k \Psi_{a_k,b_j}\,\psi_{a_k,b_j}(t) \tag{9.4}$$

Wavelet analysis of this type can be performed on discrete data using quadrature mirror filters and pyramid algorithms. It is also possible to compute the transform using a Fourier transform technique. Sometimes a_k and b_j are varied more continuously to make useful diagrams with continuous variations of scale and location. This gives up the orthogonality, but has the advantage of making pictures with more resolution, as scale and location generally vary by factors of 2 (dyadic wavelets) in orthogonal wavelets.

In using wavelets for data analysis, it is important to find a set of them that provides a data description that is best-suited to the problem at hand. If wavelet analysis in general, or the particular set of wavelets chosen, are not well-suited to the problem at hand, they may not lead to any useful insight. For the non-expert, who just wants to get a useful representation, one is probably restricted to choosing from among a library of established wavelet bases, and most probably from among those for which software is already available. This library is very well developed, and techniques are available for determining whether an appropriate representation has been chosen. Python and Matlab both have highly developed wavelet tool kits.

We focus here in these notes on discrete wavelets and the discrete wavelet transform (DWT) and their applications. Wavelets are basis sets for expansion which, unlike Fourier series, have not only a characteristic frequency or scale, but also a location. They can be orthogonal, biorthogonal, or nonorthogonal.

9.3 The Haar Wavelet

Haar (1910) and others were seeking functional expansions that were alternatives to the sine and cosine series of Fourier (1822). He sought an orthonormal system $h_n(t)$ of functions on the interval [0,1] such that for any function $f(t)$, the series,

$$f(t) = \sum_n \langle f, h_n \rangle\, h_n(t) \tag{9.5}$$

would converge uniformly. The angle brackets indicate a suitably defined inner product on the interval [0,1]. Haar began with the initial function,

$$h(t) = \begin{cases} 1.0 & [0.0,\ 0.5] \\ -1.0 & [0.5,\ 1.0] \\ 0.0 & \text{elsewhere} \end{cases} \tag{9.6}$$

Building on this basic mother wavelet, Haar defines his sequence of expansion functions according to,

$$\begin{aligned} n &= 2^j + k \quad j \geqslant 0,\ \ 0 \leqslant k < 2^j \\ h_n(t) &= 2^{j/2}\, h(2^j t - k) \end{aligned} \tag{9.7}$$

Each of these functions is supported (has non-zero values) on the dyadic interval,

$$I_n = \left[k\, 2^{-j},\ (k+1)\, 2^{-j}\right] \tag{9.8}$$

which is included in the interval [0,1] if $0 \leqslant k < 2^j$. Here j is the level, from the mother wavelet level ($j = 0$), to the smallest baby wavelets $j = j_{max}$, k is the spatial index for each level, and n is a mode index, starting with mother ($n = 1$). To complete the set, one must add the function $H_0(t) = 1$ on the interval [0,1], which we can refer to as father, the smoothest level of detail, in this case a constant.

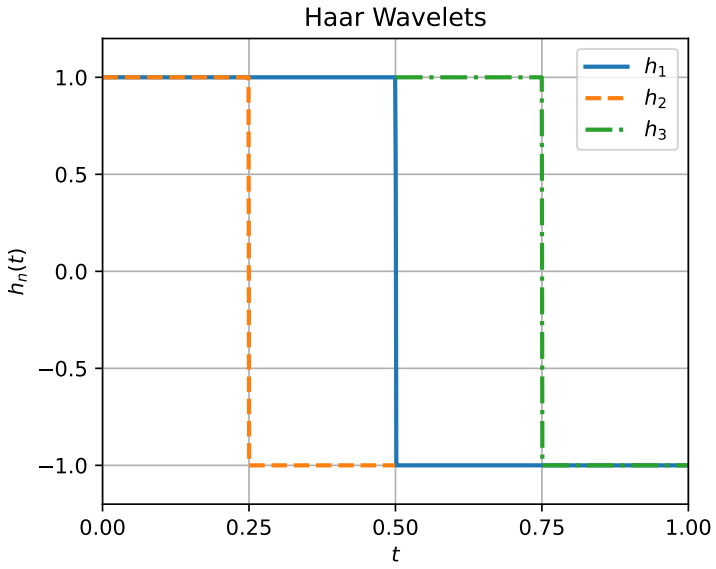

Figure 9.1 Continuous mother Haar wavelet $h_1(t)$ and her first two children.

The series $h_n(t)$ then forms an orthonormal basis on [0,1]. By looking carefully at (9.6)-(9.8) one can see that the series is the basic step function repeated on intervals that decrease in scale and increase in number by the factor of two at each level, where j is the level index and k is the number of functions at that level of detail necessary to span the interval [0,1]. Note that the mean of the Haar wavelet is zero, but that it's trend is non-zero, so that its approximation condition order is one. The 2^j in front of the Haar function in 9.7 is to normalize the functions on the interval [0, 1], but we will ignore this factor when plotting them in Figure 9.1.

