1 Random Variables and Probability

A *random*, or *stochastic*, process results in outcomes that cannot be predicted precisely. The outcome of a random process, a *random variable*, is described by its *probability* of occurrence. Probabilities range from 0, no chance, to 1, a certainty.

There are different interpretations of probability, notably the $frequentist$ and *Bayesian* views. Frequentists consider a set of repeated experiments or trials with probability expressed by the frequency of occurrence, i.e., if A occurs n_A times out of n experiments, then the probability of A is

$$Pr(A) = n_A/n. \tag{1}$$

In the frequentist view, *Pr(A)* is assumed to approach the true probability of A as $n \to \infty$. In the Bayesian approach, existing knowledge is used to assign probability beforehand, the *prior probability*, which is updated to a *posterior probability* based on the data, or *evidence*, and the application of *Bayes' theorem*. Here we will examine frequentist methodologies (e.g., confidence intervals, hypothesis testing) commonly used in the analysis of oceanographic data.

There are two related functions that assign probabilities to a random variable. The *cumulative distribution function* (*CDF*) specifies the probability that the random variable, X, is less than or equal to a specific value x,

$$F_X(x) = Pr(X \leq x). \tag{2}$$

The derivative of the *CDF* is the *probability density function* (*PDF*)

$$f_X(x) = \frac{dF_X}{dx}. \tag{3}$$

It follows that

$$Pr(a < X \leq b) = \int_a^b f(x)dx = F_X(b) - F_X(a). \tag{4}$$

Given the *PDF* of $f_X(x)$, we can compute the PDF of a random variable that is a function of X, $Y = g(X)$, provided that g is invertible, $X = g^{-1}(Y)$. The *PDF* for Y is

$$f_Y(y) = \sum_i \frac{f_X(x_i)}{|\frac{dg}{dx}|_{x_i}}, \tag{5}$$

where the x_i's are the solutions to $y = g(x_i)$.

2 Expected Value and Moments

The mean or first moment of a random variable X is

$$E(X) = \int_{-\infty}^{\infty} x f_X(x) dx = \mu, \tag{6}$$

where the *expected value*, E, of any real single-valued continuous function $g(X)$ of the random variable X is

$$E(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x) dx. \tag{7}$$

The variance or *second central moment* of X is

$$Var(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x) dx = \sigma^2, \tag{8}$$

where σ is the *standard deviation*. In general the *rth central moment* is given by

$$E((X-\mu)^r) = \int_{-\infty}^{\infty} (x-\mu)^r f_X(x) dx. \tag{9}$$

Statistical moments are comparable to the moments of a solid body. The *PDF* is analogous to the density of the body, the mean to the center of mass, and the variance to the moment of inertia.

The *skewness* of a distribution is the third *standardized moment*,

$$\gamma_1 = E\left(\left(\frac{X-\mu}{\sigma}\right)^3\right), \tag{10}$$

which measures the asymmetry of the distribution about the mean. For a unimodal distribution, a negative skewness implies that the left tail of the *PDF* is more pronounced than the right; a positive skewness has a more pronounced right tail. The *kurtosis*, or fourth *standardized moment*, measures the peakedness of a unimodal distribution.

3 The Normal and Related Distributions

A probability function with broad application is the *normal* or *Gaussian* distribution. The normal *PDF* is a function of the mean and variance,

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right). \tag{11}$$

The notation $X \sim N(\mu, \sigma^2)$ is used to indicate that X is normally distributed.

It follows from eq.(11) that the normal distribution is symmetric about the mean, all odd moments are zero, and that the mean = mode = median. Further, if

i) $X \sim N(\mu, \sigma^2)$ and a and b are real numbers, then $aX + b \sim N(a\mu + b, (a\sigma)^2)$,

ii) $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ are independent random variables, then $X \pm Y \sim N(\mu_X \pm \mu_Y, \sigma_X^2 + \sigma_Y^2)$.

iii) $X \sim N(\mu, \sigma^2)$, then a *normalized random variable* $Z = \frac{X-\mu}{\sigma}$ can be defined that has a *standard normal distribution* $Z \sim N(0, 1)$.

A number of other frequently used distributions follow from the normal distribution. Consider a set of random variables X_1, X_2, ..., X_n that are independent and identically distributed as $N(0, 1)$, and set Y equal to the sum of the squares,

$$Y = X_1^2 + X_2^2 + ... + X_n^2.$$

Then Y is considered to have a *chi-squared distribution* with n *degrees of freedom*, denoted as $Y \sim \chi^2(n)$. The *degrees of freedom* measure the number of independent pieces of information in an estimate.

If X and Y are independent random variables and $X \sim N(0, 1)$ and $Y \sim \chi^2(n)$, then

$$Z = \frac{X}{\sqrt{Y/n}}$$

has a *t distribution* with n degrees of freedom, denoted as $Z \sim t(n)$.

If Y_1 and Y_2 are independent random variables with $Y_1 \sim \chi^2(n_1)$ and $Y_2 \sim \chi^2(n_2)$, then

$$Z = \frac{Y_1/n_1}{Y_2/n_2}$$

has a *Fdistribution* with n_1 and n_2 degrees of freedom, denoted by $Z \sim F(n_1, n_2)$.

4 Central Limit Theorem

The *Central Limit Theorem* explains why the normal distribution applies to a wide range of random variables, and it provides a guideline for assessing when a normal distribution is likely to apply.

Let X_1, X_2, ..., X_n be n independent, identically distributed random variables, with the mean and variance of each random variable X_i given by μ_i and σ_i^2. Define a new random variable which is a linear combination of the X_i's,

$$Y = \sum_{i=1}^{n} a_i X_i, \tag{12}$$

where the a_i are arbitrary constants. The mean and variance of Y are

$$\mu_Y = E\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i E(X_i) = \sum_{i=1}^{n} a_i \mu_i, \tag{13}$$

$$\sigma_Y^2 = E\left((Y-\mu_Y)^2\right) = E\left(\sum_{i=1}^{n} a_i (X_i - \mu_i)^2\right) = \sum_{i=1}^{n} a_i^2 \sigma_i^2. \tag{14}$$

The Central Limit Theorem states that as $n \to \infty$, $Y \sim N(\mu_Y, \sigma_Y^2)$. In practice, the distribution of Y tends to be close to normal by $n \approx 30$. Thus, any variable that is the sum of other variables through averaging, integration, etc. tends to be normally distributed, regardless of the distribution of the summed variables.

5 Example: Distribution Functions and Ocean Waves

Ocean surface waves provide an intuitive example of how distribution functions can be used to describe various aspects of a random process. A measure of ocean surface waves is the water surface elevation (η) about the mean still water line ($z = 0$). The wave height (H) can be defined as the distance between the highest (crest) and lowest (trough) elevations over the course of one wavelength (λ) or wave period (T).