9.4 Discrete Wavelet Transforms

In working with data, we have values at discrete times not at continuous times. In transforming data from time space to wavelet space. We can do this as a matrix operation, and rather that starting with the mother wavelet, we start from the finest detail that can be resolved and work our way up to the mother and father wavelet coefficients. Since the Haar wavelet is dyadic, the whole time series length must be a power of two for this to converge on the mother and father wavelets.

Since the Haar functions are orthogonal, we can derive their coefficients α_i using the relation,

$$\alpha_i = \langle \phi_i,\, x(t) \rangle \tag{9.9}$$

where the angle bracket indicates a suitably defined inner product. It may be easier to see how this is all working by considering how (9.9) looks when expressed in matrix notation, and using the abbreviation $a = 1/\sqrt{2}$.

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \\ \alpha_6 \\ \cdot\cdot \end{bmatrix} = \begin{bmatrix} a & a & & & & & \\ a & -a & & & & & \\ & & a & a & & & \\ & & a & -a & & & \\ & & & & a & a & \\ & & & & a & -a & \\ & & & & & & \cdot\cdot\cdot \end{bmatrix} \begin{bmatrix} x(1) \\ x(2) \\ x(3) \\ x(4) \\ x(5) \\ x(6) \\ \cdots \end{bmatrix} = \begin{bmatrix} y_1(1) \\ y_2(1) \\ y_1(2) \\ y_2(2) \\ y_1(3) \\ y_2(3) \\ \cdots \end{bmatrix} \tag{9.10}$$

Note that the wavelet transform is divided into a smoothing part $[a, a]$ and a wavelet part $[a, -a]$. In the final column we have divided the coefficients into the smoothed coefficients $y_1(t)$ and the wavelet coefficients $y_2(t)$, each with a value at every other time step. We then continue the wavelet transform by reserving the wavelet coefficients as the highest level of detail, then perform the wavelet transform on the smoothed coefficients.

We can think of y_1 and y_2 as the time series of the coefficients of the even and odd Haar wavelets, respectively. These have only half the time resolution of the original series. You can think of y_1 as a low-frequency representation of $x(t)$ and y_2 as the high frequency details. Often in wavelet analysis literature, the smooth function $[a, a]$ would be called the scaling function , and the wavy one $[a, -a]$ would be called the wavelet . The projection into the coefficient space of the two Haar functions is equivalent to filtering followed by "down sampling", by taking only every other point of the filtered time series. The Haar transform is an example of a two-channel filter bank. It sorts the original series into two filtered data sets. The Haar filter functions are members of a special class of filter function pairs called a quadrature mirror filter pair. After the filtering is done the sum of the energies (or variances) in the two filtered time series is equal to the variance in the original time series.

$$|y_1|^2 + |y_2|^2 = |x|^2 \tag{9.11}$$

Since we are thinking of a wavelet transform as a filtering operation, now is a good time to think about the scaling achieved by this filtering process. Remember from chapter 8 on filtering of time series how we determine the frequency response of the filter from its coefficients. The scaling function $[a, a]$ is a filtering operation that does this,

$$y(t) = a\, x(t + \frac{\Delta t}{2}) + a\, x(t + \frac{-\Delta t}{2}) \tag{9.12}$$

The Fourier Tranform of this is,

$$Y(\omega) = X(\omega)\left(a e^{i\omega\Delta t/2} + a e^{-i\omega\Delta t/2}\right) = X(\omega)\, 2a\, \cos(\omega\Delta t/2) \tag{9.13}$$

So the response function is $R(\omega) = 2a\cos(\omega\Delta t/2)$. If you wanted a unit response at zero frequency then You would choose $a = 1/2$, but because the wavelets are normalized to have unit length $a = 1/\sqrt{(2)}$, and the response function at zero frequency is $\sqrt{(2)}$. The frequency response goes from $2a\cos(0)$ to $2a\cos(\pi/2)$ while the frequency goes from zero to $\pi/\Delta t$. Just one slow transit from maximum to zero across the Nyquist interval.

For the wavelet we have

$$y(t) = a\,x(t + \frac{\Delta t}{2}) - a\,x(t + \frac{-\Delta t}{2}) \tag{9.14}$$

and the Fourier transform is

$$Y(\omega) = X(\omega)\left(ae^{i\omega\Delta t/2} - ae^{-i\omega\Delta t/2}\right) = X(\omega)\,2a\,\sin(\omega\Delta t/2) \tag{9.15}$$

So the response functions for the Haar scaling and wavelet are,

$$R_{scaling}(\omega) = 2a\,\cos(\omega\Delta t/2) \quad R_{wavelet}(\omega) = 2a\,\sin(\omega\Delta t/2) \tag{9.16}$$

From these formulas one can see that the response functions are complements of each other, so that the amplitude that is rejected by one is the amplitude that is passed by the other. This is the required characteristic of quadrature mirror filters, and will result in the preservation of power as the expansion in these wavelets continues. The Haar wavelet representation has the advantage of very good time localization, but the frequency resolution is minimal. Discrete wavelets with more weights will be able to provide better frequency resolution at the expense of less precise time localization.