Consider a time series of n waves measured at a fixed position. A common overall measure of the size of the wave field is the significant wave height ($\overline{H}_{1/3}$), defined as the average of the 1/3 largest waves [*Sverdrup and Munk*, 1947]. The choice of 1/3 is arbitrary but has remained in use because $\overline{H}_{1/3}$ is comparable to what an observer at sea would estimate as the typical height of a wave field. In general, $\overline{H}_p$ is the average of the highest pn waves in a record, with $p \leq 1$. A related measure of a wave record of n waves is the root-mean-square (rms) wave height,

$$H_{rms} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} H_i^2}. \tag{15}$$

Let's treat the wave field as the sum of random waves generated from multiple sources. Based on the Central Limit Theorem, it follows that $\eta \sim N(0, \sigma_\eta^2)$, where σ_η is the standard

deviation of the surface elevation. For narrow-banded seas, or a superposition of waves within a narrow range of periods, *Longuet-Higgins* [1952] showed that the PDF of H is given by the *Rayleigh distribution*.

$$f(H) = \frac{2H}{H_{rms}^2} exp\left(\frac{-H^2}{H_{rms}^2}\right), \tag{16}$$

where

$$H_{rms}^2 = 8\sigma_\eta^2. \tag{17}$$

The *CDF* for wave height is then

$$F(H) = \int_0^H f(H')dH' = 1 - exp\left(\frac{-H^2}{H_{rms}^2}\right). \tag{18}$$

Given the PDF of H, we can compute the most probable wave

$$\frac{df(H)}{dH} = \frac{2}{H_{rms}^2} exp\left(\frac{-H}{H_{rms}^2}\right)\left(1 - \frac{2H^2}{H_{rms}^2}\right) = 0 \Rightarrow H_{mp} = \frac{H_{rms}}{\sqrt{2}} = 2\sigma_\eta. \tag{19}$$

and the mean wave height

$$\overline{H} = \int_0^\infty H f_H(H)dH = \frac{\sqrt{\pi}}{2} H_{rms} = 0.886 H_{rms}. \tag{20}$$

The significant wave height is

$$\overline{H}_{1/3} = \int_{H_{1/3}}^{-\infty} H f(H)dH = 1.416 H_{rms}. \tag{21}$$

Other characteristic wave heights include $\overline{H}_{1/10} = 1.800 H_{rms}$ and $\overline{H}_{1/100} = 2.359 H_{rms}$

The probability, Q , that H will exceed a threshold value H_Q is

$$Q(H_q) = Pr(H > H_Q) = 1 - F(H_Q) = exp\left(\frac{-H_Q^2}{H_{rms}^2}\right). \tag{22}$$

This can be also written as the value of H expected to be exceeded with probability Q,

$$H_Q = H_{rms}\sqrt{ln\left(\frac{1}{Q}\right)}. \tag{23}$$

For example, $Q = 1/3$, $H_{1/3} = 1.048 H_{rms}$.

Longuet-Higgins, M. S., 1955: On the statistical distribution of the heights of sea waves, *J. Mar. Res.*, 11, 245-266.

Sverdrup, H.U., and W. H. Munk, 1947: Wind, sea, and swell; theory of relations for forecasting. U. S. Navy Hydrographic Office, H. O., Publ. No. 601.

6 Joint and Conditional Distributions

The joint probability of occurrence of a pair of random variables, X and Y, is specified by the *joint cumulative distribution function*

$$F_{X,Y}(x,y) = Pr(X \leq x \ \& \ Y \leq y), \tag{24}$$

and the *joint probability density function*

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y}. \tag{25}$$

It follows that

$$Pr(X \leq a, Y \leq b) = F_{X,Y}(a,b) = \int_{-\infty}^{a} \int_{-\infty}^{b} f_{X,Y}(x,y) dy dx,$$

and

$$Pr(X \leq a, Y \text{any}) = F_{X,Y}(a,\infty) = \int_{-\infty}^{a} \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy dx.$$

The *conditional probability density* of Y given X is

$$f_Y(y|X=x) = \frac{f_{X,Y}(x,y)}{f_X(x)} \tag{26}$$

for $f_X(x) > 0$. When the occurrence of Y is not influenced by X, then

$$f_Y(y|X=x) = f_Y(y), \tag{27}$$

and

$$f_{X,Y}(x,y) = f_X(x) f_Y(y), \tag{28}$$

in which case X and Y are *independent*.

7 Covariance and Correlation

The degree to which two random variables (X and Y) vary together is measured by the *covariance*

$$C_{XY} = E\left((X-\mu_X)(Y-\mu_Y)\right) = E(XY) - \mu_X \mu_Y \tag{29}$$

where $\mu_x = E(X)$ and $\mu_y = E(Y)$. If X and Y are *independent*, then $C_{XY} = 0$ ($C_{XY} = 0$ does not necessarily imply independence).

The *correlation* is defined as the covariance normalized by the standard deviations of each of the variables,

$$\rho_{XY} = \frac{C_{XY}}{\sigma_X \sigma_Y}. \tag{30}$$

The correlation ranges from -1 to 1.

8 Joint Normal Distribution

Two random variables X_1 and X_2 are *joint normally distributed* if their sum $aX_1 + bX_2$ is normally distributed for all a and b. The joint normal PDF is

$$f_{XY}(x,y) = Aexp\left(-\frac{1}{2(1-\rho_{XY}^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\rho_{XY}\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right) \tag{31}$$

where

$$A = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{XY}^2}}, \tag{32}$$

and ρ_{XY} is the correlation coefficient. When X and Y are independent (i.e., $\rho_{XY} = 0$), $f_{XY} = f_X f_Y$.

9 Estimators

Statistical estimation involves the use of a model, or *estimator*, to predict a desired parameter or set of parameters from available data. For example, an estimate of the expected value, $\mu_X = E(X)$, is the *sample mean* given by

$$\hat{\mu}_X = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \tag{33}$$

with the caret symbol used to indicate an estimate. The sample mean varies with the sample. Thus, an estimate is itself a random variable subject to its own probability distribution, known as the *sampling distribution*.

Estimators are evaluated by how well the sample estimate represents a so-called *true value*, which might be the value obtained from all available data in a population, or from an infinite set of identically prepared experiments. Desirable properties of an estimator include:

i) on average, the estimate of some parameter, $\hat{\phi}$, should be equal to the true value, ϕ, or in other words that the estimator is *unbiased*

$$B = E(\hat{\phi} - \phi) = 0. \tag{34}$$

Otherwise B is the *bias* of the estimate.

ii) The estimator should be *efficient* in the sense that it yields a small *mean square error* (*MSE*), measured as the variance of the estimate about the true value,

$$E\left((\hat{\phi} - \phi)^2\right) = var(\hat{\phi}) + B^2. \tag{35}$$

iii) The estimator should also be *consistent*, such that $\hat{\phi} \to \phi$ as $n \to \infty$.

For the case of the sample mean (eq. 33),

$$E(\overline{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \mu_X.$$

so the estimate is unbiased. The *MSE* is

$$E\left((\overline{X} - \mu_X)^2\right) = var(\overline{X}) + B^2,$$

but the estimate is unbiased, $B^2 = 0$, so

$$MSE = var(\overline{X}) = E\left(\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu_X\right)^2\right).$$

If the X_i's are independent, then

$$var(\overline{X}) = \frac{1}{n^2}\sum_{i=1}^{n} E\left((X_i - \mu_X)^2\right) = \frac{\sigma_X^2}{n}.$$

The standard deviation of $\overline{X}$ is $\sigma_X/\sqrt{n}$, which is also referred to as the *standard error* of the sample mean.