9.5 The Pyramid Scheme of Discrete Wavelet Transforms.

Applying the Haar transform reduces the original N data point time series $x(t)$ into two time series of length $N/2$, which are y_1 and y_2 , as defined in (9.10). One of these contains the smoothed information and the other contains the detail information. The smoothed one could be transformed again with the Haar wavelets again, producing two time series of length $N/4$, with smoothed and detail information, and so on, keeping the details and doing an additional transform of the smoothed time series each time. If the original time series was some power of 2, $N = 2^n$, then this process, called a pyramid algorithm, would terminate when the last two time series were the coefficients of the time mean and the difference between the mean of the first half of the time series and the last half of the time series. The number of coefficients at the end would total N, and would contain all of the information in the original time series, organized according to scale and location, as defined by the Haar wavelet family. The original fine wavelet weights of (a,a) and (a,-a) on an interval of two time points are stretched, or dilated in factors of 2 to create a sequence of wavelets with increasingly large scale, culminating in the mother and father wavelets that span the entire time series.

Let's suppose we start with a time series of 8 data $x_n \;\; n = 1, 8$, and perform successive Haar transforms on this time series. The resulting Haar transformed vector, y_{jk} represents the k time steps corresponding to each of j levels of detail. The diagram below is intended to give some idea of how the original data vector would be transformed into a representation vector in Haar function amplitudes using the pyramid scheme. In this representation, the first 4 values are the amplitudes of the first level of detail, defined at 4 time locations. The next two values represent the wavelet transform of the smoothed data set, which has 4 smoothed values and results in two wavelet coefficients, y_{21} and y_{22} The last two values in the wavelet vector are the mother wavelet y_{31} and the father wavelet y_{32}. Because the Haar wavelet transform is orthogonal, the original time series can be reconstructed from the wavelet coefficient vector y_{jk}.

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \end{bmatrix} \Rightarrow \begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{bmatrix} \tag{9.17}$$

Let's consider the specific example of a sine wave with wavelength of 8 time steps, of which we have a total of $2^6 = 64$ data points. Figure 9.2 shows the time series and its Haar transform. The Haar transform is organized with the father and mother wavelet amplitudes on the left and the greatest level of detail in the

32 positions on the right of the transform vector. Note how the Haar coefficient is constant and large for the third level of detail, which corresponds to a period of 8 time units. It is fortuitous that the Haar wavelet of period 8 projects exactly onto the period and phase of the sine wave. If a cosine wave had been chosen, then this would not be the case and the amplitude would be spread over more Haar wavelets. The Haar wavelet, or any orthogonal wavelet, has very poor frequency resolution, as the frequency changes by a factor of two with each change in level. The coefficients for levels higher than the third level of detail are zero, since they have periods of 16, 32 and 64 time steps, and so do not project onto the wave of period 8. Similarly the mean is zero, so the father wavelet (a constant value) is zero.

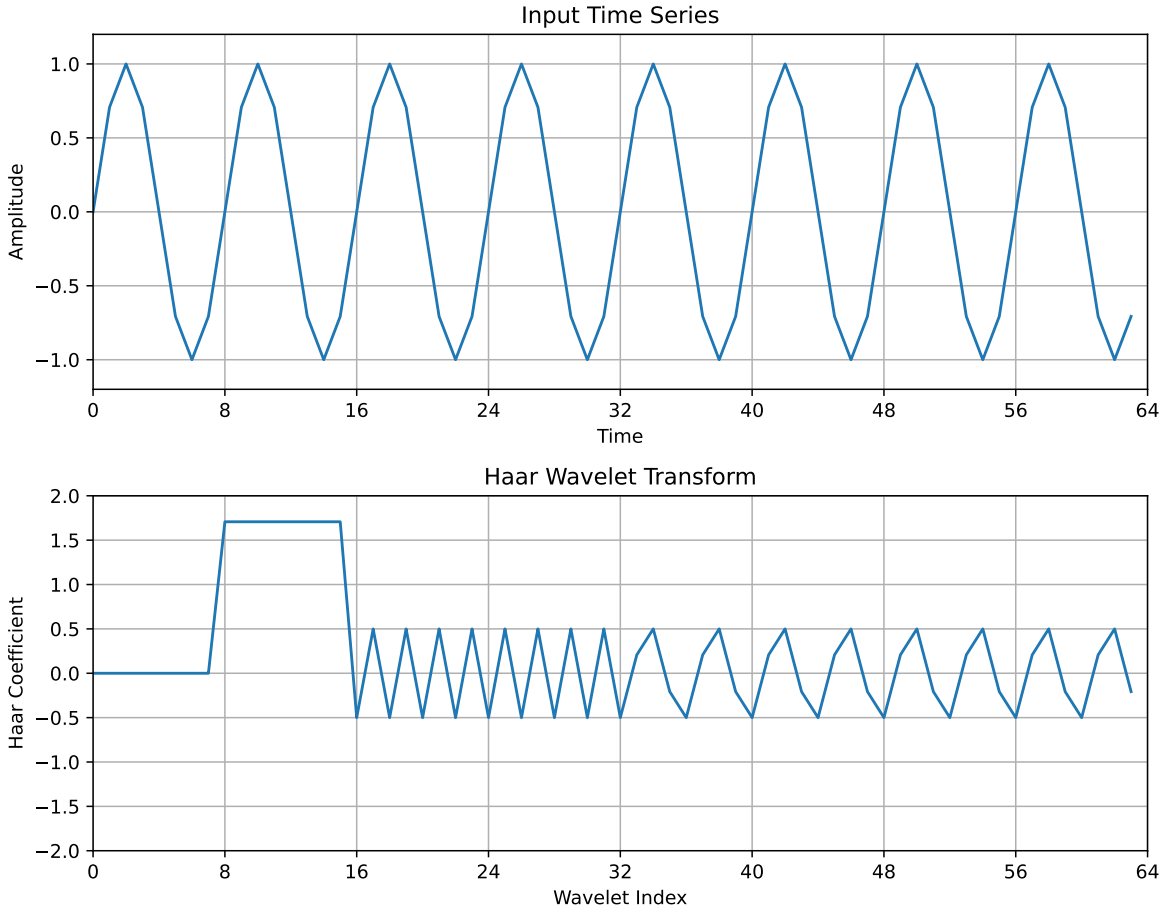

Figure 9.2 Time series of a sine wave with a period of 8 time units (top) and the Haar wavelet transform of the time series (bottom).

9.6 Daubechies Wavelet Filter Coefficients

In seeking other possible basis function sets on which we would like to expand we consider the following desirable characteristics:

(1) Good localization in both time and frequency (these conflict so we must compromise) (2) Simplicity, and ease of construction and characterization (3) Invariance under certain elementary operations such as translation (4) Smoothness, continuity and differentiability (5) Good moment properties, zero moments up to some order.

From the example of the Haar wavelet, we can see that a wavelet transform is equivalent to a filtering process with two filters that are quadrature mirror filters and divide the time series into a wavelet part, which represents the detail, and another smoothed part. Daubechies (1988, 1992) discovered an important and useful class of such filter coefficients. The simplest set has only 4 coefficients (DB2), and will serve as a useful illustration. Consider the following transformation matrix acting on a data vector to its right.

$$\begin{bmatrix} c_0 & c_1 & c_2 & c_3 & & & & & & & \\ c_3 & -c_2 & c_1 & -c_0 & & & & & & & \\ & & c_0 & c_1 & c_2 & c_3 & \cdots & & & & \\ & & c_3 & -c_2 & c_1 & -c_0 & \cdots & & & & \\ & & & & & & \cdots & c_0 & c_1 & c_2 & c_3 \\ & & & & & & \cdots & c_3 & -c_2 & c_1 & -c_0 \\ c_2 & c_3 & & & & & \cdots & & & c_0 & c_1 \\ c_1 & -c_0 & & & & & \cdots & & & c_3 & -c_2 \end{bmatrix} \tag{9.18}$$

Here we are only showing only the top two rows, the bottom two rows, and a subset of the columns. The blank spaces are occupied by zeros. The matrix is arranged in such a way that cyclic continuity of the data is assumed, much as in Fourier Analysis. Other options are possible. Dots represent where the matrix should be continued. The action of this matrix is to perform two convolutions with different, but related, filters, $[c_0, c_1, c_2, c_3]$ is the scaling filter and smooths the input if all the coefficients are positive and $[c_3, -c_2, c_1, -c_0]$ is the wavelet filter. These coefficients have been chosen such that the inner product of the smoothing and wavelet coefficients is zero, so that the two filters are orthogonal mirror filters. The pyramid algorithm can be applied, as with the Haar filter, so that successive levels of wavelet data are retained. We still have 4 unknown coefficients that we can solve for by using an approximation condition of two, and also requiring that the matrix be orthonormal. This matrix is called Daubechies-2 or DB2 because its approximation condition is 2. To ensure that it has approximation condition 2, we want to choose the coefficients of the wavelet so that their mean and trend are zero.

$$\begin{aligned} c_3 - c_2 + c_1 - c_0 &= 0 \\ 0c_3 - 1c_2 + 2c_1 - 3c_0 &= 0 \end{aligned} \tag{9.19}$$

For the transformation of the data vector to be useful, one must be able to reconstruct the original data from its smooth and detail components. This can be assured by requiring that the matrix (9.18) is orthogonal, so that its inverse is just its transpose. In discrete space, this is the equivalent of the orthogonality condition for continuous functions. The orthogonality condition places two additional constraints on the coefficients, which can be derived by multiplying (9.18) by its transpose and requiring that the product be the unit matrix. This yields two additional conditions on the coefficients, so that we now know them uniquely.

$$\begin{aligned} c_0^2 + c_1^2 + c_2^2 + c_3^2 &= 1 \\ c_3 c_1 + c_2 c_0 &= 0 \end{aligned} \tag{9.20}$$

These four equations for the coefficients (9.19 and 9.20) have a unique solution up to a left-right reversal. DB2 is only the simplest of a family of wavelet sets with the number of coefficients increasing by two each time (2, 4, 6, 8, 12, . . .). Note that the Haar wavelet would be DB1 in this family of wavelets. Each time we add two more coefficients we add an additional orthogonality constraint and raise the number of zero

moments, or the approximation condition order, by one. Daubechies (1988) has tabulated the coefficients for lots of these, and they are available in most wavelet software packages.