We can evaluate the efficiency of the sample mean versus other estimators of the mean. For example, the sample mean is a more efficient estimator than the median, which has a 56% higher *MSE*. Finally, because the sample mean $MSE \to 0$ as $n \to \infty$, the sample mean is a consistent estimate.

For the *sample variance*, let's consider a more general estimator of the form

$$s_k^2 = \frac{1}{k}\sum_{i=1}^{n}(X_i - \overline{X})^2,$$

and try to select k to optimize our estimate. The bias for this estimator is

$$B = E(s_k^2) - \sigma_X^2 \ = \left(\frac{n-1-k}{k}\right)\sigma_X^2.$$

An unbiased estimator is obtained for $k = n - 1$. Selecting $k = n$ leads to $B = -\sigma_X^2/n$. The *MSE* is

$$E = \left(\left(s_k^2 - \sigma_X^2\right)^2\right) = var(s_k^2) + B^2 = \frac{2(n-1)}{k^2}\sigma_X^4 + \frac{(n-1-k)^2}{k^2}\sigma_X^4.$$

The minimum *MSE* occurs for $k = n + 1$. This illustrates that in some instances the best estimator will depend on the properties to be optimized. The *sample variance* usually is specified as $k = n - 1$.

10 Confidence Intervals

A probabilistic measure of an estimator is given by the *confidence interval*. Confidence intervals are derived by defining a function of $\hat{\phi}$ and ϕ that has a known sampling distribution. If the probability distribution off $\varphi = g(\hat{\phi}, \phi)$ is known, then a probability statement can be constructed such that

$$Pr\left[\varphi_L < \varphi < \varphi_U\right] = 1 - \alpha. \tag{36}$$

where $0 < \alpha < 1$. This can be rewritten as

$$Pr\left[\phi_L < \phi < \phi_U\right] = 1 - \alpha, \tag{37}$$

where ϕ_L and ϕ_U are functions of φ_L, φ_U, and $\hat{\phi}$. The interval between ϕ_L and ϕ_U is the 100(1-α)% confidence interval for ϕ.

For example, define Z such that

$$Z = \frac{\hat{\phi} - \phi}{\sigma_\phi} \sim N(0, 1), \tag{38}$$

where $E(\hat{\phi}) = \phi$ (i.e., an unbiased estimate), $var(\hat{\phi}) = \sigma_\phi^2$, and Z is a standard Normal random variable. We can construct confidence intervals for Z as

$$Pr[Z_{\alpha/2} < Z < Z_{1-\alpha/2}] = 1 - \alpha. \tag{39}$$

Substituting for Z, noting that $Z_{\alpha/2} = -Z_{1-\alpha/2}$ for a standard Normal variable, and solving for ϕ gives

$$Pr[\hat{\phi} - Z_{1-\alpha/2}\sigma_\phi < \phi < \hat{\phi} + Z_{1-\alpha/2}\sigma_\theta] = 1 - \alpha. \tag{40}$$

We thus have defined a confidence interval around the estimate $\hat{\phi}$ within which we expect to find the true value ϕ with probability $1 - \alpha$.

Confidence intervals for μ_X when σ_X is known

Based on the *Central Limit Theorem*, we know that the sample mean ($\overline{X}$) approaches a Normal distribution for large n. Thus we can construct a standard Normal variable Z such that

$$Z = \frac{\overline{X} - \mu_X}{\sigma_X/\sqrt{n}}. \tag{41}$$

Recall that σ_X is the standard deviation of X, and $\sigma_X/\sqrt{n}$ is the standard deviation of $\overline{X}$. We then can construct a confidence interval for the true mean as

$$\overline{X} - Z_{1-\alpha/2}\frac{\sigma_X}{\sqrt{n}} < \mu_X < \overline{X} + Z_{1-\alpha/2}\frac{\sigma_X}{\sqrt{n}}. \tag{42}$$

Confidence intervals for μ_X when σ_X is unknown

Typically the true variance, σ_X, is not known. Confidence intervals for μ_X can be specified when X is Normally distributed. Using the sample variance (S^2) we construct a new random variable

$$t = \frac{\overline{X} - \mu_X}{S/\sqrt{n}}, \tag{43}$$

where

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2, \tag{44}$$

based on n independent X_i. W. Gosset, an employee of Guinness Breweries, published the solution for the *PDF* for t under the pseudonym *Student* in 1908, hence it became known as *Student's t-distribution*, or the *t-distribution*, with $\nu = n-1$ degrees of freedom. Recall that the random variable t has a Student's t-distribution if it is in the form

$$t = \frac{z}{\sqrt{y/n}}, \tag{45}$$

where $z \sim N(0,1)$ and y is chi-square distributed. It follows that the confidence interval for the true mean is

$$\overline{X} - t_{1-\alpha/2,\nu}\frac{S}{\sqrt{n}} < \mu_X < \overline{X} + t_{1-\alpha/2,\nu}\frac{S}{\sqrt{n}}. \tag{46}$$

Confidence intervals for σ_X^2 when μ_X is unknown

Confidence intervals for the variance are obtained by appealing to the *chi-square (χ^2) distribution*. The expected value and variance of χ^2 are $E(\chi^2) = n$ and $var(\chi^2) = 2n$. We first transform the sample variance into a χ^2 variable,

$$\frac{(n-1)S^2}{\sigma_X^2} = \sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\sigma_X}\right)^2 = \chi_\nu^2 \tag{47}$$

with $\nu = n-1$ degrees of freedom. The associated probability statement is

$$Pr\left(\chi^2_{\alpha/2,\nu} < \frac{(n-1)S^2}{\sigma_X^2} < \chi^2_{1-\alpha/2,\nu}\right) = 1-\alpha, \tag{48}$$

which yields the confidence interval for the true variance,

$$\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,\nu}} < \sigma_X^2 < \frac{(n-1)S^2}{\chi^2_{\alpha/2,\nu}}. \tag{49}$$

Note that the confidence intervals are asymmetric about S^2.

11 The correlation coefficient

The *correlation coefficient* provides a measure of the linear association between two random variables. The sample correlation coefficient between two random variable $X_i, Y_i, i = 1, 2, ..., n$ is

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{[\sum_{i=1}^{n}(X_i - \overline{X})^2]^{1/2}[\sum_{i=1}^{n}(Y_i - \overline{Y})^2]^{1/2}}. \tag{50}$$

The correlation coefficient provides a normalized measure of covariability such that $-1 \leq r \leq 1$.

Confidence intervals for the correlation coefficient can be constructed using *Fisher's z-transform*,

$$z = \frac{1}{2} ln \left(\frac{1+r}{1-r} \right). \tag{51}$$

If X and Y are n independent random variables with a joint Normal distribution, then

$$z \sim N \left(\frac{1}{2} ln \left(\frac{1+\rho}{1-\rho} \right), \frac{1}{\sqrt{n-3}} \right). \tag{52}$$

where ρ is the true correlation. Confidence intervals for ρ can be computed.

A more typical measure of the statistical significance of the correlation coefficient is based on a *hypothesis test*. A hypothesis test is used to assess whether an estimate is the result of random chance or not. The test is posed in terms of a *null hypothesis*, H_o, which typically is that the estimate is the result of random chance (e.g., the true correlation is zero). A sample distribution is needed to evaluate the null hypothesis, which is either accepted or rejected based on the *p-value*, defined as the probability of obtaining a value at least as high as the estimate given the null hypothesis. The p-value can be evaluated for the right or left tail of the sample distribution (one-sided), or both (two-sided). When the p-value falls below a specified significance level α, the null hypothesis is rejected at the $(1 - \alpha)100$ significance level. Rejection of H_o suggests that the complement of H_o, or the *alternative hypothesis*, may be true. A p-value $> \alpha$ does not necessarily mean that H_o is true, only that we cannot reject H_o given the available data. For the case of the correlation coefficient, the variable

$$t = r \sqrt{\frac{n-2}{1-r^2}} \tag{53}$$

is assumed to be t-distributed with $n - 2$ degrees of freedom.

12 Linear regression

Regression analysis is used in a variety of applications, for example to quantify relationships between two or more variables, to test causal hypotheses, to perform extrapolations and forecasts, and to identify trends in time series. A *linear regression* is an estimate of a *dependent* output variable y_i, i =1, 2, ..., n in terms of a linear superposition of j = 1, 2, ..., k input variables

$$\hat{y}_i = \sum_{j=1}^{k} b_j x_{ij}. \tag{54}$$

Coincident measurements of the input and output variables are required to obtain the b_j's.

A common regression method is *ordinary least squares* (*OLS*), which seeks to minimize the sum of squares of the residual error,

$$E^2 = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2. \tag{55}$$

The *OLS* solution is obtained by computing the b_j's that satisfy $\partial E^2/\partial b_j = 0$. The *OLS* estimate assumes that the error lies in the observed y_i. In general there are other cost functions besides E^2 that can be minimized, for example, the absolute value of the residual error or the diagonal deviation between the observations and the model function. There also are techniques to account for errors in the input variables.

To illustrate the *OLS* method, let's consider the simple case of a straight-line fit relating y_i to an input x_i plus a mean offset

$$\hat{y}_i = a + b x_i. \tag{56}$$

The residual sum of squares is

$$E^2 = \sum_{i=1}^{n} \left[y_i - (a + b x_i)\right]^2,$$

and the coefficient's that minimize the E^2 are the *slope*

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \tag{57}$$

sometimes referred to as the *regression coefficient*, and the *intercept*

$$\hat{a} = \overline{y} - \hat{b}\overline{x}. \tag{58}$$

The regression coefficient can be written as

$$\hat{b} = \frac{cov(x,y)}{\sigma_x^2}. \tag{59}$$

A measure of the quality of the fit is given by the *correlation coefficient*

$$\hat{r} = \frac{cov(x,y)}{\sigma_x \sigma_y}. \tag{60}$$

The correlation and regression coefficients are related by

$$\hat{b} = \frac{\sigma_y}{\sigma_x}\hat{r}. \tag{61}$$

The *regression coefficient* measures the change in y given a unit change in x. Scale changes in the data will alter the regression coefficient but not the correlation coefficient. The mean square error of the residual, or the unaccounted for variance, can be expressed as

$$MSE = \sigma_y^2(1 - \hat{r}^2). \tag{62}$$

The higher the correlation between x and y, the smaller the *MSE* of the *OLS* estimate.

Assuming that the residual error is *serially independent*, the standard deviations of the intercept and slope are

$$\hat{\sigma}_b = \frac{\hat{\sigma}_\epsilon}{\left[\sum_{i=1}^{n}(x_i - \overline{x})^2\right]^{1/2}},$$

$$\hat{\sigma}_a = \hat{\sigma}_\epsilon \left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right]^{1/2},$$

where

$$\hat{\sigma}_\epsilon = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\epsilon_i^2}.$$

is an unbiased estimate of the standard deviation of the residual error. Confidence intervals can be obtained by assuming that the fitted parameters and residual error are Normally distributed

$$t_\nu = \frac{\hat{b} - b}{\sigma_b},$$

$$t_\nu = \frac{\hat{a} - a}{\sigma_a},$$

where t_ν is a *Student's t* pdf with $\nu = n - 2$ degrees of freedom. The confidence interval for any specific value of $\hat{y}_i$ is given by

$$t_\nu = \frac{y_i - \hat{y}_i}{\sigma_\epsilon \left(\frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right)^{1/2}}.$$

13 Multiple linear regression

Extending *OLS* to multiple inputs yields

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & . & x_{1k} \\ x_{21} & x_{22} & . & x_{2k} \\ . & . & . & . \\ x_{n1} & x_{n2} & . & x_{nk} \end{bmatrix} \tag{63}$$

with our estimate given by $\hat{\mathbf{y}} = \mathbf{X} \cdot \mathbf{b}$, where

$$\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ . \\ . \\ . \\ b_k \end{bmatrix}. \tag{64}$$

Minimization of the residual sum of squares yields the *normal equations*

$$(\mathbf{X}'\mathbf{X})\,\mathbf{b} = \mathbf{X}'\mathbf{y}, \tag{65}$$

and the *OLS* estimate for $\mathbf{b}$ is

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{y}. \tag{66}$$

The mean square error of the residual is

$$MSE = \frac{\mathbf{y}'\mathbf{y}}{n-1}\left(1 - \mathbf{y}'\mathbf{X}\mathbf{D}^{-1}\mathbf{X}'\mathbf{y}\right) \tag{67}$$

where $\mathbf{D^{-1}} = (\mathbf{X'X})^{-1}$. We can see that the MSE decreases as the covariances between the inputs and output increase, whereas the MSE increases as the covariances between the input variables increases. The quality of the fit is measured by the *coefficient of determination*,

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}. \tag{68}$$

The *Analysis of Variance* (ANOVA) is used for significance tests of multiple regressions based on the amount of variance accounted for by the regression model. The regression model can be rewritten as

$$(y_i - \overline{y}) = (\hat{y}_i - \overline{y}) + (y_i - \hat{y}_i), \tag{69}$$

where the first term is the variation of the observed output variable about its sample mean, the second is the variation of the estimate, or model, about the mean, and the third is the residual value. Squaring this equation and summing over the observations gives

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n} 2(\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{70}$$

which can be referred to as $SST = SSM + SSE$ where SS is the sum of squares and the T, M, E refer to the total, model, and error terms. In this notation, the *correlation of determination* is

$$R^2 = \frac{SSM}{SST}. \tag{71}$$

The ANOVA is based on the sample distribution

$$F = \frac{MSM}{MSE} \tag{72}$$

where $MSM = SSM/(m-1)$, $MSE = SSE/(n-m)$, m is the number of inputs to the regression that each have n independent values, and a mean component is included in the regression inputs. F has a F *distribution* that can be used to test the null hypothesis that $b_1 = b_2 = .. = b_m = 0$. A p-value based on F that is $< \alpha$ would lead to a rejection of the null hypothesis.