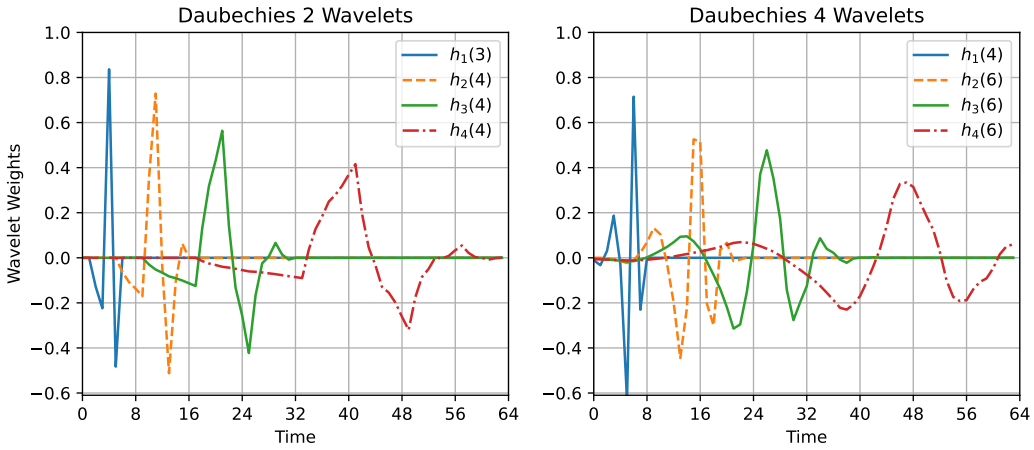

Figure 9.3 Selected examples of the Daubechies-2 and Daubechies-4 wavelets. The subscript indicates the level of approximation and the number in parentheses is the position in time. For example, only the third of the 32 level 1 Daubechies-2 wavelets is shown.

Figure 9.3 shows examples of some of the Daubechies-2 and Daubechies-4 wavelets. Note that as the number of weights is increased, the wavelets become smoother. Edge effects become increasingly important as the number of weights is increased, since the span of the longer wavelets becomes great.

9.7 Continuous, Non-orthogonal Wavelets

The frequency resolution with orthogonal wavelets is constrained to be coarse, so we may wish to use non-orthogonal wavelets in which we vary the wavelength and position of the wavelet more continuously. Some relatively famous wavelets are the Mexican Hat,

$$\psi_\sigma(t) = \frac{2}{\sqrt{3\sigma}\,\pi^{1/4}}\left(1-\left(\frac{t}{\sigma}\right)^2\right) e^{-\frac{t^2}{2\,\sigma^2}} \tag{9.21}$$

and the Complex Morlet Wavelet,

$$\begin{aligned}\psi_\sigma(t) &= c_\sigma \pi^{-1/4}\, e^{-\frac{1}{2}t^2}\left(e^{i\sigma t} - e^{-\frac{1}{2}\,\sigma^2}\right)\\ c_\sigma &= \left(1 + e^{-\sigma^2} - 2e^{-\frac{3}{4}\,\sigma^2}\right)\end{aligned} \tag{9.22}$$

Their structures are shown in Figure 9.4. The effect of the scale parameter σ on the Mexican Hat wavelet is shown in th left panel. Since the imaginary part of the Morelet wavelet is phase-shifted relative to its center location, in visual representations of data, only the real part is shown. the scale and location of the wavelet is varied to provide a representation in time-frequency space, as shown in the example below.

Figure 9.5 illustrates the use of a Morlet wavelet representation in frequency and time for the time series of Benthic $\delta 18O$ constructed by Lisiecki and Raymo (2005). Ocean water gets heavier in $\delta^{18}O$ as the lighter isotope is preferentially stored in ice sheets during ice ages. So the increase shows the growing global ice volume. As the ice volume gets larger it oscillates in time between glacial maxima with high $\delta^{18}O$ and interglacials with lower values. From the time-frequency separation in the plot, one can see that about 2.5 million years ago an oscillation with a period of about 40,000 years (2.5 cycles per 100kyr) begins to