Standard deviations, S_{b_j}, for the regression coefficients $(b_j, j = 1, 2, .., m)$ can be obtained from

$$E = MSE\,(x'x)^{-1}, \tag{73}$$

where S_{b_j} is the square root of the *jth* diagonal term of E. Assuming that the inputs are joint Normally distributed, independent random variables, then confidence intervals are obtained from

$$b_j = \hat{b_j} \pm t_{\alpha/2,n-m} S_{b_j}. \tag{74}$$

14 Serial correlation and degrees of freedom

In the statistical inferences considered so far (confidence intervals, hypothesis tests), the data (or in the case of the linear regression the residual error) have been treated as *independent*, which has simplified specifications of the mean square error and the number of degrees of freedom. In practice, an oceanographic time series is rarely a collection of independent data, as the series generally are over-sampled relative to the characteristic time scales of the variability.

To illustrate the effect of autocorrelation over time, or *serial correlation*, consider the sample mean

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{75}$$

The variance of the sample mean is

$$\begin{aligned} \sigma_{\overline{x}}^2 &= E\left((\overline{x}-\mu)^2\right) \\ &= E\left(\overline{x}^2\right) - \mu^2 \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} E(x_i' x_j') \end{aligned} \tag{76}$$

where $x_i' = x_i - \mu$. Previously we assumed that the observations were independent, in which case eq.(76) simplifies to

$$\sigma_{\overline{x}}^2 = \frac{E(x'^2)}{n} = \frac{\sigma^2}{n}. \tag{77}$$

More generally for a *stationary* time series that is not independent, eq.(76) can be expressed as

$$\sigma_{\overline{x}}^2 = \frac{1}{n}\sum_{k=1-n}^{n-1}\left(1-\frac{|k|}{n}\right)\gamma_x(k) \tag{78}$$

where

$$\gamma_x(k) = E\left(x_i' x_{i+k}'\right) \tag{79}$$

is the *autocovariance* of x at lag k. Eq.(78) can be rewritten as

$$\sigma_{\overline{x}}^2 = \frac{\sigma^2}{n^*}, \tag{80}$$

where n^* is the *effective number of degrees of freedom* for the sample mean,

$$n^* = n\left[\sum_{k=1-n}^{n-1}\left(1-\frac{|k|}{n}\right)\rho_x(k)\right]^{-1}, \tag{81}$$

and

$$\rho_x(k) = \frac{\gamma_x(k)}{\gamma_x(0)} \tag{82}$$

is the *autocorrelation function*. Serial correlation in x reduces the number of degrees of freedom (eq. 81) and increases the standard deviation of the sample mean. Note that eq.(81) is the effective number of degrees of freedom specific to the sample mean. The impact of serial correlation for different estimators will lead to different expressions for n^*.

Another measure of the number of independent points in a single time series, x, is based on the *integral time scale*, T, defined as

$$T = \sum_{k=1}^{m} \left(\rho_x(k-1) + \rho_x(k)\right) \frac{\Delta\tau}{2}, \tag{83}$$

where m is the maximum lag considered, $\Delta\tau$ is the incremental lag of the autocorrelation function, which typically equals Δt the sample period of x. T provides a measure of the *decorrelation time* of x. The effective number of degrees of freedom based on the integral time scale is

$$n^* = n\frac{\Delta t}{T}. \tag{84}$$

In practice, the calculation of the integral time scale is unsatisfactory if x contains energetic, long period components that tend to prevent eq.(83) from converging. In addition at long lags, errors in the sample autocorrelation function can lead to biased, unreliable, and inconsistent estimates of T. *Emery and Thomson* recommend evaluating eq.(83) over a range of lags to evaluate the stability of the T estimate. *Firing* [1989] sets m to the first zero crossing of the autocorrelation function.

An alternative to the integral time scale approach is obtained if the time series can be modeled as an *auto-regressive process* (e.g., *AR(1)*):

$$x_i = c + \phi x_{i-1} + \epsilon_i, \tag{85}$$

where c and ϕ are constants and ϵ is a white noise process with zero mean and constant variance, then

$$n^* = n\frac{1-\rho(1)}{1+\rho(1)}. \tag{86}$$

We again emphasize that eq. (83) is defined for a single time series, and eq. (81) for the sample mean. Each estimator will have its own effective degrees of freedom expression. For example, for the sample covariance or sample correlation, D̂avis (1977) gives the expression

$$n^* = \frac{n}{\sum_{k=1-n}^{n-1} \left[\rho_x(k)\rho_y(k) + \rho_{xy}(k)\rho_{yx}(k)\right]}, \tag{87}$$

where

$$\rho_{xy}(k) = \frac{E\left(x'_i y'_{i+k}\right)}{\gamma_x(0)\gamma_y(9)}. \tag{88}$$

Davis, R. E., 1977: Techniques for statistical analysis and prediction of geophysical fluid systems. *Geophys. Astrophys. Fluid Dynamics*, **9**, 245-277.

Firing, E., 1989: Mean zonal currents below 1500 m near the equator, 159 °W. *J. Geophys, Res.*, **94**, C2, 2023-2028

15 Monte Carlo Simulations

Estimates based on a random sample are themselves random variables with their own probability distribution, or *sampling distribution*. So far we've considered statistical inferences by relating the sampling distribution to a classical distribution. For example, we've related the sampling distribution of the sample mean to the Normal and Student's t-distributions, the sample variance to the χ^2 distribution, and variance ratios from a multiple linear regression to the F distribution. This is the classical approach, but it has limitations when the expression for the sampling distribution is not straight-forward to derive, and/or when limiting assumptions are needed to obtain the expression, such as that the sample consists of Normally-distributed and independent data.

An alternative approach for statistical inferences makes use of *Monte Carlo simulations*. Instead of deriving the sampling distribution analytically, Monte Carlo simulations allow for an empirical determination of the sampling distribution by creating a large number of synthetic data samples each with simulated random noise. The synthetic samples are run through the estimator under consideration, yielding a range of estimated values that reflect the influence of noise on the desired signal. The simulated estimates define an estimate of the underlying sampling distribution. The method has gained popularity as computing power has increased, allowing rapid generation of random numbers and repeated computation of the estimate.

The first step in a Monte Carlo simulation is to define the statistical test. This is a useful exercise as classical error analyses tend to be black box with underlying assumptions hidden. For example, we may wish to determine the standard deviation of an estimate, confidence intervals, whether the estimate stands above the range of values expected from random noise, etc.

The main consideration is how to define the underlying noise or random component of the samples. Common options include:

1) random numbers generated from a classical distribution or from a Markov chain;

2) drawing the samples from the original data set with replacement (i.e., the same datum can be selected more than once), the "signal" must be removed from the data to create an estimate of the noise, random numbers are used to scramble the index of the series;

3) use of spectral representations with random amplitude and/or phase.

The first two methods are easiest to program; however, when dealing with data series, care must be taken to account for serial correlation. The spectral method addresses serial correlation directly through the specification of the underlying spectral form. Once the method for simulating time series is specified, the estimate is repeated a large number of times, thus generating a probability distribution of the estimate..