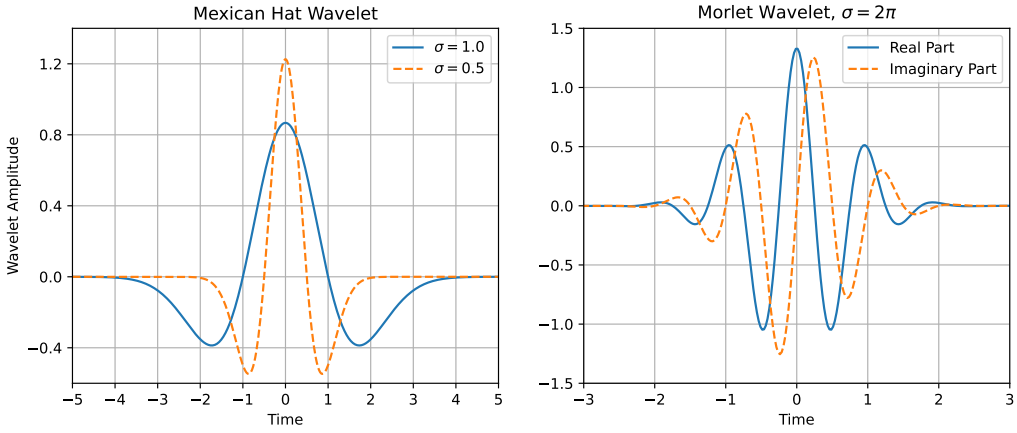

Figure 9.4 Examples of the Mexican Hat and Morlet continuous wavelets.

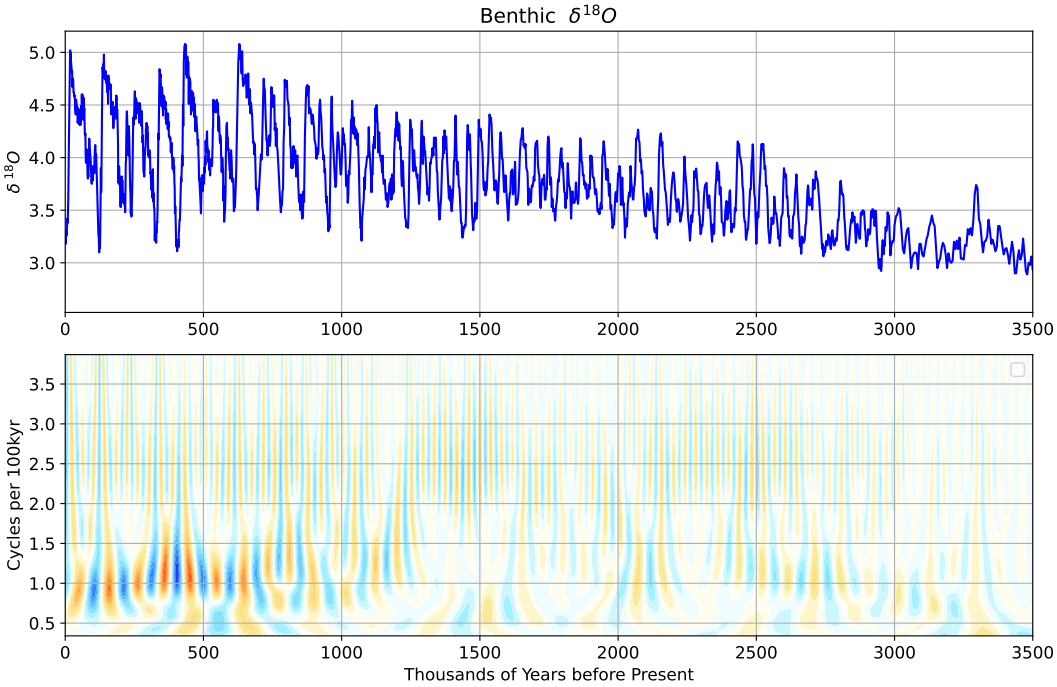

Figure 9.5 Time series of δ^{18}O from ocean sediment cores for the past 3.5 million years (top) and the Morlet wavelet transform plotted as a function of frequency and time.

occur intermittently. Later at about 1 million years ago a strong oscillation with a period of about 100,000 years begins and continues until the present. Analysis with wavelets reveals the episodic nature of these oscillations is interesting and would not be revealed by power spectral analysis, which discards all time location information in favor of maximal frequency resolution.

Dennis: *A bit more work to do here? What is missing that woud help?*

Chapter 10
Appendix A

10.1 Coherence Probability Table

Experimental coherence required to reject a null hypothesis of zero coherence at various probability levels for various degrees of freedom, n.