16 Empirical Orthogonal Functions

Empirical Orthogonal Function (EOF) analysis is used to decompose a two-dimensional dataset, typically with dimensions of space and time, into orthogonal basis functions or modes. Unlike a Fourier analysis and other decompositions in which the basis functions are specified, the EOF basis functions are determined directly from the data, or empirically. EOFs not only provide an orthogonal basis set, but the modes also are efficient in that the first mode explains the dominant covarying pattern, the second mode the next dominant pattern of the residual (mode 1 removed) that is orthogonal to the first mode, and so on. EOFs often can compress a large dataset into a small subset of modes that account for much of the overall variance.

Given a variable $h(x,t)$ that is a specified at N_x space points and N_t time points, the EOF analysis represents the data as a sum of the product of spatial and temporal functions,

$$h(x,t) = \sum_{k=1}^{N} a_k(t) e_k(x) \tag{89}$$

where $e_k(x)$ is the spatial basis function for mode k, and $a_k(t)$ is the temporal expansion function for mode k, and $N = min(N_x, N_t)$. An orthogonality condition is imposed such

that the spatial modes are orthonormal over x,

$$\sum_x e_j(x)e_k(x) = \delta_{jk}, \tag{90}$$

and that the temporal expansion functions are uncorrelated over t,

$$< a_j a_k >_t = \delta_{jk} < a_k^2 >_t, \tag{91}$$

where $< .. >_t$ represents a time average. The e_k's are the eigenfunction solutions of

$$\mathbf{Ce} = \lambda \mathbf{e}, \tag{92}$$

where $\mathbf{C}$ is the *covariance matrix* of $h(x,t)$ with elements

$$C_{mn} = < h'(x_m,t)h'(x_n,t) >_t, \tag{93}$$

where the prime indicates departures from a mean, e.g., the temporal mean at each grid point or the spatial mean at each time. The e_k's are the column vectors of $\mathbf{e}$, and λ is a diagonal matrix of eigenvalues, which represent the variance accounted for by each mode

$$\lambda_k = < a_k^2 >_t . \tag{94}$$

The eigenfunctions are ordered such that $\lambda_1 > \lambda_2 > .. > \lambda_N$. The total variance of the dataset is represented by the sum of the eigenvalues

$$\sum_x < h(x,t)^2 >_t = \sum_{k=1}^{N} \lambda_k. \tag{95}$$

For each spatial mode we compute the corresponding temporal expansion function as,

$$a_k(t) = \sum_x h(x,t)e_k(x). \tag{96}$$

The above applies when $N_t > N_x$, in which case the spatial functions are orthonormal, and the temporal expansions have the same physical unit as $h(x,t)$. If $N_x > N_t$, the dimensions can be switched such that $h(x,t)$ is represented by a set of temporal basis functions with associated spatial expansion functions,

$$h(x,t) = \sum_{k=1}^{N} a_k(x)e_k(t). \tag{97}$$

The orthonormal temporal modes,

$$\sum_t e_j(t)e_k(t) = \delta_{jk}, \tag{98}$$

are the eigenfunctions of $\mathbf{C}$, computed using spatial averages,

$$C_{ij} =< h'(x,t_i)h'(x,t_j) >_x, \tag{99}$$

and the spatial expansion functions are obtained by projecting the temporal modes on to the data,

$$a_k(x) = \sum_t h(x,t)e_k(t). \tag{100}$$

Whether $N_t > N_x$ or $N_x > N_t$, the spatial and temporal functions described by the $N = min(N_t, N_x)$ modes are equivalent. It is more efficient computationally to form C so that it has the smaller dimension.

There are a number of variations to the standard EOF analysis described above. For example, the eigenfunctions can be obtained from the correlation matrix (normalizes all data to equal variance), the cross-spectral matrix (complex eigenfunctions with amplitude and phase information), the lagged-covariance matrix (not restricted to standing patterns), etc. Since the EOFs form a basis set, they can be rotated to emphasize signals in sub-domains of the data.

The EOFs also can be obtained directly from the dataset (i.e., you do not need to compute C) using a *Singular Value Decomposition* (SVD). Given an invertible, real m x n matrix $\mathbf{A}$ with $m > n$, then the singular value decomposition of $\mathbf{A}$ is

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^\mathbf{T}, \tag{101}$$

where $\mathbf{U}$ is a m x m matrix, $\mathbf{D}$ is m x n, and $\mathbf{V}$ is n x n. $\mathbf{U}$ and $\mathbf{V}$ have orthogonal columns so that

$$\mathbf{U}^\mathbf{T}\mathbf{U} = \mathbf{I} \tag{102}$$

and

$$\mathbf{V}^\mathbf{T}\mathbf{V} = \mathbf{I}. \tag{103}$$

So if $\mathbf{A}$ is our original data array $h(x,t)$, then the columns of $\mathbf{V}$ are equivalent to the eigenfunctions obtained from the covariance matrix, $\mathbf{C}$, and the columns of $\mathbf{UD}$ are the expansion functions for each mode. $\mathbf{D}$ is a diagonal matrix whose elements are the *singular values* of $\mathbf{A}$, which are related to the eigenvalues of the covariance matrix of $\mathbf{A}$ by $d_k = \lambda_k^{1/2}$.

17 Fourier series

A *Fourier series* represents a periodic function in terms of cosine and sine basis functions. Recall that cosines and sines form an orthogonal set of functions over the interval $[-\pi, \pi]$,

such that

$$\int_{-\pi}^{\pi} sin(mt)sin(nt)dt = \pi\delta_{mn}, \tag{104}$$

$$\int_{-\pi}^{\pi} cos(mt)cos(nt)dt = \pi\delta_{mn}, \tag{105}$$

$$\int_{-\pi}^{\pi} sin(mt)cos(nt)dt = 0, \tag{106}$$

$$\int_{-\pi}^{\pi} sin(mt)dt = 0, \tag{107}$$

$$\int_{-\pi}^{\pi} cos(mt)dt = 0. \tag{108}$$

Any periodic, piecewise continuous function over the interval can be represented by a *Fourier series* given by

$$y(t) = \frac{a_o}{2} + \sum_{n=1}^{\infty} a_n cos\,(nt) + b_n sin\,(nt)\,. \tag{109}$$

A set of orthogonal cosines and sines can be obtained for any record length T by defining a change of variables $t = 2\pi t'/T$, which after substituting into eq.(109) and dropping the prime superscript yields

$$y(t) = \frac{a_o}{2} + \sum_{n=1}^{\infty} a_n cos\,(2\pi f_n t) + b_n sin\,(2\pi f_n t)\,, \tag{110}$$

where $f_n = n/T$.