DoF=n	50%	90%	95%	99%	99.9%
2	.500	.901	.951	.990	.998
3	.293	.684	.776	.901	.968
4	.206	.539	.632	.785	.901
5	.159	.437	.527	.684	.823
6	.130	.370	.450	.602	.748
7	.109	.319	.393	.536	.684
8	.094	.280	.348	.482	.627
9	.083	.250	.312	.438	.578
10	.074	.226	.283	.401	.536
11	.067	.206	.259	.370	.500
12	.061	.189	.238	.342	.466
13	.056	.175	.221	.319	.441
14	.052	.162	.206	.298	.412
15	.048	.151	.193	.280	.389
16	.045	.142	.181	.264	.370
17	.042	.134	.171	.250	.350
18	.040	.127	.162	.237	.334
19	.038	.120	.154	.226	.319
20	.036	.112	.146	.215	.305
25	.029	.091	.118	.175	.250
30	.024	.076	.098	.147	.212
35	.020	.066	.084	.127	.185
40	.018	.057	.074	.112	.162
45	.016	.051	.066	.100	.145
50	.014	.046	.060	.090	.132
60	.012	.038	.050	.075	.111
70	.010	.033	.042	.065	.096
80	.009	.029	.037	.057	.084
90	.008	.026	.033	.052	.075
100	.007	.023	.030	.045	.068
125	.006	.018	.024	.036	.054
150	.005	.015	.020	.031	.045
175	.004	.013	.017	.026	.039
200	.003	.011	.015	.023	.034

Chapter 11
Appendix B

11.1 Matrix Algebra

$$(\underset{n\times p}{\mathbf{A}}\ \underset{p\times n}{\mathbf{B}})^{\mathsf{T}} = \underset{p\times n}{\mathbf{B}^{\mathsf{T}}}\ \underset{n\times p}{\mathbf{A}^{\mathsf{T}}} \tag{11.1}$$

References

Amos, D. and L. Koopmans, 1963: Tables of the distribution of the coefficient of coherence for stationary bivariate gaussian processes. Report SCR 483, Sandia Corporation Monograph.

Bartlett, M., 1935: Some aspects of the time-correlation problem in regard to tests of significance. *J. Roy. Stat. Soc.*, **98**, 536–543.

Bretherton, C., C. Smith, and J. Wallace, 1992: An intercomparison of methods for finding coupled patterns in climate data sets. *J. Climate*, **5**, 541–560.

Bretherton, C. S., M. Widmann, V. P. Dymnikov, J. M. Wallace, and I. Blade, 1999: The effective number of spatial degrees of freedom of a time-varying field. *J. Climate*, **12 (JUL 1999)**, 1990–2009.

Coles, S., 2001: *A Introduction to Statistical Modeling of Extreme Values*. Springer Seris in Statistics, Springer Verlag London Limited.

Cressman, G. P., 1959: AN OPERATIONAL OBJECTIVE ANALYSIS SYSTEM. *Monthly Weather Review*, **87 (10)**, 367–374, doi:10.1175/1520-0493(1959)087<0367:AOOAS>2.0.CO;2.

Daubechies, I., 1988: Wavelets and quadrature filters. *Comm. Pure Appl. Math.*, **41**, 909–996.

Daubechies, I., 1992: *Ten Lectures on Wavelets,*. SIAM, 357 pp.

Dee, D. P., et al., 2011: The era-interim reanalysis: configuration and performance of the data assimilation system. *Quart. J. Royal Met. Soc.*, **137 (656)**, 553–597, doi:10.1002/qj.828, URL <Go to ISI>://WOS:000290450900001.

Flattery, T., 1971: Spectral models for global analysis and forecasting. Proceedings AWS Tech. Exch. Conf., Sept. 1970, U.S. Naval Academy, 42-54 pp.

Fourier, J. B. J., 1822: Théorie analytique de la chaleur. *Mémoires de l'Académie Royale des Sciences de l'Institut de France*, **10**.

Geisler, J. E. and R. E. Dickinson, 1976: The five-day wave on a sphere with realistic zonal winds. *J. Atmos. Sci.*, **33 (4)**, 632–641, doi:10.1175/1520-0469(1976)033<0632:tfdwoa>2.0.co;2.

Gilman, D., F. Fuglister, and J. Mitchell, J.M., 1963: On the power spectrum of "red noise". *J. Atmos. Sci.*, **20**, 182–184.

Goodman, N., 1957: On the joint estimation of the spectra, cospectrum and quadrature spectrum of a two-dimensional stationary gaussian process. Report Sci. Pap. 10, New York University, Engineering Statistics Lab.

Goodman, S. N., 2001: Of P-values and Bayes: a modest proposal. *Epidemiology*, **12 (3)**, 295–297.

Haar, A., 1910: Zur theorie der orthogonalen funktionensysteme. *Mathematische Annalen*, **69**, 331–371, doi:doi:10.1007/BF01456326.

Harris, F. J., 1978: On the use of windows for harmonic analysis with the discrete fourier transform. *Proceedings of the IEEE*, **66 (1)**, 51–83, doi:10.1109/PROC.1978.10837.

Hartmann, D. L., 1974: Time spectral analysis of mid-latitude disturbances. *Mon. Wea. Rev.*, **102**, 348–362.

Hayashi, Y., 1971: A generalized method of resolving disturbances into progressive and retrogressive waves by space fourier and time cross-spectral analyses. *J. Meteor. Soc. Japan*, **49 (2)**, 125–128.

Hayashi, Y., 1977: On the coherence between progressive and retrogressive waves and a partition of space-time power spectra into standing and traveling parts. *J. Appl. Meteor.*, **16 (April)**, 368–373, doi:10.1175/1520-0450(1977)016<0368:OTCBPA>2.0.CO;2.