The coefficients of the Fourier series are given by

$$a_n = \frac{2}{T}\int_{-T/2}^{T/2} y(t) cos\,(2\pi f_n t)\,dt, \quad k = 0, 1, 2, ... \tag{111}$$

$$b_n = \frac{2}{T}\int_{-T/2}^{T/2} y(t) sin\,(2\pi f_n t)\,dt, \quad n = 1, 2, ... \tag{112}$$

An equivalent representation to eq.(110) is the *complex Fourier series* expressed in terms of complex exponential functions

$$y(t) = \sum_{n=-\infty}^{\infty} c_n exp\,(i2\pi f_n t)\,dt. \tag{113}$$

The coefficient c_n are given by

$$c_n = \frac{2}{T}\int_{-T/2}^{T/2} y(t) exp\,(-i2\pi f_n t)\,dt, \quad n = ... - 2, -1, 0, 1, 2, ... \tag{114}$$

The complex coefficients c_n can be expressed in terms of the real coefficients for the Fourier series,

18 Discrete Fourier transform

For a discrete series sampled at regular intervals $(x_o, x_1, ..., x_{N-1})$, the *discrete Fourier transform* is defined as

$$X_n = \sum_{k=0}^{N-1} x_k exp(-i2\pi kn/N), \quad n = 0, 1, .., N-1. \tag{115}$$

If x is a time series, the Fourier transform can be considered as a link between the time (t_k) and frequency (f_n) domains. Eq.(115) also is referred to as the *Forward transform*. The complex X_n represent both the amplitude $(|X_n|/N)$ and phase $(\arctan(X_n))$ of the sinusoidal component of x at frequency $f_n = n/N$ cycles per sample.

The *inverse Fourier transform* is defined as

$$x_k = \frac{1}{N} \sum_{n=0}^{N-1} X_n exp(i2\pi kn/N), \quad k = 0, 1, ..., N-1. \tag{116}$$

The normalizations in eq.(115) and eq.(116) differ from the Fourier series described above; however, the normalizations are arbitrary as long as the forward and inverse transforms have opposite sign and the product of their coefficients equals $1/N$. The normalizations in eq.(115) and eq.(116), as well as the convention of indexing t from 0 to $N-1$, are used commonly in the *Fast Fourier Transform* (FFT), the standard algorithm for computing the discrete Fourier transform.

An important property of the Fourier transform is that both X_n and x_k are *N-periodic*, that is $X_{n+N} = X_n$ and $x_{k+N} = x_k$. It also follows that $f_{N-k} = -f_k$, thus the frequencies in the range $f_{N/2} < f_n < f_N$ correspond to the negative frequencies in the complex Fourier series. The frequency $f_{N/2} = 1/2$ (i.e., half the sample frequency) is the *Nyquist* or *cut-off frequency*, which is the highest absolute frequency that is unambiguously resolved by the Fourier transform of a discrete time series. Contributions from frequencies above the Nyquist are *aliased* in the resolved frequency range according to the N-periodic condition.

The Fourier transform applies to both real and complex series (i.e., scalars and vectors). If x is real, then $X_{N-n} = X_n^*$ where the star indicates complex conjugation.

Parseval's theorem states that

$$\sum_{k=0}^{N-1} |x_k|^2 = \frac{1}{N} \sum_{n=0}^{N-1} |X_n|^2. \tag{117}$$

Thus the variance of x is represented by the sum of the squared Fourier amplitudes.

The *convolution theorem* states that a convolution in the time domain corresponds to a product in the frequency domain and vice versa. In other words

$$\mathcal{F}[x * y] = X \cdot Y, \tag{118}$$

where $\mathcal{F}$ represents a Fourier transform, and $*$ a convolution. We've used the convolution theorem already in computing the autocorrelation function, and it is an important concept for understanding spectra and digital filtering.

19 Autospectrum estimation

The *autospectrum* describes how the variance of a series (x_k, $k = 0, 1, .., N-1$) is distributed over frequency. A basic version of the autospectrum is the *periodogram*, which plots the squared Fourier amplitude, $|X|^2/N^2$, versus f. An issue with the periodogram in this form is that the amplitude of X varies with record length, $T = N\Delta t$, or frequency bandwidth, $\Delta f = 1/T$, which makes it cumbersome to compare periodograms from time series with different lengths.

An alternative to the periodogram is the *spectral density*, or spectral amplitude per frequency bandwidth,

$$\tilde{S}(f) = \frac{|X(f)|^2}{N^2}\frac{1}{\Delta f} = \frac{|X|^2 \Delta t}{N}, \tag{119}$$

which avoids the problem of record length dependent amplitudes. $\tilde{S}$ turns out to be a poor estimate of the autospectrum, however, because it has a large mean square error (MSE). For example, if x is assumed to be Normally distributed, then the autospectral estimate can be related to a chi-square random variable with just 2 degrees of freedom (sine and cosine). At 2 degrees of freedom, the MSE of $\tilde{S}$ is on the order of the estimate $\tilde{S}$ itself. This is the case regardless of record length, i.e., increasing T does not decrease the MSE, which makes $\tilde{S}$ an inconsistent estimate of the autospectrum.

A better estimate for the autospectrum is achieved by increasing the number of degrees of freedom using some form of averaging,

$$\hat{S}(f) = \frac{1}{m}\sum_{i=1}^{m} \tilde{S}_i. \tag{120}$$

There are two common averaging methods:

1) *Segment averaging*

Divide the time series into m equal length segments, compute $\tilde{S}$ for each segment, and average these together to form $\hat{S}$. By averaging m segments, the number of degrees of freedom increases from $\nu = 2$ to $2m$, assuming that each segment is independent and the time series is stochastic. If windows are applied to each segment to suppress spectral leakage, then overlapping segments are recommended (Welchś method), generally at 50% overlap.

2) *Band averaging*
Compute $\tilde{S}$ based on the entire record and smooth in the frequency domain, typically by computing a running average of m adjacent frequencies. Block averaging over m adjacent frequencies nominally increases ν to $2m$, assuming that the underlying spectrum is white (i.e., approximately equal spectral amplitudes for all frequencies).

After using segment averaging, band averaging, or both, we can construct confidence limits for the autospectrum in the usual way:

$$Pr\left[\chi^2_{\alpha/2} < \frac{\nu\hat{S}(f)}{S(f)} < \chi^2_{1-\alpha/2}\right] = 1 - \alpha. \tag{121}$$

Note that statistical reliability is improved at the expense of frequency resolution.

For real time series, $X(f) = X(-f)^*$, in which case only positive f need be considered and the *one-sided* auto spectrum is used. This is obtained by doubling $\hat{S}(f)$ for all positive frequencies, except the mean and Nyquist.

20 Rotary spectra

Rotary spectra provide information on the distribution of energy versus frequency for a vector time series. Consider the current time series $w(t) = u(t) + iv(t)$, where u and v are east-west and north-south current components. The auto-spectrum of w is

$$\hat{S}(f) = \frac{1}{m}\sum_{i=1}^{m}\frac{|W(f)|^2\Delta t}{N} \tag{122}$$

The positive frequencies correspond to motions that rotate in the anti-clockwise direction with respect to time, and the negative frequencies in the clockwise direction.

For each frequency, we can think of the corresponding current component as a vector that traces an ellipse in time. The major and minor semi-axes of the ellipse are given by $L_M = A(f) + A(-f)$ and $L_m = |A(f) - A(-f)|$, respectively, where $A(\pm f) = |W(\pm f)|/N$. The ellipse major axis is oriented at an angle relative to the u axis given by $\theta(f) = 1/2(\epsilon(f) + \epsilon(-f))$, where $\epsilon(\pm f) = arctan(W(\pm f)))$..