Hayashi, Y., 1979: A generalized method of resolving transient disturbances into standing and travelling waves by space-time spectral analysis. *J. Atmos. Sci.*, **36**, 1017–1029.

Hendon, H. H. and M. C. Wheeler, 2008: Some space-time spectral analyses of tropical convection and planetary-scale waves. *J. Atmos. Sci.*, **65 (9)**, 2936–2948, doi:10.1175/2008jas2675.1, URL <Go to ISI>://000259378800011.

Holton, J. and G. Hakim, 2012: *An Introduction to Dynamical Meteorology (Fifth Edition)*. 3d ed., Academic Press, San Diego, 533 pp.

Leith, C., 1973: The standard error of time-averaged estimates of climatic means. *J. Appl. Meteorol.*, **12**, 1066–1069.

Liebmann, B. and C. A. Smith, 1996: Description of a complete (interpolated) outgoing longwave radiation dataset. *Bull. Amer. Meteorol. Soc.*, **77 (6)**, 1275–1277.

Lisiecki, L. E. and M. E. Raymo, 2005: A pliocene-pleistocene stack of 57 globally distributed benthic delta o-18 records. *Paleoceanography*, **20 (1)**, doi: 10.1029/2004pa001 071, doi:10.1029/2004pa001071, URL <Go to ISI>://WOS:000226581600001.

Livezey, R. E. and W. Y. Chen, 1983: Statistical field significance and its determination by monte carlo techniques. *Mon. Weather Rev.*, **111 (1)**, 46–59.

Madden, R. A. and P. R. Julian, 1971: Detection of a 40-50 day oscillation in the zonal wind in the tropical pacific. *J. Atmos. Sci.*, **28 (5)**, 702–708, doi:10.1175/1520-0469(1971)028<0702:doadoi>2.0.co;2, URL <Go to ISI>://WOS:A1971J916100002.

Marple, S. L. J., 1987: *Digital Spectral Analysis with Applications*. Prentice-Hall Signal Processing Series, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 492 pp.

Matsuno, T., 1966: Quasi-geostrophic motions in the equatorial area. *J. Meteor. Soc. Japan*, **44**, 25–43.

North, G., Bell, R. Cahalan, and F. Moeng, 1982: Sampling errors in the estimation of empirical orthogonal functions. *Mon. Wea. Rev.*, **110**, 699–706.

Nuzzo, R., 2014: Scientific method: statistical errors. *Nature*, **506 (7487)**, 150–152, doi:10.1038/506150a.

Percival, D. and A. Walden, 1993: *Spectral Analysis for Physical Applications*. Cambridge University Press, Cambridge, U.K., 583 pp.

Pratt, R., 1976: The interpretation of space-time spectral quantities. *J. Atmos. Sci.*, **33**, 1060–1066.

Prohaska, J., 1976: A technique for analyzing the linear relationships between two meteorological fields. *Mon. Wea. Rev.*, **104**, 1345–1353.

Schäfer, J., 1979: A space-time analysis of tropospheric planetary waves in the northern hemisphere. *J. Atmos. Sci.*, **36 (6)**, 1117–1123.

Taylor, G., 1921: Diffusion by continuous movement. *Proc. London Math. Soc.*, **21 (2)**, 196–212.

Thomson, D., 1982: Spectrum estimation and harmonic analysis. *IEEE Proc.*, **70**, 1055–1096.

Wallace, J. and V. Kousky, 1968: Observational evidence of kelvin waves in the tropical stratosphere. *J. Atmos. Sci.*, **25**, 900–907.

Wallace, J. M., C. Smith, and C. S. Bretherton, 1992: Singular value decomposition of wintertime sea-surface temperature and 500-mb height anomalies. *J. Climate*, **5 (6)**, 561–576, URL <Go to ISI>://A1992HX56300002.

Watt-Meyer, O. and P. J. Kushner, 2015: Decomposition of atmospheric disturbances into standing and traveling components, with application to northern hemisphere planetary waves and stratosphere–troposphere coupling. *J. Atmos. Sci.*, **72 (2)**, 787–802, doi:10.1175/jas-d-14-0214.1, URL http://journals.ametsoc.org/doi/abs/10.1175/JAS-D-14-0214.1.

Wheeler, M. and G. N. Kiladis, 1999: Convectively coupled equatorial waves: analysis of clouds and temperature in the wavenumber-frequency domain. *J. Atmos. Sci.*, **56 (3)**, 374–99.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. Academic Press, 704 pp.

Wilks, D. S., 2016: "the stippling shows statistically significant grid points": How research results are routinely overstated and overinterpreted, and what to do about it. *Bulletin of the American Meteorological Society*, **97 (12)**, 2263–2273, doi:10.1175/BAMS-D-15-00267.1.

Yanai, M., T. Maruyama, T. Nitta, and Y. Hayashi, 1968: Power spectra of large-scale disturbances over the tropical pacific. *J. Meteor. Soc. Japan*, **46 (4)**.