21 Filters

Frequency domain

Digital filters are used to suppress unwanted fluctuations in a time series, $x(t)$, from particular frequency bands. In the frequency domain, this is accomplished by computing $X(f)$, the Fourier transform of $x(t)$, multiplying by a *transfer function* $H(f)$ (also called the *frequency response function* or *admittance function*), and computing the inverse Fourier transform to obtain the filtered time series,

$$\tilde{x}(t) = \mathcal{F}^{-1}\left[X(f)H(f)\right]. \tag{123}$$

The frequencies that are included are in the *pass bands* ($|H(f)| = 1$) and those that are suppressed are in the *stop bands* ($|H(f)| = 0$). A *low-pass filter* passes energy at frequencies lower than a given *cut-off frequency*, and suppresses energy at higher frequencies. The opposite is true for a *high-pass filter*, usually computed as the original time series minus the low-pass filtered series. *Band-pass filters* can also be designed to emphasize a range of frequencies.

Filter functions with sharp transitions between pass and stop bands will result in ringing in the time domain. Recall that a product in the frequency domain corresponds to a convolution in the time domain, so that the filtered time series equals the convolution of $x(t)$ and $h(t)$, the inverse transform of $H(f)$. Rectangular-shaped filters in the frequency domain will result in sinc-shaped inverse transforms in the time domain with large side lobes. This results in the *Gibbs phenomenon*, which yields *ringing* in the filtered time series near sharp transitions or discontinuities, including the ends of the record. To minimize the ringing effect associated with ideal filters, a tapering function can be applied that creates a smooth transition in $H(f)$ from the pass band to the stop band. A cosine or cosine squared shaped taper is commonly used.

Time domain

Applying a digital filter in the time domain involves a convolution of the time series, $x(t)$, with a filter function, $h(t)$. The filter function should be normalized so that sum(h)=1. A convolution in the time domain corresponds to a product in the frequency domain, $X(f)H(f)$, where the transfer function $H(f)$ is the Fourier transform of $h(t)$. The spectral behavior of a time domain filter can be evaluated using $H(f)$.

The ends of time series require attention when applying convolution filters. If the filter function has length = $2m+1$, then typically m points are truncated at the start and end of the filtered time series. Other options include zero padding both ends of $x(t)$ by m points before filtering, specifying a mirror image of m points about the first and last point of the time series, treating the series as cyclically periodic, or adjusting the filter weights using some optimization criterion. The same concerns apply to gaps in the time series. In the case of small gaps, one option is to interpolate through the gap prior to applying the filter.

An example of a Gaussian-shaped time domain filter is the *Blackman window*, defined by

$$h(k) = 0.42 - 0.5cos\left(\frac{2\pi k}{N-1}\right) + 0.08cos\left(\frac{4\pi k}{N-1}\right), \tag{124}$$

where N is the length of the window.

22 Complex demodulation

A complex demodulation is used to specify the amplitude and phase of a finite bandwidth periodic signal. Recall that constructive and destructive interferences of the frequency components within the finite band cause the temporal modulation of the dominant periodic signal within the band. Let's consider a time series $x(t)$ with an energy peak near f_o. We can demodulate the signal near the f_o peak by first multiplying by the complex exponential

$$x_o(t) = x(t)exp(-i2\pi f_o t). \tag{125}$$

This is equivalent to shifting the energy at frequency f to $f - f_o$, so that energy at f_o appears at zero frequency (i.e., shifts band of interest to the baseband), the mean component to $-f_o$, etc. To isolate the energy at the peak of interest, now near $f = 0$, apply a low-pass filter to $x_o(t)$ with a suitable cut-off frequency to remove signals outside the frequency band of interest. The low-pass filtered time series, $x_f(t)$, is complex, and so an amplitude and phase of the modulated signal can be obtained by

$$A(t) = 2|x_f(t)|, \tag{126}$$

$$\phi(t) = arctan(x_f)). \tag{127}$$

23 Cross-Spectrum

The *cross-spectrum* measures the co-variability of two time series as a function of frequency. It is the frequency domain analog of the covariance function in the time domain. The

cross-spectrum, $S_{xy}(f)$, between time series $x(t)$ and $y(t)$ can be obtained from the Fourier transform of the cross-covariance function ($C_{xy}(\tau)$), just as the autospectrum can be computed from the auto-covariance function. More commonly, the cross-spectrum is obtained directly from the Fourier transforms of $x(t)$ and $y(t)$,

$$\hat{S}_{xy}(f) = \overline{X^*(f)Y(f)}\frac{\Delta t}{N} \tag{128}$$

where the overbar signifies segment and/or band averaging.

Unlike the autospectrum, $S_{xy}(f)$ is complex,

$$S_{xy}(f) = L_{xy}(f) - iQ_{xy}(f), \tag{129}$$

where $L_{xy}(f)$ is the *co-spectrum*, which measures in phase variability between $x(t)$ and $y(t)$ at frequency f, and Q_{xy} is the *quadrature spectrum*, which measures variability that is in quadrature or 90° out of phase.

The cross-spectrum typically is presented in polar form,

$$S_{xy}(f) = A_{xy}(f)exp\left(i\phi_{xy}(f)\right), \tag{130}$$

where

$$A_{xy}(f) = |S_{xy}(f)| = (L_{xy}^2 + Q_{xy}^2)^{1/2} \tag{131}$$

is the *cross-amplitude*, and

$$\phi_{xy}(f) = atan2\left(-Q_{xy}(f), L_{xy}(f)\right) \tag{132}$$

is the *phase*.

The *coherence* is the frequency domain analog of the correlation in the time domain,

$$\gamma_{xy}(f) = \frac{A_{xy}(f)}{\left(S_{xx}(f)S_{yy}(f)\right)^{1/2}}. \tag{133}$$

$\gamma_{xy}(f) = 0$ indicates no relationship between $x(t)$ and $y(t)$ in the frequency band of interest, $\gamma_{xy}(f) = 1$ indicates a perfect correspondence. $A_{xy}(f)$ is the frequency domain analog of the regression coefficient in the time domain. $\phi_{xy}(f)$ provides information on phase shifts, or time lags, which in the time domain is comparable to the information provided by lagged cross-correlations.

As with correlations, uncertainties for coherence estimates typically are assessed by testing the null hypothesis that the true coherence is zero. The null hypothesis is rejected if

coherence-squared estimates exceed the $(1-\alpha)100\%$ significance level, which is given in Emery and Thomson (2004) as

$$\gamma^2_{1-\alpha} = 1 - \alpha^{[2/(\nu-2)]}. \tag{134}$$

Hannah (1970) gives confidence intervals for phase as

$$|sin\left[\hat{\phi}(f) - \phi(f)\right]| \leq \left[\frac{1-\hat{\gamma}^2}{(2\nu-2)\hat{\gamma}^2}\right] t_{2\nu-2}(\alpha), \tag{135}$$

where t is the Student's t distribution. The uncertainty of phase estimates decrease as coherence levels and degrees of freedom increase.